Weizenährenerkennung mithilfe neuronaler Netze und synthetisch generierter Trainingsdaten

Lukas Lucks[1], Laura Haraké[1] und Lasse Klingbeil[2]

[1] Fraunhofer Institut für Optronik, Systemtechnik und Bildauswertung, Gutleuthausstraße 1, 76275 Ettlingen
[2] Universität Bonn, Institut für Geodäsie und Geoinformation, Nußallee 17, 53115 Bonn

Zusammenfassung This paper investigates the usability of synthesized training data for the recognition of wheat ears using neural networks in the context of semantic image segmentation. For this purpose, detailed scenes of wheat fields consisting of 3D models with high-resolution textures and defined material properties are modeled. Afterwards, photo realistic color images are synthesized, which also contain a binary image mask with the locations of the ear models. The resulting image pairs are then used as a training data for two neural networks (U-Net and DeepLab-V3+). To determine whether these data allows domain adaptation, the trained networks are evaluated using real wheat field images. The IoU value of about 69.96 shows that information transfer from the synthesized images to real images is possible.

Keywords Semantic segmentation, synthetic data, photorealistic rendering, domain adaptation

1 Einleitung

Um die Nahrungssicherheit für die wachsende Weltbevölkerung sicherzustellen, werden immer höhere Anforderungen an die landwirtschaftliche Produktion gestellt. Die Erfüllung der Anforderungen wird durch die weltweit steigende Flächenkonkurrenz erschwert. Durch diese Entwicklungen ergibt sich die Notwendigkeit,

die vorhandenen Flächen nachhaltig zu bewirtschaften und Pflanzensorten zu züchten, die eine effizientere Produktion ermöglichen. In diesem Kontext nimmt Weizen, als eine der wichtigsten Kulturpflanzen neben Mais und Reis, eine besondere Rolle ein. Um den Weizenanbau in Zukunft nachhaltig und effizient zu gestalten und eine präzisere Bewirtschaftung zu ermöglichen, ist eine ständige Analyse des Pflanzenwachstums notwendig. Je nach Wachstumsphase der Pflanze sind tägliche Erfassungungen erforderlich, welche wiederum durch die oftmals manuelle Durchführung sehr zeitaufwendig sind [1]. Von besonderem Interesse ist dabei die Erkennung der Ähren, da sich aus diesen relevante Bestandsparameter wie die Pflanzendichte oder das Reifestadium der Pflanzen bestimmen lassen.

Unter Verwendung von Kamerabildern und moderner Bildverarbeitungsalgorithmen wird versucht, diese Informationen automatisiert abzuleiten [2]. Diese Algorithmen lernen dabei mithilfe von Referenzdaten, die Ähren innerhalb der Bilder zu erkennen. Um das jeweilige domänenspezifische Wissen aus diesen Daten auf bisher unbekannte Bilder zu übertragen, ist eine große Menge an annotierten Beispielen notwendig. Diese müssen meist manuell und somit sehr zeitintensiv erstellt werden. Eine Möglichkeit diesen Aufwand zu minimieren, besteht darin, auf reale Daten zu verzichten und diese durch synthetisch erzeugte Bilder zu ersetzen. Eine synthetische Umgebung ermöglicht dabei eine einfache Modifizierung und effiziente Reproduktion der Daten sowie die schnelle Erstellung exakter Annotationen.

In diesem Paper wird untersucht, inwiefern das in den synthetisch erzeugten Bildern enthaltene Wissen mittels neuronaler Netze auf reale Bildaufnahmen adaptiert werden kann. Auf Basis quasi-prozedural erzeugter Weizenmodelle werden realitätsnahe Bilder eines virtuellen Weizenfeldes erzeugt. Diese dienen als Trainingsgrundlage für eine semantische Segmentierung, wobei unterschiedliche Netzarchitekturen verwendet werden (s. Kapitel 3). Die Übertragbarkeit der Ergebnisse auf reale Daten wird anhand realer Bildaufnahmen evaluiert (s. Kapitel 4). Der Ablauf des Verfahrens ist in Abbildung 1.1 zu finden. Ein Überblick über den Stand der Forschung in diesem Themenbereich wird im Kapitel 2 gegeben.

