JOHANNES KEPLER
UNIVERSITY LINZ

Author
Lukas Kurz
12007739

Submission
**Institute for
Computational Perception**

Thesis Supervisor
PhD **Jan Schlüter**

August 2025

Vineyard Row Detection from Satellite Imagery using Deep Learning

Bachelor's Thesis

to confer the academic degree of

Bachelor of Science

in the Bachelor's Program

Artificial Intelligence

**JOHANNES KEPLER
UNIVERSITY LINZ**
Altenberger Straße 69
4040 Linz, Austria
jku.at

Abstract

Digital farming platforms are becoming essential tools for modern vineyard management, but they face a fundamental problem: while government databases provide vineyard boundaries, they lack the internal row structure needed for practical applications like GPS-guided operations and spray tracking. Professional surveying could fill this gap, but at costs that often exceed what small vineyard operators can afford. This thesis explores whether we can automatically detect vineyard rows from publicly available aerial imagery using deep learning. Working with 79 manually annotated Austrian vineyards (containing anywhere from 1 to 196 rows each), I developed a system that combines high-resolution orthophotos from basemap.at with boundary data from the INVEKOS agricultural database. The approach uses a U-Net architecture to process 256×256 pixel image patches, treating row detection as a segmentation problem. To handle the challenge of thin linear features — vineyard rows appear as just 2-8 pixels wide in the imagery — I experimented with different target representations, ultimately finding that 9-pixel wide lines with Gaussian blur (σ=3) provided the best training signal. After testing various loss functions and architectural modifications, a relatively simple U-Net with double convolutions achieved the best results. The final system achieves 0.99 ± 0.12 meter spatial accuracy, meeting the 1-2 meter precision needed for agricultural GPS systems. More importantly, it correctly counts the number of rows in 88% of test vineyards, which is crucial for dividing fields into work sections. While the dataset is limited to Austrian vineyards and the system struggles with poorly maintained or seasonally bare fields, this work demonstrates that expensive professional surveys can be replaced with automated analysis of freely available imagery, given the requirements and contraints of digital applications. The methodology provides a foundation for making precision agriculture tools accessible to smaller operations that couldn't otherwise afford detailed field mapping.

Acknowledgements

0.1 Acknowledgment of AI Assistance

Parts of this thesis were created with the assistance of large language models (LLMs), specifically Anthropics Claude, which was used for brainstorming, refining technical phrasing, summarizing related work, and improving the clarity and readability of written sections. The use of these tools was guided by my own understanding of the subject matter, and all content was critically reviewed, edited, and validated by me to ensure accuracy, originality, and scientific integrity.

At no point was AI used to generate experimental results, conduct evaluations, or make analytical decisions without human oversight. All architectural choices, experiments, annotations, and interpretations reflect my own work and expertise.

Contents

Chapter 1

Introduction

1.1 Motivation

The digital transformation of agricultural operations has created demand for comprehensive vineyard management data that extends beyond traditional regulatory mapping requirements. While government databases such as Austria's INVEKOS [15] system track detailed vineyard boundary information for subsidy purposes, they lack internal structural details, necessary for more fine grained tracking functionalities in digital agricultural management platforms. This data gap represents a roadblock in the adoption of smart farming technologies, as current systems cannot efficiently and accurately track operational progress, optimize resource allocation, or provide meaningful analytics without detailed surveys of vineyard row structure.

Contemporary agricultural businesses increasingly rely on Software-as-a-Service (SaaS) platforms, mobile applications, and web-based management tools to coordinate farming operations, monitor treatment applications, and ensure regulatory compliance. These digital systems require accurate row count information and approximate spatial positioning to enable critical functionalities such as progress tracking across vineyard sections, prevention of double-spraying incidents, and GPS-based worker coordination. However, the absence of readily available row-level data forces agricultural operators to either forego these digital capabilities or invest in expensive professional surveying services that contradict the cost-effective, rapid-deployment model that makes digital agriculture tools attractive.

Professional surveying services, while providing centimeter-level accuracy, represent a fundamental mismatch with the operational requirements and economic constraints of digital agricultural platforms. Survey costs frequently exceed the annual subscription fees for digital management tools, while the time required for professional assessment delays the immediate deployment that characterizes successful adoption. More critically, the precision offered by traditional surveying far exceeds the accuracy requirements of digital applications, which typically operate within the 0.5-2 meter resolution constraints imposed by publicly available satellite imagery and consumer-grade GPS systems used in mobile devices, as the target application of these surveys are often GPS-guided machines that work at this level of accuracy.

The opportunity exists to leverage publicly available remote sensing data to bridge this information gap through automated analysis techniques. High-resolution orthographic imagery, such as Austria's basemap.at service providing 30cm per pixel resolution, contains sufficient detail for row detection at the accuracy levels required by digital agricultural applications. However, the extraction of meaningful row structure information from this imagery requires sophisticated computer vision approaches capable of handling the variability, complexity, and scale challenges inherent in agricultural remote sensing.

This creates a compelling use case for machine learning-based automated detection systems that can process publicly available imagery to generate the row-level information required by digital agricultural platforms. Such systems balance detection accuracy with cost- and time-efficiency, ensuring reliable performance while maintaining the rapid deployment and cost-effectiveness that characterizes successful digital solutions. The resulting technology would enable broader access to detailed vineyard mapping, enabling smaller agricultural operations to benefit from advanced digital management tools without the prohibitive costs of professional surveying services.

1.2 Objectives and Approach

The primary objective is to create a machine learning framework capable of processing high-resolution aerial imagery to extract row-level information at accuracy levels compatible with digital agriculture applications, eliminating the need for expensive professional surveying services while enabling rapid deployment of smart farming technologies.

This involves developing a deep learning model that can reliably identify vineyard row structures from 50-30cm resolution orthographic imagery, achieving spatial accuracy within the 0.5-2 meter range that aligns with the accuracy constraints of consumer-grade GPS systems used in mobile phones. This accuracy target represents a deliberate design choice that matches the practical requirements of digital management platforms rather than pursuing the centimeter-level precision offered by professional surveying services for GPS-guided applications.

A critical research goal is the optimization of row counting accuracy to support the analytical and operational functions of SaaS agricultural platforms, while necessitating minimal user intervention to correct for errors in predictions. Digital management tools require precise knowledge of row quantities to enable progress tracking, resource allocation optimization, and treatment verification workflows. The system must demonstrate high reliability in determining total row counts per vineyard section while providing sufficient spatial information to enable meaningful subdivision of vineyard areas for operational planning and tracking purposes.

