

Proceedings of the  
**OAGM Workshop 2022**

**Digitalization for  
Smart Farming and Forestry**

Oct. 18, Nov. 7–8, 2022  
University of Natural Resources  
and Life Sciences, Vienna  
Tulln, Austria

Hermann Büstmayr, Andreas Gronauer, Andreas Holzinger,  
Peter M. Roth, and Karl Stampfer (eds.)

**Proceedings of the  
OAGM Workshop 2022**

**Digitalization for Smart Farming and Forestry**

October 18, November 7 and 8, 2022

University of Natural Resources  
and Life Sciences, Vienna

Austrian Association of Pattern Recognition (OAGM)

## Editors

Hermann Bürstmayr, Andreas Gronauer, Andreas Holzinger, Peter M. Roth,  
and Karl Stampfer

## Layout

Austrian Association of Pattern Recognition  
<https://aapr.at/>

## Cover

Verlag der Technischen Universität Graz  
Coverbild von Shutterstock / Liu zishan

2023 Verlag der Technischen Universität Graz  
[www.tugraz-verlag.at](http://www.tugraz-verlag.at)

ISBN (e-book) 978-3-85125-954-4  
DOI 10.3217/978-3-85125-954-4



This work is licensed under the Creative Commons  
Attribution 4.0 International (CC BY 4.0) license.  
<https://creativecommons.org/licenses/by/4.0/deed.en>

This CC license does not apply to the cover, third party material (attributed to other sources)  
and content noted otherwise.

## Supported by



**vetmeduni**



# Contents

Cover . . . . .	i
Impressum . . . . .	ii
Table of Contents . . . . .	iv
Preface . . . . .	vii
Workshop Chairs . . . . .	viii
Program Committee . . . . .	ix
Index of Authors . . . . .	x
<b>Keynote Talks</b> . . . . .	1
Generating the unseen and explaining the seen <i>Ribana Roscher</i> . . . . .	2
Uptake & usage of Smart Farming in Austrian agriculture <i>Martin Hirt</i> . . . . .	3
<b>Full Papers</b> . . . . .	4
Redundant 1-cells in Multi-labeled 2-Gmap Irregular Pyramids <i>Majid Banaeyan, Walter G. Kropatsch, and Jiří Hladůvka</i> . . . . .	5
On the Regularising Levenberg-Marquardt Method for Blinn-Phong Photometric Stereo <i>Georg Radow and Michael Breuß</i> . . . . .	12
Image Forgery Detection and Localization Using a Fully Convolutional Network <i>David Fischinger, David Schreiber, and Martin Boyer</i> . . . . .	19
A Modular Model Combining Visual and Textual Features for Document Image Classification <i>Amer Duhan and Robert Sablatnig</i> . . . . .	26
Statistical shape modeling and analysis of the vestibular organ based on CT-images <i>Claudia Companioni, Matthias Willenbrink, Karl Fritscher, and Rainer Schubert</i> . . . . .	32

One-Pixel Instance Segmentation of Leaves <i>Julia Strebl, Eric Stumpe, Thomas Baumhauer, Lena Kernstock, Markus Seidl, and Matthias Zeppelzauer</i> . . . . .	40
Towards Uncertainty Detection in Automated Leaf Tissue Segmentation <i>Ráchel Grexová, Klara Voggeneder, Danny Tholen, Guillaume Thérroux-Rancourt, Walter G. Kropatsch, and Jiří Hladůvka</i> . . . . .	47
An unsupervised, shape-based 3d cell instance segmentation method for plant tissues <i>Alexander Palmrich, Klara Voggeneder, Danny Tholen, Guillaume Thérroux-Rancourt, Jiří Hladůvka, and Walter G. Kropatsch</i> . . . . .	54
Exploring Learning-Based Approaches for Bomb Crater Detection in Historical Aerial Images <i>Marvin Burges, Sebastian Zambanini, and Robert Sablatnig</i> . . . . .	60
<b>Student Papers</b> . . . . .	67
Automated nuclear morphometry as a prognostic marker in canine cutaneous mast cell tumors <i>Eda Parlak, Andreas Haghofer, Taryn A. Donovan, Robert Klopffleisch, Stephan Winkler, Matti Kiupel, Marc Aubreville, and Christof A. Bertram</i> . . . . .	68
Modeling the Diffusion of CO <sub>2</sub> inside Leaves <i>Yannis Sauzeau, Walter G. Kropatsch, and Jiří Hladůvka</i> . . . . .	70
<b>Application Spotlight Papers</b> . . . . .	73
Novel contactless fingerprint scanner for Legal Enforcement Agencies <i>Axel Weissenfeld, Erich Voko, Reinhard Schmid, Bernhard Strobl, Bernhard Kohn, and Gustavo Fernandez Dominguez</i> . . . . .	74
Crop row detection utilizing spatial CNN modules <i>Peter Riegler-Nurscher and Leopold Rupp</i> . . . . .	77
A Computer Vision System for Evaluation of Field Robot Operations <i>Florian Kitzler, Andreas Gronauer, and Viktoria Motsch</i> . . . . .	80
Vision-Language Models for Filtering and Clustering Forensic Data <i>Axel Weissenfeld, Bernhard Strobl, David Weichselbaum, Christopher Wimmer, and Martina Tschapka</i> . . . . .	82
Estimation of nitrogen yield in wheat using radiative transfer model inversion based on an artificial neural network <i>Lukas J. Koppensteiner and Reinhard Neugschwandtner</i> . . . . .	85
Selection of YOLOX Backbone for Monitoring Sows' Activity in Farrowing Pens with a Possibility of Temporary Crating <i>Maciej Oczak</i> . . . . .	88

Influence of Data Processing on Hyperspectral-Based Classification of Managed Permanent Grassland <i>Viktoria Motsch, Roland Britz, and Andreas Gronauer</i> . . . . .	91
<b>Scientific Spotlight Papers</b> . . . . .	94
In Defense of Information Plane Analysis <i>Mina Basirat, Bernhard C. Geiger, and Peter M. Roth</i> . . . . .	95
Computer-assisted mitotic count using a deep learning-based algorithm improves interobserver reproducibility and accuracy <i>Christof A. Bertram, Marc Aubreville, and Robert Klopffleisch</i> . . . . .	98
A Modern Approach for Early Wildfire Detection <i>Kurt Winter and Peter M. Roth</i> . . . . .	100

## Preface

The OAGM Workshop aims to bring together researchers, students, professionals, and practitioners from the fields of Computer Vision and Pattern Recognition to present and actively discuss the latest research and developments. As every year, there is a core topic which was "Digitalization for Smart Farming and Forestry" in the 2021 edition. Originally, the OAGM Workshop 2022 was planned again as an on-site event at the University of Natural Resources and Life Sciences, Vienna (Campus Tulln) in September 2022. As a result of the official restrictions by both the university and the government due to the still ongoing COVID-19 pandemic, we needed some re-organization, with the final decision not to cancel the workshop but to have an online event scheduled in three sessions (October 18, November 8, and November 9, 2022).

Consequently, it was possible to publish the conference proceedings. We thank the authors and reviewers for their contributions to this publication. We received 22 original contributions which 21 (9/9 full papers, 2/3 student papers, and 10/10 industrial and scientific spotlight papers) have been accepted. Each contribution was peer-reviewed in a double-blind process by at least two reviewers from an international program committee. One outstanding contribution will be awarded the best paper prize sponsored by OCG. In addition, there will be an IEEE Women in Engineering Award, sponsored by the Austrian Institute of Technology, for the best contribution of a female first author. We want to thank OCG and IEEE/AIT for sponsoring these awards and the project DILAG for the financial support.

We would also like to thank the invited speakers, Ribana Roscher (University of Bonn) and Martin Hirt (Austrian Chamber of Agriculture) for their presentations taking into account the scientific and application points of view.

Hermann Bürstmayr, Andreas Gronauer, Andreas Holzinger, Peter M. Roth, and Karl Stampfer  
(conference chairs)  
Tulln, November 2022

## **Workshop Chairs**

Hermann Bürstmayr (University of Natural Resources and Life Sciences, Vienna)

Andreas Gronauer (University of Natural Resources and Life Sciences, Vienna)

Andreas Holzinger (University of Natural Resources and Life Sciences, Vienna)

Peter M. Roth (University of Veterinary Medicine, Vienna, Technical University of Munich)

Karl Stampfer (University of Natural Resources and Life Sciences, Vienna)



## **Program Committee**

Mina Basirat (Graz University of Technology)  
Csaba Beleznai (Austrian Institute of Technology)  
Christof Bertram (University of Veterinary Medicine, Vienna)  
Marcus Bloice (Medical University Graz)  
Kristian Bredies (University of Graz)  
Jan Egger ( University Hospital Essen)  
Gustavo Fernandez Dominguez (Austrian Institute of Technology)  
Margrit Gelautz (Vienna University of Technology)  
Thomas Heitzinger (Vienna University of Technology)  
Martin Hirzer (AVL List GmbH)  
Jiří Hladůvka (Vienna University of Technology)  
Andreas Holzinger (University of Natural Resources and Life Sciences, Vienna)  
Martin Kappel (Vienna University of Technology)  
Manuel Keglevic (Vienna University of Technology)  
Florian Kleber (Vienna University of Technology)  
Christoph Lampert (IST Austria)  
Mathias Lux (AAU Klagenfurt)  
Hubert Mara (MLU Halle-Wittenberg)  
Maciej Oczak (University of Veterinary Medicine, Vienna)  
Pablo Rischbeck (University of Natural Resources and Life Sciences, Vienna)  
Antonio Rodriguez-Sanchez (University of Innsbruck)

# Index of authors

- Aubreville, Marc, 68, 98
- Banaeyan, Majid, 5
- Basirat, Mina, 95
- Baumhauer, Thomas, 40
- Bertram, Christof A., 68, 98
- Boyer, Martin, 19
- Breuß, Michael, 12
- Britz, Roland, 91
- Burges, Marvin, 60
- Companionì, Claudia, 32
- Dominguez, Gustavo Fernandez, 74
- Donovan, Taryn A., 68
- Duhan, Amer, 26
- Fischinger, David, 19
- Fritscher, Karl, 32
- Geiger, Bernhard C., 95
- Grešov, Rchel, 47
- Gronauer, Andreas, 80, 91
- Haghofer, Andreas, 68
- Hirt, Martin, 3
- Hladvka, Jir, 5, 47, 54, 70
- Kernstock, Lena, 40
- Kitzler, Florian, 80
- Kiupel, Matti, 68
- Klopfleisch, Robert, 68, 98
- Kohn, Bernhard, 74
- Koppensteiner, Lukas J., 85
- Kropatsch, Walter G., 5, 47, 54, 70
- Motsch, Viktoria, 80, 91
- Neugschwandtner, Reinhard, 85
- Oczak, Maciej, 88
- Palmrich, Alexander, 54
- Parlak, Eda, 68
- Radow, Georg, 12
- Riegler-Nurscher, Peter, 77
- Roscher, Ribana, 2
- Roth, Peter M., 95, 100
- Rupp, Leopold, 77
- Sablatnig, Robert, 26, 60
- Sauzeau, Yannis, 70
- Schmid, Reinhard, 74
- Schreiber, David, 19
- Schubert, Rainer, 32
- Seidl, Markus, 40
- Strebl, Julia, 40
- Strobl, Bernhard, 74, 82
- Stumpe, Eric, 40
- Tholen, Danny, 47, 54
- Throux-Rancourt, Guillaume, 47, 54
- Tschapka, Martina, 82
- Voggeneder, Klara, 47, 54
- Voko, Erich, 74
- Weichselbaum, David, 82
- Weissenfeld, Axel, 74, 82
- Willenbrink, Matthias, 32
- Wimmer, Christopher, 82
- Winkler, Stephan, 68
- Winter, Kurt, 100
- Zambanini, Sebastian, 60
- Zeppelzauer, Matthias, 40

# Keynote Talks

# Generating the unseen and explaining the seen

Ribana Roscher

Institute of Geodesy and Geoinformation

University of Bonn

## *Abstract*

*Deep generative models and explainable machine learning are two emerging areas of data science that we can use to address current challenges in agricultural and environmental sciences. Deep generative models are neural networks that are capable of learning complex data distributions. In general, they can be used for a variety of applications, such as anomaly detection, current state estimation, and prediction. Explainable machine learning, which analyzes the decision-making process of machine learning methods in more detail, is used whenever an explanation for the result is required in addition to the result. This can be done for various reasons, e.g., to increase confidence in the outcome or to derive new scientific knowledge that can be inferred from patterns in the decision process of the machine learning model. This talk addresses methods and applications from both areas and how we can take advantage of their combination.*

# Uptake & usage of Smart Farming in Austrian agriculture

Martin Hirt

Austrian Chamber of Agriculture

## *Abstract*

*In 2021 Austrian Federal Institute of Rural Education and Training conducted a survey among 1.000 farmers regarding attitudes, motivation and investment intentions towards increasing digitization in Austrian agriculture. The study aimed to provide valuable insights into actual usage and intended uptake of digital and precision farming technologies since this has been very much discussed since several years. While general attitude towards digitization seems to be quite “positive-pragmatic” (only 11% stated to be sceptical or negative), the actual usage vary largely between technology groups: Low-cost solutions in farm management like nutrient management recording are used more often than specific precision farming technologies. When asked about motivations for using digital technologies, farmers don’t argue with higher yields or performance but rather with more easier environmental recording (73%), less physical strain (65%) and increased time flexibility and leisure time (59%). Coming to the barriers of a quicker uptake, they stated mainly economic factors like doubtful cost-benefit considerations (70%), initial investments (69%) and running costs (62%). It’s interesting that even while most farmers named themselves as well-informed about new technologies in farming, a high share stated to be open for visiting further training (68%) or even individual advisory (59%) dealing with digital technologies in their specific agricultural branches.*

# Full Papers

# Redundant 1-cells in Multi-labeled 2-Gmap Irregular Pyramids

Majid Banaeyan, Walter G. Kropatsch and Jiří Hladůvka  
Pattern Recognition and Image Processing Group, TU Wien  
1040 Wien, Favoritenstr. 9/5, E193-03, Vienna, Austria  
{majid,krw,jiri}@prip.tuwien.ac.at

## Abstract

Nowadays the amount of generated digital data is growing faster and faster in a broad spectrum of application domains such as biomedical and biological imaging, document processing, remote sensing, video surveillance, etc. Processing such big data encourages efficient data structure and powerful processing algorithms. The  $n$ -dimensional generalized map is a useful structure that completely represents the topological structure of an image. Their advantages have been widely proved in the literature. Nevertheless, the main disadvantage of these structures is the high rate of memory requirement. This paper, first proposes an efficient method that implicitly encodes two of the three involutions in the 2-Gmap that dramatically reduces the amount of required memory. Second, it introduces a new formalism to define and detect redundant 1-cells (edges), in the 2-Gmap. Removing such redundant information the reduced memory is further decreased approximately by half. Finally, experiments show the advantage of the proposed method in a real database of high-resolution X-ray micro-tomography ( $\mu$ CT) and fluorescence microscopy.

## 1. Introduction

We are live in the era of *Big Data*. In 2018 it was stated "Data volumes are exploding; more data has been created in the past two years than in the entire history of the human race [9]." Nowadays, the data volume and velocity is growing even faster [15]. Processing such a huge amount of data requires efficient data structures and efficient processing algorithms. In addition, currently we are working on the *Water's gateway to heaven* project<sup>1</sup> dealing with high-resolution X-ray micro-tomography ( $\mu$ CT) and fluorescence microscopy. The size of the labeled cross slice of a leaf scan is more than 2000 in each dimension. To correctly preserve the structure of the elements in the image, in this paper we employ 2-dimensional generalized map (2-

Gmap) [13].

Although the  $n$ -Gmap is an efficient structure for describing an  $n$ -dimensional orientable or non-orientable quasi-manifold [13] it suffers from requiring a huge amount of memory storage. To remedy this problem, this paper first introduces an efficient encoding to implicitly preserve elements of the 2-Gmap without taking extra space of memory. Second and more important, it introduces a new formalism to define and detect redundant elements of the 2-Gmap structure of the multi-labeled image.

By removing the redundant elements, the resulted 2-Gmap not only has the same structure to the original one but it would be also computationally more efficient to be used in upcoming processing. In particular, to process both general and local information of the structure we use the irregular graph pyramid. Removing such redundant elements in the hierarchical structure, simplifies and speeds up the construction of the pyramid. In this paper we are dealing with *multi-labeled* images. The multi-labeled image is defined as an image consists of different connected components (CCs) where each CC has a unique label (color).

### 1.1. Irregular Pyramid

Pyramids are powerful and efficient hierarchical structures in pattern recognition that were introduced by Rosenfeld [16]. They are able to propagate local information from the base level into global and abstract information at top of the pyramid [14]. Irregular image pyramids consist of a series of successively smaller images constructed over a base image [10]. By presenting a digital image as a 4-adjacent neighborhood graph, each pixel in image  $P$  corresponds to a vertex  $v \in V$  of the graph  $G = (V, E)$ . Each edge,  $e \in E$ , of the graph encodes the neighborhood relationship between pixels. In addition, the gray-value of a pixel  $g(p)$  becomes an attribute of the corresponding vertex  $v$ ,  $g(v) = g(p)$  and the *contrast*( $e$ ) =  $|g(u) - g(v)|$  becomes an attribute of an edge  $e(u, v)$  where  $u, v \in V$ . In an irregular pyramid, in order to produce the smaller graph at the upper level, two operations are performed at each level: edge contraction and edge removal [5, 6]. The former re-

<sup>1</sup><https://waters-gateway.boku.ac.at/>

moves one edge and one vertex while preserving the connectivity of a graph and the latter removes one edge. Vertices (edges) of the current level that will be disappeared at the upper level are called *non-surviving* vertices (edges) while those that remain at the upper levels are called *surviving* vertices (edges). The decision of which vertices (and consequently which edges) must be selected as the surviving vertices (edges) is taken by introducing the *contraction kernel* (CK).

**Definition 1 (Contraction Kernel (CK))** A CK is a tree consisting of a surviving vertex as its root and some non-surviving neighbors with the constraint that every non-survivor can be part of only one CK.

An arrow over an edge is commonly used to indicate the direction of contraction, i.e., from non-survivor to survivor vertex. Using the 4-adjacent neighborhood relationship results in the plane graph. A *plane* graph is a graph embedded in the plane such that its edges intersect only at their endpoints [18]. In a plane graph, a *face* is the connected spaces between edges and vertices where its degree is the number of edges bounding the face. A face bounded by a cycle is called an *empty* face.

## 1.2. Gmap

An  $n$ -dimensional generalized map ( $n$ -Gmap) is a combinatorial data structure allowing to describe an  $n$ -dimensional orientable or non-orientable quasi-manifold with or without boundaries [13]. An  $n$ -Gmap is defined by a finite set of darts  $\mathcal{D}$  on which act  $n + 1$  involutions<sup>2</sup>  $\alpha_i$ , satisfying composition constraints of the following definition [7]:

**Definition 2 ( $n$ -Gmap)** An  $n$ -dimensional generalized map, or  $n$ -Gmap, with  $0 \leq n$  is an  $(n + 2)$ -tuple  $G = (\mathcal{D}, \alpha_0, \dots, \alpha_n)$  where:

1.  $\mathcal{D}$  is a finite set of darts,
2.  $\forall i \in \{0, \dots, n\}$ :  $\alpha_i$  is an involution on  $\mathcal{D}$
3.  $\forall i \in \{0, \dots, n - 2\}, \forall j \in \{i + 2, \dots, n\}$ :  $\alpha_i \circ \alpha_j$  is an involution.

A 2-Gmap  $(\mathcal{D}, \alpha_0, \alpha_1, \alpha_2)$  represents the structure of a set of surfaces. Darts as the fundamental elements of the 2-Gmap are linked together by involution functions. For example in Fig. 1,  $\alpha_0(21) = 22$ ,  $\alpha_1(21) = 10$  and  $\alpha_2(21) = 23$ .

**Definition 3 (i-cell)** Let  $G = (\mathcal{D}, \alpha_0, \dots, \alpha_n)$  be an  $n$ -Gmap,  $d \in \mathcal{D}$ , and  $i \in \{0, \dots, n\}$ . The  $i$ -dimensional cell (or  $i$ -cell) containing  $d$  is:

$$c_i(d) = \langle \alpha_0, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_n \rangle (d) \quad (1)$$

<sup>2</sup>self-inverse permutations

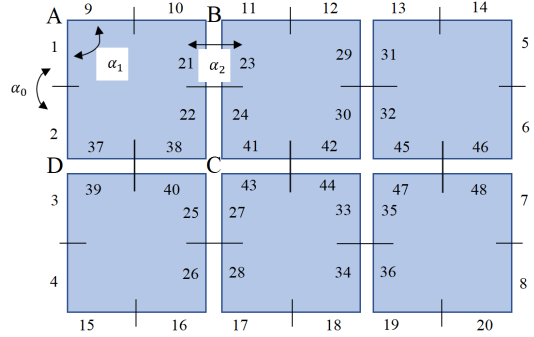


Figure 1. An example of a 2-Gmap

where

1.  $\langle \alpha_1, \alpha_2 \rangle (d)$  denotes the propagation of  $(\alpha_1^*, \alpha_2^*)^*(d)$  and identifies the 0-cell (a point), the eight darts surrounding  $C$  in Fig. 1.
2.  $\langle \alpha_0, \alpha_2 \rangle (d)$  denotes the propagation of  $(\alpha_0^*, \alpha_2^*)^*(d)$  and identifies the 1-cell consisting of the four darts between  $B$  and  $C$  in Fig. 1.
3.  $\langle \alpha_0, \alpha_1 \rangle (d)$  denotes the propagation of the orbit  $(\alpha_0^*, \alpha_1^*)^*(d)$  and identifies the 2-cell between  $A$ ,  $B$ ,  $C$  and  $D$  in Fig. 1.

Based on the definition 3, in Fig. 1,  $c_0(22) = \{22, 24, 41, 43, 27, 25, 38\}$  means this set of darts represents the 0-cell of the  $d = 22$ . In addition, the set  $\{22, 21, 23, 24\}$  and the set  $\{22, 21, 10, 9, 1, 2, 37, 38\}$  represent 1-cell and 2-cell corresponding to  $d = 22$ , respectively.

## 2. Corresponding graph of a 2-Gmap

Let  $G$  be a corresponding graph of a 2-Gmap. 0-cells and 1-cells of the 2-Gmap correspond to the vertices and edges of  $G$ , respectively. The 2-cells of the 2-Gmap correspond to the faces of degree 4 in the  $G$ . Fig. 2 shows  $G$  as the corresponding graph of the 2-Gmap of Fig. 1. Each edge of  $G$  consists of two half-edges or *darts*. There are three involutions,  $\alpha_0$ ,  $\alpha_1$  and  $\alpha_2$  encoding the relationships between darts (Fig. 3). To store the involutions one may consider an array of darts encoding each involution. However, we introduce specific encoding such that only one of these three involutions, i.e.  $\alpha_1$ , explicitly be stored in the 1D array of darts. The remaining two involutions,  $\alpha_0$  and  $\alpha_2$ , are implicitly encoded.

Assume  $G$  consists of  $M$  by  $N$  vertices containing  $n_d = 2 \times 2(M + N) + 2 \times (2MN - M - N)$  darts. The first term,  $2 \times 2(M + N)$ , indicates the number of darts in the



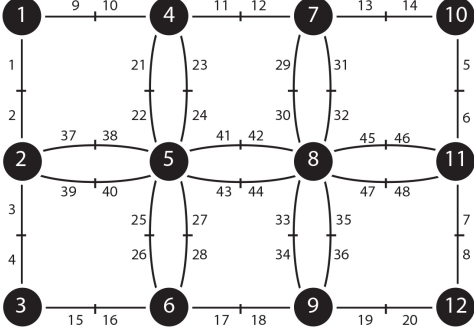


Figure 2. Corresponding graph  $G$  of a 2-Gmap

d	1	2	3	4	5	6	...	21	22	23	24	...
$\alpha_0(d)$	2	1	4	3	6	5	...	22	21	24	23	...
$\alpha_1(d)$	9	37	39	15	14	46	...	10	38	11	41	...
$\alpha_2(d)$	1	2	3	4	5	6	...	23	24	21	22	...

Figure 3. Array of darts in the 2-Gmap of Fig. 2

boundary where  $\alpha_2(k) = k$  and  $k \in [1, 2 \times 2(M + N)]$ . In Fig. 2 these darts are indicated by numbers 1 to 20. The second term,  $2 \times (2MN - M - N)$ , illustrates the remaining darts where  $\alpha_2(4k + 1) = 4k + 3$  and  $\alpha_2(4k + 2) = 4k + 4$  where  $k \in [(2 \times 2(M + N)) + 1, n_d]$ . In this manner,  $\alpha_2$  is implicitly encoded. In Fig. 2, these darts are indicated by numbers 21 to 48. Furthermore, we consider  $\alpha_0(2k - 1) = 2k$ , where  $k \in [1, n_d/2]$ . Therefore, the  $\alpha_0$  can be implicitly encoded as well.

### 2.1. Edge Classification

A multi-labeled image consists of different labels where each label represents an object (connected component). In the neighborhood graph of an input image, each connected component (CC) consists of a set of vertices with the same label (color). In this regard, we partition the edges of the neighborhood graph into two categories: intra-CC and inter-CCs as follows:

**Definition 4** *Intra-CC edge*: an edge  $e = (u, v)$  is *intra-CC* iff  $g(u) = g(v)$ .

**Definition 5** *Inter-CCs edge*: an edge  $e = (u, v)$  is *inter-CCs* iff  $g(u) \neq g(v)$ .

Based on the definitions above, the contrast of an intra-CC edge is equal to zero,  $c(e) = 0$ . We show the intra-CC edge by  $e_0 \in E_0$ . On the other hand, the contrast of an inter-CCs edge is larger than zero,  $c(e) > 0$ . The inter-CCs edge is shown by  $e_i \in E_i, i \in \mathbb{N}$ . Fig. 4 illustrates an example of multi-labeled image containing 4 CCs where each CC has a different color. We illustrate the  $E_0$  and  $E_i$  edges with

black and red color, respectively. The edges are partitioned as follows:

$$E = E_0 \cup E_i \quad (2)$$

### 2.2. Selecting the CKs

Selecting the CKs is the main procedure in building the irregular pyramid. In construction of the pyramid, a CC at the base level will be reduced into a single vertex at top of the pyramid. In other words, all vertices of a CC will be contracted through the pyramid until to reach a corresponding surviving vertex at the top level. To this aim, we select the CKs only from the  $E_0$  edges. In addition, in order to select a unique set of CKs, a **total order** is used over the indices of vertices [1, 2]. Consider the corresponding graph  $G$  of the 2-Gmap with  $M$  by  $N$  vertices. Let  $(1, 1)$  be the coordinate of the vertex at the upper-left corner and  $(M, N)$  at the lower-right corner. The vertices of  $G$  receive a unique index as follows:

$$Idx : [1, M] \times [1, N] \mapsto [1, M \cdot N] \subset \mathbb{N} \quad (3)$$

$$Idx(r, c) = (c - 1) \cdot M + r \quad (4)$$

The total order has two main properties [8]. First, any two elements (indices of vertices) are comparable. Second, every subset of vertices has one minimum and one maximum. Since the CKs are selected from  $E_0$ , a neighborhood  $\mathcal{N}(v)$  is defined as follows [1]:

$$\mathcal{N}(v) = \{v\} \cup \{w \in V | e_0 = (v, w) \in E_0\} \quad (5)$$

If the neighborhood has at least one member ( $|\mathcal{N}(v)| > 1$ ), then the surviving vertex is selected as follows [1]:

$$v_s = \operatorname{argmax}\{Idx(v_s) | v_s \in \mathcal{N}(v), |\mathcal{N}(v)| > 1\} \quad (6)$$

Because there is only one maximum number in every subset of the total order, there is only one unique surviving vertex for each non-surviving vertex.

### 3. Redundant edges in multi-labeled images

Graphs as a versatile representative tool may have many unnecessary edges [1, 2]. The definition of these unnecessary edges is different based on the specific application. In this paper, we study the redundant edges in multi-labeled images. In particular, a new formalism is defined to detect the redundant edges in the hierarchical structure of the irregular pyramid.

In constructing the irregular pyramid [6, 10], the neighborhood graph of an input image forms the base level of the pyramid. To reach the smaller graph at the upper level, a set of vertices are selected for contractions. The contraction operation reduces the number of vertices and edges in the resulting graph. The resulting graph may have empty self-loops or double edges that we define later as redundant

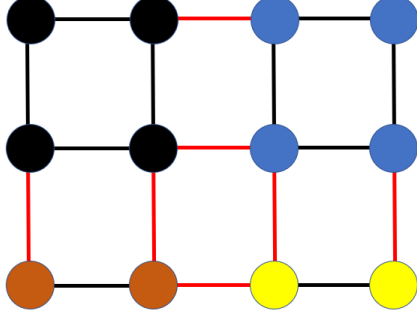


Figure 4. Edge classification in a  $3 \times 4$  multi-labeled neighborhood graph.

edges. The simplification procedure removes these redundant edges after the contractions. The edge contraction and edge removal are consecutively performed till the pyramid reaches to its top level [11, 12]. However, the simplification procedure can be performed before the contraction operation [4] where it facilitates the construction of the irregular pyramid.

In [1,4] the redundant edges in binary images are defined as follows:

**Definition 6 (Redundant-Edge (RE))** *In an empty face, the non-oriented edge incident to the vertex with lowest  $Idx$  is redundant iff:*

- The empty face is bounded by only non-oriented edges with the same contrast value.
- The empty face is bounded by non-oriented edges with the same contrast value and oriented edges.

By defining the new edge classification in Sec 2.1, the definition of redundant edges of binary images would be valid for the corresponding graph of the multi-labeled image. Fig. 5 shows all possible configurations of  $E_0$  and  $E_i$  in a face of degree 4 in the grid structure. The right column of this figure illustrates the resulting graph after the edge contraction. The RE after the contraction are either empty self loops or one of the double edges of a face of degree 2.

#### 4. Removing redundant 1-cells

In the previous section, it was shown that the RE can be predicted before constructing the pyramid. Since these RE have no rules in pyramid construction, they can be removed without harming the structure. Therefore, removing the RE reduces the memory space of the pyramid. In a binary image it is proved that up to 50% of the edges are redundant [4]. Considering the  $E_i$  edges as the category of

	All possible faces at the base level	After contractions
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		

Figure 5. The configuration of all possible redundant edges in a face of degree 4.

$E_1$  edges, it is concluded that the upper bound of the RE in the multi-labeled images is 50% as well.

Having sufficient independent processing elements, the redundant edges are removed with parallel  $\mathcal{O}(1)$  complexity [3,4]. To this aim, a set of independent edges (darts) are selected to be removed at the same time. By definition, two edges not sharing an endpoint are considered as independent edges [4]. Therefore, redundant edges or equivalently redundant 1-cells in the 2-Gmap are removed in a constant time.

	#Images	size	$ RE_\mu $	$\text{std}( RE )$
RE	120	$1350 \times 1142$	48.43%	1.04

Table 1. The amount of redundant edges (RE) in multi-labeled image.

## 5. Results

A 2-Gmap is completely defined by encoding its  $\alpha_i, i = 0, 1, 2$  involutions. We have shown in Sec .2 that by only preserving the  $\alpha_1$  darts the 2-Gmap is completely encoded. By using the canonical encoding [17], the memory consumption is equal to the size of the initial generalized map independent to the number of pyramid’s level. To build up the whole pyramid and use only darts at the base level, the history of contractions is preserved in a 1D array of darts. Two operations of the pyramid, edge contraction and edge removal, modify the  $\alpha_1$  while the  $\alpha_0$  and  $\alpha_2$  do not change in the entire pyramid. By detecting the redundant edges (darts), we put all the redundant darts on the left side of the array. These redundant darts have no role in constructing the pyramid. Fig. 6 shows an example of a 2-Gmap where the array of  $\alpha_1(d)$  encodes the entire of the 2-Gmap.

The redundant edges are illustrated by dashed-line in Fig. 6-b. These redundant edge (darts) are highlighted in the array of Fig. 6-d. By putting the redundant darts into the left side of the array, the remaining darts preserve the structure of the simplified 2-Gmap. As it was proved (Sec .4) up to 50% of the whole darts in a 2-Gmap would be redundant.

To exploit the advantage of the proposed method in a real application, we calculate the percentage of RE through a labeled 2D cross slice of a leaf scan (Fig. 7). The multi-labeled input image (Fig. 7) has six different labels illustrating different regions inside the leaf. The size of the original 2D slice is  $2560 \times 2560$  and there are 2160 slices in the volume of the 3D imaging. After cropping the unnecessary parts of the original image, the proposed algorithm was tested over 120 multi-labeled images<sup>3</sup> with the size  $1350 \times 1142$ .

Tab .1 displays the outcome of the proposed method. The first column shows number of images (#Images) of our multi-label data base. The second column displays the size of the 2D input image. The last two columns give the average amount of RE ( $|RE_\mu|$ ) along with the standard deviation ( $\text{std}(|RE|)$ ) over all images. The results show that the proposed method enormously reduces the size of the input image approximately by half.

<sup>3</sup>The images are from the *Water’s gateway to heaven* project, <https://waters-gateway.boku.ac.at/>

## 6. Conclusion

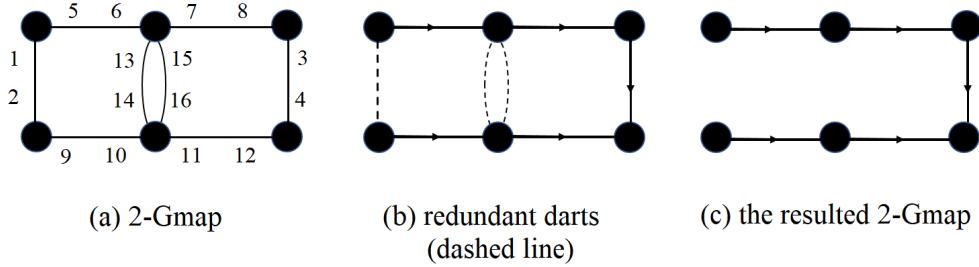
The paper presents a novel formalism to define redundant 1-cells in the 2-Gmap of a multi-labeled image. It defines a corresponding graph of the 2-Gmap and detects the redundant edges in the graph. The obtained formalism then translated into the 2-Gmap structure where the redundant 1-cells are detected. We proved that up to half of the whole 1-cells (edges) would be redundant in theory. Having sufficient processing elements by employing the set of independent edges, all the redundant edges (1-cells) are removed in constant complexity. The experiments show almost 48% of the 1-cells in the 2-Gmap are structurally redundant on average. By removing these redundant 1-cells the memory consumption is dramatically reduced. Moreover, we introduced an efficient encoding of involutions in the 2-Gmap where the two third of the involutions can be implicitly encoded. Finally, using the generalized map structure the proposed method can be extended to higher dimensional n-Gmaps.

## Acknowledgments

We acknowledge the Paul Scherrer Institut, Villigen, Switzerland for provision of beamtime at the TOMCAT beamline of the Swiss Light Source and would like to thank Dr. Goran Lovric for assistance. This work was supported by the Vienna Science and Technology Fund (WWTF), project LS19-013, and by the Austrian Science Fund (FWF), projects M2245 and P30275.

## References

- [1] Majid Banaeyan, Darshan Batavia, and Walter G. Kropatsch. Removing redundancies in binary images. In *International Conference on Intelligent Systems and Patterns Recognition (ISPR), Hammamet, Tunisia, March 24-25, 2022*, pages 221–233. Springer, 2022.
- [2] Majid Banaeyan and Kropatsch Walter G. Fast Labeled Spanning Tree in Binary Irregular Graph Pyramids. *Journal of Engineering Research and Sciences*, 1(10):69–78, 2022.
- [3] Majid Banaeyan and Walter G. Kropatsch. Pyramidal connected component labeling by irregular graph pyramid. In *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 1–5. IEEE, 2021.
- [4] Majid Banaeyan and Walter G. Kropatsch. Parallel  $\mathcal{O}(\log(n))$  computation of the adjacency of connected components. In *International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI), Paris, France, June 1-3, 2022*, pages 102–113. Springer, 2022.
- [5] Luc Brun and Walter Kropatsch. Introduction to combinatorial pyramids. In *Digital and Image Geometry*, pages 108–128. Springer, 2001.
- [6] Luc Brun and Walter G. Kropatsch. Hierarchical graph encodings. In Olivier L  zoray and Leo Grady, editors, *Image Processing and Analysis with Graphs: Theory and Practice*, pages 305–349. CRC Press, 2012.



d	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\alpha_1(d)$	5	9	8	12	1	13	15	3	2	14	16	4	6	10	7	11

(d) array of darts in memory

d	1	2	13	14	15	16	3	4	5	6	7	8	9	10	11	12
$\alpha_1(d)$	5	9	6	10	7	11	8	12	1	13	15	3	2	14	16	4

⏟
⏟

redundant darts
required darts to preserve 2-Gmap in (c)

(e) removing the array of redundant darts

Figure 6. Removing redundant darts in the canonical encoding

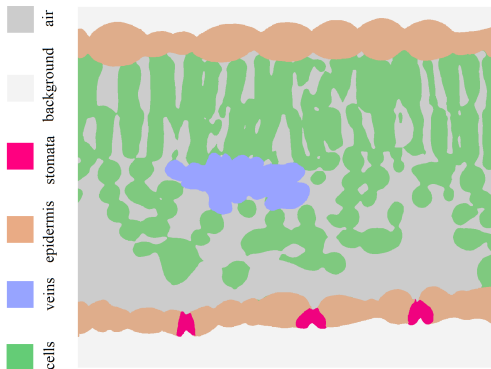


Figure 7. Multi-labeled input image

[7] Guillaume Damiand and Pascal Lienhardt. *Combinatorial maps: efficient data structures for computer graphics and image processing*. CRC Press, 2014.

[8] Brian A Davey and Hilary A Priestley. *Introduction to lattices and order*. Cambridge university press, 2002.

[9] Anand Deshpande and Manish Kumar. *Artificial intelligence for big data: complete guide to automating big data solu-*

*tions using artificial intelligence techniques*. Packt Publishing Ltd, 2018.

[10] Walter G. Kropatsch. Building irregular pyramids by dual graph contraction. *IEE-Proc. Vision, Image and Signal Processing*, Vol. 142(No. 6):pp. 366–374, 1995.

[11] Walter G Kropatsch, Yll Haxhimusa, and Pascal Lienhardt. Hierarchies relating topology and geometry. In *Cognitive Vision Systems*, pages 199–220. Springer, 2006.

[12] Walter G. Kropatsch, Yll Haxhimusa, Zygmunt Pizlo, and Georg Langs. Vision pyramids that do not grow too high. *Pattern Recognition Letters*, 26(3):319–337, Feb. 2005.

[13] Pascal Lienhardt. Topological models for boundary representation: a comparison with n-dimensional generalized maps. *Computer-aided design*, 23(1):59–82, 1991.

[14] Peter Meer. Stochastic image pyramids. *Computer Vision, Graphics, and Image Processing*, 45(3):269–294, 1989.

[15] Alessandro Negro. *Graph-powered machine learning*. Simon and Schuster, 2021.

[16] Azriel Rosenfeld and Mark Thurston. Edge and curve detection for visual scene analysis. *IEEE Transactions on computers*, 100(5):562–569, 1971.

[17] Fuensanta Torres and Walter G. Kropatsch. Canonical encoding of the combinatorial pyramid. In *Proceedings of*

*the 19th Computer Vision Winter Workshop*, pages 118–125, 2014.

- [18] R.J. Trudeau. *Introduction to Graph Theory*. Dover Books on Mathematics. Dover Pub., 1993.

# On the Regularising Levenberg-Marquardt Method for Blinn-Phong Photometric Stereo

Georg Radow and Michael Breuß

Chair of Applied Mathematics

Brandenburg University of Technology Cottbus-Senftenberg

{radow,breuss}@b-tu.de

## Abstract

*Photometric stereo refers to the process to compute the 3D shape of an object using information on illumination and reflectance from several input images from the same point of view. The most often used reflectance model is the Lambertian reflectance, however this does not include specular highlights in input images. In this paper we consider the arising non-linear optimisation problem when employing Blinn-Phong reflectance for modeling specular effects. To this end we focus on the regularising Levenberg-Marquardt scheme. We show how to derive an explicit bound that gives information on the convergence reliability of the method depending on given data, and we show how to gain experimental evidence of numerical correctness of the iteration by making use of the Scherzer condition. The theoretical investigations that are at the heart of this paper are supplemented by some tests with real-world imagery.*

## 1. Introduction

The photometric stereo (PS) problem is a fundamental task in computer vision [5]. The aim of PS is to infer the 3D shape of an object from a set of multiple images. Thereby the images depict an object from the same perspective, but the illumination direction changes throughout the images. An important information besides the illumination is the light reflectance of the object. The classic PS model [13, 14] is formulated in terms of Lambertian light reflectance. A Lambertian surface is characterised by diffuse reflectance and the independence of perceived shading from the viewing angle. The Lambertian set-up is certainly convenient for modeling, as it represents the most simple mathematical model for reflectance, and thus resulting formula and inverse problems are relatively simple. However, it is quite well known that in PS specular highlights [6] as well as non-Lambertian diffuse effects [7] may have an important impact on 3D reconstruction.

Let us also comment on some other basic characteristics of PS. Depending on the knowledge on the lighting, one discerns between calibrated and uncalibrated PS. In this work we consider only the calibrated case, where lighting directions and intensities are known. Furthermore, the final goal of PS is to obtain a depth map, such that for each relevant image pixel three-dimensional information of the depicted object is obtained. While some approaches tackle this problem directly in terms of depth values [8], the more common strategy is to divide depth computation into two sub-problems. In doing so at first a map of normal vectors is computed, from which the (relative) depth is obtained in a second step. See for instance [11] for a survey on surface normal integration. In this paper we only consider the first of the latter tasks, that is to find the normal vectors. Another aspect is sometimes the projection performed by the camera during image acquisition, often leading to orthographic or perspective models, respectively. In this work we address effectively both settings.

**Our contribution.** In this paper, we consider some theoretical aspects of practical value in the optimisation of PS when using Blinn-Phong reflectance. Here we extend in several ways upon previous work; let us especially refer to [6], where the Blinn-Phong model is employed in a similar way as here. Thereby, we consider to include the potentially most important specular parameter, the so-called shininess, as an unknown in the optimisation, which is in contrast to [6] and many other works in the field. The approximate solution of the non-linear optimisation problem arising pixel-wise is performed by the regularising Levenberg-Marquardt method, see especially [2]. As this is an iterative method, it is important to assess the influence of initialisation on the convergence and to give a rigorous bound as a stopping criterion. Furthermore as the problem is non-linear, one can observe in practical examples, that it may be difficult to minimise the underlying residual. To address this issue we investigate the use of a coarse-to-fine (CTF) scheme as well as an initialisation obtained through classical PS. We show how to explore Scherzer's criterion [3], which appeared in [4] for the first

time. This criterion is considered for theoretical purposes within the construction of the method, in order to assess the convergence property in our PS problem experimentally.

## 2. Classical Photometric Stereo

Let us reiterate the classic PS approach of Woodham [13, 14]. Given is a set of  $m \geq 3$  images  $(\mathcal{I}_1, \dots, \mathcal{I}_m)^\top =: \mathcal{I}$ , so that  $\mathcal{I} : \Omega \rightarrow \mathbb{R}^m$ , along with the corresponding lighting directions  $L_k \in \mathbb{R}^3$  with  $\|L_k\| = 1$  for  $k = 1, \dots, m$ , with associated intensities  $l_k \geq 0$ . Throughout the paper  $\|\cdot\|$  denotes the Euclidean norm or the induced spectral norm. The object to be reconstructed is depicted usually as a non-rectangular domain  $\Omega \in \mathbb{R}^2$ , which is embedded in the image domain.

The surface normal vectors  $\mathcal{N} : \Omega \rightarrow \mathbb{R}^3$  with  $\|\mathcal{N}(x, y)\| = 1$  for all  $(x, y)^\top \in \mathbb{R}^2$  and the albedo  $\rho^d : \Omega \rightarrow \mathbb{R}$  are fitted through a least squares approach, by minimising

$$\iint_{\Omega} \|\mathcal{R}^L(x, y) - \mathcal{I}(x, y)\|^2 dx dy, \quad (1)$$

with reflectance function  $\mathcal{R}^L := (\mathcal{R}_1^L, \dots, \mathcal{R}_m^L)^\top$ , consisting of components

$$\mathcal{R}_k^L := \rho^d l_k L_k^\top \mathcal{N}, \quad k = 1, \dots, m. \quad (2)$$

In practice this boils down to finding a local solution  $N \in \mathbb{R}^3$  at every sample location  $(x, y)^\top$  for the problem

$$\min_N \|LN - I\|^2, \quad L := \begin{pmatrix} l_1 L_1^\top \\ \vdots \\ l_m L_m^\top \end{pmatrix}, \quad I := \mathcal{I}(x, y). \quad (3)$$

This, in turn, leads to the computation of the normal vectors and, as a byproduct, the albedo according to

$$N = (L^\top L)^{-1} L^\top I, \quad (4)$$

$$\rho^d(x, y) = \|N\|, \quad \mathcal{N}(x, y) = N/\|N\|. \quad (5)$$

## 3. Blinn-Phong Photometric Stereo

In the general least squares approach Eq. (1), we can modify the reflectance function to account for non-Lambertian effects. To this end we investigate the Blinn-Phong (BP) model [1, 9], which has the form  $\mathcal{R}^{\text{BP}} := (\mathcal{R}_1^{\text{BP}}, \dots, \mathcal{R}_m^{\text{BP}})^\top$  with components

$$\mathcal{R}_k^{\text{BP}} := \rho^d l_k L_k^\top \mathcal{N} + \rho^s h_k \max\{0, \mathcal{H}_k^\top \mathcal{N}\}^\alpha, \quad (6)$$

$k = 1, \dots, m$ . We observe by (6) that in the BP model, diffuse reflection as in (2) is supplemented by a specular reflection term. Here  $\rho^s : \Omega \rightarrow \mathbb{R}$  denotes the specular albedo. Another material parameter is the specular sharpness

or shininess  $\alpha : \Omega \rightarrow \mathbb{R}$ . The halfway vectors  $\mathcal{H}_k : \Omega \rightarrow \mathbb{R}^3$  depend on the viewing directions  $\mathcal{V} : \Omega \rightarrow \mathbb{R}^3$  and are computed for  $k = 1, \dots, m$  as

$$\mathcal{H}_k(x, y) := H_k / \|H_k\|, \quad H_k := L_k + \mathcal{V}(x, y). \quad (7)$$

Making use of focal length  $f$ , the viewing directions  $\mathcal{V}^\perp$  and  $\mathcal{V}^<$  in the orthographic and perspective setting respectively are

$$\mathcal{V}^\perp = (0, 0, 1)^\top, \quad \mathcal{V}^<(x, y) = (x, y, f)^\top. \quad (8)$$

We reinterpret  $l_k$  as diffuse intensity of the light source and denote  $h_k \geq 0$  as specular intensity. To ensure that image intensities are only increased due to diffuse and specular terms, it is reasonable to enforce  $\rho^d, \rho^s \geq 0$ . Furthermore  $\rho^d, \rho^s \leq 1$  ensures that at most as much image intensity is added as light intensity is supplied by each light source. Finally, it is reasonable to enforce  $\alpha > 1$  to actually produce specular highlights through the specular term.