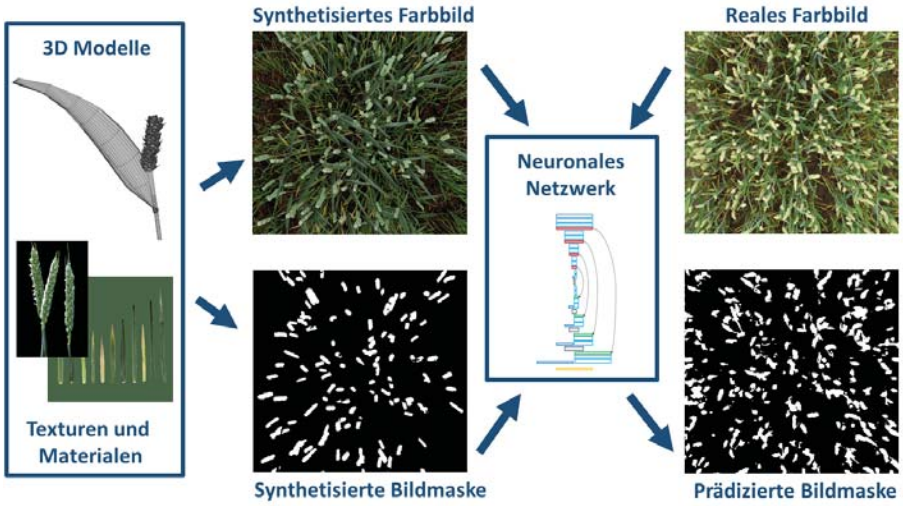

Abbildung 1.1: Übersicht der Ährendetektion. Auf Grundlage von 3D Modellen werden synthetische Farbbilder und Bildmasken erstellt. Mithilfe dieser wird ein neuronales Netz trainiert, welches die Prädiktion realer Bilder ermöglicht.

2 Stand der Forschung

Methodisch lassen sich bei der **Erkennung von Weizenähren** Deep Learning Ansätze von merkmalsbasierten Verfahren unterscheiden. Bei letzteren werden verschiedene Farb-, Textur- [1] oder Kantenmerkmale [3] definiert, welche eine pixelweise Detektion der Ähren innerhalb der Bilder ermöglichen. Dabei werden schwellwertbasierte Klassifikatoren sowie klassische Klassifikations- oder Clusterverfahren verwendet. Neuere Methoden dagegen basieren häufig auf Convolutional Neural Networks (CNNs) (s. [4] oder [5]). Weiterhin ermöglicht DeepCount [6] die Erkennung der Ähren, indem basierend auf Superpixeln diverse Merkmale berechnet und mittels eines CNNs analysiert werden. Bei [7] werden Farbinformationen mit thermalen Informationen verknüpft, um die Ähren zu identifizieren. Bei [8] wird ein semi-überwachtes Verfahren vorgestellt. Um den Annotationsaufwand für die Datengrundlage des Netzwerkes zu minimieren, wird die Idee des Aktiven Lernens auf das Deep Learning übertragen.

Die Nutzung von photorealistischen **synthetischen Datensätzen** für die semantische Segmentierung im Kontext von Computer Vision Anwendungen wird in [9] evaluiert. Die Grundlage bildet eine

Abbildung 3.1: Verwendete Weizenpflanzen- (links) und Grasmodelle (rechts) verschiedener Reifestadien. Während die Weizenmodelle in Farbe, Textur, Länge der Blätter, Ähren und in Ausprägung der Grannen variieren, sind bei den Grasmodellen lediglich die Texturen an den Reifegrad angepasst.

prozedural generierte, komplexe Szene, deren Geometrien aus der jeweiligen Perspektive physikalisch-basiert gerendert werden. Für jedes Trainingsbild wird dabei die Szene durch die von dem Benutzer definierten Parameter neu instanziiert. Andere Verfahren wie ProcSy [10] nutzen eine prozedurale Modellierungssoftware wie CityEngine® in Kombination mit Gaming-Engines, um photorealistischen Trainingsdaten zu erzeugen.

Domain Randomization beschreibt die Idee, die Verteilung der gerenderten Daten so zu variieren, dass das neuronale Netz, welches mit diesen Daten trainiert wird, robust genug ist, auch auf den realen Daten zu funktionieren. Dabei können sowohl Position, Ausrichtung oder Materialeigenschaften der zu synthetisierenden Inhalte variiert werden. Insbesondere spielen auch Beleuchtungseigenschaften, wie Intensität und Ausrichtung von Lichtquellen, als auch Renderingparameter eine große Rolle [11].