In this thesis, we try to make use of cost-effective data sources and processing methodologies that align with the economic model of digital agriculture platforms. For the orthographic imagery, we use the WMTS (Web Map Tile Service) [39] service that is publicly available from basemap.at [25], which is constructed from several aerial imagery sources and this provides a higher resolution than publically available satellite imagery, which is often in the 0.5-1m per pixel range, where as basemaps resolution reaches up to 30cm per pixel in certain areas. This approach lets us use free high resolution imagery for training the model, while using a image resolution that can also be achieved using cost-effective commercial satellite imagery services, such as the Airbus Pleiades [27] constellation, which provides similar resolution, enabling the model to be used in a commercial settings anywhere around the globe. Since there is no public data on vineyard rows available, and private sources from existing surveys are hard to come by through cooperation, as the application of this model runs counter to the business model of professional surveying services, we used a combination of vineyard area and boundary information from the publically available INVEKOS Feldstücke Dataset from data.gv.at and manual annotation through the QGIS [35] application to create a dataset of 79 vineyards containing roughly 1,400 row segments.

1.3 Technical Challenges

1.3.1 Limited Training Data and Annotation Scalability

The most fundamental challenge stems from the lack of annotated vineyard row data suitable for machine learning model development. Unlike well-established computer vision domains with extensive public datasets, agricultural row detection lacks comprehensive labeled databases, necessitating labor-intensive manual annotation of training examples. With only 79 vineyards and approximately 1,400 manually annotated row segments available, the dataset represents a fraction of the data typically required for robust deep learning performance. This problems is amplified by the varying shapes, growth stages and quality of vineyard structures, where rows can be missing, falling into disrepair, obscured by vegation, or hard to detect due to the time of the year when the imagery was taken, which can lead to a lack of leaves on the vines. These edge case are more difficult to detect, but are also less likely to occur in the training data, making it even more difficult for the model to generalize to these cases.

1.3.2 Resolution Constraints and Feature Detection Precision

Despite the relatively high 30cm per pixel resolution provided by publicly available aerial imagery, vineyard rows spanning 80-250cm create target features measuring only 2.7-8.3 pixels in width. This narrow feature detection challenge is compounded by the organic, variable nature of agricultural structures that lack the geometric regularity of artificial features such as roads or building edges. Furthermore, the manual row annotation in QGIS introduces human error and inaccuracies into the dataset, since even using domain expertise, it can be difficult to make out rows from the aerial imagery exactly, so this inbuilt inaccuracy has be to taken into account in the training and evaluation of the model.

1.3.3 Agricultural Variability and Environmental Robustness

Vineyard structures exhibit considerable variation that challenges standardized detection approaches. Rows may follow terrain contours creating curved patterns, vary in spacing due to varietal requirements or historical planting decisions, and contain structural discontinuities caused by missing plants, shadows, or maintenance activities. Temporal variations present additional complexity as vine appearance changes dramatically throughout growing seasons, from bare winter structures to dense summer foliage. Different growth stages, pruning practices, and varietal characteristics create substantial visual diversity that the model must accommodate without requiring season-specific training data or user intervention. The system must demonstrate robustness across varying imaging conditions, sun angles, and shadow patterns since aerial imagery is typically captured under diverse environmental conditions and at different times of the year.

Chapter 2

Related Work

This chapter examines the foundational and contemporary research relevant to automated linear structure detection in agricultural imagery. The literature review encompasses three domains: traditional computer vision approaches for agricultural applications, deep learning architectures for semantic segmentation, and evaluation methodologies for linear feature detection. The synthesis of these research areas provides the theoretical foundation for developing automated systems capable of detecting continuous linear patterns in high-resolution aerial imagery.

2.1 Agricultural Computer Vision

2.1.1 Traditional Computer Vision Approaches

The application of computer vision techniques to agricultural crop detection originated from fundamental image processing methodologies developed during the 1960s and 1970s. Hough [14] introduced the Hough transform algorithm for detecting linear features in images containing noise and discontinuities, providing the theoretical basis for subsequent developments in automated feature detection.

Canny [4] developed the theoretical framework for optimal edge detection through a computational approach that maximizes signal-to-noise ratio while ensuring precise localization and minimal response multiplicity. The Canny edge detection algorithm has been applied as a preprocessing technique in various agricultural computer vision systems for enhancing linear features prior to geometric detection methods. Duda and Hart [11] formalized the integration of edge detection with the Hough transform, establishing mathematical frameworks for line and curve detection in digital images that subsequently influenced various computer vision applications, including agricultural crop structure detection systems.

One such agricultural application was demonstrated by Romeo et al. [28], who developed computer vision systems specifically for crop row detection in maize fields. Their methodology utilizes perspective projection analysis combined with pixel accumulation techniques for line detection, designed to handle real-world field conditions, including varying illumination, plant growth irregularities, and geometric distortions. The research demonstrated comparative performance against established methods, particularly the Hough transform algorithm, under various field conditions including different illumination levels and crop growth stages

Kise et al. [19] extended computer vision applications to three-dimensional agricultural environments through stereovision-based crop row detection systems. Their approach provided depth information enabling more robust detection of crop structures under varying terrain conditions, demonstrating the feasibility of automated guidance systems for agricultural machinery.

2.1.2 Remote Sensing Applications

The application of satellite and aerial imagery analysis to precision agriculture has developed sophisticated methodologies for understanding agricultural environments across multiple spatial scales. Patrício and Rieder [26] conducted a comprehensive systematic review of computer vision and artificial intelligence applications in precision agriculture, analyzing 25 publications to establish theoretical frameworks for categorizing detection approaches. Their analysis identified significant opportunities for advancing precision agriculture through the integration of GPU-accelerated computing and advanced artificial intelligence techniques, such as Deep Belief Networks, in the development of robust computer vision methodologies.

Chen et al. [7] developed a least squares fitting approach for crop row detection utilizing unmanned aerial vehicle imagery, demonstrating performance improvements over traditional Hough transform methods. Their experimental evaluation achieved Crop Row Detection Accuracy (CRDA), introduced by Vidović et al. [38], values between 0.99-1.00 for cotton crops and 0.66-0.82 for wheat crops under varying nitrogen and water treatment conditions. The research established quantitative benchmarks for aerial imagery-based detection systems and demonstrated the influence of crop management practices on detection performance.