The BP model was originally proposed for computer graphics. It is not based on physical laws, but it enables to create plausible images with a still simple model compared to other possible approaches. Despite its simplicity, for use in inverse problems in computer vision, the non-linearities in Eq. (6) may pose considerable hurdles.

Let us now discuss the modeling of the components in Eq. (6) along with a few adaptations we employ. First we turn our attention to the normal vectors  $\mathcal{N}$ . One may model them through derivatives of the depth or its logarithm. In this approach we may parametrise them at a specific location through depth derivatives  $p, q$  as

$$\mathcal{N}(x, y) = \frac{N(p, q)}{\|N(p, q)\|}. \quad (9)$$

However the step of obtaining a normal vector of length 1 in Eq. (9) adds another layer of non-linearity to the model. In numerical experiments we found this approach to be not very reliable. Therefore we opt for an approach in analogy to classical PS. In Eq. (6) we replace  $\rho^d \mathcal{N} = N$  introducing the auxiliary variable  $r = \rho^s / (\rho^d)^\alpha$ . By furthermore replacing  $\alpha = 1 + \exp(a)$  we ensure that  $\mathcal{R}^{\text{BP}}$  has continuous first derivatives. Eq. (6) then takes the form

$$\mathcal{R}_k^{\text{BP}}(N, r, a) = l_k L_k^\top N + r h_k \max\{0, \mathcal{H}_k^\top N\}^{1+\exp(a)}, \quad (10)$$

with  $r, a \in \mathbb{R}$  and  $N \in \mathbb{R}^3$ .

## 4. On the Optimisation Strategy

With BP reflectance, we have to solve a non-linear least squares problem, to which end we utilise the regularising Levenberg-Marquardt (RLM) scheme [2, 3]. Writing the

underlying task in standard notation, with this algorithm one may aim to find a solution  $\vec{x}$  of the problem

$$F(\vec{x}) = \vec{y}, \quad F: \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad (11)$$

with a known differentiable function  $F$ . Let us note that the description and discussion of the RLM algorithm in [3] is in a more general setting. For simplicity we only give an overview of the algorithm based on finite dimensional spaces, as is fitting for the problem at hand.

It is furthermore assumed that the original data  $\vec{y}$  is not known, but with some  $\delta > 0$  an estimate is required on how good the given data  $\vec{y}^\delta$  approximates the original data, according to

$$\|\vec{y}^\delta - \vec{y}\| \leq \delta. \quad (12)$$

Then with some starting point  $\vec{x}_0$  the iterative rule takes the form

$$\vec{x}_{k+1} = \vec{x}_k + (F'(\vec{x}_k)^\top F'(\vec{x}_k) + \alpha_k I_n)^{-1} F'(\vec{x}_k)^\top (\vec{y}^\delta - F(\vec{x}_k)) \quad (13)$$

with Jacobian matrix  $F'$ ,  $n \times n$ -dimensional identity matrix  $I_n$  and a regularisation weight  $\alpha_k > 0$  such that with a preassigned  $\rho \in (0, 1)$  the new iterate  $\vec{x}_{k+1}$  fulfils

$$\|\vec{y}^\delta - F(\vec{x}_k) - F'(\vec{x}_k)(\vec{x}_{k+1} - \vec{x}_k)\| = \rho \|\vec{y}^\delta - F(\vec{x}_k)\|. \quad (14)$$

The *stopping criterion* of the RLM scheme depends explicitly on the noise level  $\delta$  in the given data. To stop at an iterate  $\vec{x}_k$ , it has to fulfil

$$\|\vec{y}^\delta - F(\vec{x}_k)\| \leq \tau \delta, \quad (15)$$

with a preassigned  $\tau > 2$ , fulfilling  $\rho\tau > 1$ . For numerical experiments we set  $\rho = 0.5$ ,  $\tau = 2.5$ , following [3].

The discussion of the RLM scheme in [3] relies on the strong Scherzer condition [4]. For the Jacobian matrices at two points  $\vec{x}_1, \vec{x}_2 \in \mathbb{R}^n$  there exists a matrix  $R = R(\vec{x}_1, \vec{x}_2)$  such that  $F'(\vec{x}_1) = RF'(\vec{x}_2)$  and

$$\|R - I_m\| \leq C^R \|\vec{x}_1 - \vec{x}_2\| \quad (16)$$

with some  $C^R > 0$ , which is constant for all  $\vec{x}_1, \vec{x}_2 \in \mathbb{R}^n$ . This condition imposes a certain regularity of the Jacobian matrix  $F'$ . In this context we are interested in a local approximation of  $C^R$ . For two consecutive iterations  $\vec{x}_k, \vec{x}_{k+1}$  we estimate  $R$  as a solution of  $F'(\vec{x}_k) = R(\vec{x}_k, \vec{x}_{k+1})F'(\vec{x}_{k+1})$  with minimal norm. Then we can locally approximate the constant in Eq. (16) as

$$C_k^{R, \text{loc}} = \frac{\|R(\vec{x}_k, \vec{x}_{k+1}) - I_m\|}{\|\vec{x}_k - \vec{x}_{k+1}\|}. \quad (17)$$

Since  $F$  in Eq. (11) is nonlinear, we employ a CTF framework. In doing so the data is scaled to a coarser scale, *i.e.*

to a lower resolution. The obtained result is then used as initialisation on the next finer scale, until we arrive at the original resolution.

Let us focus on the assumption Eq. (12). The noise level  $\delta$  governs the stopping criterion of the RLM scheme. If Eq. (12) is not fulfilled then the iterates may actually diverge.

At this point we make the assumption that our data  $\mathcal{I}(x, y)$  is a realisation of the BP model corrupted by additive white Gaussian noise, *i.e.* it can be modelled as

$$\mathcal{I}(x, y) = \mathcal{R}(x, y) + \varepsilon(x, y), \quad \text{for } (x, y)^\top \in \Omega. \quad (18)$$

Here  $\varepsilon(x, y)$  is a realisation of a multivariate normal distribution, such that the  $m$  components are independent and identically distributed (i.i.d.) with mean zero and standard deviation  $\sigma > 0$ , the corresponding density function is

$$f(X) = \frac{1}{\sqrt{2\pi}^m \sigma^m} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^m X_i^2\right), \quad (19)$$

*cf.* [10]. The probability that Eq. (12) holds can be computed with the following result. The proof, which is technical but straightforward, is included for the readers convenience. The following result is also related to the Chi distribution.

**Proposition 1.** *Let  $m \in \mathbb{N}$ ,  $\delta > 0$  and let  $\varepsilon$  be a realisation of an  $m$ -dimensional multivariate normal distribution with mean zero, standard deviation  $\sigma > 0$  and density Eq. (19). The probability of  $P := P(\|\varepsilon\| \leq \delta | \sigma, m)$  can be computed as follows:*

(i) *If  $m$  is even, then*

$$P = 1 - \exp\left(-\frac{\delta^2}{2\sigma^2}\right) \sum_{i=0}^{\frac{m}{2}-1} \left(\frac{\delta^2}{2\sigma^2}\right)^i \frac{1}{i!}. \quad (20)$$

(ii) *If  $m$  is odd, then*

$$P = \sqrt{\frac{2}{\pi}} \left(\frac{1}{\sigma} \int_0^\delta \exp\left(-\frac{r^2}{2\sigma^2}\right) dr - \exp\left(-\frac{\delta^2}{2\sigma^2}\right) \sum_{i=1}^{\frac{m-1}{2}} \left(\left(\frac{\delta}{\sigma}\right)^{m-2i} \prod_{j=1}^{\frac{m+1}{2}-i} \left(\frac{1}{2j-1}\right)\right)\right). \quad (21)$$

*Proof.* For any continuous probability density  $f$  we have

$$P = P(\|\varepsilon\| \leq \delta | \sigma, m) = \int_{\|X\| \leq \delta} f(X) dX. \quad (22)$$

Since the density function in Eq. (19) is radially symmetric, this simplifies to

$$P = \int_0^\delta O_m(r) f(r, 0, \dots, 0) dr, \quad (23)$$



where

$$O_m(r) = 2r^{m-1} \frac{\pi^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)} \quad (24)$$

denotes the surface area of a sphere with radius  $r$  around the origin in  $\mathbb{R}^m$ .  $\Gamma$  denotes the gamma function. Inserting Eq. (19), we write

$$P = \frac{2^{1-\frac{m}{2}}}{\sigma^m \Gamma\left(\frac{m}{2}\right)} \int_0^\delta r^{m-1} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr. \quad (25)$$

Since  $\int r \exp(r^2/(2a)) dr = a \exp(r^2/(2a)) + c$ , for  $m > 2$  the integral in Eq. (25) can be simplified by partial integration, *i.e.*

$$\begin{aligned} & \int_0^\delta r^{m-2} \cdot r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= -\sigma^2 \left[ r^{m-2} \exp\left(-\frac{r^2}{2\sigma^2}\right) \right]_{r=0}^\delta \\ &+ \sigma^2(m-2) \int_0^\delta r^{m-4} \cdot r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr. \end{aligned} \quad (26)$$

We now consider the two cases of  $m$  being even or odd.

Let  $m \in \mathbb{N}$  be even. Then repeated partial integration of the integral Eq. (25) leads to

$$\begin{aligned} & \int_0^\delta r^{m-1} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= -\sum_{i=1}^{\frac{m}{2}-1} \sigma^{2i} \prod_{j=1}^{i-1} (m-2j) \left[ r^{m-2i} \exp\left(-\frac{r^2}{2\sigma^2}\right) \right]_{r=0}^\delta \\ &+ \sigma^{m-2} \prod_{j=1}^{\frac{m}{2}-1} (m-2j) \int_0^\delta r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= -\sum_{i=1}^{\frac{m}{2}} \sigma^{2i} \prod_{j=1}^{i-1} (m-2j) \left[ r^{m-2i} \exp\left(-\frac{r^2}{2\sigma^2}\right) \right]_{r=0}^\delta \\ &= -\sum_{i=1}^{\frac{m}{2}} \sigma^{2i} \frac{2^{i-1} \left(\frac{m}{2}-1\right)!}{\left(\frac{m}{2}-i\right)!} \left[ r^{m-2i} \exp\left(-\frac{r^2}{2\sigma^2}\right) \right]_{r=0}^\delta \\ &= \sigma^m 2^{\frac{m}{2}-1} \left(\frac{m}{2}-1\right)! \\ &- \sum_{i=1}^{\frac{m}{2}} \sigma^{2i} \frac{2^{i-1} \left(\frac{m}{2}-1\right)!}{\left(\frac{m}{2}-i\right)!} \delta^{m-2i} \exp\left(-\frac{\delta^2}{2\sigma^2}\right). \end{aligned} \quad (27)$$

This formula can easily be verified for  $m = 2$ , as in this case the initial integral simplifies to the form  $\int r \exp(r^2/(2a)) dr$ . Inserting Eq. (27) and  $\Gamma(m/2) = (m/2 - 1)!$  into Eq. (25), we obtain after an index shift Eq. (20).

Now let  $m \in \mathbb{N}$  be odd. Again we use repeated partial integration on the integral in Eq. (25), until we arrive at

$$\begin{aligned} & \int_0^\delta r^{m-1} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= \sigma^{m-1} \prod_{j=1}^{\frac{m-1}{2}} (2j-1) \int_0^\delta \exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &- \sum_{i=1}^{\frac{m-1}{2}} \sigma^{2i} \frac{\prod_{j=1}^{\frac{m-1}{2}} (2j-1)}{\prod_{j=1}^{\frac{m+1}{2}-i} (2j-1)} \delta^{m-2i} \exp\left(-\frac{\delta^2}{2\sigma^2}\right). \end{aligned} \quad (28)$$

Plugging this together with

$$\begin{aligned} \Gamma\left(\frac{m}{2}\right) &= \Gamma\left(\frac{m-1}{2} + \frac{1}{2}\right) = \frac{(m-1)! \sqrt{\pi}}{\left(\frac{m-1}{2}\right)! 2^{m-1}} \\ &= \frac{\prod_{j=1}^{\frac{m-1}{2}} ((2j)(2j-1)) \sqrt{\pi}}{2^{\frac{m-1}{2}} \prod_{j=1}^{\frac{m-1}{2}} (2j)} = \frac{\sqrt{\pi}}{2^{\frac{m-1}{2}}} \prod_{j=1}^{\frac{m-1}{2}} (2j-1) \end{aligned} \quad (29)$$

into Eq. (25) leads to Eq. (21).  $\square$

## 5. Experiments

Since we focus on the computed vector fields of surface normals, it appears adequate to employ colour coding of surface normals for visual assessment, *cf.* Figure 1. For quantitative evaluation we consider here the standard AAE, where the averaging is performed over the object domain. Let us note that we use the result obtained through classical PS as an initialisation for the BP model. Throughout the experiments we computed  $\delta$  according to Proposition 1, such that Eq. (12) is fulfilled with a probability of 95%. We observed that the choice of this confidence level is not critical for the outcome of our experiments.

**Synthetic Test Example.** As a synthetic experiment for our investigations we consider the *sphere* example, see Figure 1. Let us note that we consider an orthographic setting for all the *sphere* experiments. As we observe in Figure 1, in this experiment the developed computational model and set-up enables to obtain a nearly perfect result. For optimisation we employed in total 5 input images, of which we show here just one example. For comparison, we give here the corresponding result obtained by Lambertian PS applied to analogous input images where we filtered the specular highlights by the subspace technique proposed in [15], which is supposed to make the input nearly Lambertian. As is confirmed here visually as well as quantitatively, it appears favorable (at least in this example) to explore an explicit modeling like with the proposed BP framework.

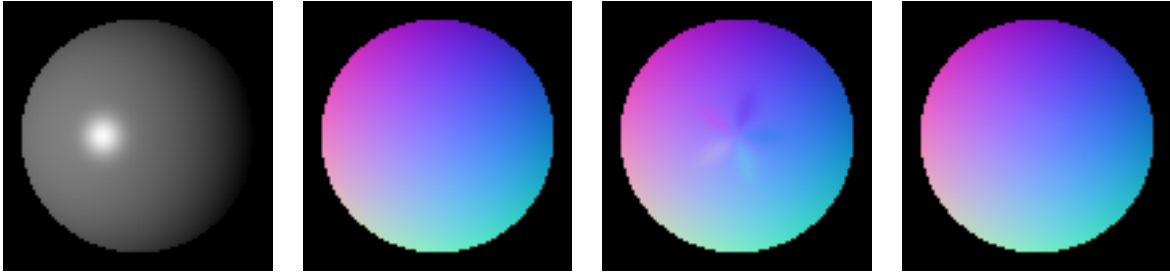


Figure 1. (left-to-right:) one of the input images of the *sphere* rendered using the BP model; colour coded vector field of ground truth normal vectors; classical PS with preprocessing [15], average angular error (AAE) 1.02; developed BP framework with CTF, AAE 0.37

Let us note that in fact this test example may not be too easy, as can be observed by the results obtained by preprocessing and Lambertian PS. The reason is that the specular highlights in the input are not perfectly distributed over the sphere and may result in distortions if not being accounted for sufficiently accurate in the model.

**Evaluation of Scherzer’s Condition.** As discussed in Sec. 4, between two iterates of the RLM scheme we observe the local approximation  $C_k^{R,loc}$  of the constant in Eq. (16) according to Eq. (17). As the Scherzer condition is an important assumption for the results in [3], we opt to add a break condition, where the algorithm stop if the estimate grows too large. In practice the algorithm is halted if we observe an iterate with  $C_k^{R,loc} \geq 2000$ . As can be seen in Figs. 2 and 3 this is usually the case at locations where specular highlights may occur, as the angle between halfway vectors and surface normals becomes small. One may interpret this result in the way, that the energy that is minimised features at highlights many small variations that makes it difficult to obtain a reliable local minimum.

We evaluated the restarting of the RLM scheme with a larger parameter  $\rho$  in Eq. (14), if it stopped before an iterate fulfils Eq. (15). This may lead to a smaller trust region and to a more stable behaviour of the algorithm. However we did in general not observe a significant increase in quality. The results displayed here were thus computed without restarting the RLM scheme, giving an account of the unstabilised version of the method.

**Real World Test Example.** In order to assess the properties and usefulness of the developed numerical BP framework, we exploit here a selected variety of examples taken from the *DiLiGent* data set [12] which gives an account of photographed real-world objects with different reflectance properties. Here we do not employ a CTF scheme, as we rely on the initialisation obtained with classical PS. Let us note that the underlying model is now (in practice, weakly) perspective.

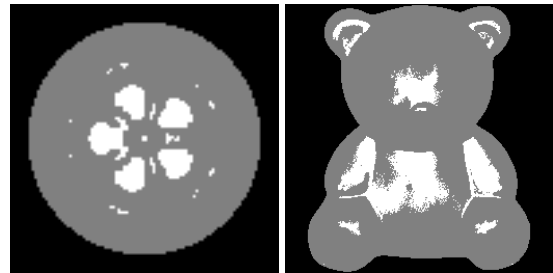


Figure 2. Algorithmic behaviour in the *sphere* experiment (left) and an example from the *DiLiGent* data set [12] (right). White depicts the locations where the RLM scheme stopped due to the  $C_k^{R,loc} \geq 2000$  criterion.

As can be visually assessed by means of Figure 3, the proposed model along with its adaptations performs very reasonably but in some details not perfect, depending on the actual example. For clarifying thereby the zones of influence of the specular terms we depict masks showing the object parts where the BP model gives an effective contribution. When taking into account the properties of the considered examples, it appears especially that the broad specularities as appearing in the input (teddy bear, goblet) may result in a certain inaccuracy. In turn, when highlights appear but are not too strong (cat, tea pot), results are quite convincing, given that the underlying reflectance in these cases is supposed to be non-linear in the diffuse reflectance as the underlying material is rough. In the tested real world setting from *DiLiGent* the results are overall of similar quality to the preprocessed Lambertian method. Therefore we conjecture that our numerical BP framework appears to be especially suited for dealing with objects with not too strong highlights, being at the same time able to tackle a certain range of diffuse reflectance of rough materials.

## 6. Conclusion

We discussed the BP reflectance in the context of PS. The augmentation of classical PS with this reflectance model is

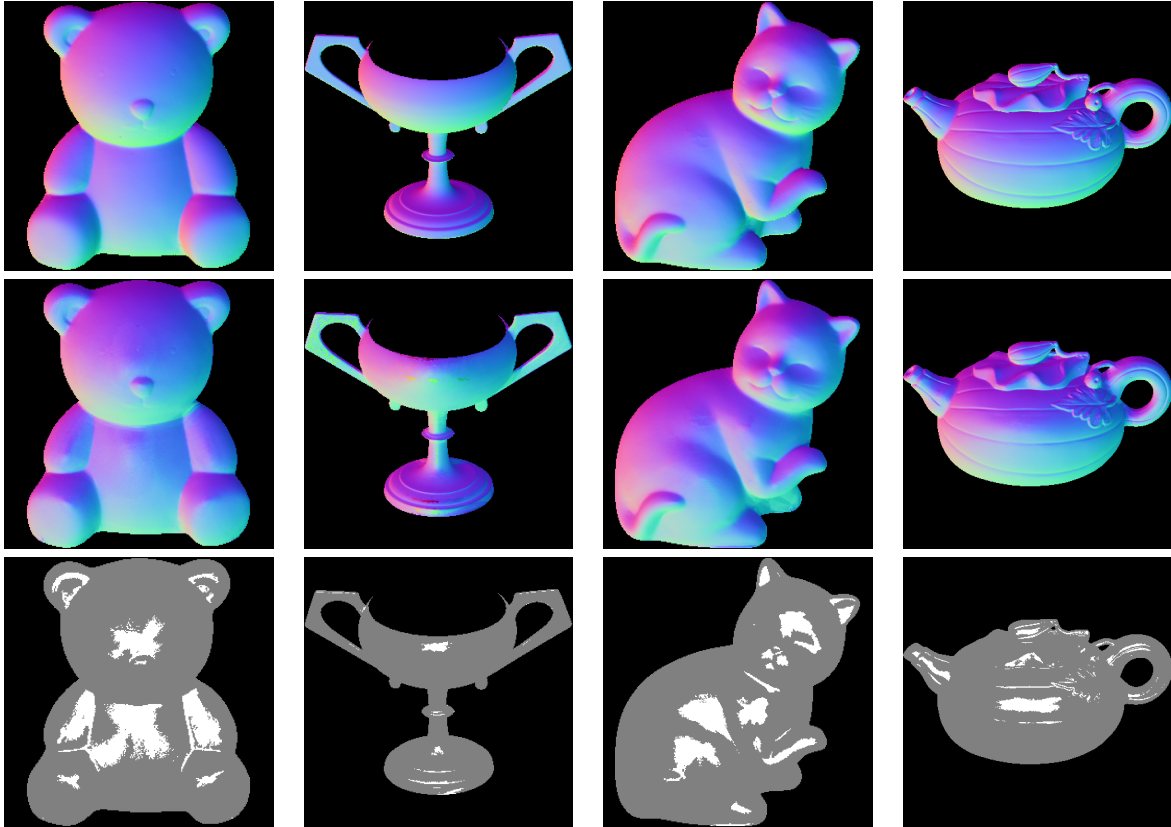


Figure 3. (left-to-right:) Examples from *Diligent* data sets. (top-to-bottom:) Visualisation of ground truth normals; normal fields based on BP (where we note the effect of the not satisfied Scherzer condition at some highlights at the goblet); mask based on half directions. White depicts locations where the maximum of the cosines between halfway vectors and the normal vector obtained with classical PS is  $\geq 0.99$ .

straightforward, but solving the arising optimisation problem is less so. This task can be tackled with the RLM scheme, which leads to satisfactory results.

The findings for the implementation of the RLM scheme may be translated to other problems, since the assumption that the data follows a normal distribution is very common. The application of the BP model to more complex data sets poses considerable hurdles, which may be adressed in future work.

## References

- [1] J. F. Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques - SIGGRAPH '77*. ACM Press, 1977.
- [2] M. Hanke. A regularizing Levenberg - Marquardt scheme, with applications to inverse groundwater filtration problems. *Inverse Problems*, 13(1):79–95, 1997.
- [3] M. Hanke. The regularizing Levenberg-Marquardt scheme is of optimal order. *Journal of Integral Equations and Applications*, 22(2):259–283, 2010.
- [4] M. Hanke, A. Neubauer, and O. Scherzer. A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numerische Mathematik*, 72(1):21–37, 1995.
- [5] B. K. P. Horn. *Robot Vision*. MIT Electrical Engineering and Computer Science. MIT Press, 1986.
- [6] M. Khanian, A. S. Boroujerdi, and M. Breuß. Photometric stereo for strong specular highlights. *Computational Visual Media*, 4(1):83–102, 2018.
- [7] G. McGunnigle, J. Dong, and X. Wang. Photometric stereo applied to diffuse surfaces that violate lambert’s law. *Journal of the Optical Society of America A*, 29(4):627, mar 2012.
- [8] R. Mecca, A. Tankus, A. Wetzler, and A. M. Bruckstein. A direct differential approach to photometric stereo with perspective viewing. *SIAM Journal on Imaging Sciences*, 7(2):579–612, 2014.
- [9] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.
- [10] S. J. D. Prince. *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012.
- [11] Y. Quéau, J.-D. Durou, and J.-F. Aujol. Normal integration: A survey. *Journal of Mathematical Imaging and Vision*, 60(4):576–593, 2017.

- [12] B. Shi, Z. Mo, Z. Wu, D. Duan, S. K. Yeung, and P. Tan. A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2018.
- [13] R. J. Woodham. Photometric stereo: A reflectance map technique for determining surface orientation from image intensity. In Ramakant Nevatia, editor, *Image Understanding Systems and Industrial Applications*, volume 155 of *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, pages 136–143. SPIE, 1978.
- [14] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):134–144, 1980.
- [15] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Asian Conference on Computer Vision (ACCV)*, volume 6494 of *Lecture Notes in Computer Science*, pages 703–717. Springer Berlin Heidelberg, 2010.

# Image Forgery Detection and Localization Using a Fully Convolutional Network

David Fischinger, David Schreiber and Martin Boyer  
Austrian Institute of Technology - AIT  
{forename}.{surname}@ait.ac.at

## Abstract

To fight the growing problem of fake news – and specifically image manipulation – we propose a simple, yet efficient neural network architecture for detecting and localizing various image forgeries on a pixel-level. Robust features for forgery detection and localization were learned and the trained model performs well, even on heavily down-scaled images, but without the excessive processing time of competitive approaches based on image decomposition and merging of the fragmental results. We provide detailed explanations regarding the creation of our training dataset comprising 1.9 million images. Finally, we compare the proposed solution against several state-of-the-art methods on four public benchmark datasets in order to demonstrate its superior performance.

## 1. Introduction

“Fake News” are a growing problem of our society. Technological progress makes it easier and faster to produce high quality forgeries of digital media material such as audio, video and images. The impact ranges from satirical memes to orchestrated political Fake News campaigns aiming to influence public opinion – and at the same time raising the hard question where to draw a line between fighting Fake News and the fundamental right of free speech. In this paper, we present a new approach for identifying forged regions in images, thereby enabling institutions such as media organizations and interested citizens to get a better indication of whether specific images may have been manipulated.

During the last decade, various approaches for detecting the main categories of image forgery were proposed: copy-move [9] splicing [11], inpainting [8] and further specific filtering, subsumed as enhancement [20]. However, these approaches frequently focus on specific features of the respective manipulation type. In recent years, more general approaches for multiple manipulation types were developed, such as [24] and [23]. Each of them promotes sophisticated and problem-specific concepts, like modeling known and unknown noise on images that result from transmis-

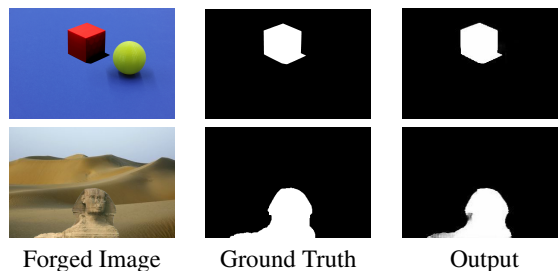


Table 1. Results of our model for image forgery detection and localization. Example images are taken from the CASIA [4] and the NIST [15] datasets.

sion to Online Social Networks (OSNs). In this paper, we present an image forgery detector which outperforms state-of-the-art approaches with a quite simple and general deep learning network architecture and a carefully constructed training dataset. To be more specific, our major contributions are as follows:

- We propose a deep learning network architecture for the task of image forgery detection and localization, capable to learn relevant features for composed image manipulations.
- We present a model that outperforms current state-of-the-art (SOTA) approaches on four public benchmark datasets.
- We present a processing time comparison with a SOTA approach showing a significant time saving, especially for larger images.
- We give a detailed description of our training dataset, as well as instructions on how to generate such a dataset.

## 2. Related Work

Many methods of detecting and localizing image forgery have been published (see, for example, the review of [21] and references therein), in order to ensure visual information authenticity. Some of these forensic techniques are designed to detect specific forms of tampering, such as splic-

ing [11], copy-move [12, 16, 22, 25–27], and inpainting [8]. Unfortunately, these forensic approaches can only be applied to detect specific tampering manipulations.

In recent years, deep learning-based methods were developed to address the problem of detecting general (compound) types of forgeries. In [28], a two-stream Faster R-CNN network is trained end-to-end to detect the tampered regions in a manipulated image. One of the two streams is an RGB stream whose purpose is to extract features from the RGB image input. The other one is a noise stream that leverages the noise features extracted in order to discover noise inconsistencies between authentic and tampered regions. Notably, [24] proposes a unified deep neural architecture called ManTra-Net, which is an end-to-end network that performs both detection and localization without extra preprocessing and postprocessing. ManTra-Net is a fully convolutional network which can handle images of arbitrary sizes and many known – and even unknown – forgery types. Furthermore, the authors design a self-supervised learning task to learn robust image manipulation features, formulate the forgery localization problem as a local anomaly detection problem, and propose a long short-term memory (LSTM) solution to assess local anomalies.

In [13], a CNN-based image forgery detection framework is proposed which makes decisions based on full-resolution information gathered from the entire image, without the need for preliminary image resizing. The framework is trainable end-to-end with limited memory resources and weak (image-level) supervision, thus allowing for the joint optimization of all parameters. The work of [29] addresses the issue of tampering localization by focusing on the detection of commonly used editing tools and operations in Photoshop. A fully convolutional encoder-decoder architecture is designed, as well as a training data generation strategy by resorting to Photoshop scripting.

The widespread availability of online social networks (OSNs), e.g., Twitter, Facebook, Whatsapp, etc., makes them the dominant channels for transmitting forged images. However, almost all OSNs manipulate the uploaded images in a lossy fashion (including format conversion, resizing, enhancement filtering and JPEG compression). The noise introduced by these lossy operations could severely affect the effectiveness of forensic methods. In a recent paper [23], the problem of OSN-shared image forgeries is tackled by employing a dedicated training scheme. A baseline detector is presented, which is based on a modified U-Net [17] as the backbone architecture. Next, an analysis of the noise introduced by OSNs is conducted, and the noise is decoupled into two parts, i.e., predictable noise and unseen noise. These are then modelled separately and the modelled noise is further incorporated into the training framework.

*Outline:* The rest of this paper is structured as follows: Section 3 describes in detail how the datasets for training

and validation were generated. In section 4, we present different models we have created, evaluate them on benchmark datasets, and describe the architecture of the best performing model in detail. In section 5, our proposed network is evaluated and compared to state-of-the-art methods. Final remarks are made in section 6.

### 3. Datasets

Currently, there are no sufficiently large training datasets publicly available for the task of image forgery detection. In the following, a detailed description is provided, of how our training dataset, which comprises 1.9 million manipulated images, was created.

#### 3.1. Training and Validation Datasets

As a source of pristine and donor images we facilitated the MS-Coco [10] 2017 training dataset containing 118K images. This public and widely used dataset encompasses a wide range of images. Our training dataset includes 4 major types of image manipulation: splicing, copy-move, removal and enhancement. The overall process for training data generation was as follows:

1. Select Pristine Image:

A pristine image  $\mathcal{I}_P$  from MS-COCO 2017 was selected randomly. For the few images with width  $W$  or height  $H$  smaller than 224 pixels, the image was resized to the size  $(\max(W, 224), \max(H, 224))$ . For 50% of the images  $\mathcal{I}_P$  in the training dataset, a proportion-preserving downscaling was executed. This avoided extracting only small portions of bigger images (like a monochrome patch depicting a part of the sky from the original image). This scaling for an image  $\mathcal{I}_P$  with size  $(W, H)$  to  $(W_{new}, H_{new})$  was done as follows:

$$\begin{aligned} W_{new} &= \max(\lfloor \frac{224 \cdot W}{\min(W, H)} \rfloor, 224) \\ H_{new} &= \max(\lfloor \frac{224 \cdot H}{\min(W, H)} \rfloor, 224) \\ \mathcal{I}_P &= \mathcal{I}_P.resize((W_{new}, H_{new})) \end{aligned} \quad (1)$$

Next, a patch of size  $(224, 224)$  pixels is randomly chosen from the image  $\mathcal{I}_P$  and used as a pristine image patch  $\mathcal{P}$ .

2. Select Donor Image:

A donor image  $\mathcal{I}_D$  from MS-COCO was selected. For the splicing operation, a random image other than the pristine image  $\mathcal{I}_P$  was selected. For the copy-move, removal and enhancement manipulations, the same pristine image was selected as a donor image






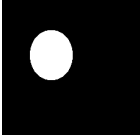








Forgery Type	Manipulation Mask Shape
Copy-Move 	triangle 
Enhance 	rounded rectangle 
Enhance 	ellipse 
Removal 	poligon 5 vertices 
Copy-Move 	ellipse + 4V polygon 
Copy-Move 	superpixel segmentation 
Splicing 	person segmentation 

Table 2. Training and Validation data: showing forged images for major manipulation types and related manipulation mask shapes

( $\mathcal{I}_D = \mathcal{I}_P$ ). Then, a donor patch  $\mathcal{D}$  of size  $(224, 224)$  was randomly cropped from  $\mathcal{I}_D$ . For enhancement and removal (inpainting) manipulations, the donor patch  $\mathcal{D}$  and the pristine patch  $\mathcal{P}$  share the same location in  $\mathcal{I}_D = \mathcal{I}_P$ .

### 3. Preprocess Donor Image Patch $\mathcal{D}$ :

Table 4 shows which preprocessing steps may be

applied to the donor image patch  $\mathcal{D}$  for each manipulation type. **Resample** rescales the height and the width image dimension independently by 70 to 130 percent. The resulting image has at least the size  $(224, 224)$ . The preprocessing step **Flip** flips the donor image horizontally with a likelihood of 50%, while **Rotate** rotates the image by either 90, 180 or 270 degrees with a likelihood factor controlled by a parameter (for the generated dataset, 30% of the donor images were rotated). **Blur** is blurring the donor image with a likelihood of 50%. In case the blurring filter is applied, either `ImageFilter.BoxBlur` or `ImageFilter.GaussianBlur` from the Python package `PIL` are used with equal probabilities. The blur radius is set randomly between 1 and 7 pixels. **Contrast** uses one of the `ImageFilters` `EDGE_ENHANCE`, `EDGE_ENHANCE_MORE`, `SHARPEN`, `UnsharpMask` or `ImageEnhance.Contrast` from the Python package `PIL`. **Noise** adds Gaussian noise with mean and standard deviation  $(\mu, \sigma) = (0, 12)$  with likelihood of 1 out of 3. The **Brightness** is changed with probability of 50% by a factor uniformly chosen from the range  $[0.5-1.5]$ . With 0.5 probability, a **JPEG-Compression** with quality factor  $10x$  for  $x \in [1, 2, 3, 4, 5, 6, 7]$  is employed. In case the manipulation type is **Removal**, an inpainting filter from OpenCV [2] is applied (either `cv2.INPAINT_TELEA` or `cv2.INPAINT_NS`) on the manipulation mask defined in the next step.

In case the chosen manipulation type is **Enhance** and none of the filters (blur, contrast, noise, brightness, jpeg compression) were applied to the donor patch  $\mathcal{D}$ , the process is repeated.

### 4. Create Binary Manipulation Mask

7 types of binary masks were used to define the region in an image where manipulations have been executed (see Tab. 3). In Table 2, various examples for created masks and the resulting forged images are shown. The Python’s image processing toolbox `scikit-image` is employed to segment the donor patch in Superpixels [1] of appropriate size, and selects one Superpixel (connected set of pixels) for the splicing manipulation. The ”person segmentation” uses the segmentation ground truth from the MS-COCO dataset. All pixels from a donor image patch  $\mathcal{D}$  marked as person are selected and used as splicing input. Masks are recalculated if their portion of the image patch is not in the range of 5% to 40%.

### 5. Generate Forged Image

Given a pristine patch  $\mathcal{P}$ , a donor patch  $\mathcal{D}$ , a manipulation  $m$  and a binary manipulation mask  $\mathcal{M}$ , the forged

Shape of Mask	Parameters	Impact
Triangle	p1, p2, p3	3 random points
Rounded Rectangle	X, Y, r	2 points for Bbox; radius of the corners
Ellipse	X, Y	2 points to define the bounding box
Polygon with 5 vertices	p1, ..., p5	sequence of 5 random points
Ellipse + Polygon with 4 vertices	X, Y, p1, ..., p4	ellipse + 4 vertex polygon
Superpixel Segmentation	[min, max]	range for number of Superpixels
Person Segmentation	-	

Table 3. Types of mask shapes generated for local image manipulation

image  $\mathcal{X}$  is given by

$$\mathcal{X} = \mathcal{M} \cdot \mathcal{P} + (1 - \mathcal{M}) \cdot m(\mathcal{D}) \quad (2)$$

meaning that each pixel of the resulting image  $\mathcal{X}$  is taken either from the pristine patch  $\mathcal{P}$  or the manipulated donor patch  $\mathcal{D}$ , depending on the binary mask  $\mathcal{M}$ . In case of a copy-move manipulation, an additional translation of the copied image part  $(1 - \mathcal{M}) \cdot m(\mathcal{D})$  towards another position in the pristine image patch is made.

Using this process, a training dataset with 1.9 million forged images was generated, comprised of 700,000 splicing, 500,000 copy-move, 400,000 enhance and 200,000 inpainting images as their main forgery type. This training dataset was used in Section 4 for model training.

Manipulation-Type	C	S	R	E
Resample	×	×	-	-
Flip	×	×	-	-
Rotate	×	×	-	-
Blur	-	-	-	×
Contrast	-	-	-	×
Noise	-	-	-	×
Brightness	-	-	-	×
JPEG-Compression	-	-	-	×

Table 4. Preprocessing steps for donor image per manipulation types: Copy-Move (C), Splicing (S), Removal (R) and Enhancement (E)

## 4. Network Architecture Evaluation

In this section we implemented several network architectures for image forgery detection and localization. The models were trained on the dataset created in Sec. 3. The problem of image forgery detection and localization is essentially a segmentation problem in which each pixel is classified as an original or a manipulated pixel. For this task, U-Nets are a well established network architecture and we present 3 variants of U-Net models with promising performance, evaluate them on 4 benchmark datasets and describe the best performing model in more detail.

### 4.1. Models and Evaluation

**MobileNet - MoNet:** This model is a modified U-Net. U-Nets consist of an encoder for downsampling and a decoder for upsampling. MobileNetV2 [19], pretrained on Imagenet, is used as an encoder. MobileNet [5] is a lightweight architecture that has already learned robust features in the context of image classification and hence allows to reduce the number of trainable parameters. For upscaling, the Tensorflow implementation of pix2pix [7] was utilized. Furthermore, 5 skip connections between output layers from downsampling and layers from the upsampling part were established.

**U-NET:** This network is one implementation of the original U-Net architecture [17].

**SE-UN:** Our improved version of U-NET architecture, which adds a recalibration with Spatial and Channel Squeeze & Excitation Blocks [18].

Table 5 shows results for the three network architectures evaluated on the benchmark datasets CASIA [4], Columbia [6], DSO [3] and NIST16 [15].

While the MobileNet implementation (MoNet) gives the best results for the metrics F1 and IoU averaged over all 4 benchmark datasets, SE-UN performs better for AUC and the pixel-wise accuracy. Since the average over all 4 metrics is higher for the latter model (0.574) compared to MoNet with a score of 0.566, we chose our U-Net variation with additional Spatial Channel Squeeze and Excitation (SE-UN) for further experiments and SOTA comparison. The architecture is depicted in more detail in Fig. 1.

### 4.2. Implementation Details

The deep learning framework Tensorflow was used for training our network. For training and detection, the images were resized to (224, 224) pixels. An Nvidia GeForce GTX 1080 Ti GPU was used for training, with batch size set to 16. We use Adam optimizer and perform 1500 steps per epoch and stop after the loss of the validation dataset did not improve for 35 epochs. Training starts with a learning rate of 0.00006, which is halved after 20 epochs without improvement.



Models	Test Datasets																				
	DSO [3]				Columbia [6]				NIST [15]				CASIA [4]				Average				
	AUC	F1	IoU	ACC	AUC	F1	IoU	ACC	AUC	F1	IoU	ACC	AUC	F1	IoU	ACC	AUC	F1	IoU	ACC	all
MoNet	.690	<b>.348</b>	<b>.227</b>	.716	.781	<b>.663</b>	<b>.568</b>	<b>.829</b>	.660	.257	.195	.833	.723	.384	.306	.878	.713	<b>.413</b>	<b>.324</b>	.814	.566
U-NET	.599	.098	.061	.835	.803	.519	.411	.802	.655	.222	.174	.897	.750	.212	.176	.924	.701	.263	.206	.864	.508
SE-UN	<b>.732</b>	.152	.108	<b>.848</b>	<b>.827</b>	.503	.428	.827	<b>.780</b>	<b>.265</b>	<b>.221</b>	<b>.921</b>	<b>.851</b>	<b>.429</b>	<b>.369</b>	<b>.929</b>	<b>.797</b>	.337	.282	<b>.881</b>	<b>.574</b>

Table 5. Comparison of three developed U-Net architectures (MoNet, U-NET, SE-UN) by AUC, F1 and IoU metrics.

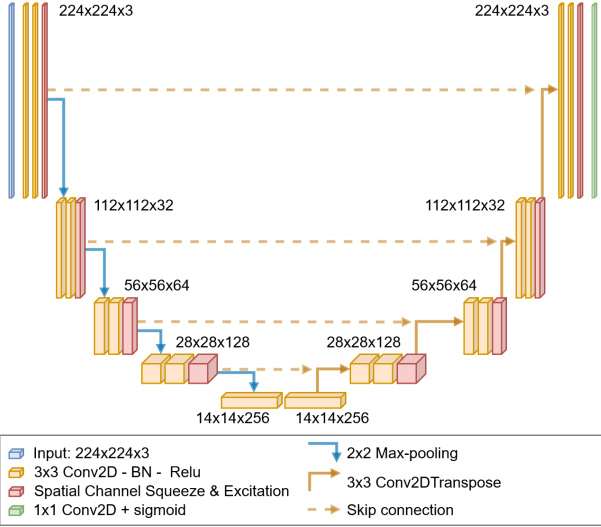


Figure 1. Our proposed network: A U-Net architecture with 4 skip connections and spatial channel Squeeze & Excitation (scSE) extension. Two (3x3)-convolutions combined with one scSE layer, a batch normalization (BN) layer and a Relu activation layer form the building blocks, followed by Max-pooling (encoder) respectively upscaling (with a Conv2DTranspose layer in the decoder part on the right side). The expected input image size  $(H, W) = (224, 224)$ .

### 4.3. Image Manipulation Classification

To investigate the capability of our networks to handle the image manipulation classification (IMC) task, we trained the encoder part of our MoNet model with an additional Softmax layer to detect one of the 4 main manipulation types (splicing, copy-move, removal, enhancement) as the outcome. We trained on a dataset created according to Sec. 3 with one million images divided into 4 classes. For an evaluation dataset with 1,200 images created similarly to the training dataset, a classification accuracy of 94,92% was achieved, thus showing the capacity of the model for the classification task.

## 5. Experimental Evaluation

### 5.1. SOTA Comparison

The proposed model SE-UN was compared with 4 state-of-the-art methods: ForSim [14], DFCN [29], ManTra-Net

[24] and OSN [23]. We used the officially released models from the latter two approaches to evaluate the methods on the four benchmark datasets CASIA V1 [4], Columbia [6], DSO [3] and NIST16 [15]. For DFCN and ForSim, we listed the results from [23]. As metric, the Area Under the Receiver Operating Characteristic curve (AUC) was chosen as it is widely used in the research field of image forgery detection. As in previous works (e.g. [23]), the ground truth mask is inverted if it sums up to more than 50% of the image. This seems in line with the principal concept of manipulation detection, although it has an insignificant impact on the overall metric scores.

As shown in Table 6, our approach performed best on the Columbia, NIST16 and CASIA datasets. Only for the DSO dataset, the ForSim achieved the highest AUC value. With an average AUC-value of 79.7 our approach outperformed OSN, the second best performing approach, by 5.3 points. Table 7 shows examples from each of the benchmark datasets, comparing the three methods with the highest average AUC values.

### 5.2. Processing Time

Our proposed SE-UN model is trained on images of size  $(224, 224)$ . Therefore, for the purpose of evaluation, images are first rescaled to this size. The learned network features are so robust, that they are capable to predict forgeries with SOTA performance even on down-scaled images. This brings significant advantages compared to other approaches ([13], [23]), which make decisions base on full resolution information gathered from whole images. In Table 8, we show a comparison with [23] of the processing time when predicting all images for each of the benchmark datasets. For datasets with images of smaller size (CASIA, Columbia), the processing time of our approach and the OSN [23] method is on the same scale. For datasets with larger images (DSO, and specifically NIST), the processing time for OSN rises rapidly with the size of the images. The reason is that, for the 564 images of the NIST dataset, this approach produces 24,996 tiles from the original images, executes forgery detection on each of these image parts, and finally merges the result for the predicted outcome per image.

Models	AUC of Test Datasets				
	DSO [3]	Columbia [6]	NIST [15]	CASIA [4]	Average
ForSim [14]	<b>.796</b>	.731	.642	.554	.681
DFCN [29]	.724	.789	.778	.654	.736
ManTra-Net [24]	.795	.747	.634	.776	.738
OSN [23]	.723	.815	.686	.751	.744
SE_UN (ours)	.732	<b>.827</b>	<b>.780</b>	<b>.851</b>	<b>.797</b>

Table 6. Comparison of our SE\_UN model with SOTA methods using AUC metric on four benchmark datasets. The highest value per column is **bold**

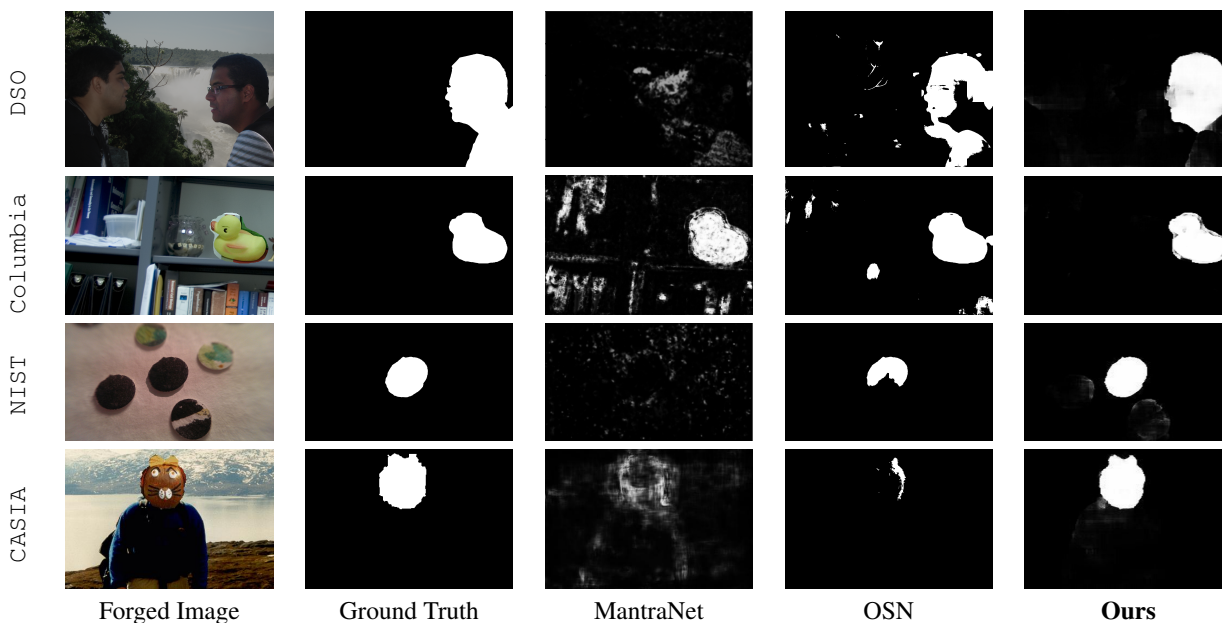


Table 7. Examples of qualitative comparison of MantraNet [24], OSN [23] and our proposed forgery detector. Each line shows one example image for each of the four benchmark datasets DSO [3], Columbia [6], NIST [15], CASIA [4]. The five columns show: the forged image (input), manipulated area (ground truth), results (output) from MantraNet, OSN and our detector.

Dataset	# Images	Format	t-OSN	t-Ours
CASIA [4]	920	jpg	169	<b>94</b>
Columbia [6]	160	tif	<b>120</b>	178
DSO [3]	100	png	701	<b>20</b>
NIST [15]	564	jpg	15250	<b>188</b>

Table 8. Processing time (t) in seconds for prediction per benchmark dataset. For datasets with huge images as NIST (images of size up-to 5616x3744 pixels) tile-based approaches considerably take longer than approaches performing pre-scaling.