In dieser Arbeit wird eine fixe Szene mit manuell aufbereiteten Modellen verwendet, deren Diversität über Randomisierung weniger Parameter und einen virtuellen Kameraflug erreicht werden kann.

3 Methoden

Bei der **Bildsynthese von Weizenpflanzen** wird auf die freie Software Blender® zurückgegriffen. Mit dieser lässt sich eine beliebig große Szene aus nur wenigen 3D Modellen quasi-prozedural zusammenstellen, ohne jedes einzelne Szenenobjekt bei Anpassungswünschen

Abbildung 3.2: Manuell aufgenommene Texturen für ein frühes Reifestadium von Weizenmodellen (links), Modell einer Weizenpflanze versehen mit Materialeigenschaften (mittig), das gleiche Pflanzmodell texturiert und physikalisch korrekt gerendert (rechts).

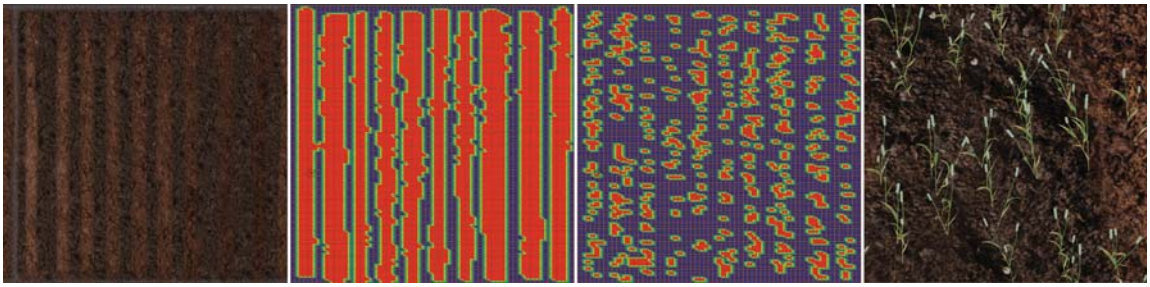

Abbildung 3.3: Von links nach rechts: Bodentextur für Weizenfeld, Platzierungskarte für Grasmodelle, Positionen für ein Weizenpflanzenmodell randomisiert auf dem Feld, Ausschnitt der Positionierung des ersten Weizenpflanzenmodells auf dem Feld.

von Geometrie- oder Materialeigenschaften bearbeiten zu müssen. Zusätzlich lässt sich die Szene unter variablen Aufnahmepositionen und Beleuchtungssituationen physikalisch korrekt rendern, sodass sich Bilder in großer Menge erstellen und beliebig reproduzieren lassen. Zu jeder photorealistischen Aufnahme wird ein Maskenbild aus gleicher Aufnahmerichtung und -höhe erzeugt. Somit ist es möglich automatisiert große Mengen an annotierten Beispieldaten zu erstellen, die für eine Bildsegmentierung genutzt werden können.

Grundlage für die **Modellierung des virtuellen Weizenfeldes** sind jeweils sechs variierende 3D Modelle von Weizenpflanzen und zwei unterschiedliche Modellgruppen von Grashalmen. Alle werden entsprechend eines Ziel-Reifegrades manuell angepasst und mit rea-

listischen Oberflächeneigenschaften versehen (Abb. 3.1). Dabei bilden Aufnahmen von realen Pflanzenblättern und Grashalmen die RGB-Textur für die Blattmodelle (Abb. 3.2). Die Materialeigenschaften der Weizenähren und -stängel sind über einen optischen Vergleich mit Aufnahmen von realen Pflanzen im jeweiligen Reifestadium festgelegt. Für die Grasmodelle werden frei verfügbare Texturen verwendet, die entsprechend angepasst sind.

Die Modellierung der Weizen- und Grasmodelle im virtuellen Weizenfeld lässt sich als quasi-prozedural beschreiben, da die Zufälligkeit der Farben und Texturen durch manuelle Auswahl beschränkt wird. Gleichzeitig kann jedoch durch die Anordnung der Pflanzen selbst eine optisch ausreichende Variabilität erreicht werden. Von jedem der sechs Weizenmodelle werden je nach gewünschter Dichte mindestens 3000 Instanzen zufällig auf dem gesamten Feld verteilt (Abb. 3.3). Dabei werden auch Höhe und Ausrichtung jeder Instanz in einem gewissen Intervall randomisiert. Die zusätzliche Verwendung der Grasmodelle trägt zu einer realitätsnahen Abbildung der Szene bei. Beispiele der synthetisierten Bilder des virtuellen Weizenfeldes sind in Kapitel 4 dargestellt und bewertet.