Barros et al. [3] conducted comprehensive analysis of multispectral vineyard segmentation utilizing deep segmentation networks, comparing SegNet, U-Net, and ModSegNet architectures across vineyard locations in Portugal. Their experimental results indicated statistically equivalent performance ($p > 0.05$) among the three architectures, with mean IoU differences < 0.02 for vine detection tasks. The analysis revealed that Near-Infrared (NIR) spectral bands generally provided sufficient information for vineyard detection, while high-definition RGB imagery achieved equivalent or superior performance compared to lower resolution multispectral combinations.

Santos et al. [31] demonstrated machine learning approaches for vineyard segmentation from satellite imagery utilizing Support Vector Machine classifiers with Local Binary Pattern descriptors. Their experimental evaluation on Douro region vineyard datasets achieved accuracy rates exceeding 90%, establishing the continued relevance of traditional machine learning approaches for structured agricultural detection tasks when combined with appropriate feature engineering and domain-specific preprocessing methodologies.

Delenne et al. [9] developed comprehensive methodologies for vineyard delineation and characterization utilizing aerial imagery without requiring prior knowledge of parcel boundaries. Their frequency analysis approach employed Fast Fourier Transform techniques to evaluate row orientation and inter-row spacing automatically. The experimental validation achieved 90% accuracy for parcel detection and 92% accuracy for classification of missing vine plant densities, demonstrating the feasibility of automated vineyard characterization from aerial platforms.

2.2 Deep Learning for Segmentation

2.2.1 Foundational Architectures

The development of semantic segmentation underwent significant advancement with the introduction of fully convolutional architectures designed to preserve spatial information while enabling dense pixel-wise prediction. Ronneberger et al. [29] introduced U-Net, an encoder-decoder architecture incorporating skip connections originally developed for biomedical image segmentation. The architecture employs a contracting path that captures context through successive convolutions and pooling operations, followed by an expanding path that enables precise localization through upsampling and concatenation with high-resolution features from the contracting path. The U-Net architecture demonstrated superior performance on biomedical segmentation tasks, achieving an IOU (Intersection over Union) of 0.9203 on the ISBI cell tracking challenge 2014 and 0.7756 on the DIC-HeLa Dataset using minimal training data (35 and 20 images respectively). The symmetric expanding path enables precise localization of continuous features through the combination of high-resolution spatial information with upsampled semantic information. Skip connections preserve fine spatial details and facilitate gradient flow during training, addressing the degradation problem common in deep encoder-decoder architectures.

Zhou et al. [43] extended the U-Net framework through UNet++, introducing nested and dense skip connections to address the semantic gap between encoder and decoder feature representations. The architecture incorporates intermediate layers with deep supervision, enabling the network to operate at multiple semantic levels simultaneously. Experimental evaluation on four medical image segmentation datasets demonstrated improvements ranging from 3.9% and 3.4% in IoU compared to the original U-Net architecture and wide U-net respectively.

2.2.2 Multi-Scale and Context-Aware Architectures

Zhao et al. [42] introduced Pyramid Scene Parsing Network (PSPNet), which incorporates pyramid pooling modules to capture global contextual information at multiple spatial scales. The pyramid pooling module applies average pooling operations at different scales (1×1, 2×2, 3×3, and 6×6) to extract features representing different sub-regions, followed by upsampling and concatenation to create a comprehensive feature representation. The architecture achieved state-of-the-art performance on the PASCAL VOC 2012 dataset with 85.4% mean Intersection over Union (mIoU), demonstrating the effectiveness of multi-scale context aggregation for semantic segmentation tasks.

Chen et al. [5, 6] developed the DeepLab series of architectures, introducing atrous (dilated) convolutions as a fundamental component for semantic segmentation. Atrous convolutions enable explicit control of feature resolution through the introduction of holes (zeros) between kernel elements, allowing the receptive field to be enlarged without increasing the number of parameters or computational complexity. The mathematical formulation for atrous convolution is given by:

$y[i] = \sum_{k=1}^{K} x[i + r \cdot k]w[k]$

where $y[i]$ represents the output signal, $x[i]$ denotes the input signal, $w[k]$ are the filter weights, K is the kernel size, and r indicates the dilation factor. Chen et al. [6] combines atrous spatial pyramid pooling with encoder-decoder architectures, achieving 89.0% mIoU on the PASCAL VOC 2012 validation set.

Recent developments in semantic segmentation have addressed multi-scale characteristics and sparse linear structures through sophisticated architectural designs. Mahara et al. [24] introduced Dense Depthwise Dilated Separable Spatial Pyramid Pooling (DenseDDSSPP) integrated with DeepLabv3+ for automated road extraction from satellite imagery. The approach incorporates dense connections and depthwise separable convolutions to reduce computational complexity while maintaining feature extraction capability, achieving IoU scores of 67.21% and 71.61%, respectively on the Massachusetts and DeepGlobe dataset for road extraction tasks.

2.2.3 Specialized Architectures for Linear Structures

The detection of linear structures in aerial imagery has motivated the development of specialized architectural modifications and novel network designs. Li et al. [21] developed transmission line detection methodologies utilizing instance segmentation with multitask neural networks, addressing the detection of continuous elongated structures in aerial imagery. Their CableNet architecture incorporates overlaying dilated convolutional layers specifically designed for linear structure detection, achieving up to 59.91-73.44% pixel and 64.65-88.25% instance-level accuracy for transmission line detection in aerial imagery on two self compiled datasets.

Khan and Singh [18] advanced automated extraction methodologies for linear features utilizing CNN-based U-Net architectures applied to satellite imagery. The research incorporated Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) preprocessing to address common degradation artifacts in remote sensing imagery. Experimental evaluation on road extraction tasks demonstrated 95.48% accuracy and 60.97% IoU accuracy, indicating the effectiveness of quality-aware preprocessing for linear feature extraction.

Teplyakov et al. [37] addressed fundamental challenges in line detection through the development of lightweight CNN architectures incorporating embedded Hough layers. The approach combines traditional Hough transform mathematical principles with modern deep learning architectures, providing global strip-like receptive fields while maintaining computational efficiency.

2.2.4 Medical Imaging Parallels for Agricultural Applications

Medical image analysis has contributed significant insights to linear structure detection through analogous challenges in segmenting continuous, elongated anatomical structures. Du et al. [10] conducted a comprehensive review of deep learning applications for vascular segmentation, analyzing publications spanning 2016-2023. The review identified multiscale feature capture and attention mechanisms as promising future directions for effective linear structure detection models.