## 6. Conclusion

In this paper, we propose a new network model for image forgery detection. The proposed approach reaches and exceeds state-of-the-art performance on various benchmark dataset. The relatively simple network architecture

learns very robust features from scratch from the presented dataset. Even on heavily down-scaled images, the detector delivers very good results, and a considerable processing time advantage for bigger sized images compared to competitors.

Our model can detect compound and unseen forgeries of postprocessed images (as included in the benchmark datasets). But still, the fact that our model achieves pixel-wise accuracy rates of 99% on a validation dataset created similarly to the training dataset, but about 88% on the benchmark datasets used for evaluation shows potential for more improvement of the detector by generating more challenging training data.

## References

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpix-

- els compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. 3
- [2] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 3
- [3] T. Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. Rezende Rocha. Exposing digital image forgeries by illumination color classification. *IEEE Trans. Inf. Forensics and Security*, 8(7):1182–1194, 2013. 4, 5, 6
- [4] J. Dong, W. Wang, and T. Tan. Casia image tampering detection evaluation database. In *IEEE China Summit Inter. Conf. Signal Info. Proc.*, pages 422–426. IEEE, 2013. 1, 4, 5, 6
- [5] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, 2017. 4
- [6] Y. Hsu and S. Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *IEEE Inter. Conf. Multim. Expo*, pages 549–552. IEEE, 2006. 4, 5, 6
- [7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 4
- [8] Haodong Li, Weiqi Luo, and Jiwu Huang. Localization of diffusion-based inpainting in digital images. *IEEE Transactions on Information Forensics and Security*, 12(12):3050–3064, 2017. 1, 2
- [9] L. Li, S. Li, Hancheng Zhu, Shu-Chuan Chu, John Roddick, and Jeng-Shyang Pan. An efficient scheme for detecting copy-move forged images by local binary patterns. *Journal of Information Hiding and Multimedia Signal Processing*, 4:46–56, 01 2013. 1
- [10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*, September 2014. 2
- [11] Siwei Lyu, Xunyu Pan, and Xing Zhang. Exposing region splicing forgeries with blind local noise estimation. *International Journal of Computer Vision*, 110:202–221, 11 2013. 1, 2
- [12] Toqeer Mahmood, Aun Irtaza, Zahid Mehmood, and Muhammad Mahmood. Copy-move forgery detection through stationary wavelets and local binary pattern variance for forensic analysis in digital images. *Forensic Science International, Elsevier*, 279:8–21, 10 2017. 2
- [13] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. A full-image full-resolution end-to-end-trainable cnn framework for image forgery detection. *IEEE Access*, PP:1–1, 07 2020. 2, 5
- [14] Owen Mayer and Matthew C Stamm. Forensic similarity for digital images. *IEEE Transactions on Information Forensics and Security*, 2019. 5, 6
- [15] National Institute of Standards and Technology (NIST). Nist nimble 2016 datasets. <https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation/>. 1, 4, 5, 6
- [16] Junlin Ouyang, Yizhi Liu, and Miao Liao. Robust copy-move forgery detection method using pyramid model and zernike moments. *Multimedia Tools and Applications*, 78:1–19, 04 2019. 2
- [17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015. 2, 4
- [18] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Recalibrating fully convolutional networks with spatial and channel 'squeeze & excitation' blocks. 4
- [19] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4
- [20] Jee-Young Sun, Seung-Wook Kim, Sang-Won Lee, and Sung-Jea Ko. A novel contrast enhancement forensics based on convolutional neural networks. *Signal Processing: Image Communication*, 63:149–160, 2018. 1
- [21] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 1
- [22] Yuan Wang, Lihua Tian, and Chen Li. Lbp-svd based copy move forgery detection algorithm. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 553–556, 2017. 2
- [23] Haiwei Wu, Jiantao Zhou, Jinyu Tian, Jun Liu, and Yu Qiao. Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*, 2022. 1, 2, 5, 6
- [24] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9535–9544, 2019. 1, 2, 5, 6
- [25] Pei Yang, Gaobo Yang, and Dengyong Zhang. Rotation invariant local binary pattern for blind detection of copy-move forgery with affine transform. In *Cloud Computing and Security*, pages 404–416, 07 2016. 2
- [26] Ibrahim A. Zedan, Mona M. Soliman, Khaled M. Elsayed, and Hoda M. Onsi. Copy move forgery detection techniques: A comprehensive survey of challenges and future directions. *International Journal of Advanced Computer Science and Applications*, 12(7), 2021. 2
- [27] Jun-Liu Zhong and Chi-Man Pun. An end-to-end dense-inceptionnet for image copy-move forgery detection. *IEEE Transactions on Information Forensics and Security*, 15:2134–2146, 2020. 2
- [28] Peng Zhou, Xintong Han, Vlad Morariu, and Larry Davis. Learning rich features for image manipulation detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2018. 2
- [29] Peiyu Zhuang, Haodong Li, Shunquan Tan, Bin Li, and Jiwu Huang. Image tampering localization using a dense fully convolutional network. *IEEE Transactions on Information Forensics and Security*, 16:2986–2999, 2021. 2, 5, 6

# A Modular Model Combining Visual and Textual Features for Document Image Classification

Amer Duhan  
TU Wien

amer1.duhan@gmail.com

Robert Sablatnig  
TU Wien

sab@cvl.tuwien.ac.at

## Abstract

*Document image classification is the classification of digitized documents. Typically, these documents are either scanned or photographed. One page of such a document is referred to as a document image. Classifying document images is a crucial task since it is an initial step in downstream applications. Most state-of-the-art document image classification models are based on a transformer network, which are pretrained on millions of scanned document images and thus require a huge amount of training resources. Additionally, this and other state-of-the-art document image classification models have well beyond 100 million parameters. In this work, we address both challenges. First, we create a model capable of competing with the current state-of-the-art models without pretraining on millions of scanned document images. Second, we create a model several times smaller than current state-of-the-art models in terms of parameters. The results show that the developed approach achieves an accuracy of 93.70% on the RVL-CDIP dataset, and a new state-of-the-art accuracy of 96.25% on Tobacco3482.*

## 1. Introduction

The increasing digitalization has led companies to digitize their processes and content [4], and organize their information to improve the search and access to relevant data [6]. Thus, paper documents are subject to digitization, and document images are the output [21]. The task of document image classification is to categorize a given document image into a set of defined classes [12].

Due to its high importance, document image classification has been explored extensively [1]. However, most of the current State-Of-The-Art (SOTA) methods have either parameters in the hundreds of millions, pretrain on a larger dataset, or both, such as [32] or [33].

Thus, we propose a multimodal system based on SOTA image and language models, which are relatively small in

their size (less than 100 million parameters). Furthermore, the amount of training data is limited to the RVL-CDIP dataset. Due to the modular nature of the architecture, we tested two model combinations to analyze their impact on the overall test set accuracy. Our experiments show that an image-only system achieves a higher test set accuracy than a multimodal system.

The contributions are the following:

- Developing a model that can compete with current SOTA models on the RVL-CDIP dataset without requiring millions of document images. Moreover, the developed model is much more efficient than the current SOTA models.
- Achieving a new SOTA on the Tobacco3482 dataset with 96.25% accuracy.

The remainder of this paper introduces the datasets in Section 2, discusses related work in Section 3, presents the methodology in Section 4, depicts the results in Section 5, and concludes the paper in Section 6.

## 2. Datasets

In the following, the two datasets used in this paper are discussed. First, the dataset on which the proposed architecture is trained and evaluated, and second on which it is finetuned and evaluated.

### 2.1. RVL-CDIP

This work is based on the RVL-CDIP [11] dataset since it was specifically created to test image classification algorithms on document images [7]. RVL-CDIP is a subset of the IIT-CDIP Test Collection (11 million documents) [20], which itself is a subset of the LTDL dataset [26] (14 million documents), that was created from public records of lawsuits against American tobacco companies [11]. The RVL-CDIP dataset contains 400,000 grayscale images with 16 classes, split evenly in an 8:1:1 ratio of training, validation, and test set.

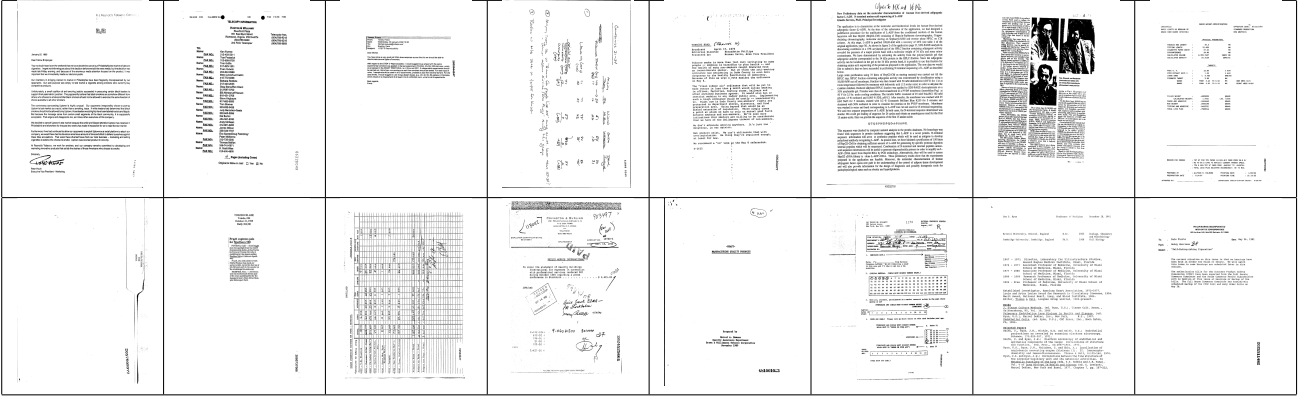


Figure 1. An example document image for each class from the RVL-CDIP dataset. From the top left image, the labels are the following: Letter, Form, Email, Handwritten, Advertisement, Scientific report, Scientific publication, Specification, File folder, News article, Budget, Invoice, Presentation, Questionnaire, Resume and Memo.

## 2.2. Tobacco3482

The Tobacco3482 [18] dataset, created from the same dataset as RVL-CDIP, the IIT-CDIP Test Collection, contains 3,482 grayscale document images. These images are split into 10 classes, which are not evenly distributed as in the RVL-CDIP dataset.

## 3. Related Work

The methods in all of the following works are tested on the RVL-CDIP test set.

Harley et al. [11], who have created the RVL-CDIP dataset, stack 5 CNNs, one of which is trained on the whole document image, and the others are trained over the header, footer, left body, and right body. These CNNs are either trained from scratch or transfer-learned from AlexNet [17]. Das et al. [6] use a similar technique. However, their CNNs are transfer-learned from VGG-16 [27]. A MLP, a class of artificial neural networks, is then found to perform as the best ensemble technique.

Afzal et al. [1] show that even though the ImageNet and RVL-CDIP datasets have different domains, a pretrained network on ImageNet, such as VGG-16, has a better accuracy score on the RVL-CDIP test set than no pretraining.

Tensmeyer and Martinez [29] train CNNs from scratch, i.e., randomly initialized. Various modifications are performed, such as changing the network depth, width, or input size. The authors show that the input size significantly impacts the performance.

Sarkhel and Nandi [25] utilize a spatial pyramid model to extract highly discriminative multi-scale feature descriptors from a visually rich document by leveraging the inherent hierarchy of its layout.

Ferrando et al. [10], Jain and Wigington [12], Audebert et al. [4], Kanchi et al. [14], and Bakkali et al. [5] combine

image and text features in a two-stream approach by utilizing a CNN for image and an embedding for text. Jain and Wigington [12] use the VGG-16 to get image features and use different methods to extract text features, representing text at the sequence, word, and character level. Audebert et al. [4] utilize the MobileNetV2 [23] for image feature extraction, which has a similar performance in terms of accuracy, compared to VGG-16 while being significantly faster. As in [12], word-level text features are generated with Fast-Text [13], a word embedding technique. Ferrando et al. [10] combine EfficientNet [28] for image features and a reduced version of BERT [8], a transformer model, for text features. Kanchi et al. [14] propose a hierarchical attention network for the textual stream, with fine-tuned BERT embeddings as input and an EfficientNet-B0 for the image stream. Bakkali et al. [5] combine NasNet<sub>Large</sub> [34] with BERT to achieve a SOTA accuracy of 97.05%, using an average ensembling for the image and text stream.

A transformer [30] architecture for document image classification is used in the work of Xu et al. [32]. This architecture is an extended version of BERT [8]. However, the model is pretrained on the IIT-CDIP Test Collection, which contains more than 11 million scanned document images. Another major difference, compared to all previous mentioned approaches, is that this method is suitable for classifying document images, and, for example, for form understanding, where the goal is to extract key-value pairs from document images. Xu et al. [33] extend [32]. The authors integrate visual information in the pre-training stage and use 2-D relative position representation for token pairs instead of absolute 2-D position embeddings, which Xu et al. [32] use to model the page layout. Just as its predecessor, this model is also suitable for other tasks outside of classifying document images.

Similarly, Powalski et al. [22], Wang et al. [31]. and

Srikar et al. [2] develop each a multimodal transformer based architecture, which performs a pretraining step. [22] simultaneously learns layout information, visual features, and textual semantics. In [31] the layout knowledge from monolingual structured documents is learned and then generalized to deal with multilingual ones. [2] combines textual, visual, and spatial features using a novel multi-modal self-attention layer.

## 4. Methodology

In this section, both streams (image and text) are elaborated, covering the preprocessing steps and the training strategy and architecture. Then, the method to combine both streams to form the final piece of the document image classification system is covered.

### 4.1. Image Stream

Compared to textual features, image features are preferred for the problem of document image classification [16]. The current SOTA CNN architecture, EfficientNet [28], is used for the image stream. The image stream and text stream are two independent parts of the whole model, which are combined in a later stage. The preprocessing steps, the training strategy, and the architecture are explained in the following.

#### 4.1.1 Preprocessing steps

In our method, the image stream consists of five EfficientNets, each focusing on a input part. The preprocessing steps partly follow the work of [11]. First, all images are resized to  $936 \times 720$ . Then, 5 regions are defined for an image; holistic, header, footer, left body, and right body. The holistic region is the whole image itself. The header is defined as the first 307 pixel rows. Similarly, the footer is defined as the last 307 pixel rows. The left body is defined as the 480 central pixel rows and the first 360 pixel columns; similarly, the right body is defined as the 480 central pixel rows and the last 360 pixel columns. A slight intersection exists between the left and right body areas with the header and footer. Finally, each image is resized to  $384 \times 384$ .

The focus on specific regions of a document follows from the fact that certain categories show a low interclass variability, as seen in Figure 1 when comparing memo and letter. While memos often have a complete address section, letters typically have a "To:" and "From:". Having a CNN to classify documents using only this region will much more likely learn those differences than a holistic CNN [11]. Similar to the header region, different CNNs are applied to each region described in the previous paragraph.

Since the document images are in grayscale, they are transformed into images with three channels, i.e., copied two times and stacked depth-wise along the third axis.

#### 4.1.2 Training strategy and architecture

The training strategy and architecture on the full dataset are inspired by [6]. The main benefit of the following training strategy is reducing computational complexity. A three-level transfer learning achieves this.

The first level of transfer learning (L1) is initializing the weights of the holistic model from the corresponding EfficientNet-B1 model, trained on the ImageNet dataset. To train the holistic model, only the classifier added on top of the EfficientNet-B1 model is trained first, and all other weights of the model are frozen, such that they are not updated during backpropagation. This model's weights are then used to initialize the same model (L2), but with all layers unfrozen, including the batch normalization layers. Now, all weights can be updated to further increase the prediction accuracy.

Next, its weights are taken to initialize the remaining four models (L3), i.e., the models for the header, footer, left body, and right body region. Like the holistic model, these four models are trained with early stopping on the validation loss and patience of 10. ReLU [9] is used as the activation function.

### 4.2. Text stream

The recent development in this field suggests that textual features are necessary to achieve SOTA results. A distilled version of BERT [8], called DistilBERT [24], is used as the backbone in our work since it is 40% smaller in size compared to BERT while retaining 97% of its language understanding capabilities. In the following sections, the preprocessing steps, as well as the training strategy, are explained.

#### 4.2.1 Preprocessing steps

The Tesseract OCR system (version 4.1.1) extracts the text from the document images. Once this is done, the next step is preprocessing the extracted text before feeding it into a neural network. This is even more important when the text is extracted from document images, instead of, for instance, scraping the text from the web. Everything that is not a letter or a digit is removed. It is ignored if the text is less than two characters long, but single-digit numbers are kept. Moreover, the text is lowercased. There are some pages where no text can be extracted by Tesseract. In this case, or where the whole extracted text of a document image is removed due to the preprocessing steps, the extracted text is set to "", i.e., a string of length zero.

#### 4.2.2 Training strategy and architecture

Following the results from the image stream, the training strategy in the text stream takes a similar approach. Only

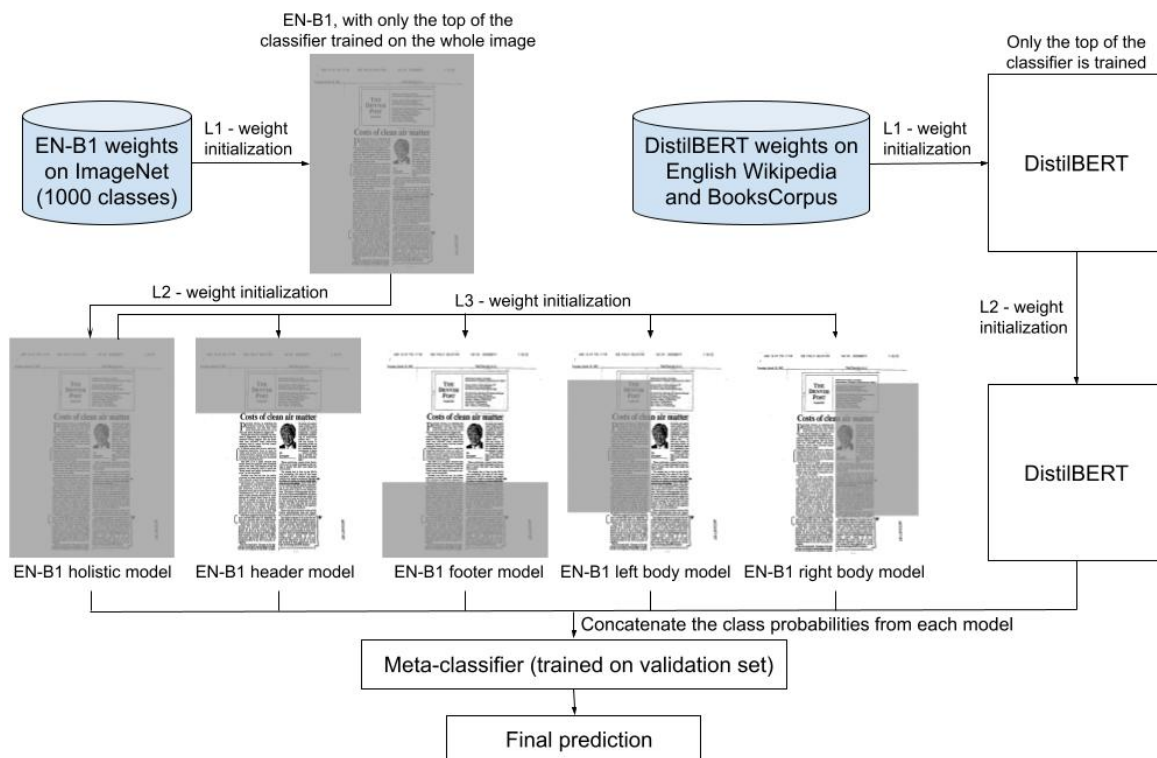


Figure 2. Proposed architecture for document image classification based on SOTA architectures, with an image stream, text stream, and utilizing different levels of transfer learning. EN = EfficientNet.

the classification head added on top of DistilBERT is trained first, with the features extracted from the base model.

DistilBERT and the original BERT model have two unique tokens: [CLS], a classification token, and [SEP], a separator token. The [CLS] token is used for classification tasks and is added in front of every sequence. Specifically, the last hidden state representation of the [CLS] token is used. This hidden state representation is then used as an input to the classification head.

Like in the image stream, the classification head is first trained, then the whole model. The final model is trained with early stopping and patience of 10, with ReLU as the activation function.

### 4.3. Stacked generalization

The last part of the system is to train a meta-classifier, which outputs the final predictions. It is adopted in document image classification models, such as in [6], [4], [10], [12], [3], and works by combining the (intermediary) output of one or more classifiers and feeding that as an input to a meta-classifier. To reduce overfitting, the meta-classifier is trained on the validation set. The goal of stacked generalization is to provide a lower generalization error than the base models. The meta-classifier is the last module of the

document image classification system, and the full architecture is shown in Figure 2.

The input for the meta-classifier are the class probabilities (i.e. the softmax output). In this work, the meta-classifier, a 3-layer neural network, combines visual and textual features by concatenating them and producing the final output of the document image classification system.

Adam is chosen as the optimizer. Moreover, an image-only system versus a multimodal system is tested.

### 4.4. Tobacco3482

The document image classification model is also finetuned and evaluated on the Tobacco3482 dataset. To make results comparable with other works, such as [11], [15], [18], [19], or [10], the dataset is split as follows. From 3,482 images, 100 images per class are randomly selected. This constitutes the training set; and the remaining 2,482 images are the test set. This process is repeated 10 times, such that there are 10 different training and test sets, from which the median test set accuracy is reported. From the 1,000 training images, 200 are used for the validation set.

The training approach first uses the pretrained models on the RVL-CDIP dataset and then finetunes on the Tobacco3482 dataset, where only the added classification head

Results						
Author	Accuracy	# Parameters	Modality	Extra training data	Tobacco3482 Accuracy	
Afzal et al. (2017) [1]	90.97	138.36	I	No	91.13	
Kang et al. (2014) [15]	-	4.21	I	No	65.35	
Kumar et al. (2014) [19]	-	-	I	No	43.27	
Das et al. (2018) [6]	92.21	691.87	I	No	-	
Audebert et al. (2020) [4]	90.60	3.64	I + T	No	87.80	
Ferrando et al. (2020) [10]	92.31	85.47	I + T	No	94.90	
Harley et al. (2015) [11]	89.80	58.35	I	No	79.90	
Jain and Wigington (2019) [12]	93.60	138.36	I + T	No	-	
Sarkhel and Nandi (2019) [25]	92.77	-	I	No	82.78	
Tensmeyer and Martinez (2017) [29]	91.03	-	I	No	-	
Xu et al. (2020) [32]	94.42	160.00	I + T	Yes	-	
Xu et al. (2021) [33]	95.64	426.00	I + T	Yes	-	
Srikanth et al. (2021) [2]	96.17	183.00	I + T	Yes	-	
Wang et al. (2022) [31]	95.68	-	I + T	Yes	-	
Powalski et al. (2021) [22]	95.52	780.00	I + T	Yes	-	
Bakkali et al. (2020) [5]	97.05	197.21	I + T	No	-	
Kanchi et al. (2022) [14]	95.48	-	I + T	Yes	95.70	
Proposed approach	93.70(I) / 93.50(I+T)	40.72	I	No	95.65(I) / 96.25(I+T)	

Table 1. Test set results on RVL-CDIP and Tobacco3482. Accuracy in %. The number of parameters (in millions) is either explicitly stated in the work, an estimation, or omitted. I = Image, T = Text.

is trained.

Additionally, a meta-classifier is trained to combine the softmax outputs on the training set of the image and text models. Similarly, an image-only and multimodal system is trained. The models are trained with Adam, ReLU, and early stopping with patience of 3.

## 5. Results

The proposed approach includes two results per dataset, each with an image-only and multimodal system. The results are depicted in Table 1.

An accuracy of 93.70% on RVL-CDIP and 96.25% on Tobacco3482 is achieved. Note that on the RVL-CDIP dataset, the image-only system achieves a higher accuracy, while on the Tobacco3482 dataset, it is the multimodal system. That is, adding textual information decreases the accuracy on the RVL-CDIP dataset, which goes against the results of other papers that have used textual information (see Table 1). The difference in the accuracy between the image-only and multimodal approach is larger on the Tobacco3482 dataset.

Most SOTA papers have used additional training data with a multimodal approach. Table 1 shows, that all papers, who have reached an accuracy of over 94%, have used an extra training data, either the full IIT-CDIP Test Collection (11 million documents) or a fraction of it, except the current SOTA [5], with 97.05% accuracy. Moreover, all papers with an accuracy of over 94% are fully based on a Transformer architecture, except [5] and [14].

The number of parameters of the proposed approach (around 41 million) is multiple times smaller than in the current SOTA methods. Even though the result on the RVL-CDIP dataset could not match them, a new SOTA has been achieved on the Tobacco3482 dataset using the multimodal

approach, beating the previous SOTA result of Kanchi et al. [14] by 0.55 percentage points. Additionally, the image-only approach missed the previous SOTA result by 0.05 percentage points.

The model is trained on a NVIDIA T4 GPU with 16GB VRAM. One epoch takes about 220 minutes for the image models on the RVL-CDIP dataset. Each image model is trained for about 14 epochs, i.e., for 70 epochs combined. The text model is trained for 8 epochs, with about 136 minutes per epoch. These numbers refer to those models, where the weights of all layers are unfrozen.

## 6. Conclusion

The goal of the proposed approach is to develop a model, which can compete with current SOTA methods and be relatively efficient, i.e., have a relatively small number of parameters. Even though the current SOTA results on the RVL-CDIP dataset could not be quite matched, the developed model is around 5 times smaller in terms of the number of parameters. On the Tobacco3482 dataset, however, a new SOTA result is achieved. Interestingly, contrary to the papers using a multimodal approach mentioned in Table 1, the textual information decreases the accuracy on the RVL-CDIP dataset.

## References

- [1] Muhammad Zeshan Afzal, Andreas Kolsch, Sheraz Ahmed, and Marcus Liwicki. Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification. *ICDAR*, 1:883–888, 2017.
- [2] Srikanth Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. DocFormer: End-to-End Transformer for Document Understanding. *ICCV*, pages 973–983, 2021.



- [3] Muhammad Nabeel Asim, Muhammad Usman Ghani Khan, Muhammad Imran Malik, Khizar Razaque, Andreas Dengel, and Sheraz Ahmed. Two stream deep network for document image classification. *ICDAR*, pages 1410–1416, 2019.
- [4] Nicolas Audebert, Catherine Herold, Kuider Slimani, and Cédric Vidal. Multimodal deep networks for text and image-based document classification. *CCIS*, 1167:427–443, 2020.
- [5] Souhail Bakkali, Zuheng Ming, Mickael Coustaty, and Marçal Rusinol. Visual and textual deep feature fusion for document image classification. *CVPRW*, pages 2394–2403, 2020.
- [6] Arindam Das, Saikat Roy, Ujjwal Bhattacharya, and Swapan Kumar Parui. Document Image Classification with Intra-Domain Transfer Learning and Stacked Generalization of Deep Convolutional Neural Networks. *ICPR*, pages 3180–3185, 2018.
- [7] Tyler Dauphinee, Nikunj Patel, and Mohammad Rashidi. Modular multimodal architecture for document classification. *arXiv:1912.04376*, 2019.
- [8] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 1:4171–4186, 2019.
- [9] Jianli Feng and Shengnan Lu. Performance Analysis of Various Activation Functions in Artificial Neural Networks. *Journal of Physics: Conference Series*, 1237(2), 2019.
- [10] Javier Ferrando, Juan Luis Domínguez, Jordi Torres, Raúl García, David García, Daniel Garrido, Jordi Cortada, and Mateo Valero. Improving accuracy and speeding up document image classification through parallel systems. In *Computational Science – ICCS 2020*, pages 387–400. 2020.
- [11] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval. *ICDAR*, pages 991–995, 2015.
- [12] Rajiv Jain and Curtis Wigington. Multimodal document image classification. *ICDAR*, 3:71–77, 2019.
- [13] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *EACL 2017*, 2:427–431, 2017.
- [14] Shrinidhi Kanchi, Alain Pagani, Hamam Mokayed, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. EmmDocClassifier: Efficient Multimodal Document Image Classifier for Scarce Data. *Applied Sciences (Switzerland)*, 12(3), 2 2022.
- [15] Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for document image classification. *ICPR*, pages 3168–3172, 2014.
- [16] Andreas Kolsch, Muhammad Zeshan Afzal, Markus Ebbecke, and Marcus Liwicki. Real-Time Document Image Classification Using Deep CNN and Extreme Learning Machines. *ICDAR*, 1:1318–1323, 2018.
- [17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NeurIPS*, 25, 2012.
- [18] Jayant Kumar and David Doermann. Unsupervised classification of structurally similar document images. *ICDAR*, pages 1225–1229, 2013.
- [19] Jayant Kumar, Peng Ye, and David Doermann. Structural similarity for document image classification and retrieval. *PRL*, 43(1):119–126, 2014.
- [20] David D. Lewis, Gady Agam, Shlomo Engelsson Argamon, Ophir Frieder, David A. Grossman, and Jefferson Heard. Building a test collection for complex document information processing. *ACM SIGIR*, pages 665–666, 2006.
- [21] Lawrence O’Gorman and Rangachar Kasturi. Executive Briefing: Document Image Analysis. 1997.
- [22] Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Palka. Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer. *ICDAR*, pages 732–747, 2021.
- [23] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CVPR*, pages 4510–4520, 2018.
- [24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv: 1910.01108*, 2019.
- [25] Ritesh Sarkhel and Arnab Nandi. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. *IJCAI*, pages 3360–3366, 2019.
- [26] Heidi Schmidt, Karen Butter, and Cynthia Rider. Building Digital Tobacco Industry Document Libraries at the University of California, San Francisco Library/Center for Knowledge Management. *D Lib Mag.*, 8(9), 2002.
- [27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR 2015*, pages 1–14, 2015.
- [28] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *ICML 2019*, pages 6105–6114, 2019.
- [29] Chris Tensmeyer and Tony Martinez. Analysis of Convolutional Neural Networks for Document Image Classification. *ICDAR*, 1:388–393, 2017.
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Jones Llion, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, pages 6000–6010, 2017.
- [31] Jiapeng Wang, Lianwen Jin, and Kai Ding. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. *ACL*, 1:7747–7757, 2 2022.
- [32] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *ACM SIGKDD*, pages 1192–1200, 2020.
- [33] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. *ACL 2021*, pages 2579–2591, 2021.
- [34] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning Transferable Architectures for Scalable Image Recognition. *CVPR*, 2018.

# Statistical shape modeling and analysis of the vestibular organ based on CT-images

Claudia Companioni Brito      Matthias Willenbrink      Karl Fritscher  
Rainer Schubert  
Private University of Health Sciences, Medical Informatics and Technology  
Hall in Tirol, Austria

## Abstract

*The human body's stable posture and movement are dictated by the precise functioning of the vestibular organ, mainly the ampulla organs in the semicircular canals. The development of electronic devices such as vestibular implants aims to improve the vestibular system's capacity by stimulating the involved vestibular nerves. We aim to describe and analyze anatomical variations of the inner ear using computationally derived statistical shape models. The models should support the design process of vestibular implants. Based on a dataset of 81 cone-beam computed tomography, this work covers constructing a statistical shape model of the semicircular canals using a recently developed novel Particle-Based Modeling approach. The method optimally places correspondence points on each surface using a gradient descent energy function. Then Principal Component Analysis is used to describe anatomical variation. The model was evaluated in terms of reconstruction accuracy, compactness, generalization, and specificity. Results obtained by the workflow based on human datasets and the average shape of a statistical model revealed a high qualitative understanding and a quantitatively comparable range. The first three principal components captured 57.7% of the cumulative variation. The analysis led to 26 principal components to account for 95% of the total shape variation captured. The shape model can be used for virtual product development and testing and to estimate the detailed inner ear shape from a clinical patient computed tomography scan. For the first time, we could describe the geometry of the human semicircular canals based on a large sample of data from living humans compared with other studies.*

## 1. Introduction

The human ear is the organ that enables hearing and balance. The anatomy of the human ear consists of three parts: the outer ear, the middle ear, and the inner ear. The vestibular

system is the apparatus of the inner ear involved in balance. It is a complex organ consisting of three semicircular canals (superior/anterior, posterior, and horizontal/lateral) and the vestibule that houses the otolith organs. The most common medical complaints [24, 26] associated with imbalance symptoms include dizziness or vertigo. Among the vestibular disorders, benign paroxysmal positional vertigo (BPPV) is the most common cause of vertigo [26] which affects females twice as often as males [18]. Vestibular implants (VI) are a new promising technology based on the experiences of cochlear implants [14]. The vestibular nerves are the subject for electrical stimulation to treat balance disorders instead of the cochlear nerve. All the conditions leading to a loss of balance can be severely debilitating and cause a decrease in the quality of life [13, 23], so even though much research is still needed, the technology has a lot of potentials. Patient-specific 3D reconstruction of the vestibular system and its substructures could improve different aspects of vestibular implantation. It can facilitate anatomical understanding for doctors and suggest modifications in the design of electrode placement for vestibular implant manufacturers. To acquire surgical skills, lots of practice and effort are needed; thus, 3D models could be used for surgery simulation and training [10]. Quantitative analysis of the anatomical semicircular canal shape from medical images is essential for diagnosing shape abnormality. In this context, statistical shape models (SSM) turned out to be very useful for investigating variations of shape within anatomical structures of the inner ear. Statistical shape models describe and analyze the human anatomy and its variations, where the parameters of the probabilistic model have been learned from data [1, 7]. It has become an indispensable tool for medical image analysis. Moreover, having a shape model of the vestibular system could be further used for segmentation applications.

### 1.1. Related work

There are many applications concerning SSM of different anatomies (i.e., segmentation of brain and cardiac

structures, orthopedics, and other non-segmentation applications). However, few works applied SSM to analyzing the inner ear, most of which focus on the cochlea [6, 12, 16, 17]. The mathematical method developed by Bradshaw et al. automatically reconstructs the semicircular ducts from high-resolution computed tomography (CT) images in living humans [2] using a 2D B-spline contour. Noble et al. [18, 19] presented a point distribution model (PDM) based on micro-CT images along with an active shape model (ASM) approach to segment and predict preoperative CT datasets. The model is based on micro-CTs of cadaveric cochlea specimens with 36  $\mu\text{m}$  voxel size. The scala tympani and scala vestibuli were manually segmented to create surface models. Point correspondences between the surfaces were generated using an image registration based on the Adaptive Bases algorithm [20]. In [16], Kjer et al. created an SSM of the human inner ear from micro-CT data. The cochlea and structures of the vestibular system were manually delineated based on 17 micro-CT scans of the human temporal bones. An initial alignment was applied to remove translational and rotational differences between the samples, followed by a multi-level B-Spline registration approach using bending energy regularization [21]. The resulting transformations were used to create a statistical PDM of the inner ear containing 16 modes of variation.

Fritscher et al. [11] introduced a framework for creating statistical shape and appearance models of the vestibular system for morphological analysis and the segmentation of the temporal bone. To find corresponding points across all subjects, a transformation consisting of two components: a global rigid transformation and a local deformable transformation, was applied [11]. The resulting deformation vector fields represented the shape variations among the training set and were the input for statistical analysis using Principal Component Analysis (PCA). Furthermore, the approach presented in [11] was extended to visualize and analyze novel multi-object models [19]. Based on the manual segmentation of 31 micro-CT datasets of temporal bones with an isotropic resolution of 15  $\mu\text{m}$ , the SSMs for the following structures were created: Perilymph, Endolymph, Bony labyrinth (approximated using a combined label of endolymph and perilymph), N. ampullaris, N. singularis, N. facialis.

Recently, another approach for reconstructing semicircular canals (SCC) uses an automatic skeletonization process [8]. This approach is based on magnetic resonance imaging (MRI) scans of 20 individuals. The method computes the geometric parameters of the SCC through a skeletonization process of a binary image. The skeletonization approach uses potential field methods, which track field lines and potential valleys in a continuous space. Most of these works mentioned above are based on specimens from deceased subjects and have limited data for experiments.

High-resolution images are obtained in cadaveric specimens after cropping the temporal bone around the bony labyrinth. The preparation and processing of ex-vivo specimens add consequent effort to acquiring the samples and potentially impact the data's quality and usability. While clinical CT images provide a less detailed representation of the inner ear, it is the best data source for living VI candidates. This study is based on existing clinical patient data, which is used in the regular routine of medical doctors that will, later on, use the VI. Therefore we wanted to take not artificially produced data but the kind of images used at the hospitals. The present work aims to develop and describe a detailed statistical shape model of the human SCC geometry based on an initial cohort of 81 subjects to serve as design decision support for a vestibular implant. Using a novel particle-based shape modeling approach facilitates the design of VI.

## 2. Materials and methods

### 2.1. Dataset

Eighty-one cone-beam computed tomography (CBCT) scans of human temporal bones used for this study were acquired from Maastricht University, from which 41 were from the left ear, and the rest were from the right ear. The age of the subjects ranged from 19 to 88 years, with an average age of 58.5 years. The group was divided into 44 men with an average age of 57.7 years and 37 women with an average of 59.3 years. The images in the dataset were acquired over a period of approximately ten years with different slice thicknesses. Some scans have a resolution of 0.4 mm and others 0.6 mm. The segmentations performed by medical experts include hearing bones, vestibular organs, and the facial nerve.

The dataset used in this work did not include abnormal anatomy. Exclusion criteria were applied due to missing CTs in the dataset, with only segmented labels available, and segmentations containing gaps in the canals. In total, 71 subjects were considered to create the SSM.

### 2.2. Statistical shape modeling

The approach presented by Cates et al. [4, 5] to establish correspondence has been used for the creation of the SSM. The general strategy of Particle-Based Modeling (PBM) is to represent correspondence points as interacting sets of particles, one for each shape, that redistribute themselves under an energy optimization and therefore describe the surface geometry [4]. The optimization function finds correspondence positions that minimize the entropy of the model. A more detailed description of the PBM method is referred to [5]. Since the vestibular system is a very complex structure, the PBM method suits well since its particle system formulation captures better-detailed areas by increasing the particle distribution rates in the higher curvature regions.

The PBM is defined as a collection of  $n$  shapes of  $k$  correspondence points. In our experiments,  $n$  is the number of SCC segmentations ( $n = 71$ ), and  $k$  represents the number of landmarks used to describe each surface. The correspondences among the SCC segmentations are determined by running the PBM method to define a set of  $k$  correspondence landmarks  $x$ ; where point  $x_i$  on shape number 1 matches to point  $x_i$  on shape 2, 3, 4, ... ,  $n$  and  $i = 1, \dots, k$ . Several experiments were carried out to set parameters for obtaining an optimal and detailed shape representation of the SCCs embedded in the bone structures of the inner ear. The final shape model of the SCC surface was constructed using 4096 correspondence particles per shape. This number of points was chosen by adding particles until the representation was able to recover anatomically plausible and accurate SCC shapes, and increasing this number did not reveal additional details. The experts quantitatively validated the final number of points for the surface representation at our research group (Institute of Biomedical Image Analysis) by visualizing and comparing the results with the given segmentations.

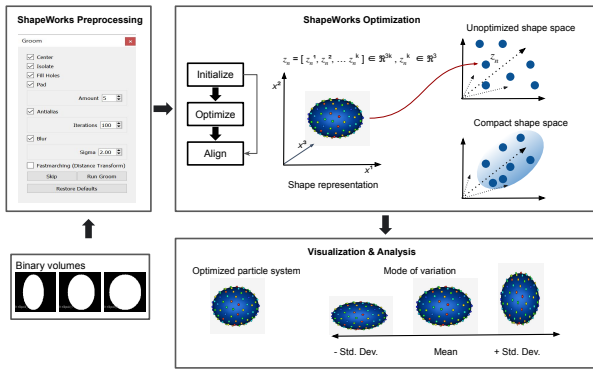


Figure 1. ShapeWorks pipeline. First, the binary segmentations need to be converted to signed DT using a set of grooming steps. After the ShapeWorks optimization stage, statistical analysis is performed using PCA. The mean and modes of shape variation are computed based on the optimized correspondence model. Image modified from [3].

This work uses an open-source distribution of the PBM algorithm called ShapeWorks [3], developed at the University of Utah. ShapeWorks is a publicly available tool with a pipeline of pre-processing steps required before computing the correspondence points. The optimization phase initializes the particle system and runs the PBM algorithm. It takes an initial set of particle positions and the processed data to the signed distance transform to construct the correspondence point model of shape Fig. 1. After the ShapeWorks Optimize step, we have a correspondence model for the population. Then, PCA was used to reduce the high dimensionality of the data matrix required to examine vari-

ation among the different SCC structures while still retaining most of the geometric information of the shapes. PCA isolated the modes of variation from the optimized correspondence particle locations. Once the  $m$  PCA modes that contain substantial variation are chosen, the model can represent every SCC shape in the set as an  $m$ -dimensional vector of scalar values. The shape variations are analyzed by examining the shape described by each principal component (PC), moving between  $\pm 2$  standard deviations from the mean in that PC.

### 3. Results

#### 3.1. Data processing

The segmentation quality of the data was not sufficient, and manual clean-up was needed before using the dataset for further processing in the construction of the SSM (Figure 2a). Small holes and voxel-islands caused by manual segmentation were removed using a connected component analysis and morphological closing operation [25]. Then, the noticeable defects not eliminated by the pre-processing algorithms were corrected by hand using the 3D Slicer toolbox. Since the focus in this work is concerned with the SCCs, only the segmented labels, including it, were considered.

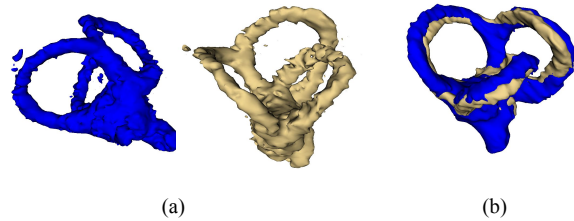


Figure 2. 3D view of two dataset samples (a) before and (b) after pre-processing. All the segmentations were mirrored and aligned. Small holes and voxel-islands caused by manual segmentation have been removed using morphological operations.

All the right inner ears were mirrored so that all datasets appear to be of a left inner ear in order to obtain consistent data (Fig. 2b). First, a transformation including mirroring was calculated using the provided fiducial points in each semicircular canal and applied to all datasets on the right side. Next, all the datasets were aligned using a two-step registration process [25]. The fiducial points were used to apply for an initial rigid point base registration. Then, an additional rigid registration was applied using the former stage as initialization to avoid dependence on fiducials from unknown precision.

All images were resampled using linear interpolation with an isotropic voxel resolution of  $0.15 \times 0.15 \times 0.15$  mm.

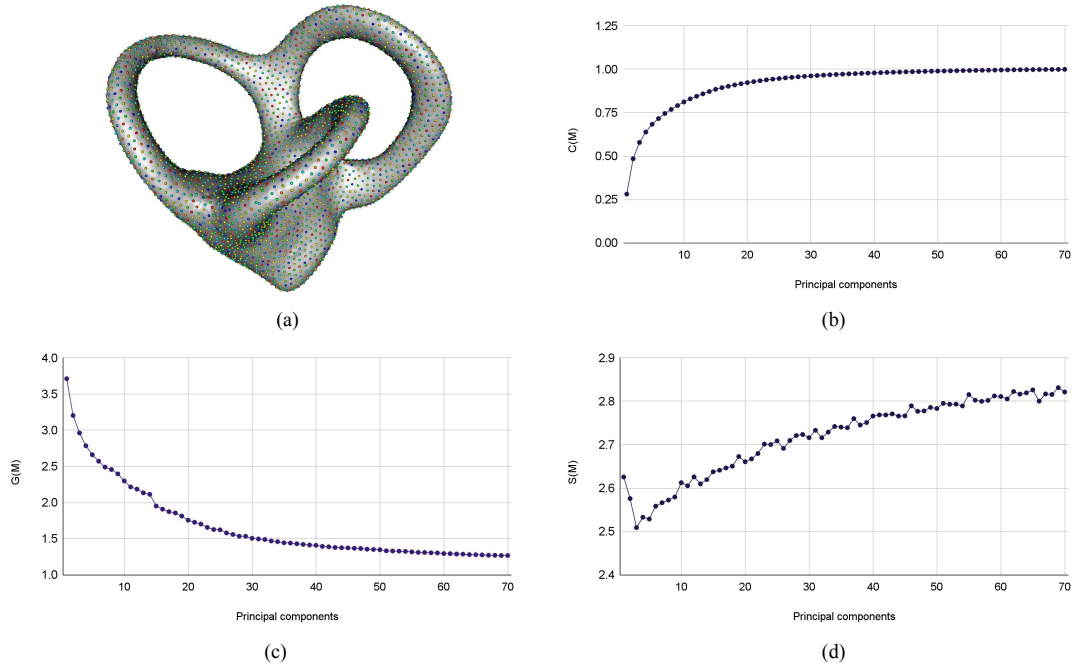


Figure 3. a) Mean shape. b) Compactness, c) Generalization, and d) Specificity of the SCC shape model.

The size was set to  $300 \times 300 \times 300$  voxel ROI spanning. At the time of writing, the selected SSM software in this study needed all the data with the same size, and the voxel spacing equals 1. Thus, the volumes were artificially scaled up by setting the spacing to  $1.0 \times 1.0 \times 1.0$  mm. More details about the registration and processing of the data can be found in [25].

### 3.2. Statistical shape model of the vestibular organ

The resulting mean shape after generating the SSM of the vestibular system using the PBM algorithm is visualized in Fig. 3a.

The vestibule's SSMs with different points were generated and analyzed, showing that poor reconstructions are observed with a smaller number of particles, especially along the canals. In our experiments, increasing the particle counts further than 4096 does not significantly improve the model's accuracy but increases the complexity of the model and computational time. The optimization routine using 4096 particles in a computer with 96 GB of RAM and an Intel Core i7 processor took approximately 7 hours. The duration of the optimization with 256, 1024, and 8192 particles was around 0.31, 1.3, and 14 hours respectively.

### 3.3. Principal component analysis

The PCA shape decomposition is able to represent 95% of the variation among SCC using 26 modes. The first three modes captured 57.7% of the cumulative variation among

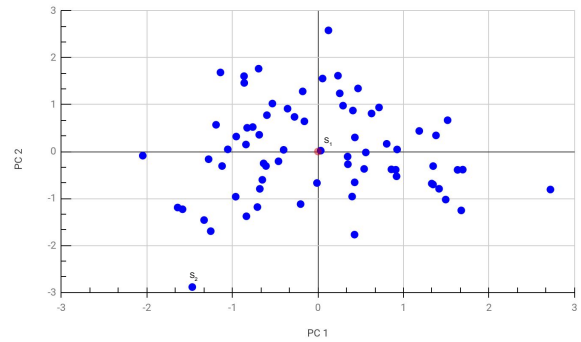


Figure 4. Distribution of input datasets with respect to PC1 and PC2. The red dot represents the mean shape.  $S_1$  is the closest shape to the mean and  $S_2$  is a random shape distant from the mean.

all shapes. Specifically, mode 1 captured 28.0% of the variation, followed by mode 2 at 20.3%, and mode 3 at 9.4%. Knowledge of the position of different datasets in PCA space is significant for identifying similar shapes and datasets that are close to the mean shape. Therefore, the PCs covering the highest amount of shape variation were used to analyze the distribution of the datasets in PCA space Fig. 4.