Das Ziel dieser Arbeit ist die **semantische Bildsegmentierung zur Ährenerkennung**, d. h. jeder Bildpixel soll dabei entweder als Ähre oder Hintergrund klassifiziert werden. Das hierfür notwendige Wissen soll mithilfe eines neuronalen Netzes aus den erzeugten synthetischen Bildpaaren adaptiert werden. Für diese Aufgabe wird sowohl das U-Net [12] als auch das DeepLab-V3+ [13] verwendet. Die Layer der Netze weisen eine klassische Encoder-Decoder-Struktur auf. Innerhalb des Encoders werden die Informationen des Eingangsbildes sukzessive verdichtet, sodass eine semantische Interpretation ermöglicht wird. Die räumliche Auflösung der einzelnen Layer nimmt mit jeder Verdichtung ab. Die durch den Encoder verlorene räumliche Information, wird durch den Decoder wiederhergestellt, sodass eine pixelweise Segmentierung des Eingangsbildes ermöglicht wird.

Der Encoder des U-Nets besitzt eine klassische Kaskade von Convolutional und Pooling Layern, deren Struktur gespiegelt im Decoder wiederzufinden ist. Dagegen besteht der Encoder des DeepLabs aus Atrous Convolutional Layern. Diese ermöglichen es, Fea-

tures mit hoher Kontextinformation zu berechnen, ohne dass die räumliche Auflösung der einzelnen Layer zu stark reduziert wird und somit ein schärferes Segmentierungsergebnis erzielt werden kann. Der Decoder des Netzwerkes besteht aus einfachen Upsampling und Convolutional Layern, die zusammen mit einigen Low-Level Features des Encoders das finale Ergebnis liefern. Als Kostenfunktion wird die binäre Kreuzentropie zum Trainieren der Netze verwendet.

4 Ergebnisse und Diskussion

Im folgenden Abschnitt werden die Ergebnisse der Bildsynthese und ihre Eignung als Trainingsdaten zur Ährendetektion erörtert.

Synthetisch generierte Trainingsdaten zur Ährenerkennung

In Abbildung 4.1 sind zwei Ergebnisse des physikalisch-basierten Renderers von Blender® dargestellt. Die gerenderten Bilder vermitteln einen photorealistischen Eindruck der Szene. Die Verteilung, Farbe, Größe und Position der Elemente sind vergleichbar mit deren Ausprägung in realen Aufnahmen. Ein erkennbarer Unterschied lässt sich allerdings bei der Darstellung des Untergrundes feststellen, da die verwendete Textur starke Glanzlichter aufweist, welche nicht gesondert aufbereitet wurden. So wirken die künstlichen Bilder an diesen Stellen dunkler als in der Realität.

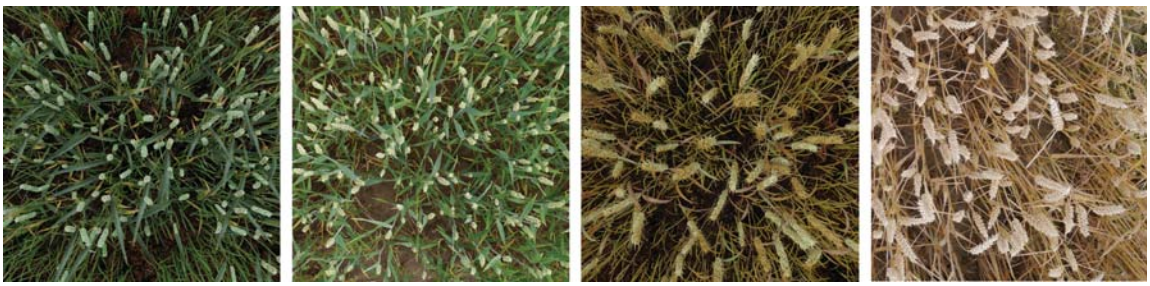

Abbildung 4.1: Ergebnis der Bildsynthese. Von links nach rechts ist jeweils ein synthetisches Bild und eine vergleichbare reale Aufnahme dargestellt.