Galdran et al. [12] demonstrated state-of-the-art retinal vessel segmentation utilizing minimalistic U-Net architectures, achieving high accuracies on the DRIVE and STARE dataset with reduced parameter counts compared to conventional approaches. The research employed a cascaded U-Net extension, called W-Net to improve linear structure detection while maintaining computational efficiency, demonstrating that architectural simplification can enhance rather than compromise performance when properly designed.

Jiang et al. [16] investigated transfer learning approaches for retinal blood vessel segmentation utilizing fully convolutional networks, addressing limited training data challenges common in specialized segmentation tasks. Cross-database evaluation demonstrated robustness across different imaging conditions, when transferring between datasets. The research established transfer learning protocols that enable leveraging larger datasets from related domains to improve performance on tasks with limited annotated data.

2.3 Evaluation Metrics and Loss Functions

2.3.1 Traditional Segmentation Metrics

Traditional segmentation evaluation encompasses two fundamental approaches: distribution-based and region-based metrics, each offering distinct advantages and limitations for linear structure segmentation. Binary Cross-Entropy (BCE) Loss represents the foundational distribution-based approach, treating segmentation as independent pixel-wise binary classification problems. The BCE formulation $\text{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$ evaluates the probability distribution at each pixel independently, where y_i represents the ground truth label and p_i the predicted probability for pixel i.

BCE Loss provides several advantages for vineyard row detection, including computational efficiency, stable gradients during training, and direct optimization of probability calibration. The pixel-wise independence assumption enables effective handling of irregular boundary shapes common in agricultural imagery, where vineyard rows may exhibit varying width and geometric irregularities. However, BCE suffers from severe limitations in class-imbalanced scenarios typical of linear structure segmentation, where background pixels vastly outnumber foreground pixels, potentially leading to biased optimization toward majority class prediction.

In contrast, region-based metrics focus on spatial overlap between predicted and ground truth regions rather than pixel-wise probability distributions. The Intersection over Union (IoU) metric, also known as the Jaccard index, measures regional overlap through $\text{IoU} = \frac{|P \cap GT|}{|P \cup GT|}$, where P represents predicted pixels and GT represents ground truth pixels. While IoU provides intuitive understanding of overall segmentation quality, it fails to capture topological properties essential for linear structure evaluation, particularly connectivity and geometric accuracy critical for vineyard row detection applications.

The Dice coefficient, mathematically equivalent to F1-score and computed as $\text{Dice} = \frac{2|P \cap GT|}{|P| + |GT|}$, offers slightly different emphasis compared to IoU by giving equal weight to precision and recall components. Dice Loss, derived as $\text{DiceLoss} = 1 - \text{Dice}$, directly optimizes for regional overlap and often competes favorably with BCE for segmentation performance, particularly in scenarios where spatial coherence is more important than pixel-wise probability calibration.

Both distribution-based and region-based approaches suffer from fundamental limitations when evaluating linear structures where small spatial displacements can dramatically impact metric values despite maintaining structural integrity. A vineyard row prediction offset by one or two pixels may achieve low IoU/Dice scores while remaining functionally equivalent for agricultural applications requiring 1-2 meter accuracy. Similarly, BCE optimization may produce well-calibrated pixel probabilities that fail to preserve essential connectivity properties of linear structures.

Pixel-wise accuracy metrics become particularly problematic in agricultural contexts where severe class imbalance exists between linear structures and background regions. Vineyard rows typically occupy 5-15% of image area, making it possible to achieve high pixel accuracy (>90%) by simply predicting all pixels as background. This limitation affects both BCE and region-based metrics, necessitating specialized approaches that focus on linear structure characteristics rather than overall pixel classification performance.

2.3.2 Advanced Loss Functions for Class Imbalance

Lin et al. [22] introduced Focal Loss to address class imbalance in dense object detection, based on the balanced cross entropy, which adds a weighting factor for class imbalanced datasets. The focal loss formulation $\mathsf{FL}(p_t) = -\alpha(1-p_t)^\gamma \log(p_t)$ introduces a modulating factor $(1-p_t)^\gamma$ that reduces the relative loss for well-classified examples, enabling the model to focus on hard to classify examples. This approach proves particularly valuable for applications where the vast majority of pixels represent background, and traditional cross-entropy loss can be dominated by easy negative examples that provide little learning signal.

Salehi et al. [30] introduced Tversky Loss for 3D medical image segmentation, providing independent control over false positives and false negatives through the formulation $\mathsf{TL} = 1 - \frac{\mathsf{TP}}{\mathsf{TP}+\alpha\mathsf{FN}+\beta\mathsf{FP}}$, where α and β control the penalty for false negatives and false positives respectively. For vineyard applications, setting $\alpha < \beta$ emphasizes recall over precision, ensuring comprehensive row detection while tolerating minor false positives that can be filtered in post-processing stages. This approach proves particularly valuable when missing row segments (false negatives) create more significant problems than false detections for downstream agricultural applications.

The focal loss approach has been extensively adapted for medical image segmentation where similar class imbalance challenges exist. Abraham and Khan [1] developed Focal Tversky Loss, combining advantages of focal loss with Tversky index control over precision-recall trade-offs. This hybrid approach addresses both class imbalance and provides fine-grained control over false positive/negative trade-offs, highly suitable for preserving thin connections in linear structures where connectivity preservation may be more important than precise boundary accuracy.

2.3.3 Topology-Preserving Metrics

The preservation of topological properties represents a critical advancement in linear structure evaluation, addressing fundamental limitations of region-based metrics. Shit et al. [34] introduced centerlineDice (clDice), a novel topology-preserving loss function calculated on the intersection of segmentation masks and their morphological skeletons. Unlike traditional Dice loss that measures volumetric overlap, clDice evaluates overlap on morphological skeletons using the formulation clDice $= \frac{2|\mathsf{cl}(P)\cap\mathsf{cl}(GT)|}{|\mathsf{cl}(P)|+|\mathsf{cl}(GT)|}$, where $\mathsf{cl}()$ represents the centerline extraction operation.

The clDice approach provides mathematical guarantees for topology preservation in binary 2D and 3D segmentation, making it particularly suitable for vineyard row detection where maintaining connectivity across gaps caused by shadows, missing plants, or imaging artifacts is essential. The soft differentiable implementation enables end-to-end training while preserving topological constraints, addressing the fundamental challenge of maintaining linear structure integrity during segmentation. It also opens up opportunities for other skeleton based losses, due to its differentiable skeletonization.