Shape variations of the first four modes were investigated to analyze the influence of specific PCs on the SSM. Fig. 5 shows the mean correspondence positions from the model

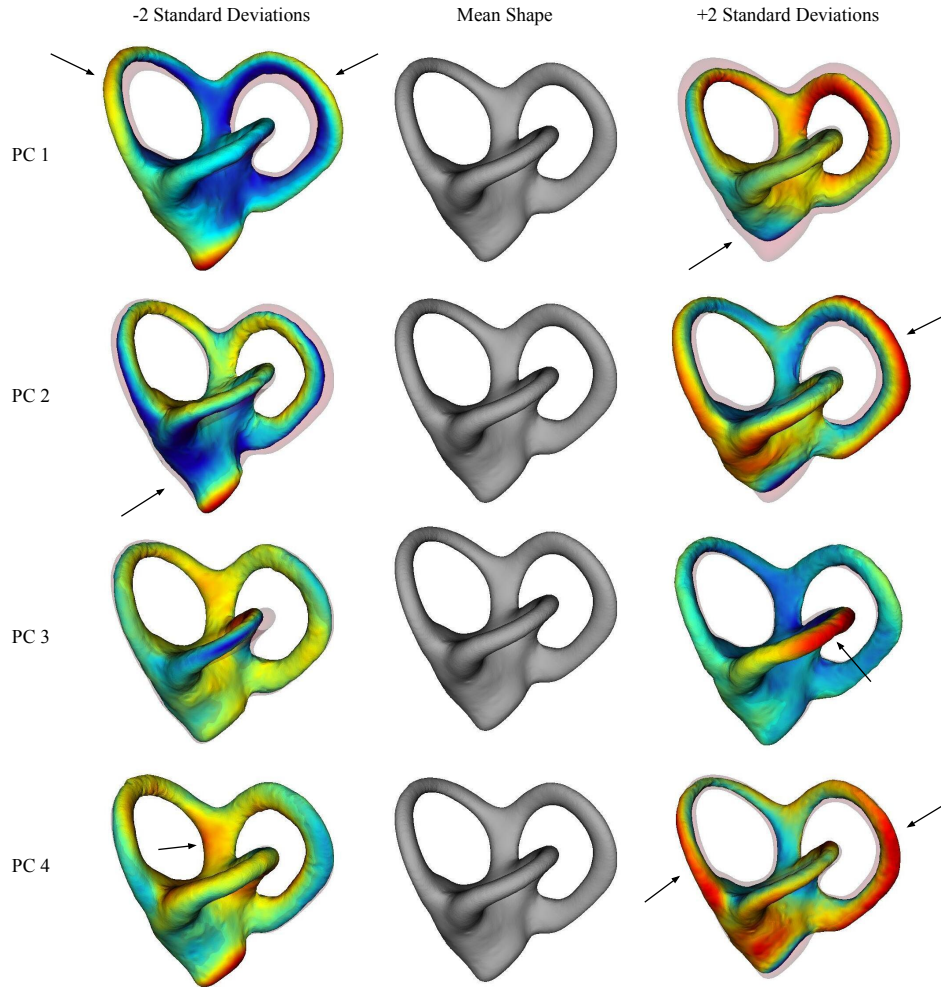


Figure 5. Influence of different PCs on the shape variation. The mean shape is in the center. In the left and right columns, deformation is represented as a color map, and the mean shape is visualized as opaque. The color maps represent the distance for any value inside the geometry with a negative value (blue) and outside the geometry with a positive value (red). The arrows highlight some relevant parts of the shape variations.

moved along each of the top four PCA modes. SCC shapes along each mode are reconstructed from the learned PBM model parameters at  $-2$  to  $+2$  standard deviations from the mean.

From Fig. 5, we can interpret that the first PC causes size changes of the SCC. Variation in a positive direction describes a shrinkage, whereas changes in the negative direction result in an enlargement of the SCC. The extent of variation in the lateral canal is less compared to the posterior and superior canals. By looking at Fig. 5, PC2 affects the area where the superior and lateral canals converge. Moving in a positive direction leads to an enlargement of the posterior canal. PC1 and PC2 cause size variations in the vestibule area. From  $-2$  to  $+2$ , the vestibule area results in

a shrinking and vice-versa. PC3 mainly captures changes in the middle part of the lateral semicircular. PC4 primarily influences the area of the posterior semicircular canal and the superior semicircular canal. Looking at Figure 6, we can tell that the shape outlier S2 from the plot in Figure 5 has a large vestibule area and a large size of the canals.

### 3.4. Shape model evaluation

A substantial part of the creation of SSM is to validate the results. Intuitively a first qualitative approach is the visual inspection of the shape instances that the model is able to create. When shapes have point-to-point correspondence, an SSM is evaluated using more objective accepted measures, namely Generalization, Specificity, and

Compactness, which are considered as useful benchmarks for measuring correspondence quality [9,22].

Figure 3(b,c,d) shows the evaluation of the SSM concerning compactness, generalization, and specificity with an increasing number of modes of variation included. Briefly, the generalization measures the model's ability to represent unseen shape instances of the class. It is performed using leave-one-out cross-validation reconstruction experiments. The generalization error is expected to decrease with an increasing number of model parameters. Specificity measures whether the model can generate instances of an object close to those presented in the training set. It is measured by generating a large number of  $N$  random instances ( $N = 1000$  in our experiments) using different modes. For every new sample, compute the distance to the closest shape in the training set. The mean distance error is expected to increase with more parameters, as the increasing number of PCs gives more flexibility to shape reconstruction. The compactness of the model is the ability to use as few parameters as possible to represent more shape instances in the training sets. Compactness is defined as the cumulative variance of the  $M$  largest modes.

### 3.5. Reconstruction accuracy

The reconstruction accuracy of the model has been evaluated by computing the mean surface distance between approximated model instances and input segmentations to ensure that each shape in the training set is well represented. The landmarks of the shape model constitute a point cloud. To represent an instance of the training data, the point cloud should cover the important part of the shape. The original mesh is obtained from the distance transform created from the initial manual segmentation and then compared to a mesh reconstructed from the predicted PBM. We compute the Hausdorff distance (mm) that takes the max of these vertex-wise distances to return a single value as a measure of accuracy [15]. The results after computing the Hausdorff distance range from 1.15 to 4.91 mm. The mean surface-to-surface distance was 2.42 mm (0.87 Std. Dev.).

## 4. Discussion and conclusion

This study aimed to explore and analyze the shape variations of the vestibular system for further application of the electrode placement for VI. To control design and implant variables, having realistic and detailed computational models of the SCC are needed, including population variability. This work describes the first stage, having the model which can be used currently for design decision support of an implant.

An important aspect in this work was the use of data sets that are acquired in clinical practice. On the other hand, the quality of the data, especially the resolution and contrast of the scans and the accuracy of the manual segmen-

tation, was a major concern. In some samples, the spacing between voxels is so low that the SCC consists of a single voxel across the entire diameter. In addition, due to the different voxel spacings for all images, resampling the volumes introduces even more artifacts. A challenge during the construction of the shape models is the methodology for creating point correspondences between the data. The initial shape model contained imperfections due to bad correspondences, which was alleviated with the application of a smoothing filter. Several experiments were carried out to establish the parameters to obtain an optimal SSM of the SCCs. Increasing the number of particles above 4096 does not significantly improve the shape representation of the model in the sense that no additional anatomical details become visible, but increases model complexity and computational time, especially when modeling such a complex structure as the vestibule and larger datasets. Therefore, a balance between a good representation and the number of particles is necessary. In general, the rest of the parameters involved in the optimization do not significantly affect the final model for this dataset.

The analysis of the shape variation based on the principal modes could help to find some outliers. Of course, it is mainly a proof-of-concept since the model is built with a small number of datasets, and therefore the representation of actual anatomy is not proved completely. Nevertheless, the accuracy tests have shown that the generated model based on the 71 segmentations approximated the shape of the vestibular system with reasonable accuracy. A lower generalization and specificity error is desirable for an ideal shape model, but Fig. 3(c, d) indicates that they move in opposite directions with an increasing number of model components. The compactness is also essential to guarantee that most of the shape variation is captured by the model using as few model parameters as possible. So how many components should be used to represent 90% or more of the shape variation is still a very interesting question when dealing with biological data, and a trade-off between these three metrics is necessary. In our model, 17 PCs are sufficient to represent 90% of the variation. For representing more than 95%, the gains in compactness and generalization are very light after 30 PCs, and there is a diminishing penalty in specificity as the number of components in the model increases. This flattening of the curve mainly occurs between 20 and 30 components. With more than 30 components used, the model constructed has the best performance, but also more noisy shape variation is introduced, and more computation is required to fit our models.

## References

- [1] Dean C. Barratt, Carolyn S.K. Chan, Philip J. Edwards, Graeme P. Penney, Mike Slomczykowski, Timothy J. Carter, and David J. Hawkes. Instantiation and registration of sta-

- tistical shape models of the femur and pelvis using 3d ultrasound imaging. *Medical Image Analysis*, 12(3):358–374, jun 2008.
- [2] Andrew P. Bradshaw, Ian S. Curthoys, Michael J. Todd, John S. Magnussen, David S. Taubman, Swee T. Aw, and G. Michael Halmagyi. A mathematical model of human semicircular canal geometry: A new basis for interpreting vestibular physiology. *Journal of the Association for Research in Otolaryngology*, 11(2):145–159, 6 2010.
- [3] Joshua Cates, Shireen Elhabian, and Ross Whitaker. Chapter 10 - shapeworks: Particle-based shape correspondence and visualization software. In *Statistical Shape and Deformation Analysis*, pages 257–298. Academic Press, 3 2017.
- [4] Joshua Cates, P. Thomas Fletcher, Martin Styner, Heather Cody Hazlett, and Ross Whitaker. Particle-based shape analysis of multi-object complexes. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, volume 5241, pages 477–485. Springer Berlin Heidelberg, 2008.
- [5] Joshua Cates, P. Thomas Fletcher, Martin Styner, Martha Shenton, and Ross Whitaker. Shape modeling and analysis with entropy-based particle systems. In *Information Processing in Medical Imaging*, pages 333–345. Springer Berlin Heidelberg, 2007.
- [6] Juan Cerrolaza, Sergio Vera, Alexis Bagué, Mario Ceresa, Pablo Migliorelli, Marius George Linguraru, and Miguel Ángel González Ballester. Hierarchical shape modeling of the cochlea and surrounding risk structures for minimally invasive cochlear implant surgery. In *Clinical Image-Based Procedures. Translational Research in Medical Imaging*, pages 59–67. Springer International Publishing, 2014.
- [7] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, jan 1995.
- [8] Iván Cortés-Domínguez, María A. Fernández-Seara, Nicolás Pérez-Fernández, and Javier Burguete. Systematic method for morphological reconstruction of the semicircular canals using a fully automatic skeletonization process. *Applied Sciences*, 9(22):4904, nov 2019.
- [9] Rhodri Huw Davies. *Learning shape: optimal models for analysing natural variability*. PhD thesis, University of Manchester Manchester, 2002.
- [10] Thomas Demarcy. *Segmentation and study of anatomical variability of the cochlea from medical images*. PhD thesis, Université Côte d’Azur, 2017.
- [11] K. D. Fritscher, P. F. Raudaschl, M. Handler, C. Baumgartner, L. Johnson, A. Schrott-Fischer, R. Glueckert, R. Saba, and R. Schubert. Towards a framework for thesegmentation and statistical shape analysis of the vestibular system using micro-ct. In *Shape Symposium 2015, Delémont, Switzerland*, 2015.
- [12] Nicolas Gerber, Mauricio Reyes, Livia Barazzetti, Hans Martin Kjer, Sergio Vera, Martin Stauber, Pavel Mistrik, Mario Ceresa, Nerea Mangado, Wilhelm Wimmer, Thomas Stark, Rasmus R. Paulsen, Stefan Weber, Marco Caversaccio, and Miguel A. González Ballester. A multi-scale imaging and modelling dataset of the human inner ear. *Scientific data*, 4(1):1–12, 2017.
- [13] Nils Guinand, Frans Boselie, Jean-Philippe Guyot, and Herman Kingma. Quality of life of patients with bilateral vestibulopathy. *Annals of Otolaryngology, Rhinology & Laryngology*, 121(7):471–477, 7 2012.
- [14] Jean-Philippe Guyot and Angelica Perez Fornos. Milestones in the development of a vestibular implant. *Current Opinion in Neurology*, 32(1):145–153, feb 2019.
- [15] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.
- [16] Hans Martin Kjer, Jens Fagertun, Sergio Vera, Miguel Angel González Ballester, and Rasmus Reinhold Paulsen. Shape modelling of the inner ear from micro-ct data. In *Shape Symposium*, volume 1992, 2014.
- [17] J. H. Noble, R. F. Labadie, O. Majdani, and B. M. Dawant. Automatic segmentation of intracochlear anatomy in conventional ct. *IEEE Transactions on Biomedical Engineering*, 58(9):2625–2632, 9 2011.
- [18] NORD. Benign paroxysmal positional vertigo. <https://rarediseases.org/rare-diseases/benign-paroxysmal-positional-vertigo>, 2020. Accessed 23 September 2021.
- [19] Patrik Raudaschl and Karl Fritscher. Statistical shape and appearance models for bone quality assessment, 2017.
- [20] G.K. Rohde, A. Aldroubi, and B.M. Dawant. The adaptive bases algorithm for intensity-based nonrigid image registration. *IEEE Transactions on Medical Imaging*, 22(11):1470–1479, nov 2003.
- [21] Julia A. Schnabel, Daniel Rueckert, Marcel Quist, Jane M. Blackall, Andy D. Castellano-Smith, Thomas Hartkens, Graeme P. Penney, Walter A. Hall, Haiying Liu, Charles L. Truwit, Frans A. Gerritsen, Derek L. G. Hill, and David J. Hawkes. A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001*, pages 573–581. Springer Berlin Heidelberg, 2001.
- [22] Martin A. Styner, Kumar T. Rajamani, Lutz-Peter Nolte, Gabriel Zsemlye, Gábor Székely, Christopher J. Taylor, and Rhodri H. Davies. Evaluation of 3d correspondence methods for model building. In *Information Processing in Medical Imaging*, volume 2732, pages 63–75. Springer Berlin Heidelberg, 2003.
- [23] Daniel Q. Sun, Bryan K. Ward, Yevgeniy R. Semenov, John P. Carey, and Charles C. Della Santina. Bilateral vestibular deficiency: quality of life and economic implications. *JAMA Otolaryngology–Head & Neck Surgery*, 140(6):527, 6 2014.
- [24] Michael von Brevern and Hannelore Neuhauser. Epidemiological evidence for a link between vertigo and migraine. *Journal of Vestibular Research*, 21(6):299–304, 2011.
- [25] Matthias Willenbrink. Segmentation of the vestibular organ with statistical shape models and deep neural networks. Master’s thesis, UMIT Private Universität für Gesundheitswissenschaften, Medizinische Informatik und Technik GmbH, 2021.



[26] Jennifer Wiperman. Dizziness and vertigo. *Primary Care: Clinics in Office Practice*, 41(1):115–131, mar 2014.

# One-Pixel Instance Segmentation of Leaves

Julia Strebl\*, Eric Stumpe\*, Thomas Baumhauer, Lena Kernstock, Markus Seidl, Matthias Zeppelzauer  
Institute of Creative Media Technologies, St. Pölten University of Applied Sciences

{firstname}.{lastname}@fhstp.ac.at

## Abstract

The segmentation of plant leaves is an essential prerequisite for vision-based automated plant phenotyping applications like stress detection, measuring plant growth and detecting pests. Segmenting plant leaves is challenging due to occlusions, self-shadows, varying leaf shapes, poses and sizes and the presence of particularly fine structures. We present a novel leaf segmentation approach that takes single pixels as input to initialize the segmentation of leaves. Additionally, we introduce a new strategy for transfer learning that we call “tandem learning” which enables the integration of previously learned network representations into a structurally different network. We evaluate different configurations of our approach on publicly available data sets and show that it yields competitive segmentation results compared to more complex segmentation approaches.

## 1. Introduction

Plant phenotyping refers to methodologies for the characterization of plants, i.e., plant architecture and composition at different scales [4]. This includes the visual assessment of plant traits to investigate plant growth, plant state and plant stress [11]. The manual assessment of these properties from visual observation is an expensive and tedious process. Phenotyping at larger scales thus requires automated methods for the quantification of plant traits. Computer vision approaches can solve plant phenotyping problems at large scales in a non-invasive manner. Thereby, automated leaf segmentation is an essential prerequisite for many downstream tasks including leaf counting, leaf/plant tracking and the detection of plant stress, diseases and pests.

Leaf segmentation is an instance segmentation problem [7], where the goal is to pixel-accurately segment objects of the same type (here leaves). Plants pose a number of challenges to this task including (i) coping with complex background (e.g., from soil visible in the images, trunks, branches etc.); (ii) handling fine structures (e.g., the stems

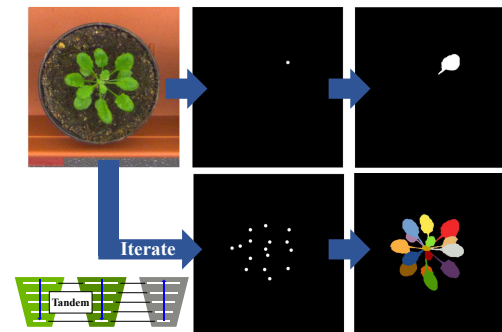


Figure 1. One-pixel instance segmentation: our approach first learns to estimate useful seed points for leaf segmentation and then segments leaves from these seed points via *tandem learning*, a more flexible form of traditional transfer learning.

of the leaves); (iii) solving occlusion problems introduced by overlapping leaves; (iv) coping with differently sized and shaped leaves (e.g., due to different ages) and different leaf poses; and (v) handling shadowing and varying reflectivity of differently oriented leaves [19].

In this paper, we present a simple and thus robust leaf segmentation approach that achieves promising segmentation results on established benchmark data sets. Our approach is anchor-free and thus makes no *a priori* assumptions about leaf size and shape and can principally learn arbitrary leaf shapes. In our approach we introduce two novel concepts for instance segmentation (see also Figure 1):

- *One-pixel segmentation*: a form of instance segmentation that requires only minimal input, i.e., a single seed pixel to segment an object instance. One-pixel segmentation makes our approach equally suitable for fully automated and interactive segmentation, which is usually hard for fully end-to-end trained methods.
- *Tandem learning*: a new form of transfer learning that helps to incorporate existing knowledge captured in a pre-trained network in a novel task that requires a *structurally different* network architecture.

We design and evaluate different configurations of our ap-

\*both authors contributed equally to this paper

proach and perform ablation studies to evaluate the influence of the individual processing steps.

## 2. Related Work

Segmentation methods can be split into anchor-based and anchor-free approaches. Here, we review both types to place our approach in context. We further review related methods that inspired our approach.

**Anchor-Based Instance Segmentation.** A common strategy for instance segmentation is the utilization of predefined anchor boxes for generating region proposals. A popular network of this category is Mask-RCNN [8]. In Mask-RCNN, first image features are extracted, followed by the prediction of object classes and Regions of Interest (RoIs), which is facilitated by the initial anchor boxes. In a second step, segmentation masks are predicted from the proposed RoIs. Huang et al. [9] introduced a separate Intersection over Union (IoU) prediction branch to Mask-RCNN to increase performance. Liu et al. [13] further improved the architecture by using a bottom-up path augmentation scheme for the extraction of image features. Other follow-up works focus on aspects such as inference speed [2] or object border refinement [10]. To get optimal results for different types of image data sets, preset anchor boxes and their dimensions have to be adapted to the dimensions of the target objects. Since our method does not require anchors it is not subject to this restriction.

**Anchor-Free Instance Segmentation.** Tian et al. [22] demonstrated an effective method for object detection that does not require the use of anchor boxes. Instead, distances to the nearest bounding box and its dimensions are directly learned and represented as a 4D feature map. This work inspired other authors to adopt this method for region proposal-based instance segmentation. Bounding box-based methods in general work best for objects with similar height and weight, but can fail for elongated objects that overlap as demonstrated in [3]. Consequently a strand of research has evolved using different working principles to avoid this issue. Bai and Urtasun [1] predict the per-pixel angle to the nearest object border, enabling the segmentation of instances through their computed watershed energy level. De Brabandere et al. [3] formulate instance segmentation as a per-pixel problem, where the discriminative loss function enforces pixels of the same object to be close in latent space. Our work falls into the group of anchor-free instance segmentation methods and uses automatically estimated seed points in combination with a trained instance model to iteratively segment leaves.

**Leaf Instance Segmentation.** Gomes and Zheng [5] adopted a standard Mask-RCNN architecture for leaf segmentation and demonstrated that leaf masks of high quality can be predicted by employing simpler strategies, such as threshold adjustment and test time augmentations. To simulate the counting process of humans, Ren and Zemel [16] utilized a recurrent neuronal network (RNN), which sequentially proposes new regions of interest based on an attention mechanism. Guo et al. [6] devised a multi-scale attention module and mask refining module to improve the segmentation quality of their instance segmentation model. Wolny et al. [24] introduced a technique, which can also deal with sparsely labelled instance annotations and is based on the pixel embedding method in [3]. Feeding perturbations of the same input image to two embedding networks, a penalty is applied if both predicted masks are not geometrically consistent, thus enforcing constraints for the embedding space leading to better segmentation accuracy. In contrast to existing methods, our network architecture is more simple and straightforward and works well with already established loss functions such as binary cross-entropy.

**Interactive Instance Segmentation.** We further draw inspiration from interactive segmentation approaches. In recent methods, users can draw positive and negative object regions to guide the segmentation process [25], or are involved in a human-in-the-loop process where they actively annotate pixels of regions which are difficult to segment [20]. Lin et al. [12] developed an approach, in which interactive segmentation is guided by multiple user clicks with a focus on the first click acting as a segmentation anchor. Our goal for the future is to advance our method for efficient and low-effort interactive segmentation, which is facilitated by our one-pixel segmentation strategy.

## 3. Approach

An overview and illustration of our approach is shown in Figure 2. Below, we describe the individual steps in detail.

### 3.1. Data Preparation

Input data for our approach are RGB images of plants (see also Section 4). Additionally, for foreground segmentation we use binary segmentation masks as ground truth. For instance segmentation, we use masks including individual leaf annotations (multi-labeled ground truth masks). The training of the leaf instance segmentation model further requires the computation of masks, which specify the center for each leaf. In these masks the center pixel is highlighted by a value of 1, while all other values are 0. These masks can easily be created from the multi-labeled ground truth masks by applying e.g., distance transform and peak detection on each instance’s area.

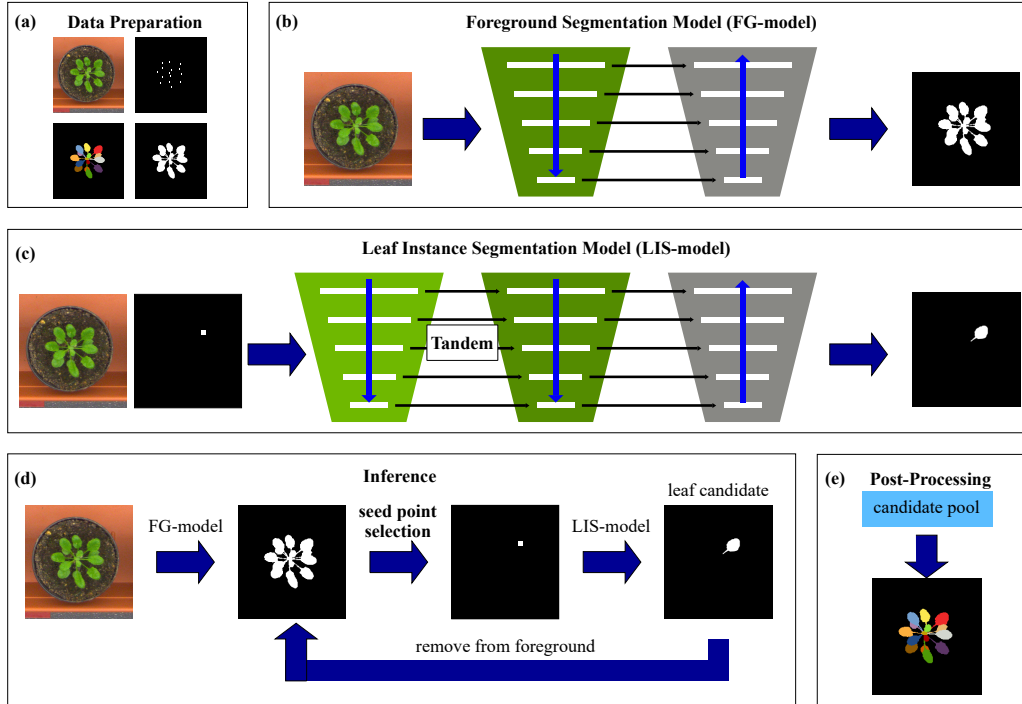


Figure 2. Overview of our leaf instance segmentation approach. First, we identify relevant image regions corresponding to leaves of the plant via semantic segmentation, see “FG-model” in (b). The result is a binary segmentation that captures the entire leaf tissue. From this segmentation we estimate potential leaf centers, which serve as seed points for *one-pixel instance segmentation*. The seed points are added as additional input channel to the leaf instance segmentation model, see “LIS-model” in (c). Using the proposed *tandem learning* scheme, a pre-trained encoder is incorporated into the LIS-model to accelerate training. The LIS-model segments one leaf at a time and is iteratively called to successively segment all leaves of the plant (d). Post-processing (e) consolidates the individual instance segments.

## 3.2. Training

### 3.2.1 Foreground Segmentation

For foreground segmentation we employ an encoder-decoder architecture with skip connections, similar to U-Net [17]. A pre-trained VGG16 backbone [21] serves as encoder [18]. The architecture of the decoder mirrors that of the VGG backbone, but instead of max-pooling layers we use up-convolutional layers (4 layers) to bring the feature maps back to the input image dimensions. In addition, the decoder receives feature maps through skip connections which are thereby incorporated in the training process. We use RGB images as input, binary segmentation masks as learning target, and binary cross-entropy as loss function.

### 3.2.2 Leaf Instance Segmentation

The LIS-model is also based on the U-Net architecture [17] from Section 3.2.1, but has two encoders A and B (see Figure 3) which are connected side-by-side in a tandem. Both encoders compute feature maps at different scales, which are concatenated with each other along the depth dimension. This architecture, which we call a “*tandem architecture*” en-

ables to combine network models (here two encoders) designed for different types of inputs.

As Encoder A we use VGG16 [21], which has been fine-tuned during foreground segmentation and takes three-channel RGB images as input. Therefore, the network is already capable of extracting meaningful plant-related features from RGB images. Encoder B receives images with a channel size of 4: the RGB channels plus the center point mask of a given leaf instance. Since no pre-trained model exists for this type of input, the model is initialized with random weights. Encoder B is further connected to the decoder in the same fashion as in the U-Net architecture [17]. All layers of the tandem network are fine-tuned/trained.

The tandem architecture should foster the integration of previously learned knowledge into a new learning task, which requires a different input (and potentially output) structure. This architecture is more flexible than standard transfer learning where usually the input is required to be equivalent and only the output layer is adapted. Additionally, it enables to combine two simple network architectures (VGG and U-Net) avoiding the need for a more complex (and more difficult to train) architecture.

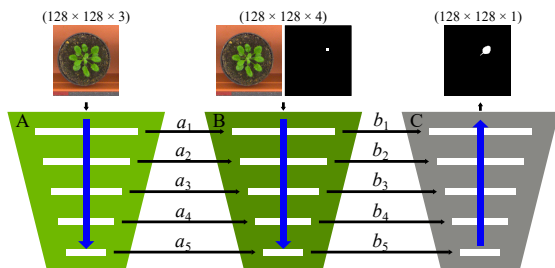


Figure 3. The concept of *tandem learning*: two encoders A and B are connected side by side. Thereby, A and B may have different input structure. Via connections  $a_i$  pre-learned information from A is shared with B. The final output is generated by decoder C.

Essential for training the network is data augmentation. Aside from conventional image transforms (see Section 4.3), we adjust the fourth input channel to make the network less dependent on the actual leaf center location. We propose two augmentation methods. First, instead of taking the leaf mask with the exact leaf center, a random pixel from the area of the leaf is taken. Second, starting from the exact center we specify a radius  $r$  that is increased by one pixel with each epoch. For augmentation, pixels are chosen at random that lie within this increasing radius. This facilitates location invariance in the optimization.

### 3.3. Inference

The goal of inference is to utilize both trained models in a combined manner to segment all leaves in an input image in absence of ground-truth. First, the foreground mask of the whole plant is computed with the FG-model. From this segmentation, we estimate potential center points automatically to initiate leaf instance segmentation. To select appropriate seed candidates, we propose two methods:

**Distance transform (DT) selection (sorted/unsorted):** First, morphological erosion is applied to the foreground mask to separate leaves that are loosely connected (i.e., touching each other). Next, the DT is computed for each connected region. The seed candidate is then selected at the location of the maximum value of the DT. Optionally, we sort the connected regions by area to start segmentation with the largest potential leaf.

**Gaussian kernel selection (sorted/unsorted):** The 2D convolution of the foreground segmentation with a Gaussian 2D kernel is computed. In the result image, pixels close to leaf borders have low values, since foreground (value 1) and background pixels (value 0) are in the effective range of the Gaussian kernel. Pixels in the center of leaves, however, yield high output values (only foreground in the effective range). We apply 2D peak detection to identify potential

leaf centers. The 2D Gaussian kernel has  $15 \times 15$  pixels and a sigma of 7. As in the first method, we optionally sort the connected regions by area.

Following the selection of seed candidates, the trained LIS-model is used to predict the leaf instance mask. Next, the segmented leaf is added to a pool of leaf candidates and the mask of this leaf is subtracted from the foreground mask. This assures that no seed candidates are selected in an already segmented area, which would lead to repeated segmentation of the same leaf. Inference repeats and keeps adding new leaf instances to the pool of leaf candidates until the foreground segmentation mask is empty.

### 3.4. Post-Processing

The result of leaf instance segmentation is a set of potentially overlapping leaf candidate regions. Noisy foreground segmentation may lead to oversegmentation (too many leaf candidates). Post-processing aims to compensate this by fusing only partially segmented leaves. We propose three strategies for consolidating leaf segments: (i) *deleting*, (ii) *merging* and (iii) *intersecting*. Thereby, all leaf candidate regions are compared via Intersection over Union (IoU) to estimate their mutual overlap. IoU is used as criterion to decide how to proceed with the two candidates as follows:

- Strategy *deleting* is based on the hypothesis that our leaf segmentation model performs better on large leaves. As soon as the IoU threshold for two candidate segments is exceeded, the smaller one is deleted.
- In *merging* two overlapping segments are joined together when their IoU is in a certain range. Hereby, we account for only partially detected leaves, i.e., cases where one leaf is over-segmented.
- In strategy *intersecting* only those leaf areas are preserved, which are supported by more than one candidate segment. This should help to increase the robustness of the segmentation.

The two latter methods facilitate the merging of leaves with a significant overlap and at the same time avoid that adjacent and touching leaves are merged.

## 4. Experimental Setup

### 4.1. Datasets

We employ publicly available data sets to facilitate performance comparisons with other methods. The first data set is subset “A1” from the *Plant Phenotyping Dataset (PPD)* introduced in [14, 15], which consists of 128 manually annotated images. To show how well our approach generalizes to other types of data and plants, we further evaluate our

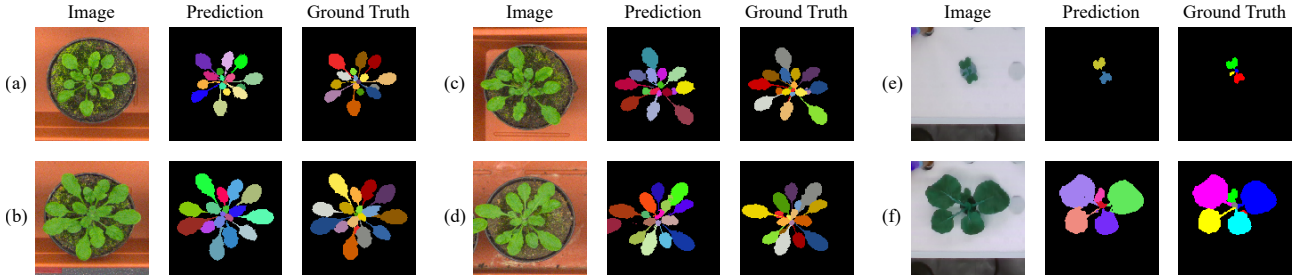


Figure 4. Instance segmentation results from our method for test images of the Plant Phenotyping Dataset (a-d) and KOMATSUNA (e,f).

method on the KOMATSUNA data set [23], that comprises 300 semi-automatically annotated images. Data has been split into 80% training and 20% testing for all experiments.

#### 4.2. Performance metrics

To assess training progress for both models in our approach we utilize Intersection over Union (IoU) and Dice similarity coefficient (DSC). As proposed by [14,15] the final instance segmentation results are measured with the Symmetric Best Dice (SBD) measure, which is particularly designed for instance segmentation problems and can cope with different but equivalent label assignments. All metrics in our experiments are averaged over three complete repetitions with different random initializations of the network weights.

#### 4.3. Parameters

Our approach has a number of hyperparameters and configuration options, which we evaluate in this paper. For object center estimation we evaluate both strategies from Section 3.3 with sorted and unsorted components. For post-processing we evaluate the three strategies from Section 3.4). For strategy deleting we apply an IoU threshold of 0.7, for merging an IoU range between 0.1 and 0.5 and for intersecting an IoU threshold of 0.5 (suitable parameter values were found via grid search in a preliminary experiment). The training parameters for the foreground segmentation network are as follows: training is conducted for 40 epochs with a learning rate of 0.00001 and batch size 20.

Downscaled RGB images of size  $128 \times 128 \times 3$  serve as the network input. For the LIS-model, training (input size  $128 \times 128 \times 4$ ) is initiated for 150 epochs with a batch size of 32 and a learning rate of 0.0001. For both models, binary cross-entropy and Adam optimizer are applied. To aid the learning process, we employ random geometrical (flipping, zooming, shifting, rotating, shearing) and color data augmentation (noise, brightness, contrast) in addition to the augmentation of leaf centers as described in Section 3.2.2.

### 5. Results

**Overall performance.** The overall instance segmentation performance of our approach in terms of SBD is shown in

Table 1. Additionally, we provide Dice and IoU for foreground segmentation and leaf instance segmentation. The highest scores for the PPD are achieved with center estimation via distance transform selection (no sorting) and post-processing via deleting strategy. Similarly, the highest scores for KOMATSUNA are achieved with distance transform selection (sorted) and deleting strategy. However, also Gaussian kernel selection and intersection strategy lead to the same peak performance, showing that the robustness of center estimation and post-processing strategy is high.

**Tandem training.** To evaluate the tandem architecture for transfer learning we perform an ablation experiment by removing the second encoder in the LIS model. The result is an average performance drop of 2.1% in SBD for the PPD and 1.8% for KOMATSUNA. We notice during our experiments that the training in tandem fashion leads to a faster and smoother convergence of the training loss compared to training without tandem. This shows that tandem learning is a suitable approach to take benefit of a previously learned representation, even if it has a different input structure.

**Instance center estimation.** Here, we evaluate the different leaf center estimation strategies from Section 3.3 systematically and the sensitivity of results to different choices. Results (see Table 2) show that Distance transform selection provides the highest performance across both data sets.

**Post-processing strategies.** Similarly, as above, we evaluate the different post-processing strategies introduced in Section 3.4. Table 2 shows their impact on overall results. We conclude that delete and intersection outperform post-processing via merging throughout all experiments.

**Performance comparison.** To objectively assess our results, we compare them with state-of-the-art results from the literature for both data sets, see Table 3 for a listing. For the PPD we achieve comparable scores to both De Brabandere et al. [3] and Ren and Zemel [16] and outperform the approaches reported in [19]. The most recent approaches still outperform our results, which may be due to the higher complexity of the approaches. An additional factor might

Data set	FG IoU	FG Dice	LIS IoU	LIS Dice	SBD
Plant Phenotyping	0.862 ( $\pm 0.0014$ )	0.928 ( $\pm 0.010$ )	0.819 ( $\pm 0.012$ )	0.882 ( $\pm 0.011$ )	0.832 ( $\pm 0.008$ )
KOMATSUNA	0.871 ( $\pm 0.006$ )	0.930 ( $\pm 0.003$ )	0.754 ( $\pm 0.033$ )	0.836 ( $\pm 0.030$ )	0.754 ( $\pm 0.005$ )

Table 1. Overall segmentation results of our approach for both evaluated data sets.

	Plant Phenotyping Dataset A1			KOMATSUNA		
	delete	merge	intersection	delete	merge	intersection
DTS unsorted	<b>0.831</b> ( $\pm 0.0032$ )	0.825 ( $\pm 0.0020$ )	<b>0.831</b> ( $\pm 0.0036$ )	0.719 ( $\pm 0.0155$ )	0.710 ( $\pm 0.0193$ )	0.712 ( $\pm 0.0169$ )
DTS sorted	0.808 ( $\pm 0.0037$ )	0.807 ( $\pm 0.0028$ )	0.807 ( $\pm 0.0034$ )	0.747 ( $\pm 0.0186$ )	0.738 ( $\pm 0.0189$ )	0.750 ( $\pm 0.0184$ )
GKS unsorted	0.787 ( $\pm 0.0029$ )	0.786 ( $\pm 0.0003$ )	0.789 ( $\pm 0.0029$ )	0.751 ( $\pm 0.0110$ )	0.739 ( $\pm 0.0095$ )	<b>0.754</b> ( $\pm 0.0119$ )
GKS sorted	0.775 ( $\pm 0.0016$ )	0.773 ( $\pm 0.0003$ )	0.775 ( $\pm 0.0006$ )	0.742 ( $\pm 0.0118$ )	0.734 ( $\pm 0.0120$ )	0.744 ( $\pm 0.0116$ )

Table 2. Systematic comparison results for different center estimation strategies (distance transform selection (DTS) sorted/unsorted, Gaussian kernel selection (GKS) sorted/unsorted) and post-processing strategies (delete, merge, intersection).

Method	PPD A1	KOMATSUNA
Scharr et al. [19] (IPK)	0.744	
Scharr et al. [19] (Nottingham)	0.683	
Scharr et al. [19] (MSU)	0.667	
Scharr et al. [19] (Wageningen)	0.711	
De Brabandere et al. [3]	0.849	
Ren and Zemel [16]	0.842	
Gomes and Zheng [5]	0.920	0.745
Guo et al. [6]	0.925	
Wolny et al. [24]	0.920	

Table 3. SBD scores of different methods on A1 subset of the Plant Phenotyping Dataset (PPD) and the KOMATSUNA Dataset.

be that the reported results stem from the leader board<sup>1</sup> and are not 100% comparable as we test our approach on a 20% subset of the training set, while the performance in the leader board refers to a separate test set (for which no labels were available for our experiments). For the KOMATSUNA data set we could identify only one approach [3] for comparison in the literature (see Table 3). The performance obtained by our approach with an SBD of 0.754 slightly outperforms the previously reported result of 0.745.

**Qualitative results** In Figure 4, exemplary segmentation results for the test sets of the Plant Phenotyping Dataset (a-d) and KOMATSUNA (e,f) are shown. Overall, most separate leaves are segmented accurately and leaf edges are very closely aligned to the ground truth. In (c) and (g) it can be seen that some very small leaves in the center are not correctly segmented. Sometimes also leaves, which are largely covered by other leaves are not segmented well (see leaves in the lower area of (b) and (d)). Examples in (e) and (f)

<sup>1</sup><https://competitions.codalab.org/competitions/18405#results>

show that leaves with different size and shapes can be segmented well. Remarkable is further that in (d) a leaf of an neighboring plant is correctly segmented, although it is not part of the annotated ground truth.

**Limitations** Our approach works slightly better for larger objects than for small ones. The reason is that large objects generate more (overlapping) segment candidates, which can be better consolidated and refined via post-processing. Our one-pixel segmentation approach functions well for the segmentation of instances that consist of a single connected region, but can fail for instances that are fragmented (e.g., a leaf that is intersected by the petiole of another leaf, and thus consists of two separate regions).

## 6. Conclusion

We have presented a novel approach for leaf instance segmentation which uses individual pixels indicating object centers as seed points for instance segmentation. Our approach yields promising results on public benchmark data sets and can compete with much more complex segmentation approaches from literature. Since our approach makes no *a priori* assumptions about the structure, shape and pose of plant leaves, it may be applicable to other instance segmentation tasks and thus may be of broader interest to the community. The same applies to the *tandem training* that we use for transfer learning. Future work will focus (i) on predicting leaf centers during foreground segmentation to replace leaf center estimation during inference and (ii) on demonstrating the broader applicability of the proposed approach for other instance segmentation tasks.

## Acknowledgments

This work was funded by the research promotion agency of the province of Lower Austria (GFF), proj. no. FTI18-005.

## References

- [1] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [3] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017.
- [4] Fabio Fiorani, Ulrich Schurr, et al. Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol*, 64(1):267–291, 2013.
- [5] Douglas Pinto Sampaio Gomes and Lihong Zheng. Leaf segmentation and counting with deep learning: on model certainty, test-time augmentation, trade-offs. *arXiv preprint arXiv:2012.11486*, 2020.
- [6] Ruohao Guo, Liao Qu, Dantong Niu, Zhenbo Li, and Jun Yue. Leafmask: Towards greater accuracy on leaf segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1249–1258, 2021.
- [7] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020.
- [8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [9] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6402–6411, Los Alamitos, CA, USA, 2019. IEEE Computer Society.
- [10] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [11] Zhenbo Li, Ruohao Guo, Meng Li, Yaru Chen, and Guangyao Li. A review of computer vision technologies for plant phenotyping. *Computers and Electronics in Agriculture*, 176:105672, 2020.
- [12] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13339–13348, 2020.
- [13] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.
- [14] Massimo Minervini, Andreas Fischbach, Hanno Scharf, and Sotirios A. Tsaftaris. Plant phenotyping datasets. <http://www.plant-phenotyping.org/datasets>, 2015.
- [15] Massimo Minervini, Andreas Fischbach, Hanno Scharf, and Sotirios A Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters*, 81:80–89, 2016.
- [16] Mengye Ren and Richard S. Zemel. End-to-end instance segmentation with recurrent attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [19] Hanno Scharf, Massimo Minervini, Andrew P French, Christian Klukas, David M Kramer, Xiaoming Liu, Imanol Luengo, Jean-Michel Pape, Gerrit Polder, Danijela Vukadinovic, et al. Leaf segmentation in plant phenotyping: a collation study. *Machine vision and applications*, 27(4):585–606, 2016.
- [20] Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1687–1697, 2021.
- [21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [22] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [23] Hideaki Uchiyama, Shunsuke Sakurai, Masashi Mishima, Daisaku Arita, Takashi Okayasu, Atsushi Shimada, and Rin-ichiro Taniguchi. An easy-to-setup 3d phenotyping platform for komatsuna dataset. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.
- [24] Adrian Wolny, Qin Yu, Constantin Pape, and Anna Kreshuk. Sparse object-level supervision for instance segmentation with pixel embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4402–4411, 2022.
- [25] Matthias Zeppelzauer, Georg Poier, Markus Seidl, Christian Reinbacher, Samuel Schuster, Christian Breiteneder, and Horst Bischof. Interactive 3d segmentation of rock-art by enhanced depth maps and gradient preserving regularization. *Journal on Computing and Cultural Heritage (JOCCH)*, 9(4):1–30, 2016.



# Towards Uncertainty Detection in Automated Leaf Tissue Segmentation

Ráchel Grexová<sup>‡</sup>, Klara Voggeneder<sup>✉</sup>, Danny Tholen<sup>✉</sup>, Guillaume Théroux-Rancourt<sup>✉</sup>,  
Walter Kropatsch<sup>‡</sup>, and Jiří Hladůvka<sup>‡</sup>

<sup>✉</sup>Institute of Botany, University of Natural Resources and Life Sciences, Vienna

{klara.voggeneder, daniel.tholen, guillaume.theroux-rancourt}@boku.ac.at

<sup>‡</sup>Pattern Recognition and Image Processing Group, Vienna University of Technology

{rachel.grexova, walter.kropatsch, jiri.hladuvka}@tuwien.ac.at

## Abstract

*In order to use segmented volumetric data for subsequent analyses, it is important to detect and understand, where the segmentation is reliable and where it is uncertain. This is especially critical in deep learning segmentation which relies on manually annotated ground truth. Especially in applications using medical and biological data, ground truth annotations are often sparse, imbalanced, and imprecise.*

*We propose to utilize 2.5D orthogonal ensembles not only to arrive at dense segmentation but, more importantly, to indicate areas of high prediction fidelity and areas of uncertainty.*

*Our ensemble achieved accuracy above 95% in the high fidelity areas of a volume of a poplar leaf segment. This accuracy was achieved not only for a fresh leaf sample similar to the training data, but also for a severely dehydrated sample. Well-represented classes contained large areas of high prediction fidelity and exhibited high validation metrics. By contrast, under-represented classes tend to contain large areas of uncertainty.*

*Indication of uncertainty could be used as a basis to revise the predictions by domain experts. This is in turn expected to improve and/or enlarge the ground truth and allows for training of higher-quality segmentation models.*

## 1. Introduction

Segmentation is crucial step for further biological [17] or biomedical analysis. Traditional approaches of image segmentation rely on homogeneity criteria such as intensity values (threshold) or large gradient magnitude (border line) [12]. Since MRI, CT or  $\mu$ -CT images are blurred, contain noise or have low contrast, it is more difficult to design such criteria in medical [18] or biological image segmentation. In these fields deep learning is increasingly gaining popularity [8] as the features are learned automatically. The

automatic feature learning is beneficial, but the filters important for the segmentation remain unknown, which makes it difficult to interpret and improve the results [15].

Deep-learning approaches rely on large ground truth training sets. Limited annotated data is a remaining challenge in medical imaging [5], but even more in botany and agriculture, where annotated image libraries are missing [13]. Moreover, any manual annotations are subject to inter- and intra-observed variability. In turn, such ground truth annotation often may become unreliable in hard-to-annotate areas.

In 2015 U-Net was introduced [14] and it has become one of the most commonly used architectures in (bio-)medical segmentation [18]. It was originally used for 2D transmitted light microscopy images. Since then it was used for nearly all major imaging modalities such as CT, MRI and X-ray [15]. The drawback of using 2D convolution for 3D data such as MRI, CT or  $\mu$ -CT is the lack of volumetric context [2]. There have been several extensions of U-Net [15], the 3D U-Net [20] being one of them. Due to high requirements on GPU memory of 3D convolutions [1] volumetric data is usually divided into smaller patches [5]. To overcome the drawbacks of 2D and 3D U-Nets, there have been several attempts to combine these approaches and run the 2D U-Net networks in parallel on several 2D projections of a 3D volume in order to incorporate some volumetric context at computationally efficient cost. This kind of ensemble U-Net is called 2.5D U-Net [15]. Usually the 3D volume is divided into 2D images along three orthogonal axes and then three U-Net models are trained and used for prediction separately. With fusion of the three predictions the final segmentation result is produced [4, 6, 11, 19]. Another possibility is to use random 2.5D U-Net with multiple 2D projections [2].