Zu jedem synthetischen Farbbild ist eine passende Bildmaske der Ähren vorhanden. Auf Basis dieser Bildpaare werden zwei Trainings-

datensätze $T_{grün}$ und T_{gelb} erstellt, die jeweils Ähren eines frühen (Ähren mit grüner Färbung) und eines späten Reifestadiums (Ähren mit gelber Färbung) beinhalten. Jeder Datensatz besteht aus 250 Bildern mit einer Größe von 1531 x 1149 Pixeln. Durch die Vereinigung beider Datensätze wird ein weiterer Datensatz $T_{grün} \cup T_{gelb}$ erstellt, der Bilder beider Reifegrade beinhaltet.

Übertragbarkeit der synthethischen Daten auf reale Daten

Tabelle 1: Gütemaße für die Ährenerkennung basierend auf verschiedenen syntethischen Datensätzen.

Datensatz	IoU	Gesamt-Genauigkeit [%]	Präzision [%]	Sensitivität [%]
U-Net				
$T_{grün}$	46.21	87.71	64.75	61.73
T_{gelb}	43.67	86.65	63.08	58.66
$T_{grün} \cup T_{gelb}$	47.03	88.21	63.78	64.16
DeepLab				
$T_{grün}$	63.52	92.23	82.40	73.49
T_{gelb}	52.00	91.27	84.16	57.63
$T_{grün} \cup T_{gelb}$	69.96	93.88	86.84	78.25

Basierend auf den erzeugten Trainingsdaten werden die Gewichte der Netze gelernt. Aufgrund der begrenzten Speicherkapazität der GPUs werden die Bilder in insgesamt 7500 Patches einer Größe von 256 x 256 Pixeln unterteilt. Um eine möglichst große Robustheit der Netze zu erreichen, werden die Trainingsdaten generalisiert, indem die einzelnen Patches zufällig rotiert und vertikal oder horizontal gespiegelt werden. Die Datensätze werden jeweils zu 70 % als Trainingsdaten und zu 30 % als Testdaten verwendet. Für die verwendeten Netze werden bei den jeweiligen Testdatensätzen Gesamtgenauigkeiten von über 95 % erzielt.

Um die Übertragbarkeit der synthetischen Trainingsdaten zu analysieren, werden die trainierten Netze auf einen Datensatz bestehend aus 20 realen Bildern angewendet. Die Weizenpflanzen in den Auf-

nahmen weisen dabei unterschiedliche Reifegrade auf. Zu jedem Bild ist eine manuell erstellte Referenzmaske vorhanden.

Die Ergebnisse der Analyse sind in Tabelle 1 zusammengefasst. Als Maß für die Ähnlichkeit zwischen den prädizierten Masken und den Referenzmasken dient der Jaccard-Koeffizent (auch als Intersection over Union (IoU) bezeichnet). Zusätzlich ist die Gesamtgenauigkeit der Segmentierung, sowie die Präzision und die Sensitivität angegeben. Letztere beschreibt den Anteil der korrekt als Ähre erkannten Pixel gegenüber aller prädizierten Ährenpixel. Die Präzision liefert eine Aussage darüber, wie viele der in den Referenzmasken enthaltenden Pixel tatsächlich detektiert wurden.

Beim Vergleich der Ergebnisse der Datensätze fällt auf, dass $T_{grün}$ und T_{gelb} deutlich niedrigere Werte erzielen als bei ihrer Vereinigung $T_{grün} \cup T_{gelb}$. Es zeigt sich, dass die Modellierung verschiedener Reifestadien zu einer besseren Erkennung der Ähren führt. Des Weiteren fällt auf, dass die Werte der Präzision für $T_{grün}$ und $_{gelb}$ zwar ungefähr gleich sind, die Sensitivität bei T_{gelb} aber deutlich geringer ausfällt. Diese Unterschiede können dadurch erklärt werden, dass die verschiedenen Reifegrade bei den realen Bildern nicht gleichmäßig verteilt sind, sondern überwiegend Aufnahmen von grün gefärbten Pflanzen untersucht wurden. Die verschiedenen Netze beeinflussen das Ergebnis maßgeblich. Während die Ergebnisse des DeepLabs eine gute Erkennungbarkeit der Ähren belegen (der maximale IoU beträgt 69.96), weist das U-Net mit einem maximalen IoU von 47.03 eine deutlich geringe Erkennungsrate auf. Die beschriebenen Effekte lassen sich auch visuell in Abbildung 4.2 für die Auswertung von $T_{grün} \cup T_{gelb}$ erkennen. Zusätzlich ist für jedes Bild der jeweilige IoU angegeben. An diesem lässt sich erkennen, dass nicht alle Reifegrade mit derselben Güte erkannt werden. Die untersuchten realen Bilder weisen unterschiedlichste Reifestadien auf, wohingegen die synthetischen Datensätze sich nur auf zwei manuell modellierte Reifegrade stützen. Diese Diskrepanz ist vermutlich die Ursache für die variierende Erkennungsrate in den realen Aufnahmen.