Recent advances by Shi et al. [33] introduced Centerline Boundary Dice (cbDice), extending clDice by incorporating boundary radius information to address diameter imbalance issues common in agricultural applications. The cbDice formulation cbDice $= \frac{2|\mathsf{cl}_{boundary}(P)\cap\mathsf{cl}_{boundary}(GT)|}{|\mathsf{cl}_{boundary}(P)|+|\mathsf{cl}_{boundary}(GT)|}$ maintains topological integrity while improving geometric detail recognition.

2.3.4 Distance-Based and Geometric Metrics

Hausdorff distance provides essential geometric evaluation capabilities for linear structure assessment, measuring the maximum distance between predicted and ground truth boundaries. Karimi and Salcudean [17] developed novel loss functions optimizing Hausdorff distance during training, enabling direct optimization of geometric accuracy metrics crucial for agricultural applications. For linear structures, Hausdorff distance captures worst-case boundary displacement, critical for evaluating thin tubular structure accuracy where small geometric errors can significantly impact downstream applications.

The directed Hausdorff distance $h(P, GT) = \max_{p \in P} \min_{gt \in GT} ||p - gt||$ measures the maximum distance from predicted boundary points to the nearest ground truth boundary points, while the symmetric Hausdorff distance $H(P, GT) = \max(h(P, GT), h(GT, P))$ ensures bidirectional geometric consistency. These metrics prove particularly valuable for vineyard row detection where geometric accuracy directly impacts agricultural machinery guidance and precision application systems require spatial accuracy within 1-2 meter tolerances.

While Hausdorff distance addresses geometric accuracy, topological correctness requires additional mathematical frameworks. Clough et al. [8] introduced groundbreaking topological loss functions using persistent homology, enforcing topological constraints during training to ensure proper connectivity and prevent topological errors like disconnections or spurious branches. For linear structures, persistent homology captures the essential topological features including connected components, loops, and holes that traditional pixel-wise metrics fail to preserve. This approach addresses common failure modes in traditional loss functions where disconnected predictions or spurious connections can dramatically impact agricultural application effectiveness.

2.3.5 Domain-Specific Metrics

Agricultural applications require specialized evaluation metrics that align with practical deployment requirements rather than purely academic performance measures. Yuan and Xu [41] developed GapLoss specifically for road segmentation in remote sensing images, addressing connectivity preservation across occlusions and shadows common in aerial imagery. The GapLoss algorithm identifies potential gap pixels using morphological skeletons and distance transforms, then applies exponential weighting $w_{gap} = \exp(-d_{gap}/\sigma)$ based on proximity to existing structures, enabling targeted optimization of connectivity preservation.

The mathematical formulation GapLoss $= CE_{loss} + \lambda \sum(w_{gap} \times CE_{pixel})$ combines traditional cross-entropy with gap-specific weighting, achieving improvements in road connectivity metrics directly applicable to vineyard row detection where maintaining continuity across shadows and missing plants is critical for accurate row counting and spatial analysis.

2.3.6 Hybrid and Combination Approaches

The complexity of vineyard row detection from aerial imagery necessitates hybrid evaluation approaches that combine multiple complementary metrics to provide comprehensive performance assessment. Ma et al. [23] conducted a large-scale analysis of segmentation loss functions across medical imaging tasks, demonstrating the performance of Dice Loss and Dice extensions, as well as combination losses that integrate multiple metrics.

Yeung et al. [40] introduced Unified Focal Loss (UFL), providing a hierarchical framework generalizing Dice and cross-entropy based losses for handling class imbalanced datasets. Similar work has been done by Taghanaki et al. [36], who developed a combo loss formulation, that combines a weighted sum of Dice Loss and a modified cross-entropy loss, enabling simultaneous optimization of pixel-wise classification and regional overlap. This approach addresses the limitations of traditional metrics by providing a combined framework that captures both distribution-based and region-based characteristics.

Chapter 3

Methodology

3.1 Data Structure and Representation

The development of an effective vineyard row detection system necessitates careful consideration of data representation, processing strategies, and augmentation techniques. This section presents the methodological framework employed to transform raw geospatial data into suitable formats for deep learning model training and evaluation.

3.1.1 Raw Data

The research utilizes three primary data sources to construct a dataset for vineyard row detection. Vineyard boundaries are derived from the publicly available INVEKOS (Integrated Administration and Control System) [15] dataset, accessed through Austria's open data portal (data.gv.at) [13]. These boundaries are provided as shapefiles containing precise polygon geometries that distinguish individual vineyard parcels across Austrian wine-growing regions.

Aerial imagery consists of RGB orthophotos obtained from the basemap.at [25] Web Map Tile Service (WMTS) [39], which provides imagery with spatial resolution up to 30 cm per pixel in available regions. The orthophotos represent geometrically corrected aerial imagery that maintains consistent scale and projection, enabling accurate spatial analysis and feature extraction.

The ground truth vineyard row data comprises manually annotated row geometries created through visual interpretation of basemap.at orthophotos using QGIS [35] software. This annotation process involved digitizing individual vineyard rows as linear features, capturing both straight and curved row configurations representative of Austrian viticulture practices.

3.1.2 Input Data Design

The input data architecture employs a four-channel structure designed to provide both RGB information and spatial context for row detection. The first three channels contain standard RGB values derived from aerial orthophotos, while the fourth channel incorporates a binary vineyard polygon mask where pixels within vineyard boundaries receive a value of 1, and all other pixels are assigned a value of 0.

Spatial resolution remains fixed throughout the processing pipeline, ensuring that each image patch corresponds to consistent real-world dimensions. This approach facilitates meaningful comparison across different vineyard sites and enables the model to learn spatial relationships at a consistent scale.

The patch-based sampling strategy extracts multiple fixed-size patches from each vineyard using a sliding window approach with configurable overlap. This methodology addresses the challenge of variable vineyard sizes while maintaining consistent input dimensions for the neural network. Patches are filtered based on relevance criteria to exclude regions with insufficient vineyard content.

Where the target 30 cm per pixel resolution is not natively available, Lanczos upsampling [20] is applied to achieve consistent spatial resolution across the dataset. Target representations are generated as binary masks indicating row presence, derived from manual annotations with configurable line thickness parameters and optional Gaussian blur to account for positional uncertainties and make gradients smoother and thus learning easier.

Additionally, GIS representations of vineyard rows are maintained as MultiLineString [32] geometries with precise coordinate information, enabling accurate evaluation metrics that assess both segmentation quality and geometric accuracy.