In this paper we utilize the 2.5D-like approach in order to localize the high fidelity predictions and to flag voxels with uncertain predictions. The aim is on one hand to address the problem with limited ground truth data typical for

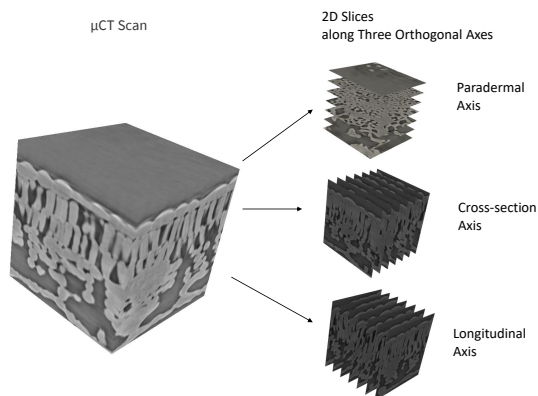


Figure 1.  $\mu$ -CT scan viewed as three orthogonal stacks of images.



Figure 2. Example of ground truth cross-section slice of scan time 3 showing all 6 available labels.

biological and (bio-)medical image segmentation. On the other hand we aim to build a tool that can enlighten how to fix errors of the predictions. The uncertain regions can be further reviewed by domain experts. This could enlarge the labeled data set, while significantly decreasing the manual labour. We present an approach that serves as the initial step for human-in-the-loop interactive segmentation.

## 2. Data

$\mu$ -CT scans of a hybrid poplar leaf were taken at the TOMCAT beamline at the Swiss Light Source of the Paul Scherrer Institute (Villigen, Switzerland) using acquisition protocols similar to [17]. The leaf was allowed to wilt and scanned in five different scan times. The first scan was done immediately after the leaf strip was prepared and placed into a holder. The other four scans were done after 10, 20,

25, and 30 minutes, while the leaf was dehydrating. While only minor differences in leaf structure were apparent during scan times 1-4, large differences were noticeable at time 5 and the cells were visibly shrunken.

The  $\mu$ -CT scans were divided into stacks of 2D slices along the three orthogonal axes (Fig. 1). Sparse set of 2D images were manually segmented into 6 classes, i.e., cells, veins, epidermis, stomata, background air, and intercellular airspaces (inner air).

This resulted in 10 to 25 segmented slices for each scan time and each axis. Two of the six classes have been heavily underrepresented: veins (5%) and the small pores on the surface, called stomata ( $\approx 1\%$ ).

## 3. Methodology

In this section we summarize the methodology of segmenting 2D slices along three orthogonal axes, orthogonal axes ensemble used for the selection of 3-consistent voxels and their evaluation.

### 3.1. 2D Segmentation Using U-Net

For 2D segmentation we divided the data into the training and validation sets. For the training set, we used scanning times 1, 2, and 4. For the validation set, the time 3 and (the challenging) time 5 were used with the aim to validate the models on a slightly different-looking dataset.

The models were trained and predicted using 3 different resolutions, i.e.  $1024 \times 1024$ ,  $512 \times 512$ ,  $256 \times 256$ . The models were trained using U-Net [14] architecture. As shown in Fig. 3 one model was trained for the paradermal axis and one for the cross- and longitudinal-section.

In order to address the problem with limited labeled ground truth data-set we used data augmentation [3]. We applied transformation functions such as random crop, flip, rotation both on the  $\mu$ -CT slices and their corresponding labeled ground truth slices simultaneously.

### 3.2. Orthogonal Axes Ensemble

The outputs of the three 2D predictions are aggregated in one 3-channel volume with 3 label predictions per voxel (see Figure 3). The number of unique labels per voxel splits the voxels into three categories:

1. all three models predicted consistently (3-consistent voxels);
2. two models were consistent, but inconsistent with the remaining one;
3. all three models were mutually inconsistent.

As no clear consensus is found for voxels of categories 2 and 3, we declare them as uncertain and call for a manual inspection. We'll discuss this later in section 5.

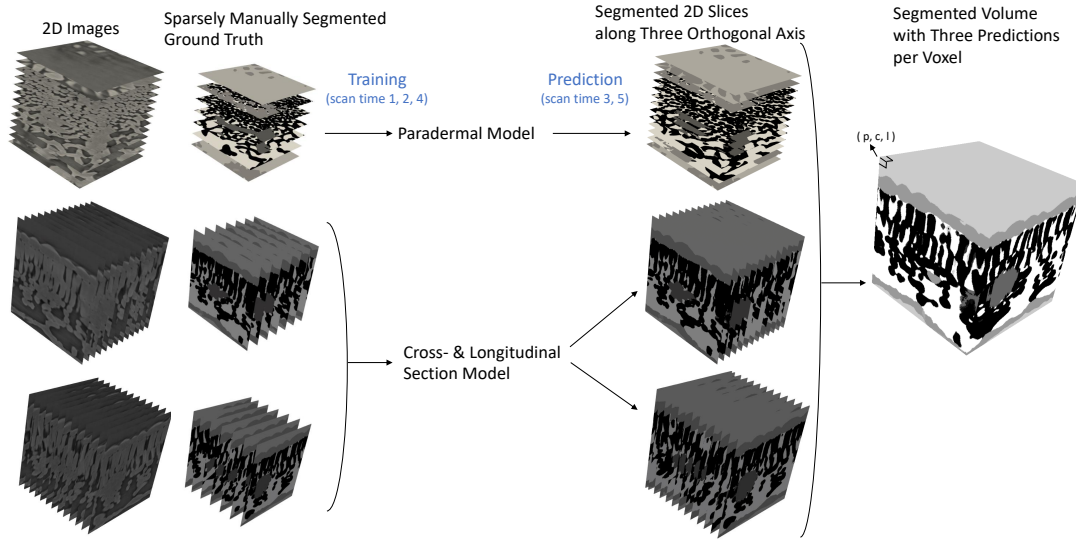


Figure 3. Training and prediction along the 3 orthogonal axes and their aggregation.

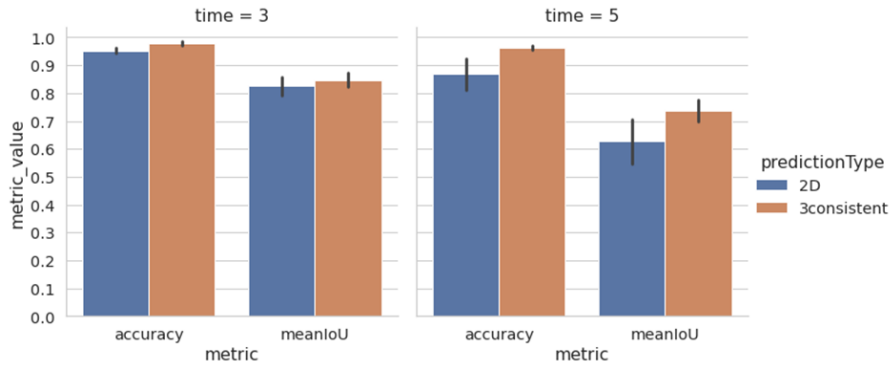


Figure 4. Mean IoU and accuracy for the test set (times 3 and 5). 2D predictions (averaged over all orthogonal axes and resolutions) compared to average of 3-consistent predictions. Black bars represent standard deviation.

In the following we are interested in how reliable are the predictions of category 1 with respect to the ground truth. To do so we compute and compare several metrics for both the ensemble and the three axis-wise 2D predictions.

### 3.3. Validation Metrics

Five spatial overlap-based metrics [16] are used for validation.

**Pixel Accuracy** (PA) is a basic metric used for segmentation evaluation. It is the ratio of correctly predicted pixels to the total number of pixels. [9]

**Precision** is used only for each label class separately:

$$precision = \frac{TP}{TP + FP} \quad (1)$$

where TP is the true positive fraction and FP is the false positive fraction [9]. Precision values indicate whether over-segmentation occurs [10].

**Recall** Similar to precision, recall is used only for each label class separately:

$$recall = \frac{TP}{TP + FN} \quad (2)$$

where TP is the true positive fraction and FN is the false negative fraction [9]. Recall values indicate whether under-segmentation occurs [10].

**Intersection over Union** (IoU, a.k.a the Jacard index [7]) is used both in the per-class and the image-mean variants.

**IoU** for individual class is defined as

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN} \quad (3)$$

where A is the mask of the class label in the ground truth image and B is the mask of the class label in the predicted image.  $|A \cap B|$  is the intersection and  $|A \cup B|$  is the union [9] [16].

**Mean IoU** is defined as the mean for the IoUs of the individual classes [9]. The mean IoU was calculated as

$$meanIoU = \frac{1}{n} \sum_{c=1}^n (IoU)_c \quad (4)$$

where  $n = 6$  is number of classes and  $(IoU)_c$  is the IoU for class  $c$ .

## 4. Results and Discussion

The metrics in 2D predictions were sufficiently high for the well represented classes, i.e. cells, epidermis, background air and intercellular airspace. In scan time 3, IoU, precision, and recall values were usually higher than 90%. In scan time 5 metrics were usually higher than 70% (Fig. 5). Since in scan time 5, the poplar leaf was much more dehydrated than during other scan times, it was expected that the predictions would be less accurate. Indeed the accuracy and mean IoU were lower for scan time 5 than for scan time 3 (Fig. 4).

For the under-represented classes, i.e., stomata and veins the precision was higher than recall. The recall was especially low for stomata for both scan times. This indicates under-segmentation. Therefore IoU was also low for stomata. The low IoU for underrepresented classes can explain why mean IoU is lower than accuracy for the 2D predictions.

After selection of 3-consistent voxels both accuracy and mean IoU increased in both scan times. The amount of voxels of this category was lower in time 5 ( $\approx 80\%$ ) than in time 3 ( $\approx 90\%$ ). The increase of the metrics values was higher for the scan time 5 than for scan time 3. Even though the average accuracy was 95.26% for scan time 3 and 86.82% for scan time 5, after selection of 3-consistent voxels the average accuracy was comparable, i.e. 97.83%, 96.16%, respectively (see Fig. 4). Mean IoU remained lower for scan time 5 than for scan time 3. The difference in the amount of uncertainty voxels for scan time 3 and 5 is demonstrated on an example in Figures 7b and 7d by yellow color.

Except for stomata in time 3 and both stomata and veins in time 5 the metrics increased class-wise. For the under-represented classes the recall was low and it got even lower

for the 3-consistent voxels (Fig. 5). Therefore also IoU was lower.

The low recall for stomata for 2D predictions can be observed in Figure 6 (f)-(h). Only some of the stomata were predicted by the particular models, but along each of the orthogonal axis it was predicted differently. In paradermal axis (f) only around half of the stomata were predicted, but when they were predicted it usually corresponded to the ground truth. This corresponds to small recall, but higher precision (see Fig. 5a). Additionally one of the stomata was predicted around hole visible in  $\mu$ -CT scan (a) and ground truth (e) near the stoma. Such an air gap between stoma and epidermis is highly unusual. In cross- (g) and long- (h) sections the number of predicted stomata is higher, but the shapes are slightly deformed. Therefore, as it is visible in Fig. 6 (b) - (d) the uncertainty depicted with yellow is high in stomata regions and 3-consistent voxels forms only small portion of stomata voxels in ground truth. Additionally around the uncertainty the 3-consistent voxels differs from the ground truth. This explains the decrease of recall after orthogonal axes ensemble. A similar pattern is visible for veins and stomata in scan time 5 (see Fig. 7c and Fig. 5b).

For well represented classes the metric values were for 3-consistent voxels usually above 90% for both scan times. In scan time 3 most voxels labeled as cells, intercellular airspace and background were labeled as their corresponding class in all 2D predictions. This is illustrated in Fig. 7b. In scan time 5 precision was significantly lower for inner air and stomata than in scan time 3 (see Fig. 5b). This indicates over-segmentation of these classes. For the 3-consistent voxels the precision significantly increased.

## 5. Conclusion and Future Work

We presented an approach that utilizes 2.5D orthogonal axis ensembles and detects areas of confidence and uncertainty. The validation metrics were higher for the 3-consistent voxels in comparison to the 2D predictions. For well represented classes, i.e. cells, epidermis, background and inner air, they were usually above 90% even for scan time 5, that was significantly more dehydrated in comparison to the training data-set.

Uncertainty areas tend to correlate with the under-represented classes, i.e. stomata and veins. Here, small recall was typical in 2D predictions, indicating under-segmentation of these classes. For the classes with large uncertainty areas, under-segmentation remained also for the orthogonal axes ensemble.

In future work this approach could be used as an initial step in a human-in-the-loop segmentation, where the uncertainty areas can be revised.

In Figure 8 we show an example of prediction by orthogonal axes ensemble overlaid by uncertainty (yellow)

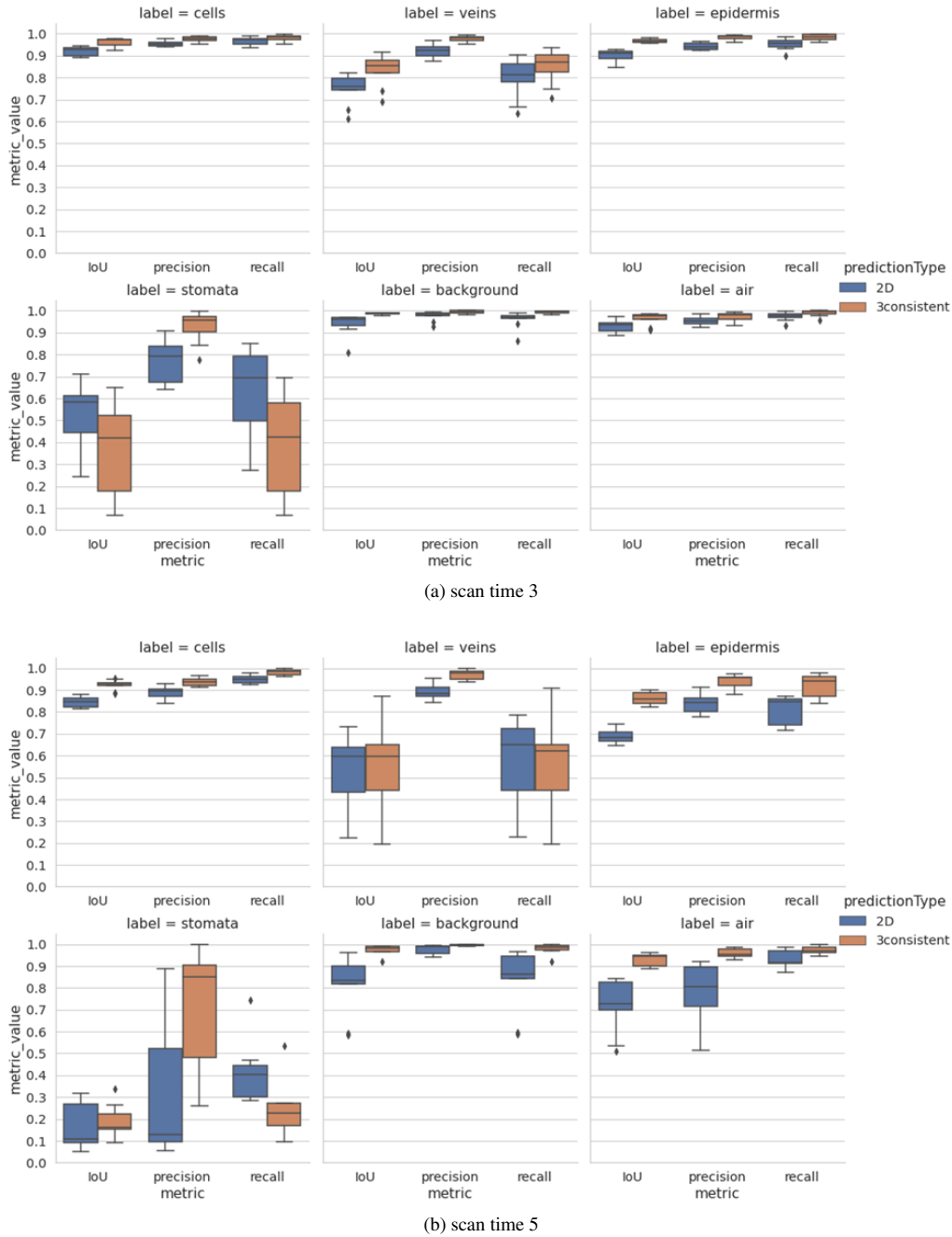


Figure 5. Label classes comparison of 2D predictions and 3-consistent voxels for scan time 3 (a) and scan time 5 (b).

for a slice *without* the ground truth. An increasing opacity can become a part of an interactive tool for revision of predictions irrespective of absence/presence of a ground truth. Such a revision can in turn enrich the training set.

In Figure 7a shows 3-consistent voxels of the veins surrounded by yellow uncertainty area and several orange spikes. Because it is hard even for a human expert to distin-

guish cells closely appressed to the veins, such cells were annotated as veins. Our approach actually correctly annotated these cells, leading to the orange spikes. This shows our approach can help to identify areas that are hard to manually label to improve the ground truth data.

**Acknowledgments** We acknowledge the Paul Scherrer Institut, Villigen, Switzerland for provision of beamtime at the TOM-

CAT beamline of the Swiss Light Source. The computational results have been partially achieved using the Vienna Scientific Cluster (VSC). This work was supported by the Vienna Science and Technology Fund (WWTF) project LS19-013 and by the Austrian Science Fund (FWF) projects M2245 and P30275.

## References

- [1] Christoph Angermann and Markus Haltmeier. Random 2.5D U-net for Fully 3D Segmentation. In Hongen Liao, Simone Balocco, Guijin Wang, Feng Zhang, Yongpan Liu, Zijian Ding, Luc Duong, Renzo Phellan, Guillaume Zahnd, Katharina Breininger, Shadi Albarqouni, Stefano Moriconi, Su-Lin Lee, and Stefanie Demirci, editors, *Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting*, Lecture Notes in Computer Science, pages 158–166, Cham, 2019. Springer International Publishing.
- [2] Christoph Angermann, Markus Haltmeier, Ruth Steiger, Sergiy Pereverzyev, and Elke Gizewski. Projection-Based 2.5D U-net Architecture for Fast Volumetric Segmentation. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–5, July 2019.
- [3] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Alumentations: Fast and Flexible Image Augmentations. *Information*, 11(2):125, Feb. 2020. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.
- [4] Lin Han, Yuanhao Chen, Jiaming Li, Bowei Zhong, Yuzhu Lei, and Minghui Sun. Liver segmentation with 2.5D perpendicular UNets. *Computers & Electrical Engineering*, 91, May 2021.
- [5] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*, 32(4):582–596, Aug. 2019.
- [6] Ke Hu, Chang Liu, Xi Yu, Jian Zhang, Yu He, and Hongchao Zhu. A 2.5D Cancer Segmentation for MRI Images Based on U-Net. In *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, pages 6–10, July 2018.
- [7] Paul Jaccard. The Distribution of the Flora in the Alpine Zone.1. *New Phytologist*, 11(2):37–50, 1912.
- [8] Priyanka Malhotra, Sheifali Gupta, Deepika Koundal, Atef Zaguia, and Wegayehu Enbeyle. Deep Neural Networks for Medical Image Segmentation. *Journal of Healthcare Engineering*, 2022:e9580991, Mar. 2022. Publisher: Hindawi.
- [9] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [10] Fernando C. Monteiro and Aurélio C. Campilho. Performance Evaluation of Image Segmentation. In Aurélio Campilho and Mohamed S. Kamel, editors, *Image Analysis and Recognition*, Lecture Notes in Computer Science, pages 248–259, Berlin, Heidelberg, 2006. Springer.
- [11] Gabriele Piantadosi, Mario Sansone, Roberta Fusco, and Carlo Sansone. Multi-planar 3D breast segmentation in MRI via deep convolutional neural networks. *Artificial Intelligence in Medicine*, 103:101781, Mar. 2020.
- [12] Bernhard Preim and Charl P Botha. *Visual computing for medicine: theory, algorithms, and applications*. Newnes, 2013.
- [13] Devin A. Rippner, Pranav V. Raja, J. Mason Earles, Mina Momayyezi, Alexander Buchko, Fiona V. Duong, Elizabeth J. Forrestel, Dilworth Y. Parkinson, Kenneth A. Shackel, Jeffrey L. Neyhart, and Andrew J. McElrone. A workflow for segmenting soil and plant X-ray computed tomography images with deep learning in Google’s Colaboratory. *Frontiers in Plant Science*, 13, 2022.
- [14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing.
- [15] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, 9:82031–82057, 2021.
- [16] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, Aug. 2015.
- [17] Guillaume Théroux-Rancourt, Matthew R. Jenkins, Craig R. Brodersen, Andrew McElrone, Elisabeth J. Forrestel, and J. Mason Earles. Digitally deconstructing leaves in 3D using X-ray microcomputed tomography and machine learning. *Applications in Plant Sciences*, 8(7):e11380, 2020.
- [18] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K. Nandi. Medical image segmentation using deep learning: A survey. *IET Image Processing*, 16(5):1243–1267, 2022.
- [19] Jie Wei, Yong Xia, and Yanning Zhang. M3Net: A multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation. *Pattern Recognition*, 91:366–378, July 2019.
- [20] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Lecture Notes in Computer Science, pages 424–432, Cham, 2016. Springer International Publishing.

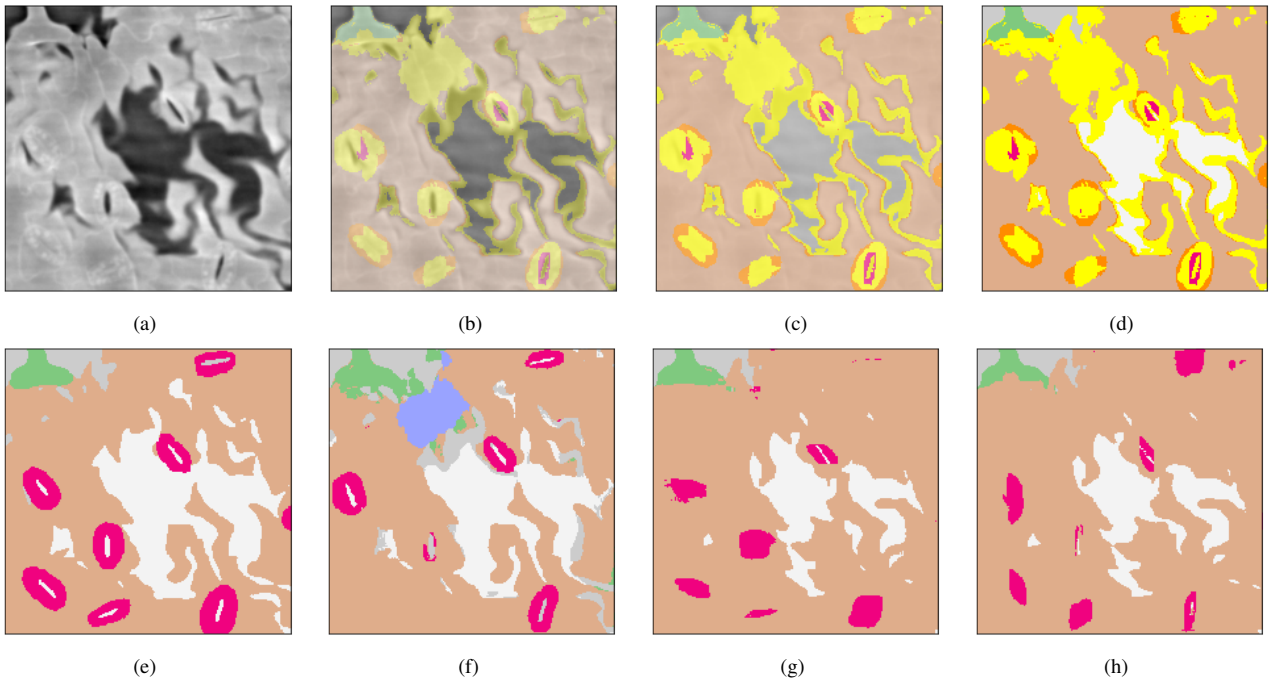


Figure 6. Top row: slice of scan (a) overlaid by ensemble predictions, using increasing opacity (b)-(d). Yellow indicates uncertainty and requests human revision. Bright orange indicates mismatch between predictions and ground truth. Bottom row: Labels by human expert (e) and predictions along the three orthogonal axes (f)-(h).

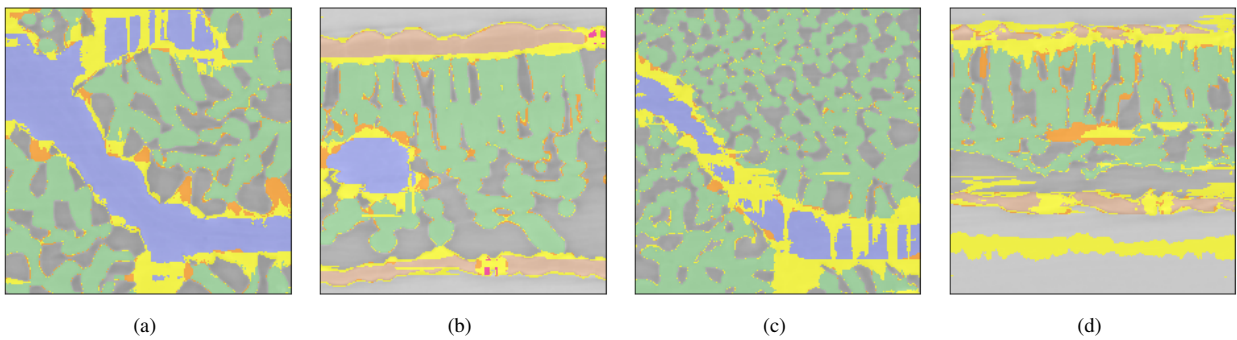


Figure 7. Selected slices from scan time 3 (a),(b) and 5 (c),(d) overlapped by ensemble predictions and uncertainty.

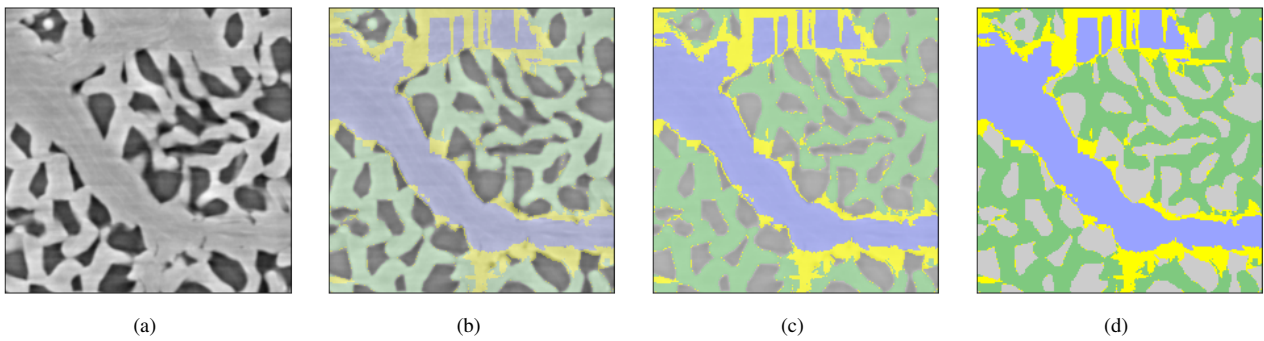


Figure 8. Slice of scan (a) overlaid by ensemble predictions, using increasing opacity (b)-(d).

# An unsupervised, shape-based 3d cell instance segmentation method for plant tissues

Alexander Palmrich<sup>‡</sup>, Klara Voggeneder<sup>✉</sup>, Danny Tholen<sup>✉</sup>, Guillaume Théroux-Rancourt<sup>✉</sup>,  
Jiří Hladůvka<sup>‡</sup>, and Walter Kropatsch<sup>‡</sup>

<sup>✉</sup>Institute of Botany, University of Natural Resources and Life Sciences, Vienna

{klara.voggeneder, daniel.tholen, guillaume.theroux-rancourt}@boku.ac.at

<sup>‡</sup>Pattern Recognition and Image Processing Group, Vienna University of Technology

{alexander.palmrich, jiri.hladuvka, walter.kropatsch}@tuwien.ac.at

## Abstract

We present a segmentation method for tissue images that uses the shape of image foreground to infer the location of individual cells. The method works in arbitrary dimension and is suited for volumetric scans. It is unsupervised, but allows a user to specify parameters to correct for the presence of noise and to steer the segmentation behavior. After describing the algorithm and its limitations, we analyze its complexity (linear in voxel count) and evaluate the quality of the segmentation result by applying it to a leaf x-ray micro-tomography scan.

keywords: instance segmentation, distance transform, skeleton, watershed, volumetric image, shape, unsupervised

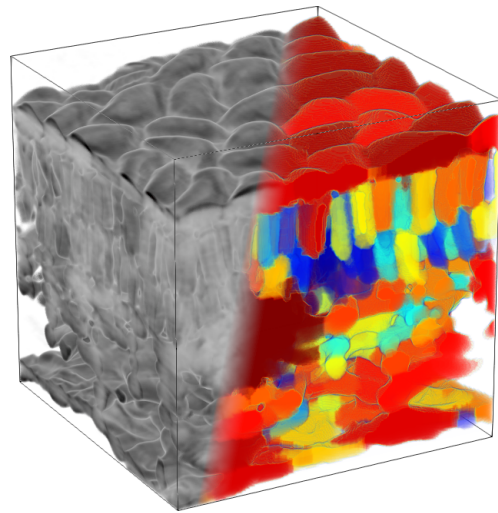


Figure 1.  $256^3$  voxel X-ray micro-tomography of leaf tissue and its automated cell instance segmentation result

## 1. Introduction

After obtaining a high resolution 3d image of biological tissue, further processing of that data may require identifying the tissue's constituent cells, i.e. cell instance segmentation. Even if a volumetric scan only allows for differentiation of tissue from background (as in Fig. 1, where the air space in-between cells is considered background), a human expert is still often able to identify single cells. They can infer the position and size of single cells from the shape of the foreground-to-background surface. Due to the large amount of data in a 3d image and the resulting work load, a full segmentation by experts is often not feasible. Our method aims to formalize and automate the human shape-based approach to segmentation. In contrast to other recent methods [8, 14] we do not rely on a neural network that needs training ("supervised"), but employ classical geometrical methods from pattern recognition ("unsupervised").

Throughout the article, we will use 2d images to illustrate the concepts involved. The actual method operates in

arbitrary dimension however, not just on slices of 2d images.

## 2. Limitations

Our method has the following requirements:

- The image must be an n-d grid (e.g 3d voxels) of grey values.
- The image must fit into RAM.
- The size of the smallest isolated cells to be captured must be known a priori or estimated at run time, to not discard such cells during noise handling.
- Cell shapes must be extractible after a suitable thresholding operation.



- Cell-to-cell interfaces must be smaller in diameter than cell bodies.

### 3. Algorithm

Our segmentation method can be summarized as following:

1. Extract foreground and background via thresholding.
2. Clean both from noise via morphological filtering.
3. Compute distance transform of foreground from background.
4. Assign labels to local maxima of the (smoothed) distance transform.
5. Use watershed to grow the labels.
6. Merge labels, if their border is nearly as wide as their thickest part.
7. Discard labels that are too small.
8. Grow remaining labels again using watershed.

We will now discuss these individual points in more detail.

The distance transform is a great tool to capture shape information, but it is highly sensitive to noise [2]: Even tiny spots of background located in the foreground can severely distort the transform, rendering it useless. This is why careful handling of noise and image artifacts is required. Rather than blurring the source image to reduce noise effects and also discarding potentially important high-frequency data (see Figure 2), we deal with the noise morphologically: After thresholding, we search for connected components of background and *delete* (i.e. assign as foreground) those, whose voxel count is below some specified value. This is a parameter we must specify, and why we must know in advance the size of the smallest structures we wish to capture.

The seed labels for watershed are constructed from both the local maxima of the distance transform (possibly a smoothed distance transform to merge close maxima), and the connected components of the background.

To use watershed, we need a height map that steers the growth of seed labels. Classical watershed segmentation [3] uses the image gradient magnitude as height, which allows for regions to grow fast where the source image has uniform brightness. This however ignores the shape information we have already extracted via the distance transform, and as such is prone to growing labels along image artifacts and noise, which strongly affect the gradient. Instead we choose to use a convex combination of image gradient magnitude and negative distance transform for height value: This allows for region growth to happen from cell centers outwards, uniformly approaching the background, independently from cell size. All the while the image gradient can inform label growth about subtle differences in brightness

that were lost during thresholding. The weight parameter in the convex combination allows the user to balance shape with brightness information.

After the first watershed segmentation, we are left with an over-segmentation. The core insight into how a human expert segments is now modeled by the *constriction factor*  $c$ : Each label remembers its radius at the thickest spot, i.e. the maximum of the distance transform within that label. At the interface between two neighboring labels, we scan for the maximum of the distance transform along the interface. This yields the local thickness of both labels at their border. Now, if this thickness value is close to the bigger of the two involved maxima, then the labels don't constrict much at their interface, and we assume that they actually belong to the same biological cell. Hence we merge them. Exactly how much of a constriction should warrant keeping both labels? This is another parameter to be specified. In order to keep the parameter scale-free, we model it as the ratio

$$c = \frac{\text{distance maximum at border}}{\text{distance maximum within both labels}}$$

From our experience, keeping labels separate for  $c < 0.75$  seems to work best.

Even after merging labels, there might be structures present that are too small in volume to reflect biological cells. An effective way to merge those tiny labels with their bigger neighbors is to discard them and use the result as the seed for another iteration of watershed (using the same height map as before). This grows the remaining labels, filling the holes just created. The result is a labeled image, with different labels for connected components of the background, labels for individual cells wherever their shape allows for separation, and with a single label per cluster of cells where they are packed tightly.

### 4. Runtime & Complexity

The building blocks of our method are thresholding, connected component search, distance transform, local maxima search, gradient magnitude, and watershed. All of these components have implementations with linear complexity  $\mathcal{O}(n)$  where  $n$  is the number of voxels in the data set [4,6,9].

We have implemented our algorithm in a Jupyter notebook, employing numpy, Skimage and numba for efficient computation. With this approach, running the method on a  $256^3$  image takes  $< 5$  minutes on an old laptop (Intel Core i3-3110M with 8GB RAM), including user input.

### 5. Validation

To assess the quality of the resulting segmentation, we apply it to a volumetric X-ray micro-tomography scan of a poplar leaf (see Figure 1). This image was downsized to get an isotropic voxel edge length of  $0.325\mu\text{m}$ , and then

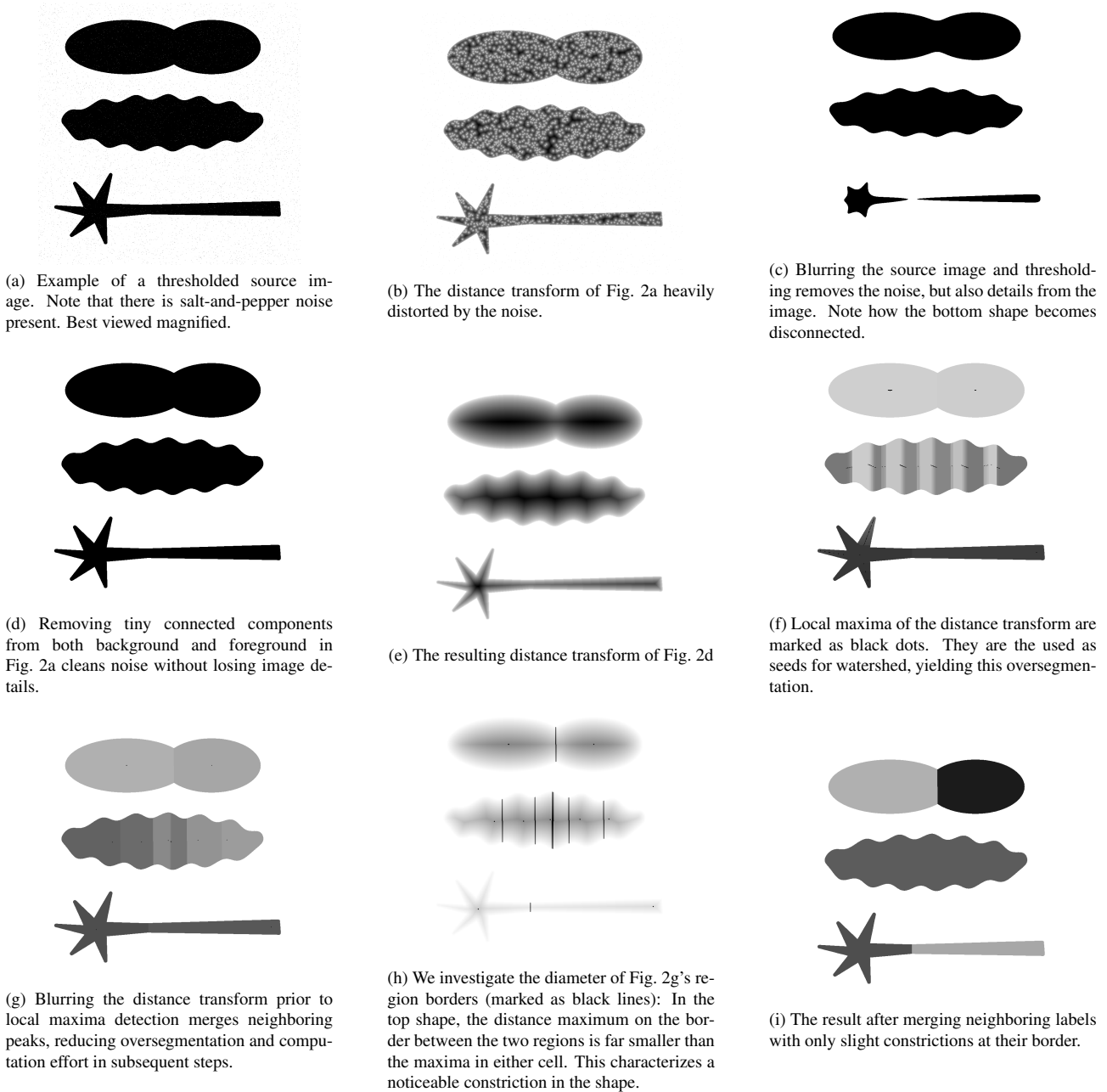


Figure 2. 2d illustration of steps involved in the algorithm

cropped to  $256^3$  voxels. Ground truths of all cells present in three paradermal and four transversal 2d slices were hand labeled, with the exception of the within-vein cells which are too densely packed and were assigned a single label.

To compare the human-generated ground truth with the automated segmentation, we use the metric

$$\text{error} = 1 - F_1 = 1 - \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

suggested in [1] as error measure, as well as the information-based measures described in [11]: The variation of information (*voi*) of a segmentation with respect to the ground truth can be understood as a measure of oversegmentation, whereas *voi* of the ground truth with respect to another segmentation quantifies under-segmentation. We refer to these measures as *splits* and *merges*, respectively. From the results listed in Table 1, we recognize that both

over- and under-segmentation occurs, but mostly in moderate less-than-one-bit amount. Error values are high in the palisade (slice 5) due to low recall, even though the segmentation looks promising upon visual inspection. Correct segmentation of the water vein from its neighboring cells proves difficult for our shape-based approach, as can be observed on slices 6 and 7.

## 6. Conclusion

We have demonstrated a linear-time method for unsupervised seeded watershed segmentation of images in arbitrary dimension. Human interaction is required only to select a few parameters. The seeds for seeded watershed segmentation are auto-generated.

The novelty of our method compared to established watershed segmentation methods [7, 10, 15] lies in the following aspects:

1. A careful morphological pre-processing regime to compensate for the distance transform's sensitivity to noise.
2. Operating on level sets of the distance transform instead of explicitly constructing a shape skeleton [5, 12, 13].
3. The height map employed during watershed uses information from both image gradient and the shape of the foreground-background surface.
4. Over-segmentation resulting from watershed is corrected using a scale-independent, isotropic shape criterion that models human expert behavior.

## 7. Future Work

We intend to explore further shape-based instance segmentation methods and aim to improve the quality of our result. The local variation of the constriction factor near label borders may offer another suitable criterion for merging labels and correcting for watershed over-segmentation.

## 8. Acknowledgments

We acknowledge the Paul Scherrer Institut, Villigen, Switzerland for provision of beamtime at the TOMCAT beamline of the Swiss Light Source. The computational results have been partially achieved using the Vienna Scientific Cluster (VSC). This work was supported by the Vienna Science and Technology Fund (WWTF) project LS19-013 and by the Austrian Science Fund (FWF) projects M2245 and P30275.

## References

- [1] Ignacio Arganda-Carreras, Srinivas C. Turaga, Daniel R. Berger, Dan Cireşan, Alessandro Giusti, Luca M. Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M. Buhmann, Ting Liu, Mojtaba Seyedhosseini, Tolga Tasdizen, Lee Kamentsky, Radim Burget, Vaclav Uher, Xiao Tan, Changming Sun, Tuan D. Pham, Erhan Bas, Mustafa G. Uzunbas, Albert Cardona, Johannes Schindelin, and H. Sebastian Seung. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in Neuroanatomy*, 9, 2015.
- [2] Dominique Attali, Jean-Daniel Boissonnat, and Herbert Edelsbrunner. Stability and computation of medial axes: a state-of-the-art report. *Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration*, 01 2009.
- [3] Serge Beucher and Christian Lantuéjoul. Use of watersheds in contour detection. volume 132, 01 1979.
- [4] Pedro Felzenszwalb and Daniel Huttenlocher. Distance transforms of sampled functions. *Theory of Computing*, 8, 08 2004.
- [5] Mathieu Gaillard, Chenyong Miao, James Schnable, and Bedrich Benes. *Sorghum Segmentation by Skeleton Extraction*, pages 296–311. 01 2020.
- [6] Lifeng He, Xiwei Ren, Qihang Gao, Xiao Zhao, Bin Yao, and Yuyan Chao. The connected-component labeling problem: A review of state-of-the-art algorithms. *Pattern Recognition*, 70:25–43, 2017.
- [7] Wala'a Jasim and Rana Mohammed. A survey on segmentation techniques for image processing. *Iraqi Journal for Electrical and Electronic Engineering*, 17:73–93, 12 2021.
- [8] Wenbo Jiang, Lehui Wu, Shihui Liu, and Min Liu. Cnn-based two-stage cell segmentation improves plant cell tracking. *Pattern Recognition Letters*, 128:311–317, 2019.
- [9] Anton Kornilov and Ilia Safonov. An overview of watershed algorithm implementations in open source libraries. *Journal of Imaging*, 4:123, 10 2018.
- [10] Aurélie Leborgne, Julien Mille, and Laure Tougne. Extracting noise-resistant skeleton on digital shapes for graph matching. pages 293–302, 12 2014.
- [11] Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.
- [12] Xiaopeng Sun, J. Pan, and Xiaopeng Wei. 3d mesh skeleton extraction using prominent segmentation. *Comput. Sci. Inf. Syst.*, 7:63–74, 02 2010.
- [13] Seung tak Noh, Kenichi Takahashi, Masahiko Adachi, and Takeo Igarashi. Skelseg: Segmentation and rigging of raw-scanned 3d volume with user-specified skeleton. In *Graphics Interface*, 2019.
- [14] Adrian Wolny, Lorenzo Cerrone, Athul Vijayan, Rachele Tofanelli, Amaya Vilches-Barro, Marion Louveaux, Christian Wenzl, Soeren Strauss, David Wilson-Sánchez, Rena Lymbouridou, Susanne Steigleder, Constantin Pape, Alberto Bailoni, Salva Duran-Nebreda, George Bassel, Jan Lohmann, Miltos Tsiantis, Fred Hamprecht, Kay Schneitz, and Anna Kreshuk. Accurate and versatile 3d segmentation of plant tissues at cellular resolution. *eLife*, 9, 07 2020.
- [15] Nida M. Zaitoun and Musbah J. Aqel. Survey on image segmentation techniques. *Procedia Computer Science*, 65:797–806, 2015. International Conference on Communications, management, and Information technology (ICCMIT'2015).