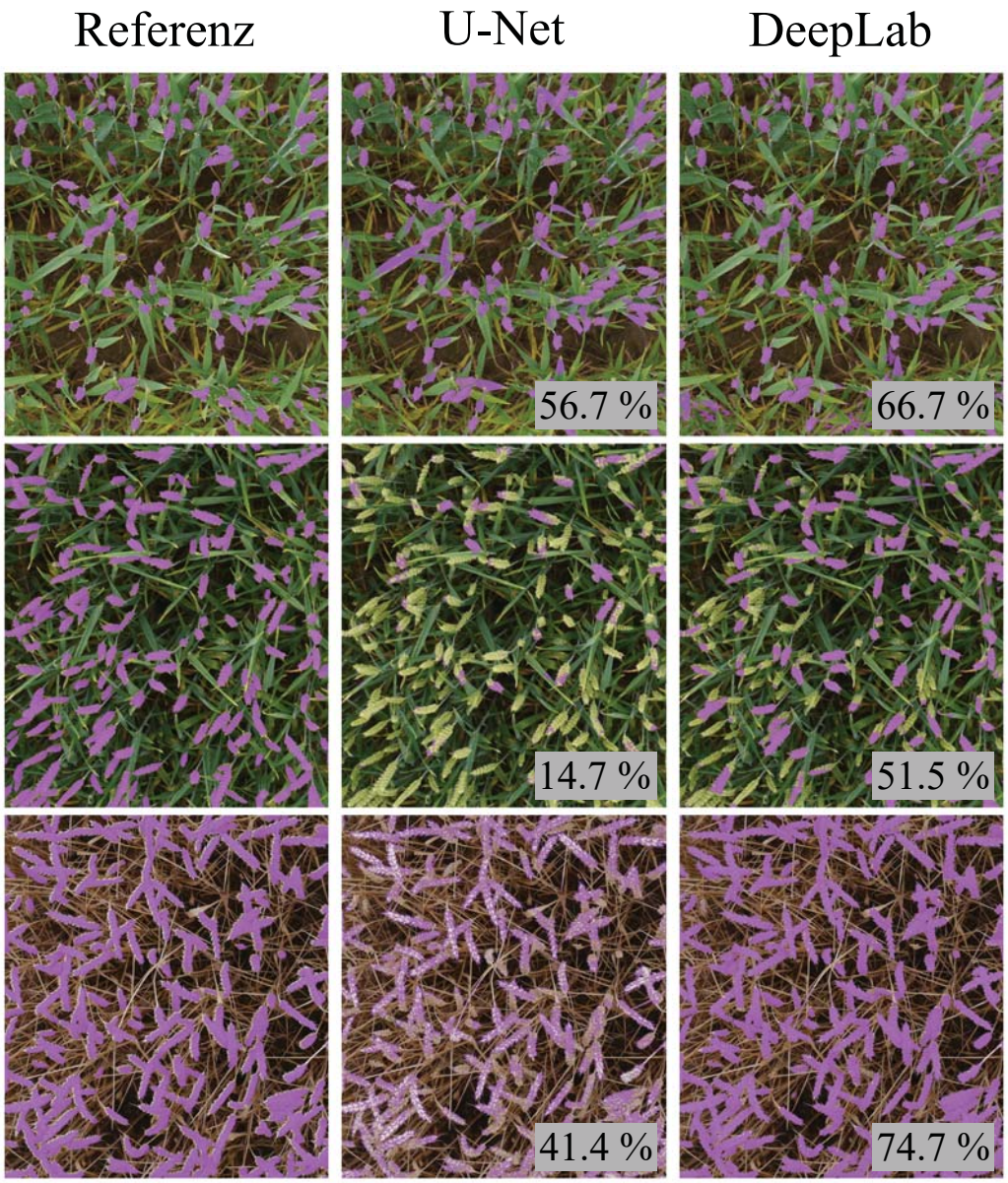

Abbildung 4.2: Ergebnis der Ährenerkennung für verschiedene reale Bildaufnahmen. Die erkannten Ähren sind violett markiert. Zusätzlich ist der IoU-Koeffizent jedes Bildes angegeben.

5 Fazit und Ausblick

Das Ziel der Arbeit war es, Weizenähren innerhalb von Farbbildern mittels neuronaler Netze zu erkennen. Anstelle manuell annotierter Trainingsdaten wurde dabei auf synthetisch erzeugte Daten zurückgegriffen. Die Ergebnisse zeigen, dass die Information aus

synthetisierten Daten auf reale Daten transferiert werden kann. Bestehende Abweichungen sind vor allem auf die nur geringe Anzahl an manuell modellierten Reifegraden innerhalb der Trainingsdaten zurückzuführen. In zukünftigen Arbeiten sollte daher eine automatisierte Modellierung verschiedener Wachstumsphasen angestrebt werden, um so ein größeres Spektrum an Informationen innerhalb der Trainingsdaten zu generieren.

Literatur

1. Y. Zhu, Z. Cao, H. Lu, Y. Li, and Y. Xiao, "In-field automatic observation of wheat heading stage using computer vision," *Biosystems Engineering*, vol. 143, pp. 28 – 41, 2016.
2. E. David, S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng, S. Liu, N. Kirchgessner, G. Ishikawa, K. Nagasawa, M. A. Badhon, C. Pozniak, B. de Solan, A. Hund, S. C. Chapman, F. Baret, I. Stavness, and W. Guo, "Global wheat head detection (gwhd) dataset: a large and diverse dataset of high resolution rgb labelled images to develop and benchmark wheat head detection methods," 2020.
3. C. Zhou, D. Liang, X. Yang, H. Yang, J. Yue, and G. Yang, "Wheat ears counting in field conditions based on multi-feature optimization and twsvm," *Frontiers in Plant Science*, vol. 9, p. 1024, 2018.