3.1.3 Processing Pipeline

The data processing pipeline transforms raw geospatial data into training-ready datasets. Initial data matching performs spatial intersection operations between row geometries and vineyard boundaries to eliminate annotation errors that extend beyond vineyard parcels and produce only relevant row geometries and vineyards.

Image acquisition utilizes WMTS tile fetching and stitching to achieve complete vineyard coverage. Patch generation subsequently applies sliding window extraction with configurable overlap percentages (translated to stride) and implements relevance filtering to retain only patches containing sufficient vineyard content.

Lastly, vector geometries are converted to rasterized training masks through a parameterized process that allows adjustment of line thickness, blur parameters or other line representations. This flexibility enables optimization of target representations based on model architecture requirements and resolution constraints.

Dataset splitting is performed at the vineyard level to prevent data leakage, ensuring that patches from the same vineyard do not appear in both training and testing sets. This ensures realistic performance estimates by testing generalization to unseen vineyard sites rather than unseen patches from known vineyards.

3.1.4 Augmentation

Data augmentation strategies are key to address the limited training data availability while maintaining the geometric integrity essential for row detection tasks. Online augmentation performs transformations during training to increase dataset diversity without requiring additional storage.

Geometric augmentations include rotations of $\pm 90^\circ$ and both vertical and horizontal flips. These transformations preserve the linear characteristics of vineyard rows while providing orientation invariance. More complex rotations were initially considered but subsequently removed due to substantial computational performance impacts that compromised training efficiency.

3.2 Model Architectures

The architectural design for vineyard row detection builds upon established encoder-decoder frameworks while testing domain-specific modifications to address the unique challenges of linear feature segmentation in agricultural imagery. Different base architectures are systematically evaluated to identify optimal configurations for this task.

3.2.1 Base Architectures

The initial baseline architecture employs a convolutional neural network (CNN) with encoder-decoder structure similar to SegNet (Badrinarayanan et al. [2]), with convolutions in the encoding pathway and transpose convolutions in the decoding pathway. The encoder pathway progressively reduces spatial dimensions, with alternating convolutions and pooling operations, while increasing feature depth, enabling the extraction of hierarchical representations from local edge features to global row patterns. The decoder pathway reconstructs high-resolution segmentation maps from these features using upsampling, and transposed convolutions.

An iteration of this model introduces double convolutional layers, similar to the original U-Net design, to enhance feature extraction capabilities. This approach allows for more complex feature interactions. Double convolutional layers used a padding strategy to maintain spatial dimensions, ensuring that the output size matches the input size, which made shape calculations and skip connection implemention easier. This also eliminates the need for transposed convolutions in the decoder pathway, as normal convolutions with the same padding can be used in the decoding pathway to reduce channel dimensions.

Skip connections are added to facilitate gradient flow and preserve spatial information, enabling the model to learn both local and global features effectively. These connections link corresponding encoder and decoder layers, allowing high-resolution features to be combined with deeper semantic representations, and combined with double convolutions reproduce the original U-Net architecture.

Atrous (dilated) convolutions were evaluated to expand receptive fields without increasing computational complexity, to capture long-range dependencies inherent in vineyard row structures, where individual rows may extend across entire image patches.

Multi-scale kernel configurations were also evaluated, meant to address the challenge of detecting rows at different scales within the same image. Parallel convolutional branches with kernel sizes of 3x3, 5x5, and 7x7, as well with dilated convolutions, should capture features at multiple spatial scales, enabling robust detection of rows with varying widths and spacing.

3.2.2 Activation Functions and Normalization

Multiple activation function variants are evaluated to determine optimal non-linearity for row detection tasks. Rectified Linear Unit (ReLU) serves as the baseline activation, while Exponential Linear Unit (ELU), Leaky ReLU, and Swish (SiLU) activations are assessed for their potential to improve gradient flow and feature representation.

$$\text{ReLU:} \quad f(x) = \max(0, x)$$

$$\text{ELU:} \quad f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(e^x - 1) & \text{if } x \leq 0 \end{cases}$$

$$\text{Leaky ReLU:} \quad f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{if } x \leq 0 \end{cases}$$

$$\text{SiLU:} \quad f(x) = x \cdot \sigma(x), \quad \text{where} \quad \sigma(x) = \frac{1}{1 + e^{-x}}$$

Normalization strategies include both Batch Normalization and Layer Normalization implementations. Batch Normalization normalizes activations across the batch dimension, while Layer Normalization operates across feature dimensions within individual samples. The choice of normalization significantly impacts training stability and convergence rates.

3.2.3 Decoder Architecture and Upsampling

The decoder pathway reconstructs spatial resolution through upsampling operations that restore the original input dimensions. Three primary upsampling approaches are evaluated: bilinear interpolation, transposed convolutions, and unpooling operations. Bilinear interpolation provides computationally efficient upsampling with smooth transitions, formulated as:

$$I(x, y) = \frac{1}{(x_2 - x_1)(y_2 - y_1)} \sum_{i=1}^{2} \sum_{j=1}^{2} f(x_i, y_j)(x_{3-i} - x)(y_{3-j} - y)$$

Transposed convolutions enable learnable upsampling through trainable parameters, potentially capturing domain-specific upsampling patterns optimal for row reconstruction. Unpooling operations utilize stored pooling indices to restore spatial information, preserving exact locations of maximum activations from the encoder pathway.

3.2.4 Architecture Parameters

Model depth and width, meaning number of layers, as well as increase in channels per layer, are systematically varied to determine optimal network capacity for the available dataset size. Layer count investigations examine architectures ranging from shallow 3-layer networks to deep 8-layer implementations, while channel increase factors control feature map growth throughout the encoder pathway. Kernel size configurations balance receptive field coverage with computational efficiency. Standard 3×3 kernels provide efficient local feature extraction, while larger kernels capture extended spatial relationships at increased computational cost. Double convolution layers, characteristic of U-Net architectures, are evaluated against single convolution alternatives to assess their impact on feature extraction quality and segmentation performance.

3.3 Loss Functions

The selection and combination of loss functions critically influence model training dynamics and final performance for vineyard row detection. This research evaluates multiple loss function categories to address class imbalance, spatial accuracy, and topological preservation requirements.

3.3.1 Distribution-Based Loss Functions

Binary Cross-Entropy (BCE) loss serves as the fundamental distribution-based loss function, measuring the divergence between predicted probabilities and ground truth binary labels:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where y_i represents the ground truth label and $\hat{y}_i$ denotes the predicted probability for pixel i.