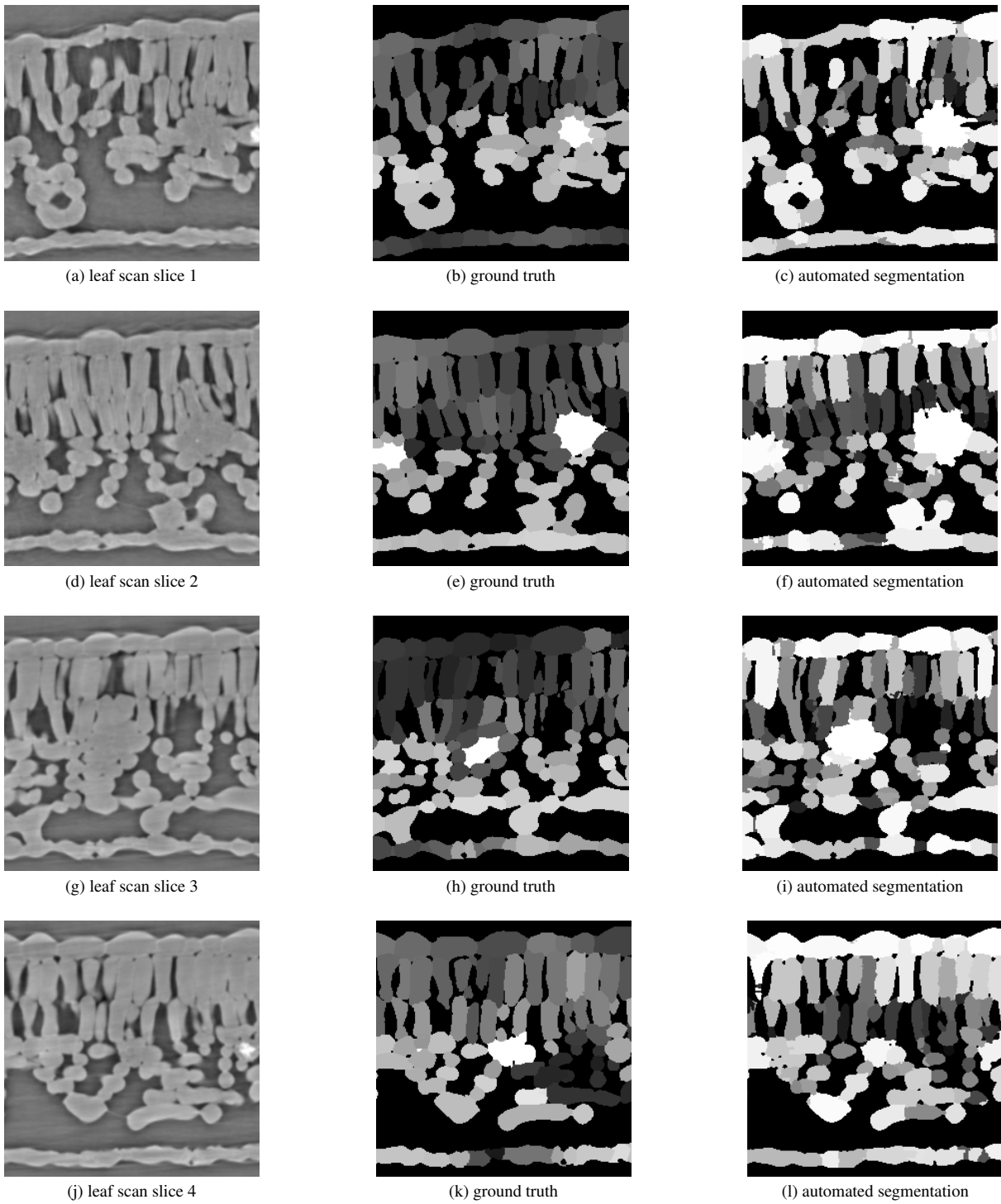


Figure 3. Transversal slices (perpendicular to the surface and to the leaf midrib)

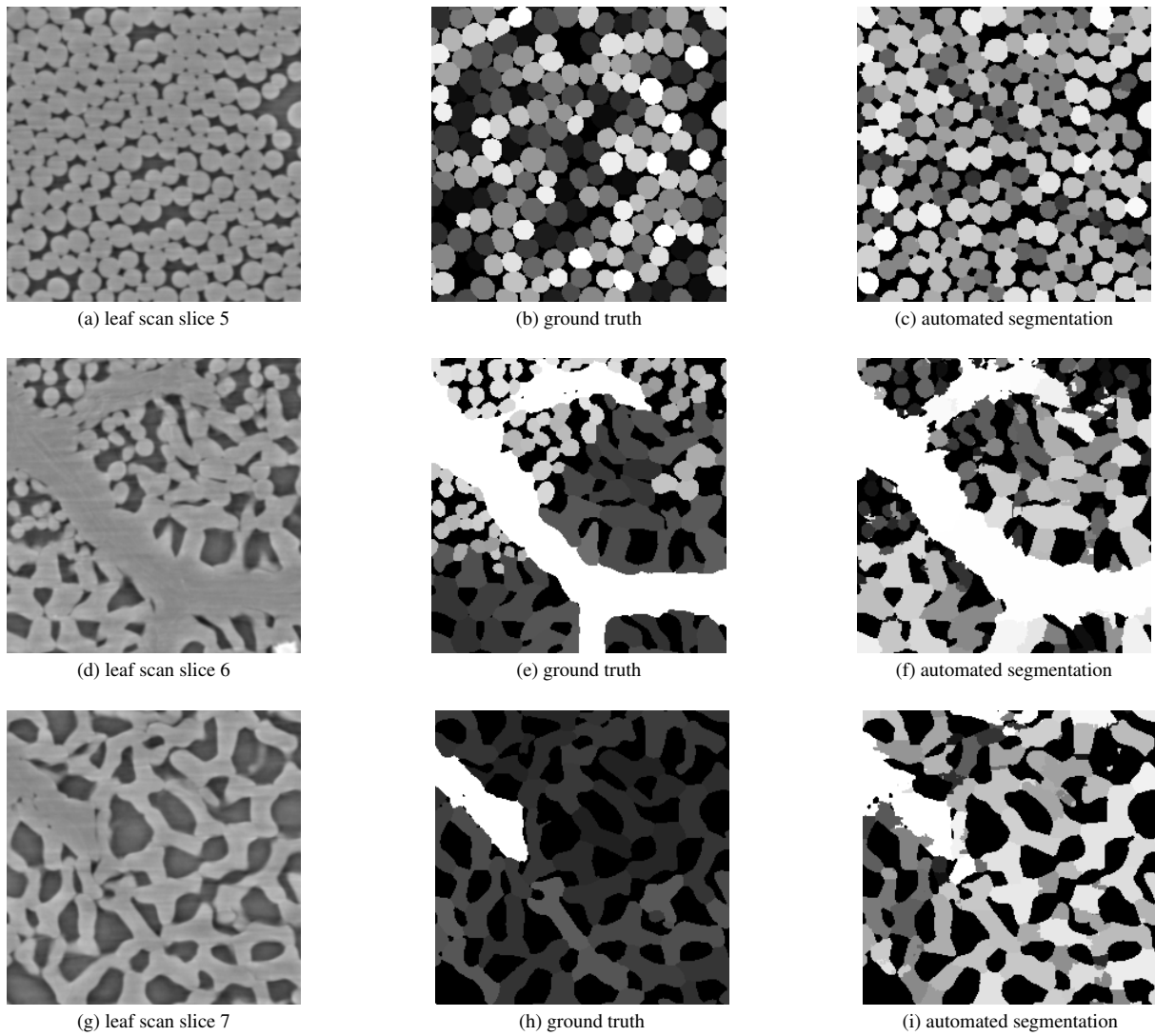


Figure 4. Paradermal slices (parallel to the leaf surface)

slice no.	error [%]	precision [%]	recall [%]	splits [bit]	merges [bit]
1	6.3	98.8	89.0	0.6	0.7
2	4.5	98.0	93.1	0.7	0.5
3	12.2	88.6	87.0	0.8	0.8
4	14.3	91.8	80.4	0.9	0.8
5	40.1	93.0	44.1	0.6	1.2
6	37.6	52.9	76.0	1.5	0.9
7	15.9	88.2	80.4	1.2	0.7

Table 1. Segmentation error metrics

# Exploring Learning-Based Approaches for Bomb Crater Detection in Historical Aerial Images

Marvin Burges, Sebastian Zambanini, Robert Sablatnig  
Computer Vision Lab, TU Wien  
1040 Vienna, Austria

{mburges, zamba, sab}@cvl.tuwien.ac.at

## Abstract

Many countries were the target of air strikes during World War II. The heritage of these attacks is still present today, as numerous unexploded bombs are uncovered yearly in Central Europe. While these bombs pose a significant explosion hazard, they can be inferred from the existence of craters. Therefore, analyzing aerial images from World War II surveillance flights allows for preliminary risk estimation. In this paper, we train and evaluate 12 different object detector architectures and compare them to a crater detection algorithm on our custom historical aerial dataset. We show that modern detectors, in combination with a large enough historical aerial crater dataset, can outperform a current method for crater detection, achieve a precision of 0.6 and a recall of 0.6 on our dataset, and can process large remotely sensed images within seconds, rather than minutes. Additionally, pretraining and different dataset extensions are evaluated and discussed.

## 1. Introduction

Although the last air raids of World War II happened more than 70 years ago, Unexploded Ordnances (UXOs) still pose a significant explosion hazard for European construction projects [11]. Specialized companies provide a preliminary risk estimation by reviewing and interpreting aerial images from World War II surveillance flights over the area of interest. To generate these risk estimations, historical aerial images have to be georeferenced, and all objects that indicate increased combat activity have to be marked. Currently, both the georeferencing task and the search for increased combat activity are performed manually by specialists. The goal of our work is to automatically generate “explosive ordnance maps” from selected images, by detecting increased combat activity. These “explosive ordnance maps” indicate whether an area is likely to be contaminated and therefore may contain UXOs, while in un-

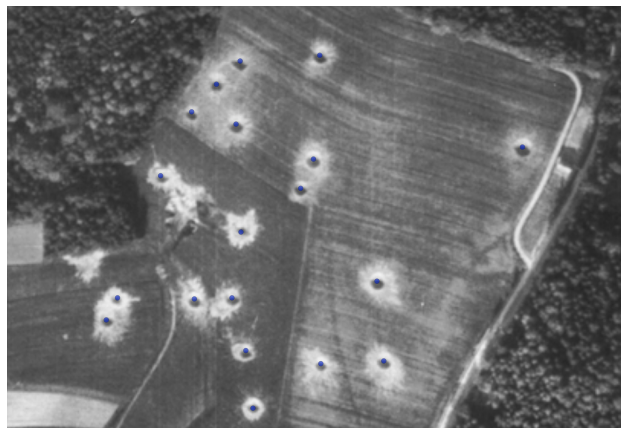


Figure 1. An example detection result of the best performing detector on a rural area. Blue point: Prediction, best viewed in color.

contaminated areas, UXOs are unlikely [11]. To achieve this goal, we survey the performance of existing fully-automatic object detectors in detecting increased combat activity by training them on our custom dataset consisting of historical aerial images.

Specifically, we focus on the detection of bomb craters, as they are the most abundant type of warfare-related object visible in aerial images. Furthermore, they represent direct evidence for the presence of an UXO, as it is assumed that 10 - 15 % of all bombs dropped during World War II did not explode [2]. We concentrate on the detection of warfare-related objects instead of the prediction and segmentation of potentially contaminated areas due to explainability reasons. As this is a task with potentially dangerous consequences, the network results always have to be verified by an expert. An example for this task is given in Fig. 1, where the automatic detections on a image from the H2OPM dataset [25] are given. In this work, we compare Convolutional Neural Networks (CNNs)-based detectors on the task of automatic detection of bomb craters and compare our findings with a crater detection method from the litera-

ture [2]. We specifically focus our comparison on network architectures that can be retrained on consumer hardware, as we intend to focus on domain adaptation in future work, which can require an end-user to retrain on newly obtained data of a different domain. We will include the best performing network in a plugin for the geographic information system *QGIS*<sup>1</sup>, for which we will release the code<sup>2</sup> and the weights<sup>3</sup> for the best performing networks can be used directly with the original YoloV5 implementation of Glenn Jocher *et al.* [8].

The remainder of this paper is structured as follows. First, in Sec. 2 the related work on object detection in historical aerial images and related domains is presented, followed by the data used for this paper in Sec. 3. The experimental setup is described in Sec. 4 and results and discussion are given in Sec. 5. Finally, a summary and potential future work is presented in Sec. 6.

## 2. Related Work

As the amount of related work on the topic of crater detection in historical aerial images is limited, we include selected methods for mars and moon crater detection, which are visually similar to bomb craters.

Brenner *et al.* [2] developed an approach to automatically detect craters in historical aerial images using a machine learning approach based on a CNN. They use a sliding window in combination with DenseNet [10] to extract candidate crater positions from the image and then use post-processing to refine the detections. These post-processing steps include a spatial proximity prior, as bombs are dropped in clusters thus, “lonely” bombs are likely a false positive, non-cluster suppression, as due to the overlap of the sliding window, bombs should be detected multiple times, as outliers are detections that are not part of a cluster detection and a non-maxima suppression to reduce multiple detections. Overall, their approach achieved a precision/recall of 90.7%/91.3% with the same amount of craters and background images in the test set. However, with a more realistic distribution of around 1:250, the precision is reduced to 4%. In [11], Kruse *et al.* assume that multiple images of the same area exist, based on which they propose an approach that combines the individual detection results of a stochastic approach based on marked point processes. This increased the F1-score from 39 % (based on single images) to 67 % (based on multiple images). In [12] Kruse *et al.* further evaluate this method by examining the influence of random number generation. They also compare their approach to a Faster RCNN object detector [17] trained on

their dataset. The results show that the CNN can outperform their approach if the correct threshold is selected. However, they also note that in a scenario where only a limited amount of training data is available, their approach delivers superior results.

Wu *et al.* [23] propose a Crater Detection Algorithm (CDA) called SUNnet 3+, that is based on the UNET architecture [18] and detects craters in the digital elevation model of Mars. The CDA proposed in [9] aims at detecting lunar craters in images in real-time by a crewed lunar lander during the landing procedure, based on a modified YoloV4 [1] architecture.

The listed publications show that, while learned crater detectors have been evaluated before, only one publication trains an object detector on this task. Brenner *et al.* [2] extracts regions via a sliding window and classifies the patches via a CNN and Clermont *et al.* [4] use a blob detector, also in combination with a CNN as classifier. Only in [12] Kruse *et al.* train and evaluate an object detector and show its potential. The primary challenge, frequently mentioned, was the lack of training data [4, 11, 12]. However, due to an industry partnership, we have access to a dataset suitable for training. Hence, this paper evaluates different object detectors trained on crater samples. We expect that by training object detectors with learned regions proposals we can outperform current crater detection methods.

## 3. Data

To the best of our knowledge, no dataset for crater detection in historical aerial images is publicly available. Therefore, we use our dataset, which covers both urban and rural areas. These images originate from finished projects in which experts georeferenced the historical aerial images and annotated the bomb craters. In total, 111 images are georeferenced, and a total of 19,506 craters have been marked in the images. Note, the analysis per image is only performed within the Region Of Interest (ROI). As a result, we had to ignore all regions outside of the ROI, as no ground truth data is available outside of the ROI. The images were made between 1943 and 1945 and contain craters with a minimum size of 1m, an average size of 8m, and a maximum size of 17m. The minimum image size is  $2,274 \times 2,388$ , the average is  $11,626 \times 10,864$ , and the maximum image size is  $16,714 \times 16,973$ . All images are split into  $960 \times 960$  images with an overlap of 10% for training for a total of 3,711 images. The Ground Sampling Distance (GSD) for all images is normalized to 0.25m, and they cover an area of 505 km<sup>2</sup>. An example of a rural image can be seen in Fig. 2a, one of an urban image in Fig. 2b, both images are from the H2OPM dataset [25] and are also part of our dataset. Additionally, we experimented on 12 panchromatic Martian satellite images [6] with a crater size of less than 5 km and a total amount of

<sup>1</sup><https://qgis.org/en/site/>

<sup>2</sup>[https://github.com/mburges-cvl/QGIS\\_Plugin\\_for\\_OAGM\\_2022](https://github.com/mburges-cvl/QGIS_Plugin_for_OAGM_2022)

<sup>3</sup><https://owncloud.tuwien.ac.at/index.php/s/AxarN33AnClCDhA,PW:”oagm2022”>

around 42,000 craters (Fig. 2c). These images were also split into  $960 \times 960$  images with an overlap of 10% for training, which resulted in 384 images. We also generated a synthetic dataset, which is based on the XVIEW dataset [13] where 16,931 cut-out craters from our training dataset are imprinted onto the XVIEW images. An example of the synthetic data is presented in Fig. 2d. We discarded the original classes of the XVIEW dataset, converted the images to grayscale and to a GSD of 0.25, and augmented the crater patches before imprinting them in the image. The augmentation consists of rotation, horizontal flipping, and Fourier Domain Adaptation (FDA) [24], with the target being the XVIEW-image. This resulted in 17,234  $960 \times 960$  images with a total of 603,190 craters and a similar GSD to the original project images.

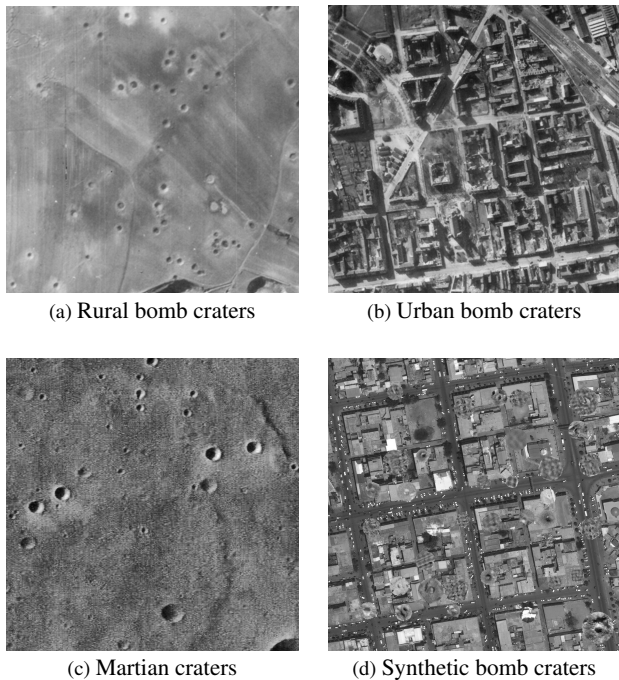


Figure 2. Four example images from our dataset. a) and b) Images from company projects, c) Martian satellite image from [6] and d) Image from the XVIEW-Dataset [13] with synthetic craters.

## 4. Experimental Setup

In this section, we present the experiments conducted for this paper. We start by evaluating different object detectors on our historical crater dataset, then compare the results with the crater detection method proposed by Brenner *et al.* [2]. We continue with the pretraining and synthetic data results and finish with the evaluation metrics.

### 4.1. Evaluation of State-of-the-Art Object Detectors

For this work, we evaluated 12 different object detection architectures. An overview of the networks can be seen in Tab. 1, which highlights the trainable parameters as well as whether the networks use one stage or two stages. Single-stage networks perform classification and regression on dense anchor boxes without generating a sparse ROI set, while two-stage networks first generate sparse region proposals, which are then further regressed and classified in a second stage [14]. We chose these specific networks as they are capable of running on consumer hardware and because these networks have been repeatedly used for few-shot learning. We trained all networks on the historical aerial dataset with an image size of  $960 \times 960$  pixels and a batch size of 2 (32 for YoloV5n), until the Average Precision (AP) started plateauing, which was between 10 - 50 epochs depending on the dataset size (larger datasets required more epochs). For the data augmentation, we relied on the Albumentations framework [3]. We used common augmentations like blur, random brightness and contrast changes, rotation, translation, and histogram equalization, but we also used the mosaic augmentation method proposed in [1] as well as FDA. All networks were pretrained on COCO before being finetuned on our crater dataset and used the same anchors (if applicable).

Table 1. Evaluated networks with their respective parameter count in million and their architecture style.

Name	Parameters (M)	Stages
YoloV5n [8]	1.9	One
YoloV4 [1]	27.6	One
YoloV7 [22]	37.1	One
Faster RCNN R 50 [17]	41.3	Two
Faster RCNN R 101 [17]	60.2	Two
YoloV3 [16]	65.3	One
ScaledYoloV4(-p5) [21]	70.2	One
EfficientDet(-d7x) [19]	76.8	One
YoloV5x [8]	86.7	One
YoloX [7]	99.0	One
Faster RCNN X 101 [17]	104.4	Two
YoloR(-d6) [20]	151.0	One

### 4.2. Comparison with State-of-the-Art Crater Detection Methods

To the best of our knowledge, no historical aerial crater detection frameworks for comparison are publicly available. We therefore chose to compare the best performing detector from Sec. 4.1 to the detector proposed by Brenner *et al.* in [2] on our crater data. However, as we did not have access to the original version, the approach was re-implemented. Similar to the original implementation, we used a 40-layer DenseNet with an input size of  $32 \times 32$  pixels and trained it



on a binary classification problem. We trained the network for 100 epochs on the 42,172 crater patches extracted by a sliding window from the training set with roughly the same amount of negative examples. Differently to the original approach, however, we use patches of size  $80 \times 80$  pixels instead of  $20 \times 20$  pixels to achieve a similar window size of  $20m \times 20m$  as our dataset has a GSD of 0.25m instead of 1m. These patches are then resized to the network input size of  $32 \times 32$  pixels. We evaluated the approach in two ways. First the classification way, where we extracted all 1,614 crater patches from the validation set as well as the same amount of negative samples (resulting in a 1:1 ratio of positive and negative samples). Second in the (more realistic) object detection way, where the network is applied as a sliding window to the validation images (resulting in roughly a 1:250 ratio of positive and negative samples). In a final experiment, we compared the run-time of the approach from Brenner *et al.* with YoloV5n and YoloV5x on one example image with the size of  $10,644 \times 10,042$ .

### 4.3. Pretraining and Synthetic Data

We also experimented with different strategies to improve the training result. One idea was to pretrain the network. In our case, we chose COCO, the Mars dataset described in Sec. 3 and the XVIEW dataset. Additionally, we experimented with increasing the raw crater dataset size by adding a combination of the Martian and synthetic data to our historical aerial crater dataset. The intention with the Martian data was to add more crater variants, while the idea for the synthetic data was to add more urban structures to the dataset. We again chose the best performing detector from Sec. 4.1 and trained it on the different dataset combinations.

## 5. Results and Discussion

In this section, we present the results of the experiments. We start by presenting and discussing the quantitative results in Sec. 5.1 and finish with a qualitative analysis of two example images in Sec. 5.2. Both show the difficulty related to detecting craters in historical aerial images.

### 5.1. Quantitative Results

The trained object detectors presented in Sec. 4.1 were evaluated on the test set of the historical crater dataset described in Sec. 3. The results can be seen in Fig. 3. The graph shows the precision recall curve for all networks presented in Tab. 1. It is visible that YoloV5n, YoloV5x, and YoloR perform similarly and better than the other tested networks. One can also see that YoloV3 and Faster RCNN R 50 perform significantly worse. This is due to the coarse search grid of YoloV3, which hinders the detection performance of small objects in the images, as stated by Pham *et al.* [15]. Similarly, the Faster RCNN family has issues with

small objects, as has been shown by Eggert *et al.* [5]. This could be a possible explanation for why the Faster RCNN family is outperformed overall by the Yolo variants and why Faster RCNN R 50, in particular, is underperforming. Overall it is shown that the majority of the tested networks have similar results, which could be due to the training set being comparatively small for an object detection dataset (compared to COCO, for instance). Overall, YoloV5n is one of the best performing networks and also has the lowest amount of trainable parameters which makes the training less prone to overfitting.

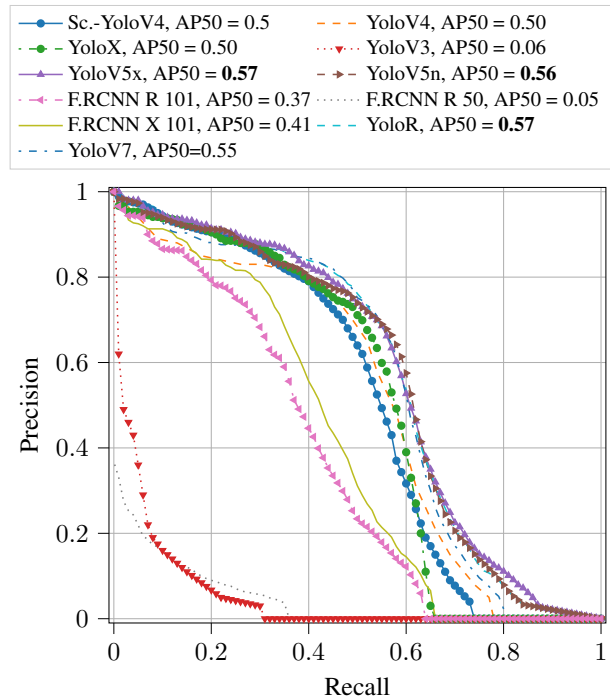


Figure 3. Precision-Recall-Curves for the different network architectures trained on the crater dataset.

The comparison with the approach of Brenner *et al.* is presented in Tab. 2. It shows that their approach achieves a precision of 0.91 and a recall of 0.87 during evaluation with a 1:1 distribution of positive and negative samples on our dataset, however their real-world performance is significantly worse. We were unable to achieve any meaningful detections with the approach, which is also reflected in a precision of 0.05 with a recall of 0.01 during the evaluation with a 1:250 distribution. To further show the edge YoloV5n has, we present the time for a detection of an image with a, for this task, common size of  $10644 \times 10042$  in Tab. 3. It is visible that YoloV5n is 58 times faster than the approach proposed by Brenner *et al.* with a precision of 0.6 and a recall of 0.6.

We further explored YoloV5n on different dataset com-

Table 2. Precision and recall for Brenner *et al.* on a synthetic (1:1) and a more realistic (1:250) distribution of positive to negative training patches, compared to YoloV5n. YoloV5n does not use a sliding window to generate patches but instead is applied to the whole image.

Approach	Precision	Recall	Positive-Negative-Ratio
Brenner <i>et al.</i>	0.91	0.87	1:1
Brenner <i>et al.</i>	0.05	0.01	1:250
YoloV5n	<b>0.6</b>	<b>0.6</b>	1:250

Table 3. Runtime comparison of the best performing detector and the approach from Brenner *et al.* Measurement: A 10,644 × 10,042 example image, split into 144 960 × 960 image patches (with overlap) for YoloV5. GPU: Nvidia T500 (Mobile).

Approach	Runtime
QGIS + Brenner <i>et al.</i>	361s
QGIS + YoloV5n	<b>6.2s</b>

binations. The results are presented in Fig. 4, in addition to YoloV5n, we present YoloV5x trained on the combined datasets. We chose to add YoloV5x to the comparison, as it performed similar to YoloV5n but had more parameters and might benefit more from the additional data. Overall it is visible that training on the Martian dataset alone does not result in a suitable detector. While it can detect craters in rural areas (i.e. fields), it has a high amount of false positives in urban areas. Training on synthetic data only resulted in a detector that was unable to detect any crater correctly. The curve was thus omitted, while training on any combination of the synthetic and Martian data with our historical dataset resulted in a similar performance. This is likely due to the fact that the Martian domain is too trivial as it only contains well-defined craters in rock and sand. The only challenges with this dataset are overlapping crater or crater contained in a larger crater. Both are rare in historical images and therefore do not contribute much to the overall precision. A likely reason why the synthetic dataset does not contribute to a better performance is because the issue with urban areas is not the false-positive rate on human-made structures but the high irregularity of the craters. But, as we insert cut-out historical craters into XVIEW images we do not increase crater verity.

Lastly, we experimented with different pretraining strategies: no pretraining, COCO pretrained weights, and XVIEW pretrained weights. In Fig. 5 one can see the AP plotted for 20 training epochs. It is visible that after 20 training epochs, all networks achieve similar results, which was also verified with the test set where all networks again performed closely. A similar effect can be seen in the validation loss, which is presented in Fig. 6. It is apparent that after 16 epochs, the loss of all methods is similar. How-

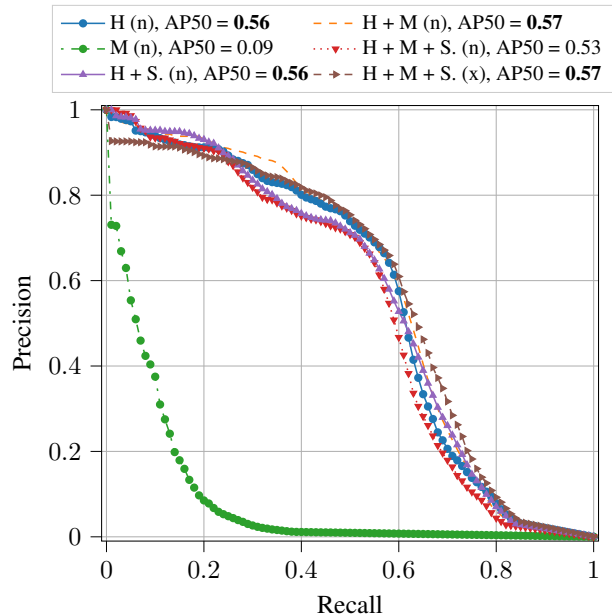


Figure 4. Precision-Recall-Curve for the different training dataset combinations. (n) refers to YoloV5n, and (x) refers to YoloV5x. H = Historical, M = Martian and S = Synthetic.

ever, it is also visible in both plots that the network pre-trained on the COCO dataset requires only about 5 epochs to achieve peak performance. This, considering previous results, means that YoloV5n can be rapidly retrained on and finetuned on a new domain.

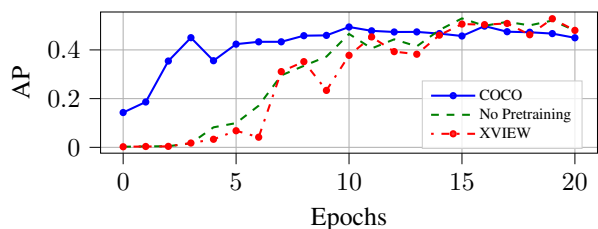


Figure 5. Pretraining evaluation: AP per epoch during training.

## 5.2. Qualitative Results

We present two detection results of YoloV5n in Fig. 7 and Fig. 1, which show a rural, but snow-covered, area and a field. While YoloV5n can detect all craters flawlessly in Fig. 1, which shows a barren field, it struggles to catch all craters in Fig. 7, which similarly presents a field, however with snow coverage and fresh craters as well as older craters. It was also observed, that in an urban environments YoloV5n is performing even worse than in domain shifted images (i.e snow). This shows that the detector has

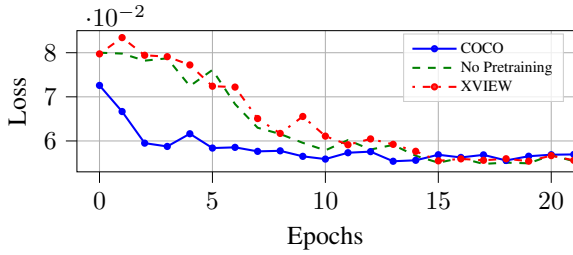


Figure 6. Pretraining evaluation: Bounding box loss per epoch during training.

a very high recall and precision for more straightforward tasks, like the detection of craters in rural areas (Fig. 1). However, it struggles with more complicated tasks, such as the detection in an urban domain. The primary challenge here is the fact that craters are highly irregular due to interference with artificial structures. In contrast to this, in rural areas and especially in barren fields, craters tend to look similar. Additional challenges are due to the low image quality, like low contrast and noise. A further issue is a domain change (i.e., fresh craters and old snow-covered craters), where training images are rare for the new domain. This can be seen in Fig. 7, where the detector can detect 5 out of 13 craters with snow coverage, which are sparsely represented in the dataset.

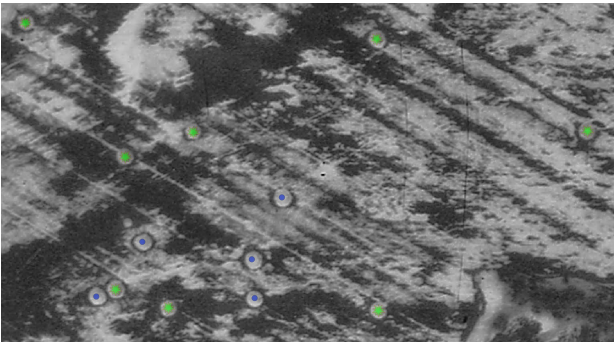


Figure 7. Snowy rural bomb crater detections from [25]. Blue points: Detections, Green Star: Predictions, best viewed in color.

A solution for this problem could be an interactive approach, where an end-user could improve the detections by removing false positives or adding false negatives, and the network then retrains based on the changes. Another possibility could be that the user preemptively marks one or a few craters in the image, which the network then uses as additional domain information during the detection.

## 6. Conclusion and Future Work

In this paper, 12 State-Of-The-Art detectors were compared on the task of detecting craters on 111 historical aerial

images, and the best detector was then compared to another approach from the literature [2]. The detections obtained by these detectors can be used in the predominantly manual process of generating an “explosive ordnance map”, which indicate areas that could be contaminated with UXOs, a significant explosion hazard for construction processes in Europe. We showed that, while the tested detectors were unable to achieve sufficient accuracy to be used fully automatically, the best detector, YoloV5n achieves a precision of 0.6 with a recall of 0.6 in real-world use cases. Furthermore, it only requires 6 seconds to process an image of average size (10,644 × 10,042), while the approach from the literature requires about 360 seconds for the same image and only has a precision of 0.05 and a recall of 0.01. This combination of accuracy and speed allows for a quick and sufficiently correct preliminary overview over an area, which then can be manually finetuned to an “explosive ordnance map” by an expert. YoloV5n also only requires between 5 - 7 epochs for training when pretrained on COCO, which allows for rapid retraining of the network. Another possibility to exploit YoloV5n would be to use it as a base detector for an interactive learning method or a few-shot learning strategy, where an expert corrects the detections and finetunes the network. This idea will be explored in the future. A further insight is that training on synthetic or Martian data does not significantly improve the detection accuracy of the network, primarily due to the fact that the biggest challenges are irregular craters or unseen crater variants.

In general, we see learning-based approaches in favor of algorithmic approaches like [12]. Our and the results of related work demonstrate the difficulty of crater detection in historical images, which makes the use of semi-automatic approaches inevitable for practical reasons. Learning-based approaches offer the needed flexibility to allow for an on-the-fly adaptation to specific image domains, which will be the direction of our future work.

## Acknowledgments

This work was supported by the Austrian Research Promotion Agency (FFG) under project grant 880883. Acquisition of historical aerial imagery: Luftbilddatenbank Dr. Carls GmbH; Sources of historical aerial imagery: National Archives and Records Administration (Washington, D.C.) and Historic Environment Scotland (Edinburgh).

## References

- [1] A. Bochkovskiy, C-Y. Wang, and H-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. arXiv, abs/2004.10934, 2020.
- [2] S. Brenner, S. Zambanini, and R. Sablatnig. Detection of bomb craters in wwii aerial images. In *Proceedings of the OAGM Workshop*, pages 94–97, 2018.
- [3] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. Albuumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020.
- [4] C. Clermont, D. and Kruse, F. Rottensteiner, and C. Heipke. Supervised detection of bomb craters in historical aerial images using convolutional neural networks. *ISPRS - International Society for Photogrammetry and Remote Sensing*, XLII-2/W16:67–74, 2019.
- [5] C. Eggert, S. Brehm, A. Winschel, D. Zecha, and R. Lienhart. A closer look: Small object detection in faster r-cnn. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 421–426, 2017.
- [6] J. Francis, A. and Brown, T. Cameron, R. Crawford Clarke, J. Dodd, R. and Hurdle, M. Neave, J. Nowakowska, V. Patel, A. Puttock, O. Redmond, A. Ruban, D. Ruban, M. Savage, W. Vermeer, A. Whelan, P. Sidiropoulos, and J-P. Muller. A multi-annotator survey of sub-km craters on mars. *Data*, 5(3), 2020.
- [7] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. arXiv, abs/2107.08430, 2021.
- [8] J. Glenn. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, 2022.
- [9] T. Hu, C. Zhao, Z. Qian, L. He, and M. Ni. Crater obstacle recognition and detection of lunar landing based on yolo v4. In *2021 33rd Chinese Control and Decision Conference (CCDC)*, pages 1748–1752, 2021.
- [10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Los Alamitos, CA, USA, 2017. IEEE Computer Society.
- [11] C. Kruse, F. Rottensteiner, and C. Heipke. Using redundant information from multiple aerial images for the detection of bomb craters based on marked point processes. *ISPRS - International Society for Photogrammetry and Remote Sensing*, V-2-2020:861–870, 2020.
- [12] C. Kruse, D. Wittich, F. Rottensteiner, and C. Heipke. Generating impact maps from bomb craters automatically detected in aerial wartime images using marked point processes. *ISPRS - International Society for Photogrammetry and Remote Sensing*, 5:100017, 2022.
- [13] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord. xview: Objects in context in overhead imagery. arXiv, abs/1802.07856, 2018.
- [14] X. Lu, Q. Li, B. Li, and J. Yan. Mimicdet: Bridging the gap between one-stage and two-stage object detection. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 541–557, Cham, 2020. Springer International Publishing.
- [15] M-T. Pham, L. Courtrai, C. Friguet, S. Lefèvre, and A. Bausard. Yolo-fine: One-stage detector of small objects under various backgrounds in remote sensing images. *Remote Sensing*, 12(15):2501, 2020.
- [16] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv, abs/1804.02767, 2018.
- [17] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
- [18] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.
- [19] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [20] C. Wang, i-H. Yeh, and H-Y. M. Liao. You only learn one representation: Unified network for multiple tasks. arXiv, abs/2105.04206, 2021.
- [21] C-Y. Wang, A. Bochkovskiy, and H-Y. M. Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13029–13038, 2021.
- [22] C-Y. Wang, A. Bochkovskiy, and H-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv, abs/2207.02696, 2022.
- [23] Y. Wu, G. Wan, L. Liu, Y. Jia, Z. Wei, and S. Wang. Fast and accurate crater detection on martian surface using sun et 3+. In *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, volume 6, pages 683–687, 2022.
- [24] Y. Yang and S. Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.
- [25] Sebastian Zambanini. Feature-based groupwise registration of historical aerial images to present-day ortho-photo maps. *Pattern Recognition*, 90:66–77, 2019.

# Student Papers

# Automated nuclear morphometry as a prognostic marker in canine cutaneous mast cell tumors

Eda Parlak

University of Veterinary Medicine Vienna  
Veterinärplatz 1, 1210 Vienna, Austria

Eda.Parlak@vetmeduni.ac.at

Taryn A. Donovan

Schwarzman Animal Medical Center  
510 E 62 St, New York, NY 10065, USA

Taryn.Donovan@amcnyc.org

Stephan Winkler

University of Applied Sciences Upper Austria  
Softwarepark 11, 4232 Hagenberg, Austria

Stephan.Winkler@fh-hagenberg.at

Marc Aubreville

Technische Hochschule Ingolstadt  
Esplanade 10, 85049 Ingolstadt, Germany

Marc.Aubreville@thi.de

Andreas Haghofer

University of Applied Sciences Upper Austria  
Softwarepark 11, 4232 Hagenberg, Austria

Andreas.Haghofer@fh-hagenberg.at

Robert Klopffleisch

Freie Universität Berlin  
Robert-von-Ostertag-Str. 15, 14163 Berlin, Germany

Robert.Klopffleisch@fu-berlin.de

Matti Kiupel

Michigan State University  
4125 Beaumont Road, Lansing, MI 48910, USA

kiupel@msu.edu

Christof A. Bertram

University of Veterinary Medicine Vienna  
Veterinärplatz 1, 1210 Vienna, Austria

Christof.Bertram@vetmeduni.ac.at

## Abstract

*The prognosis of canine cutaneous mast cell tumors (ccMCT) is evaluated by various histologic parameters including the variability in size and shape of tumor nuclei (nuclear pleomorphism). Traditionally, nuclear pleomorphism is estimated by pathologists. However, a more precise measurement could be achieved by automated morphometry, which was investigated in this study. Eighty-six annotated images from ccMCT were used to develop a nuclear segmentation model, which yields an IoU of 0.79 on the test set. The prognostic value was determined on 96 ccMCT cases with known patient outcomes by two-fold cross-validation. Several features of nuclear size and shape were extracted from the segmentation mask and the ideal combination and thresholds of these features were determined by an XGBoost model independently for the two dataset splits. Tumor-related death was predicted on the left-out data set part with an AUC of 0.82 and 0.86, respectively. This study shows a high prognostic value of algorithmic nuclear morphometry in ccMCT. Future studies should compare the algorithm with estimates by pathologists.*

## 1. Introduction

Canine cutaneous mast cell tumors (ccMCTs) are one of the most frequent skin tumors in dogs. These tumors are potentially malignant and histologic examination of the tumor cells is important to prognosticate patient outcome. Among other cellular criteria, the variability of nuclear size and shape (nuclear pleomorphism) has a well-known prognostic relevance for ccMCT and many other tumor types. Traditionally, nuclear pleomorphism is estimated by pathologists into vaguely defined categories. An alternative to the subjective assessment is the precise measurement of nuclei in digital images (nuclear morphometry). The manual measurement of nuclear size by pathologists has already been investigated in previous studies [2]; however, routine use is not to be expected since it is quite time-consuming (10 minutes per measurement were reported in this study). In comparison, fully-automated morphometry using deep learning-based algorithms would be a very practical solution, assuming that the nuclei can be accurately segmented. Deep learning methods for nuclear segmentation have been extensively researched previously [3].

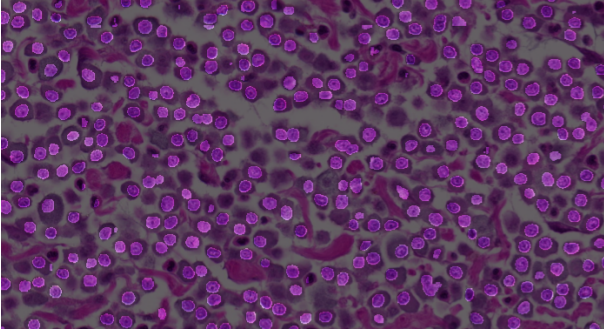


Figure 1. Algorithmic segmentation mask of tumor nuclei as an overlay on the histologic section of a ccMCT.

## 2. Material and Methods

### 2.1. Development of a segmentation model

Hematoxylin & Eosin-stained sections of 65 ccMCTs were digitized by a whole slide imaging scanner at  $0.25 \frac{\mu m}{px}$ . For development of the ground truth dataset, 86 representative regions of  $0.1185 \text{ mm}^2$  were selected and the boundaries of 41,145 tumor nuclei were annotated using the software Slide Runner 2.0.0 [1]. The dataset was split for training (N = 62 cases), validation (N = 11 cases) and testing (N = 13 cases) of a UNet++ model [3]. Small objects and connected nuclei were removed.

### 2.2. Evaluation of prognostic relevance

For evaluation of the prognostic value, 96 additional cases of ccMCT with known tumor-specific survival of the patient were collected. The cases were split into two parts for two-fold cross-validation. Of each case, up to 5 regions ( $0.1185 \text{ mm}^2$ ) were extracted and used for analysis.

An algorithm was developed that post-processes the derived nuclear segmentation mask (see above) by computing the eccentricity and solidity (nuclear shape) as well as area and diameter (nuclear size) for each nucleus. For each of these features, the standard deviation, variance, mean and median value were calculated. The ideal combination of these features was determined with an XGBoost model for each of the two cross-validation folds.

## 3. Results

Evaluated on the segmentation validation data set, the model yielded an intersection over union (IoU) of 0.788 and a Dice score of 0.772 (see Fig. 1).

Tumor-related death was predictable on the validation sets with AUC values of 0.82 and 0.86, with the accuracy being 79.2% and 93.8%, respectively (Fig. 2 and Table 1).

Dataset split	AUC	Accuracy	Sensitivity	Specificity
1	0.82	79.2%	66.7%	81.0%
2	0.86	93.8%	71.4%	97.6%

Table 1. Prognostic value (tumor-related death) of the nuclear morphometry algorithm.

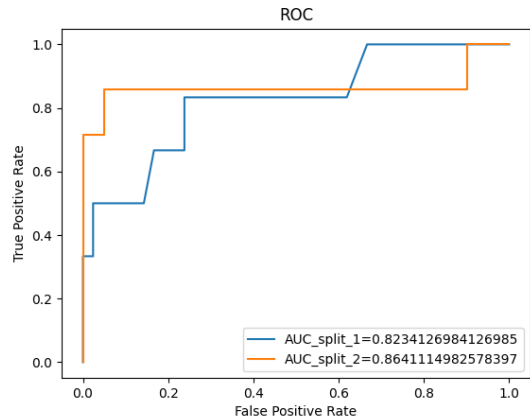


Figure 2. Receiver Operating Characteristic curves for tumor-related death of dogs with ccMCT based on algorithmic nuclear morphometry.

## 4. Discussion

Our results show that automated nuclear morphometry based on a deep learning model can provide an accurate prognosis in ccMCT. Due to its high time-efficiency and reproducibility, this methods seem promising for routine diagnostic use. Future studies should compare the prognostic value of automated nuclear morphometry with the pathologist's estimates and with other prognostic tests, such as the mitotic count, in large study populations. The influence of different image properties (such as between different whole slide image scanners), tumor types and image artifacts on algorithmic performance need to be evaluated.

## References

- [1] Marc Aubreville, Christof Bertram, Robert Klopffleisch, and Andreas Maier. SlideRunner. In *Bildverarbeitung für die Medizin 2018*, pages 309–314. Springer, 2018.
- [2] Mafalda Casanova, Sandra Branco, Inês Berenguer Veiga, André Barros, and Pedro Faísca. Stereology in grading and prognosis of canine cutaneous mast cell tumors. *Veterinary Pathology*, 58(3):483–490, 2021.
- [3] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. volume 11045 LNCS, 2018.

# Modeling the diffusion of CO<sub>2</sub> inside leaves

Yannis Sauzeau<sup>✉</sup>, Walter Kropatsch<sup>‡</sup>, and Jiří Hladůvka<sup>‡</sup>  
<sup>✉</sup>UFR SFA, University of Poitiers

yannis.sauzeau@etu.univ-poitiers.fr

<sup>‡</sup>Pattern Recognition and Image Processing Group, Vienna University of Technology

{walter.kropatsch, jiri.hladuvka}@tuwien.ac.at

## Abstract

*Propagation of fluids or gasses in closed compartments, like CO<sub>2</sub> in green plants, is described by diffusion equation. This partial differential equation is usually solved iteratively and, especially in higher dimensions, tends to be computationally intensive.*

*In this work, we propose to cast the  $n$ -dimensional problem to 1D diffusion. First, we apply a constrained distance transform to compute, for every voxel, its distance to the closest stoma. Second, we cast the iterative computation of CO<sub>2</sub> concentration to the evaluation of closed-form, polynomial functions. This in turn allows us to restrict the computation of CO<sub>2</sub> concentration to places of interest, e.g., to the close vicinity of the epidermis or cell walls where photosynthesis takes place.*

## 1. Introduction

To study gas exchanges, the diffusion equation is widely used [2, 7, 8]. The diffusion equation we choose is the heat equation, described in 1D by the following formula:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \quad (1)$$

where  $u(x, t)$  is the concentration at position  $x$  in time  $t$  and  $\alpha$  is the diffusion coefficient.

### 1.1. Iterative Solution in 1D

The majority of solutions use an iterative method using a finite difference scheme [4,5]. This method in 1D is defined as follows:

$$u(x, t+1) = u(x, t) + \alpha \sum_{n \in \Gamma(x)} (u(n, t) - u(x, t)) / |\Gamma(x)| \quad (2)$$

where  $\Gamma(x)$  is the set of neighbors of pixel  $x$ .

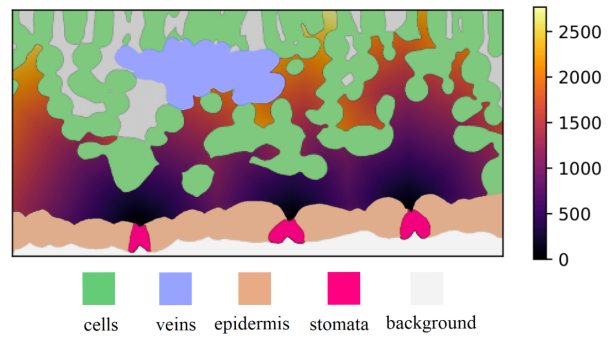


Figure 1. Constrained distance transform in air seeded at 3 stomata from 2D cross-section of a poplar leaf.

This formulation implicitly includes Neumann boundary condition [3] with a flow of 0. This condition assumes that the total gas volumes does not change by the diffusion, e.g. for all iterations  $t$  the total gas volume is constant Eq. (3).

$$\sum_x u(x, t) = \sum_x u(x, 0) \quad (3)$$

### 1.2. Constrained Distance Transform

To see the distance between two areas of interest in an image, we can use the constrained distance transform [9]. It is initialized by setting all elements of the constrained region  $R$ <sup>1</sup> to  $\infty$  and setting some seed points (stomata) to zero. Then the elements of the region repeatedly recompute their values with Eq. (4) until convergence. This algorithm can be performed with a logarithmic complexity [1].

$$d(x) = \min\{d(x), \min_{n \in \Gamma(x)} d(n) + 1\} \quad (4)$$

where  $\Gamma(x)$  is the set of neighbors of pixel  $x$ .

In Fig. 1, the distance transform assigned *every* air-pixel its distance from the closest stoma. Of uttermost interest,

<sup>1</sup>CO<sub>2</sub> diffuses in  $R$



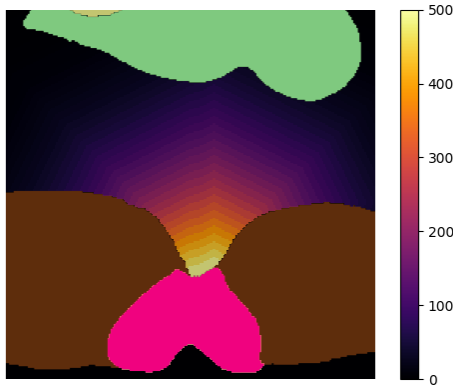


Figure 2. Eq. (5) used to describe the diffusion of CO<sub>2</sub> from the central stoma of Fig. 1.

however, are pixels next to cells (green), where photosynthesis takes place. In the following, we aim to show how to skip the calculation in areas that are of lesser interest.

## 2. Polynomial Basis Function

Consider one 1D sequence of 4-connected pixels without self-intersection. The out-of-leaf half ( $x \leq 0$ ) is initialized with high concentration of CO<sub>2</sub>,  $H$ , and the inner-leaf ( $x > 0$ ) with low concentration of CO<sub>2</sub>,  $L$ , the situation before the stomata open. Iterating with Eq. (2) we can see that the diffusion can be described by a polynomial in the diffusion coefficient  $\alpha$  of degree  $t$  with coefficients  $c(t, k, x)$ :

$$u(x, t) = H - (H - L) \sum_{k=0}^t c(t, k, x) \alpha^k \quad (5)$$

Deriving the coefficients of the polynomial (Tab. 1 shows the coefficients for the first 5 time steps), we arrived at the following closed-form involving binomial coefficients for negative arguments<sup>2</sup> [6].

$$c(t, k, x) = (-1)^{1+k+x} \binom{t}{k} \binom{2k-1}{k+x-1} \quad (6)$$

## 3. Results

To study the diffusion of CO<sub>2</sub> in the leaf from stomata to the leaf cells, we first compute the distance transform  $d(x)$  of each pixel  $x \in R$  in the airspace  $R$ . Afterward, we compute the coefficients  $c(t, k, x)$  with  $t = \max_{x \in R} d(x)$ . Once we have the coefficients, we can compute the concentration  $u(x, t)$  by Eq. (6). The result is identical to the iterative solution and is visualized in Fig. 2.