4. T. Alkhudaydi, D. Reynolds, S. Griffiths, J. Zhou, and B. Iglesia, "An exploration of deep-learning based phenotypic analysis to detect spike regions in field conditions for uk bread wheat," *Plant Phenomics*, vol. 2019, pp. 1–17, 07 2019.
5. J. Ma, Y. Li, K. Du, F. Zheng, L. Zhang, Z. Gong, and W. Jiao, "Segmenting ears of winter wheat at flowering stage using digital images and deep learning," *Computers and Electronics in Agriculture*, vol. 168, p. 105159, 2020.
6. P. Sadeghi-Tehran, N. Virlet, E. Ampe, P. Reyns, and M. Hawkesford, "Deepcount: In-field automatic quantification of wheat spikes using simple linear iterative clustering and deep convolutional neural networks," *Frontiers in Plant Science*, vol. 10, p. 1176, 09 2019.
7. Z. Grbović, M. Panić, O. Marko, S. Brdar, and V. Crnojevic, "Wheat ear detection in rgb and thermal images using deep neural networks," 10 2019.

8. S. Ghosal, B. Zheng, S. Chapman, A. Potgieter, D. Jordan, X. Wang, A. Singh, A. Singh, M. Hirafuji, S. Ninomiya, B. Ganapathysubramanian, S. Sarkar, and W. Guo, "A weakly supervised deep learning framework for sorghum head detection and counting," vol. 2019, 06 2019.
9. A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger, "Procedural modeling and physically based rendering for synthetic data generation in automotive applications," *CoRR*, 2017.
10. S. Khan, B. Phan, R. Salay, and K. Czarnecki, "Procsy: Procedural synthetic dataset generation towards influence factor studies of semantic segmentation networks," in *CVPR Workshops*, 2019.
11. S. I. Nikolenko, "Synthetic data for deep learning," 2019.
12. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.
13. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder ith atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.