Focal Loss (Lin et al. [22]) addresses class imbalance inherent in row detection tasks, where row pixels constitute a small fraction of total image pixels. The focal loss formulation incorporates a modulating factor to reduce the contribution of well-classified examples:

$$\mathcal{L}_{Focal} = -\alpha_t(1 - p_t)^{\gamma}\log(p_t)$$

where α_t represents the class-specific weighting factor, γ controls the modulating strength, and p_t denotes the predicted probability for the true class.

3.3.2 Region-Based Loss Functions

Dice coefficient loss emphasizes overlap between predicted and ground truth regions, proving particularly effective for segmentation tasks with sparse positive classes:

$$\mathcal{L}_{Dice} = 1 - \frac{2|P \cap G| + \epsilon}{|P| + |G| + \epsilon}$$

where P represents predicted pixels, G denotes ground truth pixels, and ϵ prevents division by zero.

Intersection over Union (IoU) loss, also known as Jaccard loss, measures the ratio of intersection to union between predicted and ground truth regions:

$$\mathcal{L}_{IoU} = 1 - \frac{|P \cap G|}{|P \cup G|}$$

Tversky loss (Salehi et al. [30]) generalizes Dice coefficient by introducing asymmetric penalties for false positives and false negatives, which is also called the Tversky index. This is particularly useful for imbalanced datasets where false negatives may be more detrimental than false positives:

$$\mathcal{L}_{\text{Tversky}} = 1 - \frac{|P \cap G|}{|P \cap G| + \alpha|P \setminus G| + \beta|G \setminus P|}$$

where α and β control the penalties for false positives and false negatives, respectively.

3.3.3 Topology-Preserving Loss Functions

Centerline Dice (clDice) (Shit et al. [34]) loss specifically addresses the connectivity requirements essential for linear structure segmentation. The clDice score evaluates the overlap between predicted centerlines and ground truth centerlines, ensuring topological preservation. The final loss function is defined as a weighted loss of the standard Dice coefficient and the centerline Dice coefficient, in both cases using the soft version of the coefficient, utilizing probabilities instead of binary masks:

$$T_{\text{prec}}(S_P, V_L) = \frac{|S_P \cap V_L|}{|S_P|}, \quad T_{\text{sens}}(S_L, V_P) = \frac{|S_L \cap V_P|}{|S_L|}$$

$$\text{clDice}(V_P, V_L) = 2 \times \frac{T_{\text{prec}}(S_P, V_L) \times T_{\text{sens}}(S_L, V_P)}{T_{\text{prec}}(S_P, V_L) + T_{\text{sens}}(S_L, V_P)}$$

$$L_c = (1-\alpha)(1-\text{softDice}) + \alpha(1-\text{softclDice})$$

where $\mathcal{S}(\cdot)$ denotes the skeletonization operation that extracts centerlines from binary masks.

3.3.4 Combination Loss Functions

Unified Focal Loss (Yeung et al. [40]) combines focal loss with Dice coefficient to address both class imbalance and spatial overlap:

$$L_{mF}(p_t) = \delta(1-p_t)^{\gamma} \cdot L_{BCE}(p, y)$$

$$L_{mFT} = \sum_{c=1}^{C} (1 - mTI)^{\gamma}$$

$$L_{sUF} = \lambda L_{mF} + (1-\lambda) L_{mFT}$$

where L_{mF} represents the modified focal loss, L_{mFT} denotes the modified Tversky loss, and λ is a hyperparameter controlling the balance between the two components.

Combo Loss (Taghanaki et al. [36]) combines both distribution-based BCE and region-based Dice loss to balance pixel-wise accuracy with regional overlap:

$$\mathcal{L}_{\text{Combo}} = \alpha \mathcal{L}_{\text{BCE}} + (1-\alpha)\mathcal{L}_{\text{Dice}}$$

3.4 Evaluation Metrics

Since standard loss functions or scores often do not capture the specific requirements of vineyard row detection, a separate set of evaluation metrics was defined, to better assess model performance on important aspects such as segmentation quality, geometric accuracy, and topological preservation. As some of these metrics are CPU bound or not compatible with batch processing, they are computed only on the test set while training.

First, both **IoU and Dice score** are computed to evaluate segmentation quality, measuring the overlap between predicted and ground truth regions.

The next important criteria is the displacement error of predicted rows from ground truth rows. As output predictions are thicker propability masks, we first skeletonize the predicted rows to obtain centerlines, and then compute the **Hausdorff distance** between the predicted and ground truth centerlines. This metric quantifies the maximum distance between any point on the predicted centerline and the nearest point on the ground truth centerline, providing a measure of spatial accuracy.

Finally, reusing the already calculated centerlines, we label (blob counting) the predicted centerlines and create the **Row Count Accuracy** metric for comparing the number of predicted and ground truth rows.

$$\text{Row Count Accuracy} = \max\left(0, 1 - \frac{|\mathrm{C}_{\text{pred}} - \mathrm{C}_{\text{true}}|}{\max(\mathrm{C}_{\text{true}}, 1)}\right)$$

This metric counts the number of connected components in the predicted centerlines, which should match the number of ground truth rows, and provides a measure of topological preservation. Through experimentation, it was also seen that this metric is a robost indicator for the model's ability to generalize to unseen vineyard layouts, as it is less sensitive to small segmentation errors, and is not affected by overly broad predictions, which would trick the Hausdorff distance, in a sense combining both segmentation quality and topological preservation into a single metric.

Chapter 4

Experiments

4.1 Experiment Setup

4.1.1 Dataset Characteristics

Figure 4.1: Example vineyard sites used in the datasets.

The manually curated dataset comprises 79 vineyard sites distributed throughout Austria. Dataset construction leveraged publicly available resources from the Austrian INVEKOS Feldstücke dataset obtained from data.gv.at, combined with high-resolution aerial imagery from the basemap.at WMTS service.

Figure 4.2: Demonstration of a vineyard image with annotated rows.

The training dataset encompasses diverse vineyard configurations representative of Austrian viticulture practices. The annotated vineyards exhibit substantial variation in row density, with individual vineyards ranging from 1 to 196 rows (mean $\mu = 18$ rows; see Figure 4.3).This substantial variability reflects the diverse scale and management practices characteristic of Austrian viticulture, ranging from small family operations to large commercial enterprises.