Another point of interest is to compute only the diffusion values for the pixels corresponding to the leaf cells border.

<sup>2</sup>Explaining the colored entries in Tab. 1

Indeed, to know the concentration of the leaf cells, we don't need to compute the diffusion values for the other parts of the leaf.

## 4. Further Work

This method can be extended to higher dimensions. Computation using Eq. (5) can be further optimized. Approximately half of the coefficients are equal to zero. With the symmetry of coefficients, only the upper half of the coefficients (for  $x > 0$ ) needs to be numerically (pre)computed. To sum the coefficients efficiently, we can use logarithmic coefficients  $\log(c(t, k, x)\alpha^k)$  and then make an exponentiation of the sum.

## 5. Conclusions

The paper presented a new approach to apply 1D diffusion on a pre-computed constrained distance transform. The method is based on polynomials and can be used to compute the concentration levels at specific times or for a specific pixel. One of the purposes is to study the CO<sub>2</sub> concentrations at locations that contribute to the photosynthesis. With this method, we can restrict the computations only to those parts of the leaf where the photosynthesis is likely to happen.

## Acknowledgments

The computational results have been partially achieved using the Vienna Scientific Cluster (VSC). This work was supported by the Vienna Science and Technology Fund (WWTF) project LS19-013.

## References

- [1] Majid Banaeyan, Carmine Carratù, Walter G. Kropatsch, and Jiří Hladůvka. Fast distance transforms in graphs and in Gmaps. In *IAPR Joint International Workshops on Statistical Techniques in Pattern Recognition (SPR 2022) and Structural and Syntactic Pattern Recognition (SSPR 2022)*, Montreal, Canada, August 26-27, 2022, page (in print). Springer, 2022.
- [2] Paola Diomedede, Mauritius C. M. van de Sanden, and Savino Longo. Insight into CO<sub>2</sub> dissociation in plasma from numerical solution of a vibrational diffusion equation. *The Journal of Physical Chemistry C*, 121(36):19568–19576, 2017.
- [3] Ivan Hlaváček, Jan Chleboun, and Ivo Babuška. In Ivan Hlaváček, Jan Chleboun, and Ivo Babuška, editors, *Uncertain Input Data Problems and the Worst Scenario Method*, volume 46 of *North-Holland Series in Applied Mathematics and Mechanics*, chapter X: Domains With Uncertain Boundary, pages 357–390. North-Holland, 2004.
- [4] Shou hui Zhang and Wen qia Wang. A stencil of the finite-difference method for the 2d convection diffusion equation and its new iterative scheme. *International Journal of Computer Mathematics*, 87(11):2588–2600, 2010.

Table 1. Coefficients  $c(t, k, x)$  for the very first time steps ( $t = 0 \dots 5$ ).

t	0			1			2			3				4				5				
k \ x	0	0	1	0	1	2	0	1	2	3	0	1	2	3	4	0	1	2	3	4	5	
4	1	1	0	1	0	0	1	0	0	0	1	0	0	0	-1	1	0	0	0	-5	9	
3	1	1	0	1	0	0	1	0	0	-1	1	0	0	-4	7	1	0	0	-10	35	-36	
2	1	1	0	1	0	-1	1	0	-3	5	1	0	-6	20	-21	1	0	-10	50	-105	84	
1	1	1	-1	1	-2	3	1	-3	9	-10	1	-4	18	-40	35	1	-5	30	-100	175	-126	
0	0	0	1	0	2	-3	0	3	-9	10	0	4	-18	40	-35	0	5	-30	100	-175	126	
-1	0	0	0	0	0	1	0	0	3	-5	0	0	6	-20	21	0	0	10	-50	105	-84	
-2	0	0	0	0	0	0	0	0	0	1	0	0	0	4	-7	0	0	0	10	-35	36	
-3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	5	-9	

- [5] A. Ibrahim and A. R. Abdullah. Solving the two dimensional diffusion equation by the four point explicit decoupled group (edg) iterative method. *International Journal of Computer Mathematics*, 58(3-4):253–263, 1995.
- [6] Maarten Kronenburg. The binomial coefficient for negative arguments. pages 1–6, May 2011.
- [7] David F. Parkurst. Diffusion of CO<sub>2</sub> and other gases inside leaves. *New Phytologist*, 126(3):449–479, 1994.
- [8] H. Rahn, O.D. Wangenstein, and L.E. Farhi. Convection and diffusion gas exchange in air or water. *Respiration Physiology*, 12(1):1–6, 1971.
- [9] Tilo Strutz. The distance transform and its computation - an introduction, 06 2021.

# Application Spotlight Papers

# Novel contactless fingerprint scanner for Legal Enforcement Agencies

Axel Weissenfeld, Erich Voko, Bernhard Strobl, Bernhard Kohn, Gustavo Fernández Domínguez  
AIT Austrian Institute of Technology, Center for Digital Safety & Security,  
Giefinggasse 4, 1210 Vienna, Austria

axel.weissenfeld, erich.voko, bernhard.strobl, bernhard.kohn, gustavo.fernandez@ait.ac.at

Reinhard Schmid  
Bundeskriminalamt, Bundesministerium für Inneres BMI  
Vienna, Austria

Reinhard.Schmid@bmi.gv.at

## Abstract

*Biometric recognition systems integrated into mobile devices have gained acceptance during recent years. Authorities are particularly interested in mobile contactless solutions due to many reasons: officers can acquire data wherever they are, solutions are generally easy to use, hygiene and no latent data is present. This paper presents a new mobile contactless fingerprint sensor which uses a liquid lens integrated with a TOF sensor. The device was used by the national police to acquire data of refugees. Matching results show promising results, while police officers expressed their satisfaction about the developed prototype.*

## 1. Introduction

Contactless (CL) and mobile-embedded biometric recognition systems have made considerable progress and gained acceptance in recent years. Advantages of CL devices are high-user acceptance, no latent data is present in the acquisition device, hygienic reasons, less effort to acquire data, usability and speed. On the other hand, mobile devices also present many advantages: portability, more productivity and higher efficiency of the end-user, and the possibility to acquire data on different locations, i.e. unconstrained capturing environment. This paper presents a new mobile CL fingerprint (FP) sensor which uses a liquid lens integrated with a Time-of-Flight (TOF) sensor. A liquid lens substitutes a static optic glass lens and it has not any mechanical parts inside. Liquid lenses are controlled cells containing a transparent fluid capsule. Changing the shape of the cells, the focal length is changed within milliseconds. Despite the advantages of liquid lenses ((i) no moving parts inside the lens, (ii) one lens can deliver different focal lengths, (iii) good image quality, and (iv) speed), not much

work was reported. Oku and Ishikawa [6] reported a high-speed liquid lens and its applications in different computer vision applications. Tsai et al. [10] used a liquid lens to acquire finger images. Their approach applied a strong illumination and a small distance between the lens and the fingertip to minimize environmental distortions. Recently, Jun and Won [4] used a liquid lens together with chromatic aberration to improve accuracy in depth-measurement of real 3D objects. The main contribution of this paper is the introduction of an operative mobile CL device aimed for police use. The device uses a liquid lens combined with a TOF sensor. The fully functional CL prototype is a mobile FP capturing tool aiming to optimize the process carried out by national police officers. To show the feasibility of the developed prototype, results of a matching comparison on real FP data are presented. This paper is organized as follows: Section 2 describes the developed prototype, the processing of the finger images and acquired data. Results are discussed in Section 3. Concluding remarks are summarised in Section 4.

## 2. Processing Chain

### 2.1. Developed device

We developed the capture device under various requirements: (i) to be mobile and contact-less, (ii) to record high quality images, (iii) easy to use, and (iv) the hardware costs should be as low as possible. Active illumination of the fingers during recording is a key feature to achieve a sufficiently strong contrast between the ridges and valleys of the fingertips. We use a daisy chain to connect 64 colour LEDs arranged in a U-shape with 45 degrees. The arrangement of the light emitting elements ensures a uniform illumination and provide sufficient illumination to all four fingers of a single hand. The camera sensor delivers grey-scale



Figure 1. Left image: contactless fingerprint capture device. Right image: device usage.

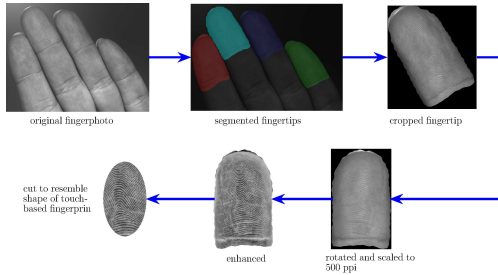


Figure 2. Overview of the processing pipeline.

images ( $3052 \times 2015$  pixels @ 10 fps). The sensor is connected via a USB3 interface to the processor. The minimum distance between the sensor and finger is 105 mm and the maximum is 175 mm. The camera sensor is supplemented by a TOF sensor to measure the distance of the fingers to the sensor. Thus, the most suitable focal plane can be targeted to capture sharp fingertips. We integrated a digital variable-focus liquid lens into the device which enables the acquisition of the fingers at pre-elected focal planes and within 5 ms a new focal plane can be reached. Fig. 1 shows the developed device. In order to capture images consisting of the detailed topology of the fingers we configured the camera in fast exposure modes with low apertures. In a pre-processing step, a flat-field correction [8] is carried out to cancel the effects of image artifacts caused by variations in the pixel-to-pixel sensitivity of the sensor, and the images are rectified to correct for lens distortions [11].

## 2.2. Image processing

The processing pipeline is depicted in Fig. 2. The fingertips are segmented using a Mask R-CNN [3], which is a two-stage pipeline for instance segmentation. Manually labeled fingertips images are used to fine tune the pre-trained model. Then, the images are cropped, rotated to a finger upright position, and scaled to 500 DPI (to be compliant to FBI-standards [2]). Because the device works with a small depth of field many images are blurred. Thus, we perform a sharpness evaluation based on edge pixels [1] and images with sufficient sharpness are enhanced to increase the con-

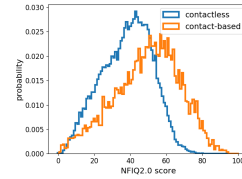


Figure 3. Probability density functions of NFIQ2.0 scores of contactless and contact-based fingerprints.

trast between ridges and valleys by applying a histogram equalization [7]. We use the well-established NFIQ2.0 [9] standard for quality estimation. NFIQ2.0 was developed for touch-based fingerprints and therefore likely not optimal for CL fingerprints.

## 3. Results

Data was acquired by national police officers using two devices: the developed prototype and a contact-based (CB) fingerprint device<sup>1</sup>. Data (all 10 fingers) of 481 people was acquired. Based on the sharpness value (at least 0.2) [5], the best 6 images of each finger were selected. To acquire 10 fingers the acquisition time is between 45 seconds and 120 seconds in case of the CB sensor, and between 8 seconds and 30 seconds in case of the CL prototype. Recording sessions took place in a national refugee registration center; thus, the dataset is diverse: people came from 4 different continents, and the gender distribution is 68.52% male, 31.32% female and 0.16% not indicated. In order to give an impression of the quality of the captured fingerprints, the distributions of NFIQ2.0 scores are shown in Fig. 3. The biometric performance is evaluated employing the ID-Kit SDK 8.0.1.50. We obtained equal error rates (EER) for different NFIQ2.0 thresholds in both cases, matching CL data against CL data (CL  $\rightarrow$  CL) and CL data against CB data (CL  $\rightarrow$  CB). When using a NFIQ threshold  $\geq 20$ , obtained EERs are CL  $\rightarrow$  CL =  $1.1e-04\%$  and CL  $\rightarrow$  CB =  $2.7e-04\%$  which are very good values in terms of performance recognition. The end user also expressed their satisfaction with the solution: it is very simple to use, the scanner turns on automatically when the hand is held over it and capturing data seems to be easier and better than using flatbed sensors.

## 4. Conclusion

This work presented a new mobile CL fingerprint sensor integrating a liquid lens and a TOF sensor, and its usability in a real setting scenario. Real fingerprint data was acquired by national police officers who expressed their satisfaction

<sup>1</sup>Optical fingerprint scanner IDEMIA TP 5300 scanner with 1000 DPI, <https://www.idemia.com/palmpoint-scanner>

with the developed prototype. Acquired data was used to perform a fingerprint matching against data of an official database acquired using a CB device. Promising and encouraging results were obtained showing the feasibility of the prototype for operational police use.

## Acknowledgements

This work was partially supported by the AIT Strategic Research Program 2022. We gratefully acknowledge the continuous support of BMI during the recording sessions and the whole study.

## References

- [1] Jorge Caviedes and Sabri Gurbuz. No-reference sharpness metric based on local edge kurtosis. In *Proceedings. International Conference on Image Processing*, volume 3, pages III–III. IEEE, 2002.
- [2] American National Standard for Information Systems. Nist special publication 500-290: Data format for the interchange of fingerprint, facial & other biometric information. US Department of Commerce, Technology Administration, National Institute of Standards, and Technology, 2011.
- [3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [4] Gyu Suk Jung and Yong Hyub Won. Compact and fast depth sensor based on a liquid lens using chromatic aberration to improve accuracy. *Optics Express* 15780, 29(10):15786–15801, 2021.
- [5] Christof Kauba, Dominik Söllinger, Simon Kirchgasser, Axel Weissenfeld, Gustavo Fernández Domínguez, Bernhard Strobl, and Andreas Uhl. Towards using police officers’ business smartphones for contactless fingerprint acquisition and enabling fingerprint comparison against contact-based datasets. *Sensors*, 21(7):2248, 2021.
- [6] Hiromasa Oku and Masatoshi Ishikawa. High-speed liquid lens for computer vision. In *IEEE International Conference on Robotics and Automation*, pages 2643–2648. IEEE, 2010.
- [7] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38(1):35–44, 2004.
- [8] James Anthony Seibert, John M. Boone, and Karen K. Lindfors. Flat-field correction technique for digital detectors. In *Medical Imaging 1998: Physics of Medical Imaging*, volume 3336, pages 348–354. SPIE, 1998.
- [9] Elham Tabassi, Martin Olsen, Oliver Bausinger, Christoph Busch, Andrew Figlarz, Gregory Fiumara, Olaf Henniger, Johannes Merkle, Timo Ruhland, Christopher Schiel, and Michael Schwaiger. Nist fingerprint image quality 2, 2021.
- [10] C. Tsai, P. Wang, and J. Yeh. Compact touchless fingerprint reader based on digital variable-focus liquid lens. In *SPIE Optical Engineering and Applications*, volume 9193, pages 173–178. SPIE, 2014.
- [11] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

# Crop row detection utilizing spatial CNN modules

Peter Riegler-Nurscher  
Josephinum Research  
Wieselburg, Austria

p.riegler-nurscher@josephinum.at

Leopold Rupp  
CFS Cross Farm Solution GmbH  
Stoitzendorf, Austria

lr@cfsolution.at

## Abstract

*Mechanical weed control is becoming increasingly important over conventional methods, not least because of environmental challenges. Precise guidance of the hoeing machine along the crop rows is necessary to be able to work efficiently. In this work, the use of deep learning methods for crop row detection is presented and evaluated on a custom data set. Recent advances in the task of vision based lane detection, like Spatial CNN (SCNN) and Recurrent Feature-Shift Aggregator (RESA), can potentially be applied to crop row detection as well. These methods are expected to improve the detection of the crop rows, especially in the case of strong weed growth and challenging environmental conditions, compared to the state of the art.*

## 1. Introduction

There is a steadily increasing demand on the food market for food produced according to organic farming standards. Likewise, the proportion of organically farmed agricultural areas or organically managed arable farms in Europe is growing continuously. The change from chemical to mechanical weed control can only remain economical with a high degree of automation. Row hoeing equipment for weed removal often uses duck-foot shares that must be guided precisely along the row to prevent crop damage. State of the art camera systems used for row guidance have limitations, due to the robustness of conventional row detection algorithms especially with strong weed cover [6].

### 1.1. Related Work

Plant row detection in robotics as well as in marketed row guidance systems is based on conventional methods like line detection and color thresholds [4]. However, first approaches for convolutional neural network (CNN) based row recognition have already been presented [2]. Recent advances in the task of vision based lane detection could potentially be applied to crop row detection as well. These methods can be categorized into segmentation-based, point-based and curve-based lane detection methods [3]. Point-based methods directly output points whereas curve-based

methods output curve parameters. Accordingly, other loss functions are used during training for point- and curve-based methods compared to segmentation based methods. Point- and curve-based methods are not discussed in detail in this paper and are also not included in the evaluation. Segmentation-based methods, like Spatial CNNs (SCNN) [7] and Recurrent Feature-Shift Aggregator (RESA) [9], output segmentation masks. A threshold is applied to the output to get sample discrete points on the lines. SCNN uses a spatial CNN module to model spatial relationships more efficiently than MRF or CRFs. The module is integrated after the top hidden layer. It preserves the continuity of long, thin structures over discontinuities. The RESA proposed in [9] utilizes spatial information by shifting sliced feature map. RESA is more computationally efficient than SCNN and also introduces an up-sampling decoder, the so called Bilateral Up-Sampling Decoder (BUSD). It is composed of two branches, a coarse grained branch and a fine detailed branch.

Methods for lane detection have not been utilized for crop row detection before. In this work, a new approach for crop row detection based on CNNs is presented. Additionally, the SCNN and RESA methods with different backbone CNNs are investigated for this task with a custom data set.

## 2. Method

Twelve different architectures for crop row detection were tested. As backbones ResNet [5] architectures of different sizes (ResNet18, ResNet34 and ResNet101), as well as a VGG16 [8] architecture are used. The up-sampling is done by a DeepLab [1] architecture except for the RESA method where the Bilateral Up-Sampling Decoder is used. The models take an input size of  $800 \times 288$  pixel and outputs the segmentation mask in the same resolution. For augmentation, the training images were randomly flipped, rotated and a color jitter as well as random lighting were added. The backbone CNNs were pretrained on ImageNet. The implementation is built upon the framework introduced in [3]. All models were trained 120 epochs on the data set introduced in the following Section 3. At 120 epochs, convergence was observed for all variants.

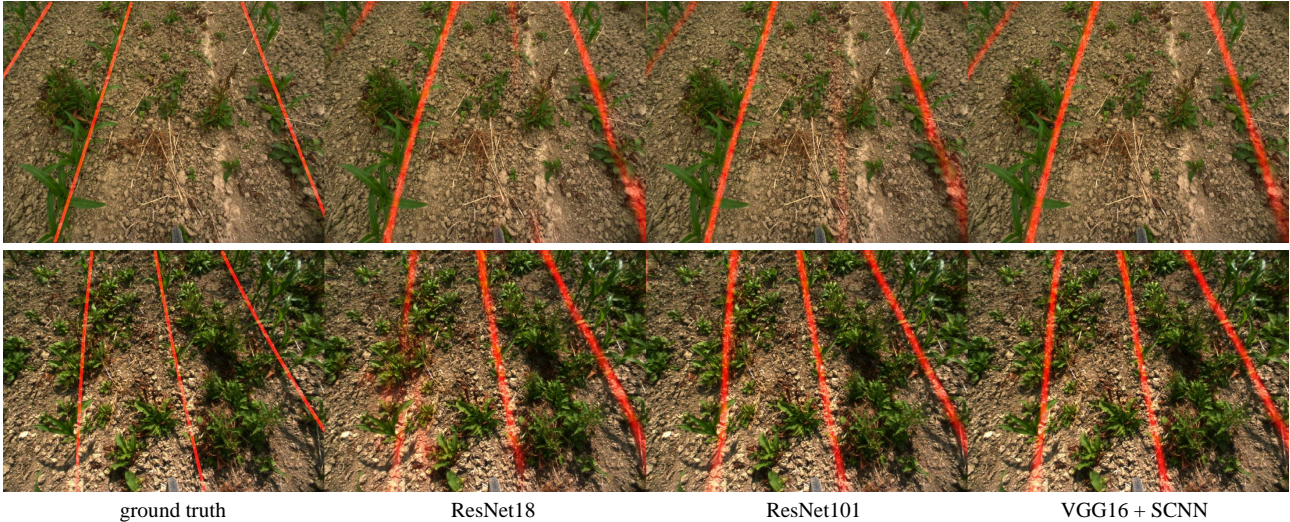


Figure 1. Segmentation result shown in red of different models (side by side) for two example test images (among one another).

### 3. Evaluation

For training and evaluation of the different methods, a custom data set was created. The data set consists of 3870 images of maize rows captured in the seasons 2021 and 2022 under various lighting conditions. The RGB images have a resolution of  $1600 \times 1200$  pixels. The images are labelled with our custom labelling tool. A row can be defined by clicking a minimum of 2 points within a row. The row is afterwards interpolated by a polynomial of degree 2 from which regularly sampled points are stored. A segmentation mask is automatically generated by drawing curves with width of 16 pixel. All architectures have an input size of  $800 \times 288$  pixel, therefore all images are resized to this resolution. The data set is split into 3475 images for training and 395 test images.

#### 3.1. Results

The segmentation accuracy of the different methods on the test images is presented in Table 1. Figure 1 shows two examples of the test set with the segmentation results for ResNet18, ResNet101 and VGG16 with SCNN. The numbers presented, as well as the sample images, show an advantage in the use of SCNNs for detecting crop rows. The superiority of RESA over SCNN on lane detection data sets [3] could not be achieved in crop rows. Although, improvements of RESA over the baseline model (just Backbones with Deeplab) are visible. Likewise, it is recognizable that larger models, like ResNet101, achieve better detection rates without the use of spatial modules.

It might be assumed that the model is implicitly distinguishing crops from weeds based on test images with strong weed cover. However, this needs to be investigated in more detail.

	ResNet18	ResNet34	ResNet101	VGG16
Baseline				
Accuracy	66.83	67.63	<b>67.90</b>	66.98
IoU	44.07	45.48	<b>44.82</b>	44.53
SCNN				
Accuracy	68.15	69.18	68.58	<b>70.33</b>
IoU	44.19	44.93	44.83	<b>45.59</b>
RESA				
Accuracy	66.51	<b>68.06</b>	56.93	N/A
IoU	43.58	<b>44.30</b>	40.66	

Table 1. Accuracy and Intersection over Union (IoU) of the crop row segmentation based on 395 test images.

### 4. Conclusion and Outlook

The work demonstrates the ability of CNNs for semantic segmentation to detect crop rows. Especially the SCNN but also the RESA method could improve the detection compared to the baseline methods. When selecting a method for steering a hoeing machine, however, the computational load should also be taken into account where SCNN has its drawbacks. Currently, we are working on integrating the models into a machine to steer along the rows. This allows for end to end evaluation of the system and an assessment of acceptable model errors. Future work could also focus on point- and curve-based methods.

### Acknowledgments

This research was funded by FFG (Austrian Research Promotion Agency) under grant 880610 (DeepRow).



## References

- [1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR*, 2015.
- [2] Rashed Doha, Mohammad Al Hasan, Sohel Anwar, and Veera Rajendran. Deep learning based crop row detection with on-line domain adaptation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '21*, page 2773–2781, 2021.
- [3] Zhengyang Feng, Shaohua Guo, Xin Tan, Ke Xu, Min Wang, and Lizhuang Ma. Rethinking efficient lane detection via curve modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 17062–17070, 2022.
- [4] Chuangxin He, Qingtai Chen, Zhonghua Miao, Nan Li, and Teng Sun. Extracting the navigation path of an agricultural plant protection robot based on machine vision. In *Proceedings of the 40th Chinese Control Conference, CCC*, pages 3576–3581, 2021.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778, 2016.
- [6] Oskar Kress, Sabine Staub, and Simon Brell. Abschlussbericht Forschungsprojekt "Beikrautregulierung in Ökobetrieben mit Gemüsekulturen unter besonderer Betrachtung von moderner RTK-Steuerungs-, Ultraschall- und Kamertechnik inkl. Arbeitswirtschaft und Kosten". Bayerische Landesanstalt für Weinbau und Gartenbau (LWG), 2019.
- [7] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 7276–7283, 2018.
- [8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR*, 2015.
- [9] Tu Zheng, Hao Fang, Yi Zhang, Wenjian Tang, Zheng Yang, Haifeng Liu, and Deng Cai. Resa: Recurrent feature-shift aggregator for lane detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3547–3554, 2021.

# A Computer Vision System for Evaluation of Field Robot Operations

Florian Kitzler, Andreas Gronauer, Viktoria Motsch  
Department of Sustainable Agricultural Systems, Institute of Agricultural Engineering  
University of Natural Resources and Life Sciences, Vienna  
Peter-Jordan-Straße 82, 1190 Vienna, Austria

florian.kitzler@boku.ac.at

## Abstract

*The usage of field robots is increasing as more commercial products become available on the market. Among other measures, they can be used for seeding and mechanical weed control, using the geolocation of each individual seedling. The weed control process is performed without visual recognition of the plants. The precision of such weed control robots depends on the quality of the localisation, plant emergence, and soil properties. In order to evaluate the field robot operation accuracy, we developed a cost-effective, long-term autonomously working computer vision evaluation system based on two RGB cameras for pre- and post-weed control image capture. Our system was successfully tested to collect image data of the hoeing precision of a FarmDroid FD20 field robot.*

## 1. Introduction

Recently, robotics and automation is playing an important role in smart farming technologies. Available field robots can be used for a variety of operations, e.g. for mechanical weed control using duck foot share and active hoes [1], a tube stamp [4] or a side-shifting frame [5]. Some of the field robots use computer vision to identify different plant species, such as [4], while others rely on a global navigation satellite system (GNSS) for high precision localisation [5]. Those systems store the location of each individual seedling and operate the area around the expected crops. The precision of such a system depends on the accuracy of the sowing, the regularity of the emergence of the crops, and the precision of the localisation [2]. The objective of this work is to develop a computer vision system to evaluate the weed control precision of a field robot.



Figure 1. Field robot FarmDroid FD20 with mounted computer vision evaluation system (CVES) on the field.

## 2. Material and Methods

### 2.1. Computer vision evaluation system

Our computer vision evaluation system (CVES) consists of a front and back camera, see Figure 1. It was used to collect RGB images before and after the weed control took place. We refer to those images as the pre-weed control (Figure 2, top left) and post-weed control (Figure 2, bottom left) images, respectively. The system consists of two single-board computers (Raspberry Pi - Model 3B+) with an integrated RGB camera (Raspberry Pi camera v2) with a resolution of  $2596 \times 1944$  px. To mount the cameras top-down at a height of 1 m, we built an aluminum carrier, that could be adjusted to the robot construction. A GNSS module (Emlid Reach M+) with an antenna above the back camera was used for RTK GNSS localization. Consumer-grade power banks were used to supply the CVES, resulting in operation times of about 8 hours.

### 2.2. Field robot

The used field robot FarmDroid FD20 [1] has seeding modules, duck foot share for inter-row and active hoes for

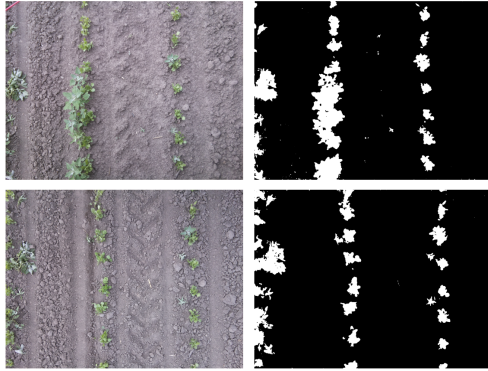


Figure 2. Representative RGB images (left) and vegetation segmentation (right) of the front camera (pre-weed control image, top) and the back camera (post-weed control image, bottom) at the same location during weed control of the anise plot at 2022/05/24.

intra-row hoeing. It is equipped with two GNSS antennas and receives real-time kinematic (RTK) correction signals. The field robot is powered by batteries that are loaded with solar panels and achieves to work for up to 24 hours, fully autonomously. The operating speed is limited to a maximum of  $0.26 \text{ m s}^{-1}$ . The weeding is based solely on the localization and the stored seeding points of the crops.

### 2.3. Field trial

Data was collected at a plot in Fuchsenbigl, Lower Austria, Austria in 2022. The field robot was used for seeding a 3.2 ha plot of anise (*Pimpinella anisum*). After seeding, our CVES was mounted. The first weed control was triggered manually based on the emergence of weeds and on a regular basis in the following weeks. Before the first weed control a test run of the CVES was performed. At selected weed control dates, we activated our CVES by connecting it with the power supply and validated the status. If the setup phase was successful, the system worked autonomously and the weed control process of the field robot continued.

## 3. Results

### 3.1. Evaluation based on total plant cover

Exemplary data analysis has shown that a decrease in the total plant cover can be observed between the pre- and post-weed control image of the same site. Here the plant cover dropped from 9.0% (pre-weed control, Figure 2 top right) to 6.4% (post-weed control, Figure 2 bottom right). We used decision tree classifiers based on color index maps for the vegetation segmentation [3] and analyzed all images of pre- and post-weed control in a given operating area, leading to a total of 9,113 images from two measurement dates. Results are given in Table 1.

Images	Date	Camera	m PC	sd PC
2,027	2022/05/06	Pre	0.58%	0.22%
1,808	2022/05/06	Post	0.44%	0.19%
2,649	2022/05/24	Pre	4.19%	2.36%
2,629	2022/05/24	Post	4.10%	2.27%

Table 1. Comparison of pre- and post-weed control total plant cover (PC) for two measurement dates, mean (m) and standard deviation (sd) are given in %.

## 4. Conclusion

We successfully developed a cost-effective, long-term autonomously working system to capture geolocated RGB images with the field robot FarmDroid FD20. A plant cover based evaluation of the whole area has limited power, see Table 1. Therefore, more advanced methods, such as a three-class semantic segmentation method are currently being developed. They should be used to distinguish between crops, weeds, and killed weeds but need more labeling effort. Our CVES is very flexible to use and can be adapted for usage with other field robots or robotic platforms.

## Acknowledgments

The authors acknowledge the funding of the project by SONNENTOR Kräuterhandels GesmbH.

## References

- [1] Farmdroid FD20 field robot. <https://farmdroid.dk/en/product/>. Accessed: 2022-07-25.
- [2] Hans W Griepentrog, Michael Nørremark, Henning Nielsen, and BS Blackmore. Seed mapping of sugar beet. *Precision Agriculture*, 6(2):157–165, 2005.
- [3] Florian Kitzler, Helmut Wagentristl, Reinhard W. Neugschwandtner, Andreas Gronauer, and Viktoria Motsch. Influence of selected modeling parameters on plant segmentation quality using decision tree classifiers. *Agriculture*, 12(9), 2022.
- [4] Frederik Langsenkamp, Fabian Sellmann, Maik Kohlbrecher, Arnd Kielhorn, Wolfram Strothmann, Andreas Michaels, Arno Ruckelshausen, and Dieter Trautz. Tube stamp for mechanical intra-row individual plant weed control. *Proceedings of the 18th World Congress of CIGR, Beijing, China*, pages 16–19, 2014.
- [5] Michael Nørremark, Hans W Griepentrog, Jon Nielsen, and H Tangen Sjøgaard. The development and assessment of the accuracy of an autonomous gps-based system for intra-row mechanical weed control in row crops. *Biosystems engineering*, 101(4):396–410, 2008.

# Vision-Language Models for Filtering and Clustering Forensic Data

Axel Weissenfeld, Bernhard Strobl  
AIT Austrian Institute of Technology, Center for Digital Safety & Security,  
Giefinggasse 4, 1210 Vienna, Austria

axel.weissenfeld, bernhard.strobl @ait.ac.at

David Weichselbaum, Christopher Wimmer, Martina Tschapka  
david.t.weichselbaum@gmail.com, cedwimmer@gmx.at, martinatschapka@gmail.com

## Abstract

With image- and video-capable devices in the hands of a majority of the population worldwide, the amount of media data keeps growing. Hence, the search of specific images and clustering of datasets is of great importance to extract the relevant information, e.g. search for a specific person by legal enforcement agencies (LEAs). This paper presents a new tool which uses vision-language models to filter and cluster forensic data. The tool provides a GUI, which enables a flexible search by accepting textual as well as image input, to search large amounts of data in near real-time.

## 1. Introduction

The search for a specific person in images and videos is an important task in forensics and part of Content-Based Image Retrieval (CBIR) [3]. Unfortunately, a manual search is very time-consuming and a fully automatic search is usually not applicable. As a result, critical evidences might literally be hidden in plain sight, among an overwhelming number of images and videos.

In this work we do propose a system based on vision-language models assisting an operator to quickly filter and cluster image data by searching for pedestrian attributes. Pedestrian attributes are humanly searchable semantic descriptions such as gender, hair length, clothing style, or facial features and can be used as soft-biometrics in visual surveillance.

Pedestrian attributes recognition (PAR) is often approached as a metric learning problem [11], where one seeks to retrieve images containing the person attributes (Fig. 1). This is challenging in the sense that images captured by different cameras often contain significant intra-class variations caused by the changes in background, viewpoint, human pose, etc.

The developed solution shall enable the operator to filter

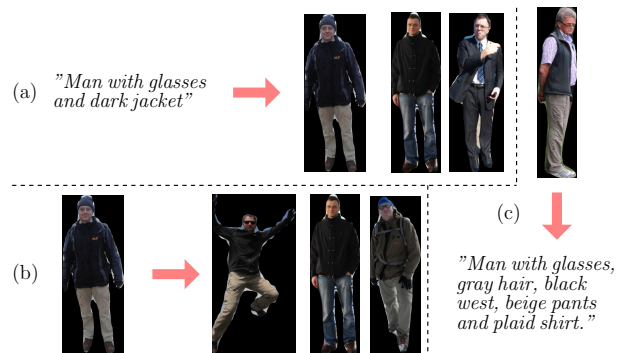


Figure 1. Vision-Language models enable a flexible search: (a) text→image retrieval, (b) image→image retrieval, (c) image→text retrieval

and cluster image data by person attributes (Sec. 2). Some results are presented in Sec. 3.

## 2. Vision-Language Models for Filtering and Clustering

More recently, CBIR systems have been extended by multimodal inputs such as image-text pairs, which we denote as Vision-Language (VL) models [2, 6, 10]. In contrast to prior models that are trained on images with class annotations, VL models are directly trained on image-text pairs to group relevant text vectors matching to the meaningful image content vectors. Recently some very large models such as CLIP [9], ALIGN [5], and BASIC [8] were trained, which achieve large robustness even on challenging datasets and a high accuracy with zero-shot classification. For instance, CLIP is a contrastive approach to learn image representations from text, with a learning objective which maximizes similarity of correct text-image pair embeddings.

VL models allow a textual search in image data by en-

tering keywords or sentences. The possibility of extensive textual input is especially beneficial for PAR analysis, since different attributes can be combined. For example, if an operator searches for a man with a black backpack, the operator can enter "man with a black backpack". The model returns a confidence score (probability) that an image contains the searched attributes. The free text input simplifies the use of the system. There are also numerous application variations such as providing reference image as input or extracting semantic information from images (Fig. 1). Moreover, VL models also produce robust embeddings, which are indispensable to accurately cluster forensic datasets.

Optionally persons in the image data can be segmented (instance segmentation) and extracted from the original image. While object detection identifies objects in the image data, segmentation assigns an object class to each pixel. For the instance segmentation used here, a network architecture called Mask R-CNN is used [4]. The Mask R-CNN model was fine-tuned on the OpenImages<sup>1</sup> and Coco<sup>2</sup> datasets to segment persons. A single segmented person is the input to our VL model as illustrated in Fig. 1.

The developed tool for clustering and filtering image data enables LEAs to search their data in a targeted and focused manner, but also to conduct general screenings of large data sets (working with over 100.000 files) before a clear investigation target is defined. The VL model generates 512-dimensional embedding vectors of the image-text input. The tool automatically groups content into meaningful clusters using unsupervised machine learning [1] and arranges the input images by a nonlinear down projection. The tool also provides a simple GUI to search for specific persons by providing natural language search terms or an image or by selecting special trained classifiers. In addition to the reference text or reference image search, the third alternative for the analysis of image data is the implementation of particularly trained MLP (multi-layer perceptron) networks that were trained on several classes. By choosing a certain classifier inside the application, the provided image data is automatically classified. For each class, a distinct single-label classifier was trained.

### 3. Results

For generating the results we used the ViT-B/16 model [9]. Fig. 2 depicts the clustering result of a fraction of the PA-100k dataset as well as the search result for the reference image shown on the left. Similar results can be received if a textual description is the input.

The VL model reliably classifies pedestrian attributes as illustrated in Fig. 3. For illustration purposes the corresponding attention parameters are saved per residual at-

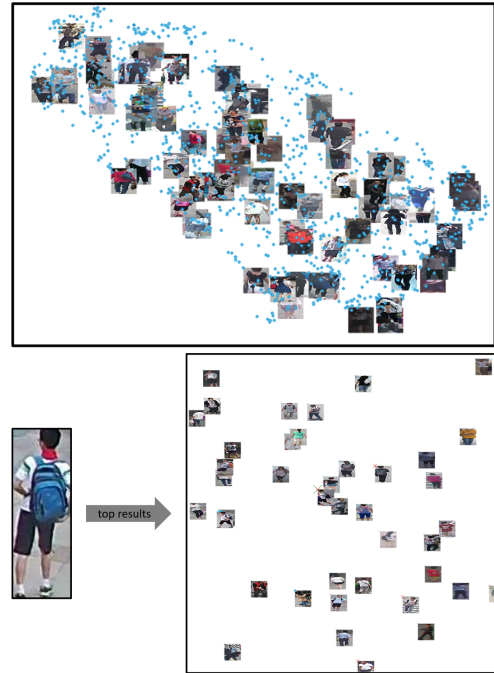


Figure 2. Top: Clustered and displayed result of a fraction of the PA-100k dataset. Bottom: Found top results for a sample image of the PA-100k dataset (provided by [7] under the CC BY 4.0 license<sup>3</sup>). Using a visual indication (red crosses) in the presented two-dimensional space, the software identifies the most relevant images to the user's particular search operation.

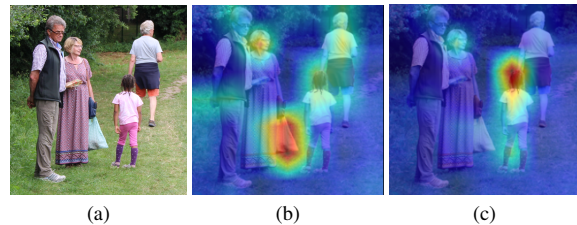


Figure 3. Searching in (a) for a bag and child. Resulting heatmaps are displayed in (b) and (c), respectively.

tention block during the forward pass of an image and a backpropagation is conducted following the forward pass computation, with respect to the known output vector. After multiplying these two values (attention value and gradient) in the respective layers and the respective subspace, the computed information of all residual attention blocks is superimposed as a heatmap over the input image.

The filtering can be very efficiently executed and enables near real-time searches, since the embeddings are highly compact and only a dot product between text and image or image and image embeddings need to be carried out.

<sup>1</sup><https://opensource.org/licenses/MIT>

<sup>2</sup><https://cocodataset.org>

<sup>3</sup><https://creativecommons.org/licenses/by/4.0/>

## 4. Conclusions

This work presents a new tool to search through forensic data. The tool is highly flexible because of the used VL models, which enable a clustering as well as search using pedestrian attributes or a reference image. Promising and encouraging results were obtained showing the feasibility of the tool for operational use by LEAs.

Note, that the solution is not limited to persons.

## Acknowledgments

This research work has been supported by Vienna Business Agency.

## References

- [1] Yoshua Bengio, Aaron C Courville, and Pascal Vincent. Un-supervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, 1(2665):2012, 2012.
- [2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.
- [3] Venkat N Gudivada and Vijay V Raghavan. Content based image retrieval systems. *Computer*, 28(9):18–22, 1995.
- [4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [5] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.
- [6] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [7] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2017.
- [8] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.
- [9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [10] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [11] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

# Estimation of nitrogen yield in wheat using radiative transfer model inversion based on an artificial neural network

L. J. Koppensteiner  
Institute of Agronomy, BOKU  
Konrad Lorenz-Straße 24, 3430  
lukas.koppensteiner@boku.ac.at

R. W. Neugschwandtner  
Institute of Agronomy, BOKU  
Konrad Lorenz-Straße 24, 3430  
reinhard.neugschwandtner@boku.ac.at

## Abstract

The objective of this study was to estimate nitrogen yield in wheat based on hyperspectral reflectance measurements with a handheld spectroradiometer. To do so, the radiative transfer model PROSAIL was inverted and an artificial neural network applied. The model was trained and tested using a simulated dataset and field experimental data. Results of the simulated dataset show that the inversion of PROSAIL based on an artificial neural network was successful. Furthermore, estimations of nitrogen yield compared to experimentally collected data feature high  $R^2$  and low RRMSE. The technique proposed in this study is a promising tool to collect information on nitrogen yield of wheat canopy in a quick and non-destructive way with low calibration requirements. This can be utilized by practical farmers for field monitoring and site-specific nitrogen fertilization as well as scientists and breeders for quick and non-destructive data collection in field experiments. Additionally, this approach can be adapted for different crops and varying sensors, e.g., multi- and hyperspectral UAV-mounted sensors as well as satellite data.

## 1. Introduction

Remote sensing allows quick and non-destructive measurements of canopy characteristics. Commonly, vegetation indices are applied, however, this approach usually requires continuous calibration and cannot use all available spectral data for analysis. Radiative transfer models (RTMs) are a promising alternative to vegetation indices. These models describe the interaction between solar radiation and vegetation canopy [1]. Compared to vegetation indices, RTMs generalize well, have low calibration needs and allow analysis of all available spectral data [2].

The objective of this study was to estimate nitrogen yield in wheat (*Triticum aestivum* L.) using an artificial neural network-based inversion of the RTM PROSAIL.

## 2. Materials and methods

The RTM PROSAIL simulates the spectral reflectance of vegetation canopy from 400 to 2500 nm in 1 nm increments using information on leaf characteristics, canopy architecture, viewing geometry and other effects (Figure 1). Simulations in the RTM PROSAIL (version 5B) were conducted using the package hsdar (version 1.0.3) in R programming language (version 4.1.1).

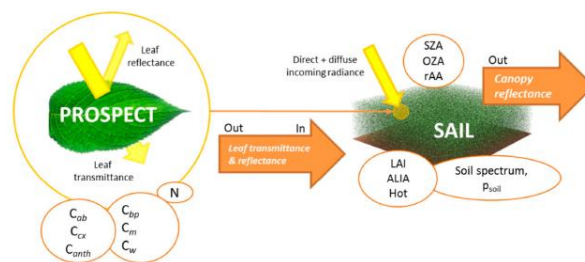


Figure 1: Calculation of canopy reflectance using the coupled PROSPECT + SAIL model (PROSAIL) [2]. N: leaf structure index (unitless).  $C_{ab}$ : chlorophyll a + b content ( $\mu\text{g cm}^{-2}$ ),  $C_{cx}$ : carotenoid content ( $\mu\text{g cm}^{-2}$ ),  $C_{anth}$ : anthocyanin content ( $\mu\text{g cm}^{-2}$ ),  $C_{bp}$ : brown pigment content (unitless),  $C_m$ : dry matter content ( $\text{g cm}^{-2}$ ),  $C_w$ : water depth (mm), LAI: leaf area index ( $\text{m}^2 \text{m}^{-2}$ ), ALIA: average leaf inclination angle ( $^\circ$ ), Hot: hot-spot parameter ( $\text{m m}^{-1}$ ), soil spectrum (% reflectance),  $p_{soil}$ : soil brightness factor (unitless), SZA: sun zenith angle ( $^\circ$ ), OZA: observer zenith angle ( $^\circ$ ) and rAA: relative azimuth angle ( $^\circ$ ).

A simulated dataset consisting of 100 000 observations was created for model training and testing. Each observation included a random set of PROSAIL input parameters drawn from uniform distributions of the PROSAIL input parameters within wheat-specific ranges from literature [3, 4]. Furthermore, spectral reflectance for background soil was varied among observations in the simulated dataset. To do so, available data on soil reflectance by the ICRAF-ISRIC Soil MIR Spectral Library of the International Soil Reference and Information Centre (ISRIC) were used [5]. The simulated dataset was divided into a train and test set in a 9:1 ratio.

Field experiments were conducted at the Experimental

Farm Groß-Enzersdorf of the University of Natural Resources and Life Sciences, Vienna, in the seasons 2019/20 and 2020/21. Data on nitrogen yield (NY,  $\text{g m}^{-2}$ ) were collected in approximately 14-day intervals from March until harvest in July in both seasons. Destructive plant sampling was conducted on  $0.6 \text{ m}^2$  per plot. Plant material was dried, weighed, milled and analyzed for N concentration according to the Dumas combustion method [6] using an element analyzer (vario MAX cube CNS, Elementar Analysensysteme, Germany). Resulting N concentration values were multiplied by above-ground dry matter to calculate NY. Measurements on canopy reflectance in the field experiment were conducted with the spectroradiometer FieldSpec Handheld 2 (ASD Inc., USA). This sensor provides hyperspectral reflectance data from 325 to 1075 nm in 1 nm increments.

An artificial neural network (ANN) was set up to achieve the inversion of the radiative transfer model PROSAIL. Model inputs were viewing geometry, background soil reflectance and canopy reflectance from 400 to 1075 nm in 1 nm increments. The spectral resolutions of soil reflectance, simulated PROSAIL canopy reflectance and spectral measurements from the field experiments were matched. Model outputs were the PROSAIL parameters  $N$ ,  $C_{ab}$ ,  $C_{cx}$ ,  $C_{bp}$ ,  $C_m$ ,  $C_w$ , LAI, ALIA and Hot. The ANN consisted of three dense layers with 128 neurons each, ReLU activation function, loss function “mean absolute error” and optimizer “Adam”. Training epochs were set to a maximum of 500 with early stopping at 50 to avoid overfitting. Google Colaboratory, an available Keras implementation (version 2.8.0) in Python (version 3.6), was used to set up the ANN. Experimentally measured NY was estimated using predictions of  $C_{ab} \times \text{LAI}$ . The model performance was evaluated using the simulated test dataset and field experimental data.

The accuracy of model predictions compared to measured values was evaluated using regression coefficients and coefficients of determination ( $R^2$ ) in regression analysis. Furthermore, root mean square error (RMSE) and relative root mean square error (RRMSE) were calculated for model testing.

### 3. Results

Figure 2 presents the results of predicted LAI and  $C_{ab}$  compared to true values in the simulated test dataset. The parameters LAI and  $C_{ab}$  show high  $R^2$ , i.e., above 0.9, and low RRMSE (LAI: 17.3%,  $C_{ab}$ : 8.5%).

The relationship between measured and predicted  $C_{ab}$  is linear, while LAI shows a quadratic fit. When LAI was  $0 \text{ m}^2 \text{ m}^{-2}$ ,  $C_{ab}$  could not be estimated. For observations with LAI below  $0.5 \text{ m}^2 \text{ m}^{-2}$ ,  $C_{ab}$  predictions show slight underestimation at high  $C_{ab}$ .

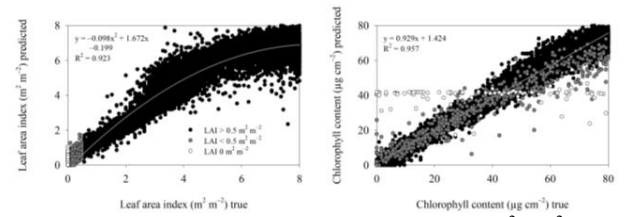


Figure 2: Estimation of leaf area index (left,  $\text{m}^2 \text{ m}^{-2}$ ) and chlorophyll content (right,  $\mu\text{g cm}^{-2}$ ) of the simulated test dataset.

Predicted  $C_{ab} \times \text{LAI}$  was calibrated using experimental data on NY from 2020/21. The calibrated predictions of NY were validated using experimental data from 2019/20 (Figure 3). In both seasons,  $R^2$  values were high, i.e., above 0.8. In the experimental validation data of 2019/20, the deviation from the  $45^\circ$  line was low. At high NY, predictions show a slight underestimation.

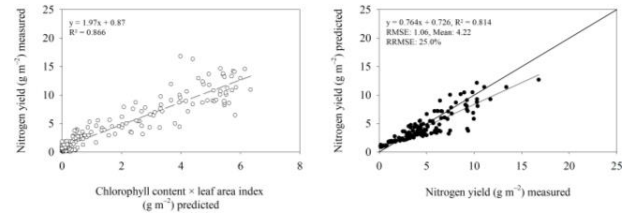


Figure 3: Calibration of predicted chlorophyll content  $\times$  leaf area index (left,  $\text{g m}^{-2}$ ) with measured nitrogen yield ( $\text{g m}^{-2}$ ) of the field experiment in 2020/21 as well as validation of calibrated predictions on nitrogen yield (right) with respective measurements of the field experiment in 2019/20.

### 4. Discussion

Results on predicted LAI and  $C_{ab}$  based on the simulated test dataset showed, that the ANN based inversion of the RTM PROSAIL was successful. The quadratic relationship between true and predicted LAI indicates, that LAI estimations saturate at high values, i.e., above  $4 \text{ m}^2 \text{ m}^{-2}$ . No leaf area is present for observations with  $\text{LAI} = 0 \text{ m}^2 \text{ m}^{-2}$ . As a result, leaf characteristics, such as  $C_{ab}$ , cannot be estimated. When LAI is low, e.g., below  $0.5 \text{ m}^2 \text{ m}^{-2}$ , the effect of background soil on reflectance measurements is large and thus affects the estimation of leaf characteristics. This results in an underestimation of high chlorophyll concentrations, when LAI is low.

Results on estimations of  $C_{ab} \times \text{LAI}$  compared to measured NY are promising, because of their high  $R^2$  in both seasons as well as the low deviation from the  $45^\circ$  line and the low RRMSE of experimental validation data in 2019/20. This indicates high predictability of NY based on our model as well as high stability among seasons.

### Acknowledgments

This work was supported by the project “DiLaAg – Digitalization and Innovation Laboratory in Agricultural Sciences” of the private foundation “Forum Morgen, Austria” and the Federal State of Lower Austria.



## References

- [1] Monteith, J.L., 1965. Light distribution and photosynthesis in Field Crops. *Annals of Botany*. 29, 17–37.
- [2] Berger, K., Atzberger, C., Danner, M., D'Urso, G., Mauser, W., Vuolo, F.; Hank, T., 2018. Evaluation of the PROSAIL model capabilities for future hyperspectral model environments: a review study. *Remote Sensing*. 10, 85–110.
- [3] Kong, W.P., Huang, W.J., Zhou, X.F., Song, X.Y., Casa, R., 2016. Estimation of carotenoid content at the canopy scale using the carotenoid triangle ratio index from in situ and simulated hyperspectral data. *Journal of Applied Remote Sensing*. 10, 026035.
- [4] Danner, M., Berger, K., Woche, M., Mauser, W., Hank, T., 2017. Retrieval of biophysical crop variables from multi-angular canopy spectroscopy. *Remote Sensing*. 9, 726.
- [5] Van Reeuwijk, L.P., 2002. Procedure for Soil Analysis, sixth edition. Wageningen: International Soil Reference Information Centre.
- [6] Winkler R., Botterbrodt, S., Rabe, E., Lindhauer, M.G., 2000. Stickstoff-/Proteinbestimmung mit der Dumas-Methode in Getreide und Getreideprodukten [Nitrogen and protein determination using the Dumas method in cereal and cereal products]. *Getreide Mehl Brot*. 54:86–91. German.

# Selection of YOLOX Backbone for Monitoring Sows' Activity in Farrowing Pens with a Possibility of Temporary Crating

Maciej Oczak

Precision Livestock Farming Hub and Institute of Animal Welfare Science, the University of  
Veterinary Medicine Vienna  
Veterinärplatz 1, 1210 Vienna, Austria  
Maciej.Oczak@vetmeduni.ac.at

## Abstract

*Activity monitoring of sows in farrowing pens is an important application of computer vision in Precision Livestock Farming. One example with a benefit for welfare of sows is farrowing prediction in pens with a possibility of temporary crating. In 2 experiments we tested various YOLOX backbones to estimate the generalization ability of the models on seen and unseen farrowing pens and animals. Models performed better on known pens and animals (~0.9 mAP) in comparison to unknown (~0.8 mAP). Results suggest that it is better to include some images of sows in the training set from the environment where the algorithm will be implemented. However, mAP as high as 0.8 suggests that on many farms it might be not necessary to re-train the model. Speed of inference of YOLOX models was ranging from 21 fps (YOLOX-x) to 42 fps (YOLOX-nano) on recorded videos. This should be sufficient to monitor activity level of sows in the farrowing compartment of production unit of VetFarm Medau (20 pens).*

## 1. Introduction

It is common practice in modern intensive pig husbandry to confine sows in farrowing crates including at least a few days before the onset of farrowing. The main reason for this practice is to improve piglet survival rate by protecting newborn piglets from fatal or injurious crushing by the mother sow [1]. However, the confinement of sows in crates has a negative impact on the sows' welfare, such as limited freedom of movement. Farrowing pens with a possibility of temporary crating offer a good compromise between the needs of the farmer, the sow and the piglets [2]. However, due to lack of precision in estimation of expected time of farrowing based on average length of gestation, there is a risk that farmer will keep the sows confined in crates in a period of nest-building, few hours

before the start of farrowing, to protect the piglets from crushing.

Automated detection of increase in sow activity with the use of sensor technology makes it possible the prediction of the onset of farrowing [3]. This could be useful in practical conditions to shorten surveillance intervals by farm staff, and the pen with a possibility of temporary crating could be prepared for an optimal farrowing [4].

To detect a sow in a farrowing pen we decided on application of YOLOX from YOLO series of object detection algorithms. YOLOX is a state-of-the-art object detector surpassing YOLOv3, one of the most widely used detectors in industry [5]. We hypothesize that YOLOX will provide an optimal trade-off between the speed and accuracy for real-time applications.

The objective of this study was to select an optimal backbone of YOLOX for real-time measurement of activity of sows, considering generalization ability of the model in unseen farrowing pens and on unseen animals.

## 2. Methodology

### 2.1. Animals and housing

Images with sows in farrowing pens were collected at the pig research and teaching farm (VetFarm) of the University of Veterinary Medicine Vienna, Vienna, Austria. Dataset 1 was collected between June 2014 and May 2016, while dataset 2 between December 2021 and July 2022. In total, images of 78 Austrian Large White sows and Landrace × Large White crossbreeds sows were recorded. These sows were housed in four types of farrowing pens. Out of 78 sows, 11 were kept in SWAP (Sow Welfare and Piglet Protection) pens (Jyden Bur A/S, Vemb, Denmark), 11 in trapezoid pens (Schauer Agrotronic GmbH, Prambachkirchen, Austria), 11 in wing pens (Stewa Steinhuber GmbH, Sattledt, Austria) and 45 in BeFree pens (Schauer, Prambachkirchen Austria). None of the animals included in the experiment were confined in a farrowing crate from the introduction to the farrowing pen

until the end of farrowing.

## 2.2. Video recording

Behaviour of sows was video recorded from introduction to the farrowing pens until weaning with 2D cameras in order to create a data set that could be annotated. Each pen in dataset 1 (SWAP, trapezoid and wing) was equipped with one IP camera (GVBX 1300-KV, Geovision, Taipei, Taiwan). In dataset 2 each IP camera (GV-BX2700, Geovision) was installed with a view on 2 farrowing pens (BeFree). Additionally, infrared spotlights (IR-LED294S-90, Microlight, Moscow, Russia) were installed in order to allow night recording. The videos were recorded with 1280x720 pixel resolution, in MPEG-4 format, at 30 fps.

## 2.3. Datasets

Out of 11 232 hours of recorded videos 15 242 images were selected for annotation and training of object detection models. To reduce correlation between sampled images K-means algorithm [6] was applied on recorded videos. For the 1<sup>st</sup> dataset 14 242 images were selected from videos recorded in SWAP, trapezoid and Wing pens. For the second dataset 1000 images were selected from videos recorded in BeFree pens.

Only one object class, a sow, was annotated on both datasets using CVAT and COCO annotator software packages (Fig. 1).

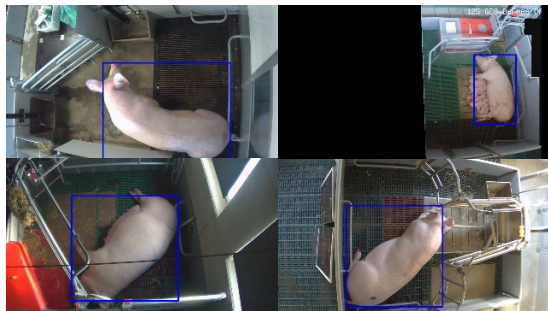


Figure 1. Annotated images with sows: top left – SWAP; top right – BeFree (one of two pens under camera view is masked); bottom left – trapezoid, bottom right – wing.

## 2.4. Experiments

We designed 2 experiments to test various backbones of YOLOX algorithm (YOLOX-nano, YOLOX-tiny, YOLOX-s, YOLOX-m, YOLOX-l, YOLOX-x) in terms of generalization ability and inference speed. We used MMDetection framework to train, validate and test the models [7]. Training was set to 50 epochs and was done on RTX Titan.

In both experiments out of total 15 242 images, 9969 (65.4%) were selected for the training set, 4273 (28%) for the validation set and 1000 (6.6%) for the test set. In experiment 1 training and validation sets included images from dataset 1, while test set from dataset 2. Thus, in experiment 1 it was possible to test the generalization ability of YOLOX backbones on new unseen farrowing pen (BeFree) and sows. In experiment 2 all 4 pen types and sows were represented in training, validation and test sets.

## 3. Results

Results of both experiment 1 and 2 revealed, as could be expected, that more complex backbones of YOLOX (YOLOX-m, YOLOX-l, YOLOX-x) had better mAP in both validation sets and test sets (Fig. 2). Higher mAP was achieved for these models after shorter training than for

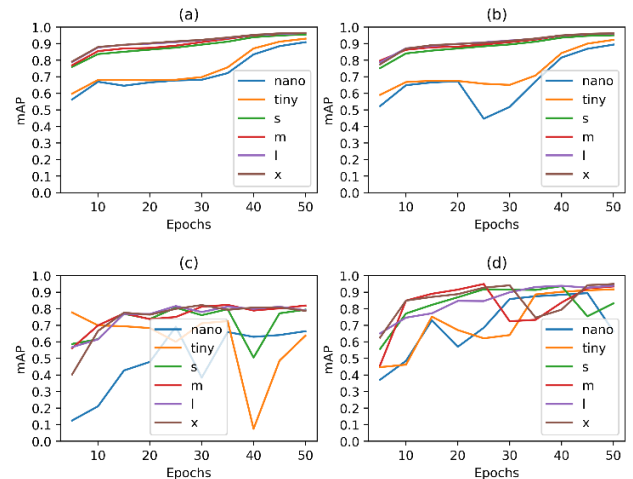


Figure 2. Performance metric mAP on a) validation set – experiment 1; b) validation set - experiment 2; c) test set – experiment 1; d) test set – experiment 2.

simpler models. Performance of models in experiment 1 was generally worse than in experiment 2 in the test set i. e. ~0.8 mAP vs 0.9 mAP for YOLOX-m, YOLOX-l and YOLOX-x. This suggests that for practical implementation of YOLOX for activity monitoring it is better to include some images of sows in the training set from the environment where the algorithm will be implemented. However, mAP as high as 0.8 suggests that on many farms it might be not necessary to re-train the model. Further validation of YOLOX with reference data on activity level of sows is needed to verify it.

Speed of inference of YOLOX models was ranging from 21 fps (YOLOX-x) to 42 fps (YOLOX-nano) on recorded videos. With assumption that 1 fps is sufficient to monitor activity level of sows, even with the most complex YOLOX-x backbone, it would be possible to monitor the whole farrowing production unit at VetFarm Medau with one RTX Titan (20 pens).

## References

- [1] R. King, E. Baxter, S. M. Matheson and S. A. & Edwards, "Temporary crate opening procedure affects immediate post-opening piglet mortality and sow behaviour.," *animal*, pp. 13(1), 189-197, 2019.
- [2] J. N. Marchant, A. R. Rudd, M. T. Mendl, D. M. Broom, M. J. Meredith, S. Corning and P. H. & Simmins, "Timing and causes of piglet mortality in alternative and conventional farrowing systems," *Veterinary record*, vol. 147, no. 8, pp. 209-214, 2000.
- [3] M. Oczak, K. Maschat and J. Baumgartner, "Dynamics of sows' activity housed in farrowing pens with possibility of temporary crating might indicate the time when sows should be confined in a crate before the onset of farrowing," *Animals*, vol. 10, no. 1, p. 6, 2019.
- [4] I. Traulsen, C. Scheel, W. Auer, O. Burfeind and J. Krieter, "Using acceleration data to automatically detect the onset of farrowing in sows," *Sensors*, vol. 18, no. 1, p. 170, 2018.
- [5] Z. Ge, S. Liu, F. Wang, Z. Li and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv*, no. 2107.08430, 2021.
- [6] T. A. D. Pereira, L. Willmore, M. Kislin, S. Wang, M. Murthy and J. Shaevitz, "Fast animal pose estimation using deep neural networks," *Nature methods*, vol. 16, no. 1, pp. 117-125, 2019.
- [7] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu and Z. Zhang, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv*, no. 1906.07155, 2019.

# Influence of Data Processing on Hyperspectral-Based Classification of Managed Permanent Grassland

Viktoria Motsch, Roland Britz, Andreas Gronauer

University of Natural Resources and Life Sciences, Vienna,

Department of Sustainable Agricultural Systems, Institute of Agricultural Engineering

Peter-Jordan-Straße 82, 1190 Vienna, Austria

viktoria.motsch@boku.ac.at

## Abstract

The botanical composition of grassland stands can be determined using a combination of hyperspectral imaging and machine learning. Data processing before machine learning can significantly improve overall model performance. Specific preprocessing variants, such as smoothening and derivation of the spectrum, were found to be beneficial for classifying grassland species groups in detached models using hyperspectral data from permanent grassland obtained under laboratory conditions. Compared to extensively preprocessed data, raw spectral data yielded no statistically decreased performance in most cases.

## 1. Introduction

Grassland vegetation typically comprises grasses, herbs, and legumes which represent different functional traits [14] and feed values; knowledge of their relative proportions offers several advantages for site-specific management and livestock feeding. Remote sensing is a non-destructive method used for the reproducible sensing of large areas [16] as detected spectral signatures may vary depending on the species group. Machine learning models based on hyperspectral data can be used for species group classification [4, 5]. For this, data preprocessing might be a substantial step in enhancing model performance. The use of derivatives together with spectral data is a common technique [10, 18] as removes background signals and visualizes spectral curve shape differences that might not be evident in the spectra [7]. Smoothing operations such as Savitzky-Golay filtering are frequently applied [6, 8] as well as data standardization or normalization (see Fig. 1). A systematic review under laboratory conditions can reveal the influence of the vast number of data processing variants in combination with machine learning on the spectral-based classification of permanent grassland vegetation.

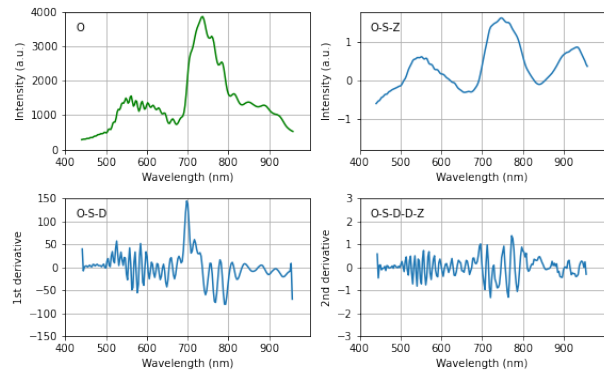


Figure 1. Representative reflectance spectrum and different preprocessing variants for a single red clover (*Trifolium pratense* L.) sample. Left upper corner denotes preprocessing variant.

## 2. Materials and Methods

The dataset used throughout this study is described in detail by Britz *et al.* [5]. Briefly, an in-house hyperspectral imaging setup was used under standardized laboratory conditions. In total, 5768 plant samples were acquired at two Austrian grassland sites. Each sample was derived from an individual plant, manually annotated and labeled according to species group (grass, herb, or legume).

### 2.1. Data Preprocessing

For each sample, a total of 100 pixels were drawn randomly stratified. All samples were grouped based on their species group, then randomly stratified and assigned a chunk number from 1 to 5. Further, data was pre-processed using different combinations of Savitzky-Golay-smoothening (function `savgol` with a filter length of 5 and quadratic filter from R package `pracma` 2.3.3 [3]), derivation, and Z-standardization (see Tab. 1). In total, 27 preprocessing variants were generated and analyzed.

Step	Variant
1.	O O
2.	Z S S D D D D S S S S D D D D D D D S S S S S S S S
3.	Z Z S S D D D D D D D S S S S D D D D D D D
4.	Z Z S S Z S S D D D D D D D D S S S
5.	Z Z Z S S Z S S D D D
6.	Z Z Z S

Table 1. Preprocessing variants generated from original data (O). D = derivative, S = Savitzky–Golay filter, Z = Z-standardization.

## 2.2. Machine Learning Algorithms

Multi-Layer Perceptron (MLP), Random Forest (RF), and Partial Least Squares Discriminant Analysis (PLS-DA) models were trained for species group classifications. The class weights were normalized to compensate for unbalanced classes. Final training was performed, 5-fold cross-validated, and performance metrics were calculated based on validation parts not used for training. Details on machine learning algorithms can again be found in Britz *et al.* [5].

Briefly, MLP networks were trained using Python, PyTorch [13], Tune [9] included in Ray [12] and hyperopt [2]. The architecture is a fully connected layer followed by batch normalization and a rectified linear unit activation function (ReLU). After another fully connected layer with a ReLU, the final layer is connected to the three output classes. Cross-entropy loss with class weights was used together with a stochastic gradient descent optimizer. Hyperparameters for each variant were searched using an ASHA and in total 100 hyperparameter combinations per dataset variant and group were evaluated. The five hyperparameter combinations, having achieved the highest accuracy per dataset variant and group, were retrained with 5-fold cross-validation for 120 epochs. Then, the model with the highest cross-validated accuracy found at any epoch is depicted in the results. RF classifiers were trained using the function ranger from the ranger package [17] with mtry of 40, SF of 1 and 400 trees, resulting in reasonable accuracy and computation time for training. PLS regression was performed using the cpls function from the pls package [11] with 64 components. Subsequently, linear DAs with the lda function from MASS package [15] were performed.

## 3. Results and Discussion

MLP achieved cross-validated accuracies of 96.9 % for species group (grass, herb, or legume) classification. While MLP and PLS-DA performed well across a wide range of preprocessing variants and showed a high generalization ability, this was not true for RF (see Fig. 2). The main reason for this is that RF usually uses only a few predictors at the tree level to form a decision boundary [1], which makes it more sensitive to data variations than MLP and PLS-DA.

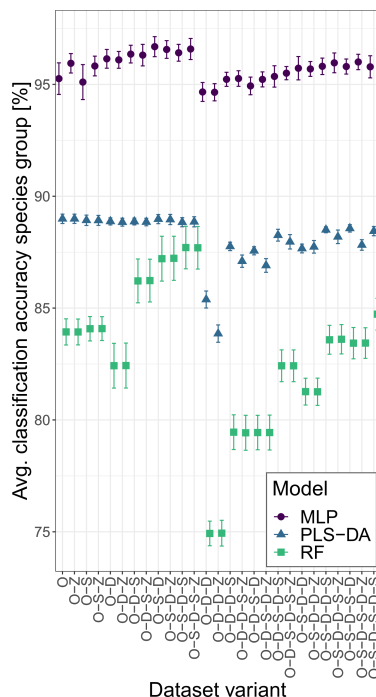


Figure 2. Mean species group classification accuracy based on the preprocessing variant for multilayer perceptron (MLP), partial least squares discriminant analysis (PLS-DA), and random forest (RF) models. X-axis abbreviations (preprocessing steps from bottom to top): O = original data, D = derivative, S = Savitzky–Golay filter, Z = Z-standardization. Error bars indicate standard deviation, 5-fold cross-validated.

In general, similar trends in classification accuracy could be observed depending on the preprocessing variant. Variants differing only in subsequent Z-standardization showed no significant differences independent of model type. Preprocessing steps that do not lead to increased accuracy should be avoided for the sake of simplicity. Preprocessing variants including a Savitzky–Golay filter before a derivation work particularly well for data with low spectral band distances. Here, differences between successive spectral channels may be slight compared to random noise [7]. Other variants can also benefit from Savitzky–Golay filtering as a noise reduction technique. Interesting preprocessing variants that performed well, independent of the model type, included the combination S-D without a second D. In particular for RF but also in other models, variants containing a derivation (D) without prior Savitzky–Golay filter (S) mainly performed worse than variants with a combination of S and D. This underlines the usefulness of spectral gradients in combination with smoothing for machine learning applications. However, for MLP and PLS-DA, even the original dataset variant (O) generated models that were not significantly different from the best statistical model.

## References

- [1] Houman Abbasiyan, Chris Drummond, Nathalie Japkowicz, and Stan Matwin. Robustness of Classifiers to Changing Environments. In Atefeh Farzindar and Vlado Kešelj, editors, *Advances in Artificial Intelligence*, pages 232–243, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
- [2] James Bergstra, Daniel Yamins, and David Cox. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.
- [3] Hans W. Borchers. *pracma: Practical Numerical Math Functions*, 2021. R package version 2.3.3.
- [4] Roland Britz, Norbert Barta, Andreas Klingler, Andreas Schaumberger, Alexander Bauer, Erich M Pötsch, Andreas Gronauer, and Viktoria Motsch. Hyperspectral-based classification of managed permanent grassland with multilayer perceptrons: Influence of spectral band count and spectral regions on model performance. *Agriculture*, 12(5):579, 2022.
- [5] Roland Britz, Norbert Barta, Andreas Schaumberger, Andreas Klingler, Alexander Bauer, Erich M Pötsch, Andreas Gronauer, and Viktoria Motsch. Spectral-based classification of plant species groups and functional plant parts in managed permanent grassland. *Remote Sensing*, 14(5):1154, 2022.
- [6] P. D. Dao, Y. He, and B. Lu. Maximizing the quantitative utility of airborne hyperspectral imagery for studying plant physiology: An optimal sensor exposure setting procedure and empirical line method for atmospheric correction. *International Journal of Applied Earth Observation and Geoinformation*, 77:140–150, May 2019.
- [7] Tanvir H. Demetriades-Shah, Michael D. Steven, and Jeremy A. Clark. High resolution derivative spectra in remote sensing. *Remote Sensing of Environment*, 33(1):55–64, jul 1990.
- [8] T. Fricke and M. Wachendorf. Combining ultrasonic sward height and spectral signatures to assess the biomass of legume-grass swards. *Computers and Electronics in Agriculture*, 99:236–247, Nov. 2013.
- [9] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. Tune: A Research Platform for Distributed Model Selection and Training. *arXiv preprint arXiv:1807.05118*, 2018.
- [10] F. Locher, H. Heuwinkel, R. Gutser, and U. Schmidhalter. Development of Near Infrared Reflectance Spectroscopy Calibrations to Estimate Legume Content of Multispecies Legume-Grass Mixtures. *Agronomy Journal*, 97(1):11–17, Jan. 2005.
- [11] Bjørn-Helge Mevik and Ron Wehrens. The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2):1–23, 2007.
- [12] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. Ray: A Distributed Framework for Emerging AI Applications. *CoRR*, abs/1712.05889, 2017.
- [13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [14] J. Schellberg and L. da S. Pontes. Plant functional traits and nutrient gradients on grassland. In *16th Symposium of the European Grassland Federation "Grassland Farming and Land Management Systems in Mountainous Regions"*, volume 16, pages 470–483, Gumpenstein, Austria, Aug. 2011. Grassland Science in Europe.
- [15] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag GmbH, New York, fourth edition, 2002. ISBN 0-387-95457-0.
- [16] M. Wachendorf, T. Fricke, and T. Möckel. Remote sensing as a tool to assess botanical composition, structure, quantity and quality of temperate grasslands. *Grass and Forage Science*, 73(1):1–14, Mar. 2018.
- [17] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.
- [18] H. Yu, B. Kong, G. Wang, H. Sun, and L. Wang. Hyperspectral database prediction of ecological characteristics for grass species of alpine grasslands. *The Rangeland Journal*, 40(1):19–29, 2018.

## Scientific Spotlight Papers



# In Defense of Information Plane Analysis

Mina Basirat  
Graz University of Technology  
mina.basirat@ist.tugraz.at

Bernhard C. Geiger  
Know-Center GmbH  
geiger@ieee.org

Peter M. Roth  
Vetmeduni Vienna, TU Munich  
peter.m.roth@vetmeduni.ac.at

## Abstract

In this paper, we tackle the problem of analyzing neural network training via information plane analysis. The key idea is to describe the mutual information between the input and a hidden layer and a hidden layer and the target over time. Even though this is a reasonable approach, previous works showed inconsistent or even contradicting interpretations. Since the mutual information cannot be computed analytically, the authors applied different kinds of estimators, often not describing the mutual information very well. Taking these findings into account, we want to show that despite this theoretical limitation information planes allow at least for a geometric interpretation. Thus, enabling us to analyze different aspects of neural network learning for real-world problems.

## 1. Introduction and Problem Statement

One prominent approach to analyze neural network training is information plane (IP) analysis [7]. Building on the idea of the information bottleneck principle [8], the main idea is to describe and analyze the mutual information between the layers of a neural network over time. In particular, we are interested in the plane described by the mutual information  $I(X;T)$  between the input  $X$  and the activation values of a hidden layer  $T$  and the mutual information  $I(Y;T)$  between  $T$  and the target variable  $Y$ .

This is illustrated in Fig. 1 for two examples. From Fig. 1a, two phases can be observed, cf. [7]: first, a phase in which both  $I(X;T)$  (expansion) and  $I(Y;T)$  (fitting) are increasing, and, second, a compression phase during which  $I(X;T)$  is decreasing again, whereas  $I(Y;T)$  is increasing only slightly. The compression phase was interpreted as the hidden layer  $T$  discarding irrelevant information about the input  $X$ , and was causally connected to generalization. In contrast, Fig. 1b shows only fitting as an increase of  $I(Y;T)$ .

Even though these examples show that IPs appear to be an appealing way to analyze learning behaviors of NNs, we are facing the problem that the literature on IP analysis

reports partially contradicting interpretations, cf. [2, 6, 7]. This, however, can be explained by the fact that the mutual information can often not be computed analytically, and different kinds of estimations for the mutual information terms  $I(X;T)$  and  $I(Y;T)$  are applied. Thus, similar to recent findings [2, 3], we would like to demonstrate that IPs represent geometric rather than information-theoretic phenomena. In this way, we are still able to use this technique to analyze NN training if the estimates are interpreted correctly.

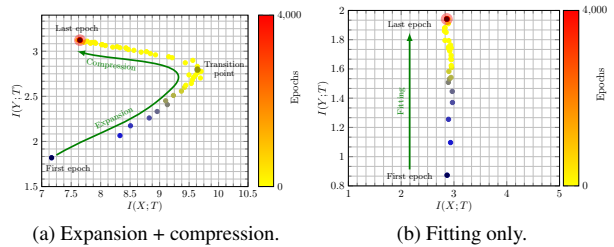


Figure 1. Information planes reveal different behavior during neural network training.

## 2. Geometric Interpretation of Information Planes

To this end, we create an IP from the plugin estimates for mutual information between the uniformly discretized activation value  $\hat{T}$  and the network input  $X$  or class label  $Y$ , respectively. Introducing both fixed and adaptive binning schemes for obtaining  $\hat{T}$ , we get the estimators  $\hat{I}(Y;\hat{T})$  and  $\hat{I}(X;\hat{T}) = H(\hat{T})$ . In this way, we argue that the correct interpretation of  $H(\hat{T})$  yields an insight into the geometric compression of the activation  $T$ , both in absolute (e.g., describing the diameter of the set of all activations of a dataset) and relative (e.g., clustering of activations of a dataset) terms. To allow for a more intuitive interpretation, we additionally show a 2D visualization of latent space. For more details on the theoretical background and the applied binning approaches, we would like to refer to [1].

### 3. Illustrative Results

In the following, we show an example demonstrating that information planes can be a valuable tool to analyze and interpret neural network learning. To this end, we train a bottleneck network (100-100-2-100) for the well-known *MNIST* dataset [4] and *Brightness MNIST (BMNIST)* [5], a modified version of *MNIST*, where the illumination of the images has been increased. In this way, the contrast of the images is decreased and, thus, the classes are pushed closer together in the image space. The bottleneck model was chosen to make the geometric interpretation more apparent. In fact, in both cases, we finally obtain a similar classification result in terms of accuracy: 96.62% for *MNIST* and 95.25% for *BMNIST*. However, when looking at the corresponding information planes in Fig. 2 (*MNIST*) and Fig. 3 (*BMNIST*) reveals that the learning behavior is different. To make the temporal character of the trajectories more apparent, the first and the last epoch are highlighted by a black point and a large circle, respectively.

For *MNIST*, using adaptive binning (see Fig. 2b), we can recognize a fitting phase, i.e.,  $\hat{I}(Y; \hat{T})$  is increasing over time, indicating a growth of the class separability. In addition, using fixed binning (see Fig. 2a), we can recognize a geometric compression with an absolute scale for  $\hat{H}(\hat{T})$  from the first to the last epoch for the last two layers. Indeed, as can be seen in Fig. 4, where we plot the two-dimensional latent space, the absolute scale reduces from approx.  $47 \times 69$  to approx.  $7 \times 7$  during training.

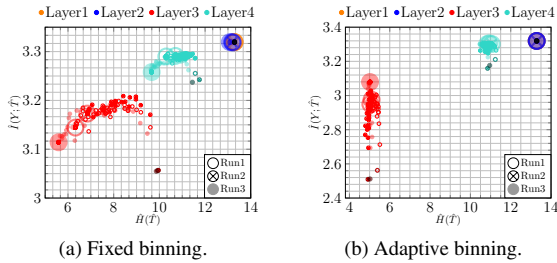


Figure 2. IPs for *MNIST*: (a) fixed and (b) adaptive binning.

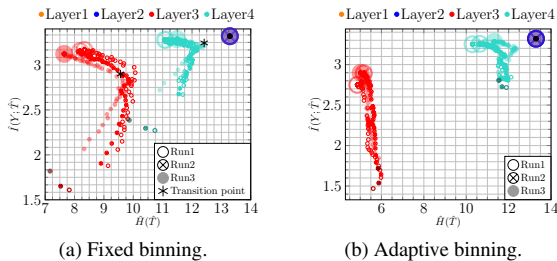


Figure 3. IPs for *BMNIST*: (a) fixed and (b) adaptive binning.

In contrast, for *BMNIST*, the IP analysis shown in Fig. 3 reveals that the learning behavior is different due to a different initial setting. Due to reduced contrast in the images, the classes are mapped to highly overlapping regions in the beginning (see Fig. 5a); for the original *MNIST* dataset, this is not the case (see Fig. 4a). Thus, during NN training the data points in the latent space have to be pushed apart according to their class label. In this way, we can recognize a fitting phase, i.e., increasing  $\hat{I}(Y; \hat{T})$ , for adaptive binning (see Fig. 3b) and an expansion phase for fixed binning (see Fig. 3a). Simultaneously, the data points are pushed apart and occupy a larger volume in the latent space (increased from  $10 \times 8$  to  $22 \times 17$ ), as can be seen in Fig. 5.

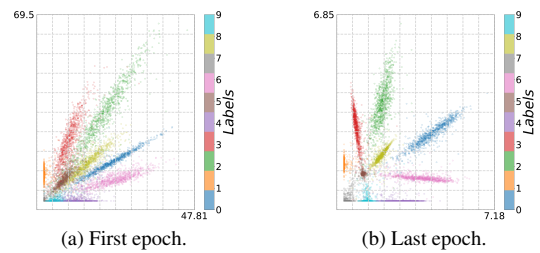


Figure 4. 2D plots for *MNIST*: (a) first and (b) last epoch.

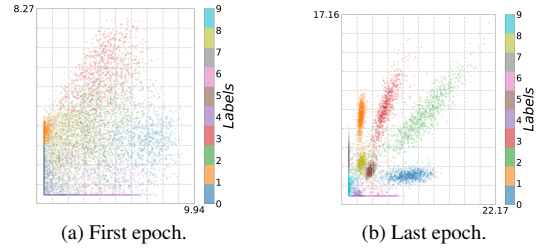


Figure 5. 2D plots for *BMNIST*: (a) first and (b) last epoch.

### 4. Discussion and Conclusion

To overcome the known issues of IP analysis, we demonstrated that the IP represents geometric rather than information-theoretic effects, which we showed based on an illustrative example. To support these findings, we built a bottleneck architecture (i.e., using a two-dimensional layer), which allows us to directly relate the information covered by IPs to the geometric structure of the latent space. **For more technical details and a more thorough evaluation, we would like to refer to [1].**

### Acknowledgment

This work was partially supported by the FFG Bridge project *SISDAL*, the BMBF (*International Future AI Lab "AI4EO"*), the project *iDev40*, and the Austrian COMET Program (*Know-Center*).

## References

- [1] Mina Basirat, Bernhard Geiger, and Peter M. Roth. A geometric perspective on information plane analysis. *Entropy*, 23(6):711, 2021.
- [2] Bernhard Claus Geiger. On information plane analyses of neural network classifiers – a review. *arXiv:2003.09671*, 2020.
- [3] Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In *International Conference on Machine Learning*, pages 2299–2308, 2019.
- [4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [5] Norman Mu and Justin Gilmer. MNIST-C: A robustness benchmark for computer vision. *arXiv:1906.02337*, 2019.
- [6] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018.
- [7] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.
- [8] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999.

# Scientific Spotlight: Computer-assisted mitotic count using a deep learning-based algorithm improves interobserver reproducibility and accuracy

Christof A. Bertram  
University of Veterinary Medicine Vienna  
Veterinärplatz 1  
1210 Vienna, Austria  
Christof.Bertram@vetmeduni.ac.at

Marc Aubreville  
Technische Hochschule Ingolstadt  
Esplanade 10  
85049 Ingolstadt, Germany  
Marc.Aubreville@thi.de

Robert Klopffleisch  
Freie Universität Berlin  
Robert-von-Ostertag-Str. 15  
14163 Berlin, Germany  
Robert.Klopffleisch@fu-berlin.de

Enumeration of tumor cells undergoing cell division (mitotic figures) is a very practicable method to quantify tumor proliferation as it can be determined in histological sections with routine staining methods. The mitotic count (number of mitotic figures per 2.37 mm<sup>2</sup> tumor area) has been shown to correlate strongly with patient outcome in several humans and animals tumors types. Tumors with a higher amount of proliferating cells are associated with a more aggressive tumor behavior and thus are more likely to result in death of the patient. Therefore, this prognostic test is routinely conducted by pathologists for many tumor types. The diagnostic task of the mitotic count is to find the tumor region with the highest density of mitotic figures (hotspot) and to count all mitotic figures within this area. Both subtasks of the mitotic count are, however, problematic for human experts as:

- Mitotic figures can only be spotted at high magnification and a tumor section may comprise of thousands of fields of view exceed the human mental capacity and time availability to screen the entire tumor.
- Mitotic figures can easily be overlooked due to the high complexity of histological images.
- Mitotic figures can be difficult to distinguish from other cell structures (such as necrotic cells) with similar morphological appearance.

Subsequently, marked observer variability is well known for the mitotic count. In order to improve the accuracy and reproducibility of the mitotic count, computer-assistance using deep learning-based algorithms with verification by pathologists have been proposed. Whereas most previous

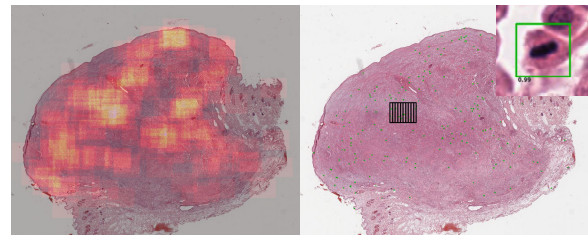
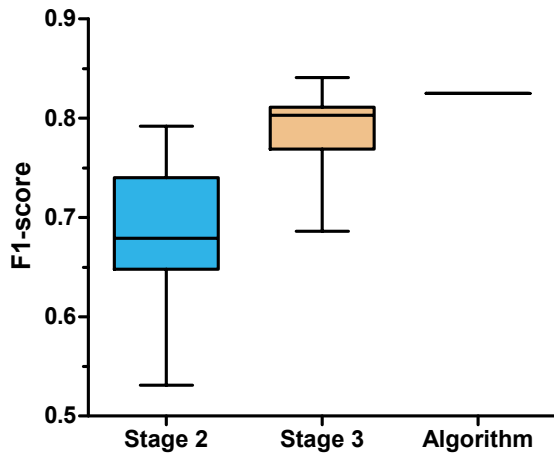


Figure 1. Algorithmic heatmap of mitotic density in the tumor (left image) is based on the algorithmic predictions (right image; green boxes). Based on current recommendations, the area with the highest density (black box in the right image) should be selected for the mitotic count.

studies have focused on developing mitotic figure algorithms, there are only few studies that evaluate the implementation of those algorithms into a diagnostic workflow.

In our study [1], we compared the performance between the routine method (without computer-assistance, stage 1), with computer-assisted mitotic counts using algorithmically preselected hotspot tumor areas (stage 2, Fig. 1) and visualisation of mitotic figures candidates within this hotspot tumor area (stage 3, inset Fig. 1). The deep learning-based algorithm was developed with a dataset of 32 mast cell tumor cases comprising 48,880 mitotic figure annotations. The three mitotic count approaches were conducted by 23 pathologists in 50 cases of canine mast cell tumors. A ground truth for the mitotic figures in the hotspot location of stage 2 and 3 was created by a pathologist assisted by immunohistochemistry for phosphohistone H3, which is a specific staining for mitotic figures.



Antoine Assenmacher, Kathrin Becker, Mark Bennett, Sarah Corner, et al. Computer-assisted mitotic count using a deep learning-based algorithm improves interobserver reproducibility and accuracy. *Veterinary Pathology*, 59(2):211–226, 2022.

Figure 2. Comparison of the mitotic figure detection performance (F1-score) by 23 pathologists in the same tumor area without (stage 2) and with (stage 3) visualization of mitotic figure candidates detected by a deep learning-based algorithm. The performance of the algorithm without review by a pathologists exceeds most study participants. Data taken from [1]

The experiment found that pathologists had higher mitotic counts in stage 2 with the preselected tumor area than in stage 1 with area selection by each pathologists. Our work demonstrates that algorithms are superior in analysing the mitotic density in large tumor sections. The ability to identify and classify individual mitotic figures was compared between stage 2 (no further computer assistance) and stage 3 (visualization of mitotic figure candidates). The F1-score was higher in stage 3 for all 23 pathologists with an average increase of 10.7 percentage points (Fig. 2). Most notably the number of false negative mitotic figures was reduced by 37.4% proving the tremendous benefits of highlighting mitotic figure candidates in the images.

In conclusion, this study [1] demonstrates the benefits of computer-assisted mitotic counts for a routine diagnostic workflow. The reproducibility and accuracy of identifying hotspot tumor locations and detecting individual mitotic figures was markedly improved. Further benefits could be an improved diagnostic efficiency, which was not systematically evaluated in our study. Further studies are needed to improve robustness of mitotic figure algorithms to different sources of domain shift, particularly image from different scanners and tumor sections with suboptimal tissue quality, in order to allow a widespread application of the software solutions.

## References

- [1] Christof A Bertram, Marc Aubreville, Taryn A Donovan, Alexander Bartel, Frauke Wilm, Christian Marzahl, Charles-

# A Modern Approach for Early Wildfire Detection

Kurt Winter

IQ Technologies for Earth and Space

winterk@iq-technologies.berlin

Peter M. Roth

Vetmeduni Vienna, TU Munich

peter.m.roth@vetmeduni.ac.at

## Abstract

Wildfire is a constant threat to wildlife, vegetation, and society in history. Thus, detecting such fires in an early stage is of high relevance, raising the need for automatic approaches building on visual object detection, namely to detect smoke. To this end, typically feature-based approaches have proven to work well in the past. However, the goal of this work was to evaluate whether or not modern approaches building on neural networks would be beneficial in this context. To this end, we generated a new dataset, allowing us to train and evaluate neural-network-based smoke detectors. In addition, we demonstrate that each of the approaches has benefits and shortcomings, however, also that a carefully designed fusion strategy can improve the detection results in practice.

## 1. Introduction and Problem Statement

As also recent events in Australia, the USA, Russia, Germany, and even Austria show, wildfires have massive consequences for nature, wildlife, and the human population. Due to climate change, socio-economic changes, and general population development, the wildfire situation is likely to become worse [6]. Besides prevention, the best way to minimize damage to nature and wildlife is to early detect wildfires. However, the flames are often not directly visible in an early stage, requiring to apply indirect approaches to detect smoke. The most common approach is human inspection. Indeed, fire watchers are sitting on fire watchtowers and looking out for smoke plumes in the distance.

However, the detection by humans is very time-consuming, monotonously and thus tiring, and very expensive. An alternative to traditional smoke detection methods is given by terrestrial visual detection systems such as *IQ FireWatch*<sup>1</sup>, building on three different camera sensors: a monochrome sensor for the detection in daylight, an RGB sensor, which provides a better view for the human eye, and a sensor working in the near-infrared (NIR) spectrum for

<sup>1</sup><https://www.iq-firewatch.com/>.

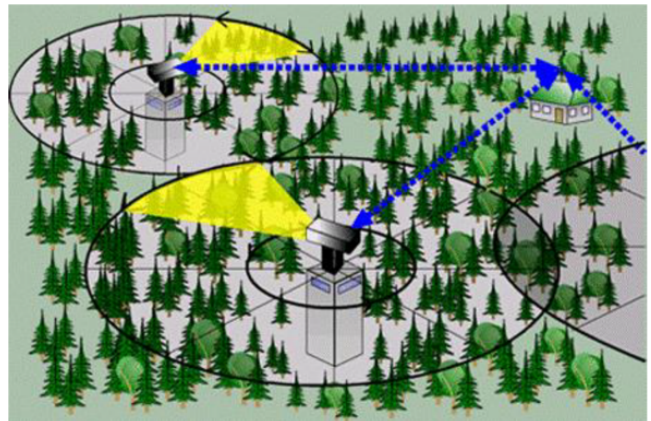


Figure 1. IQ FireWatch sensor system in practical use.

night vision. One sensor system can reliably cover a radius of 15 kilometers in a 360 degrees view, which is illustrated in Figure 1.

Even though data from different sensors is available, in this work we focus on high-resolution monochrome images, which are characterized by higher light sensitivity, which is beneficial when detecting smoke [2]. To this end, we compare the *F-Shell detector* [1] building on handcrafted features, to Faster R-CNN [3] using learned features. For that purpose, we created a new benchmark dataset for smoke detection. In addition, we evaluated how these approaches can be combined effectively for real-world scenarios

## 2. Smoke Detection

To detect the smoke, in this work we considered three approaches: the feature-based *F-Shell detector* [1], the neural-network-based Faster R-CNN [3], and a combination of both.

*F-Shell* follows a three-stage process: defining candidate regions, feature extraction, and classification, where three queues are run in parallel on a sequence of images. Using a sophisticated background subtraction to identify the regions of interest, these are described by (a) region properties such as shape or size, (b) by correlation to distinguish between

moving objects and smoke clusters, and (c) by texture properties. Finally, an alarm is raised if at least one of the three queues yields a response.

Similarly, R-CNN builds on two stages: First, the region proposal network predicts regions of interest class-agnostically. Second, these proposals are cropped and finally classified. Yielding the best trade-off of speed and accuracy, we finally decided to build our system on an InceptionV2 backbone [5]. We pre-trained it using the COCO dataset and finetuned it with our newly generated dataset.

The finally obtained results for the individual detectors are summarized in Table 1 (*F-Shell* and *FRC*). Showing a similar accuracy (acc.), we see differences in other practically relevant metrics such as true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), false negative rate (FNR), and thus in precision (prec.).

Detector	Acc.	TPR	FPR	TNR	FNR	Prec.
F-Shell	0.80	0.68	0.07	0.92	0.31	0.90
FRC	0.81	0.83	0.19	0.80	0.16	0.81
COMB	0.82	0.80	0.14	0.85	0.19	0.84

Table 1. Detection results of individual detectors.

Thus, the idea was to combine both approaches to get the best tradeoff for all of these parameters. In particular, we applied the *The COMBINATOR (COMB)* [4] to combine the individual results, which additionally takes into account the confidence and the complementarity coefficients of each detector. As can be seen from Table 1 (*COMB*), in this way, a higher number of detections can be provided while still maintaining a decrease in false alarms.

### 3. Discussion and Conclusion

In this paper, we tackled the problem of wildfire detection in the context of the *IQ FireWatch* system. In particular, we investigated smoke detection using high-resolution monochrome images using two different approaches: *F-Shell* and *Faster R-CNN*. Since both approaches have pros and cons, we finally proposed a combination of both, i.e., using *The COMBINATOR*, providing a reasonable trade-off in practice. For more details, we would like to refer to [7].

Future work will include establishing a larger dataset allowing for both training and evaluation and further combinations of the different approaches.

### Acknowledgment

Peter M. Roth was partially supported by the BMBF *International Future AI Lab "AI4EO"*.

### References

- [1] Thomas Behnke, Hartwig Hetzheim, Herbert Jahn, Jörg Knollenberg, and Ekkehard Kührt. Verfahren und Vorrichtung zur automatischen Waldbranderkennung.
- [2] Ayan Chakrabarti, William T. Freeman, and Todd Zickler. Rethinking color cameras. In *Proc. IEEE Int'l Conf. on Computational Photography*, 2014.
- [3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.
- [4] Floris De Smedt, Kristof Van Beeck, Tinne Tuytelaars, and Toon Goedemé. The combinator: Optimal combination of multiple pedestrian detectors. In *Proc. Int'l Conf. on Pattern Recognition*, 2014.
- [5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.
- [6] Richard S. Vachula, James M. Russell, and Yongsong Huang. Climate exceeded human management as the dominant control of fire at the regional scale in california's sierra nevada. *Environmental Research Letters*, 14(10):104011, 2019.
- [7] Kurt Winter. Decision-making by a combination of feature-based and ai-based algorithms for early wildfire detection. In *International Conference on Automatic Fire Detection*, 2021.