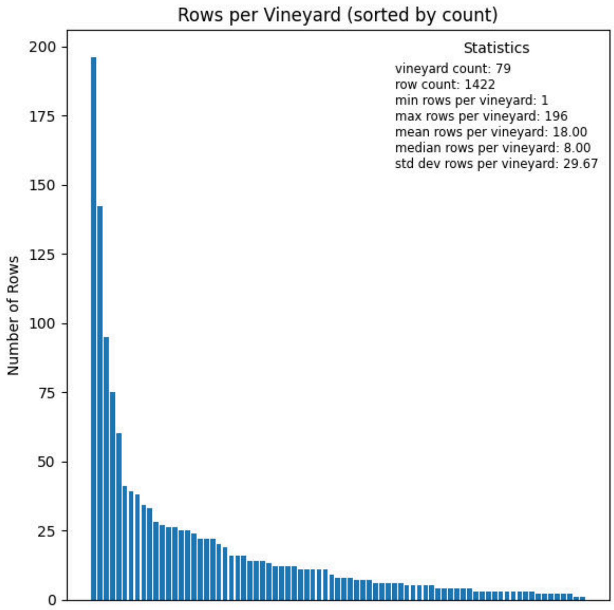


Figure 4.3: Distribution of row counts across the annotated vineyards.

The vineyard sites exhibit considerable variation in geometric properties, including parcel size, row orientation, row spacing, and topographic characteristics. Row spacing measurements range from 80cm to 250cm, reflecting different cultivation practices and varietal requirements across the study area.

The preprocessing pipeline implements a sliding window approach to extract training patches from complete vineyard imagery. Three distinct patch sizes were evaluated to assess the impact of spatial context on model performance: 256×256 pixels ($38.22m \times 38.22m$ at WMTS Zoom Level 20), 512×512 pixels ($76.44m \times 76.44m$), and 1024×1024 pixels ($152.88m \times 152.88m$). Patch size influences multiple factors, including model performance, computational cost, memory requirements, and training dataset size. Smaller patches provide more localized context and a larger dataset, potentially improving row detection accuracy in densely planted vineyards, while larger patches capture broader spatial relationships, which may enhance generalization across diverse vineyard layouts.

Ground truth annotations are converted from vector geometries to rasterized binary masks through a parameterized process that controls line thickness and applies optional Gaussian blur for improved target representation. This rasterization process, in combination with the selected window size, directly affects class distribution as demonstrated in Table 4.1. Class imbalance is calculated as the ratio of positive class pixels to negative class pixels, where positive pixels represent vineyard row locations.

$$\text{Class Imbalance} = \frac{\text{Number of positive pixels}}{\text{Number of negative pixels}}$$

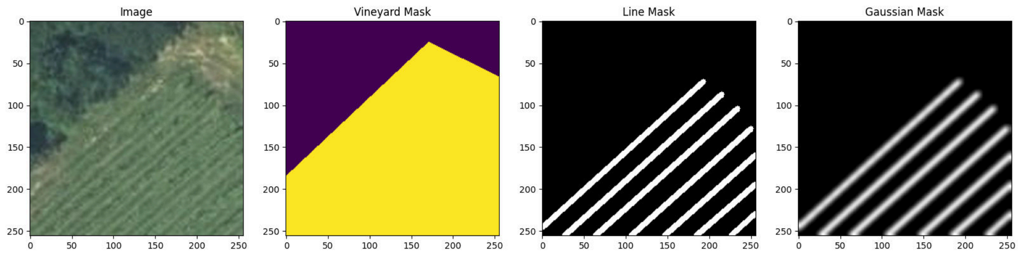


Figure 4.4: Example of a target mask generated from the rasterized ground truth annotations.

Table 4.1: Impact of patch size and rasterization parameters on class distribution and dataset characteristics. Gaussian kernel (gk), sigma (σ), and line thickness parameters control target mask generation.

Size	**Rasterization**	**Train**	**Test**	**Class Imbalance**
256×256	gk: 5, σ: 3, thickness: 9	56177	12773	0.39
	thickness: 5			0.17
	thickness: 1			0.02
512×512	gk: 5, σ: 3, thickness: 9	13968	2557	0.33
	thickness: 5			0.14
	thickness: 1			0.02
1024×1024	gk: 5, σ: 3, thickness: 9	2875	488	0.25
	thickness: 5			0.11
	thickness: 1			0.01

4.1.2 Implementation Details

Experiments were conducted on two computational platforms: a cloud-based GH200 system equipped with 96 GB VRAM, 4 TB storage, 432 GB RAM, and 64 CPU cores, and a local machine featuring an NVIDIA GTX 1080Ti GPU with 11 GB VRAM. The selected hardware configuration provides ample computational resources for training multiple deep convolutional neural networks while maintaining fast training times for iterative experimentation across multiple hyperparameter configurations. Mixed precision training was initially employed to optimize memory utilization; however, it was subsequently disabled due to stability issues with LayerNorm calculations, as well as a lack of noticeable performance improvements on the used model architectures. The AdamW optimizer with default parameters was used to optimize the model parameters, which is a variant of the Adam optimizer that incorporates weight decay regularization. Batch size selection was varied between 8, 16, 32, 64 depending on the available memory and size of input patches, which constrained the maximum batch size that could be used, although a batch size of 64 was preferred whenever possible, for its normalizing effect and computational efficiency. Depending on the experimental configuration, training was conducted for 20, 30, 50, or 100 epochs.

4.2 Hyperparameter Tuning

4.2.1 Target Representation Comparison

The analysis begins with a systematic comparison of different target representations, which constitutes a crucial aspect of the training process. The baseline configuration employed a SegNet-like architecture with the following parameters: batch size = 64, learning rate $\alpha = 0.0003$, training duration = 30 epochs, AdamW optimizer, and Binary Cross-Entropy (BCE) loss function.

Table 4.2: Ranges of hyperparameters evaluated in the target representation comparison.

Parameter	Values	Description
Zoom Level	`{19, 20 , 21}`	WMTS Zoom Levels
Window Size	`{128, 256, 512, 1024}`	In pixels
Line Thickness	`{1,3,5,9,15}`	In pixels
Overlap/Stride Factor	`[0.5-0.9]`	Fraction of overlap between windows
Gaussian Kernel Size	`{3,5,7,9,11}`	In pixels
Gaussian Kernel Sigma	`[1.0-10.0]`	Standard deviation in pixels

The evaluated parameters included WMTS zoom level (which determines spatial resolution in pixels per square meter), window size (patch dimensions), overlap fraction (stride determination), and target representation methodology. The target representation is further controlled by varying line width, Gaussian kernel size and sigma, which are used to control the thickness of the lines in the binary mask or the spread of the Gaussian heatmap. The results are shown in Figure 4.5.

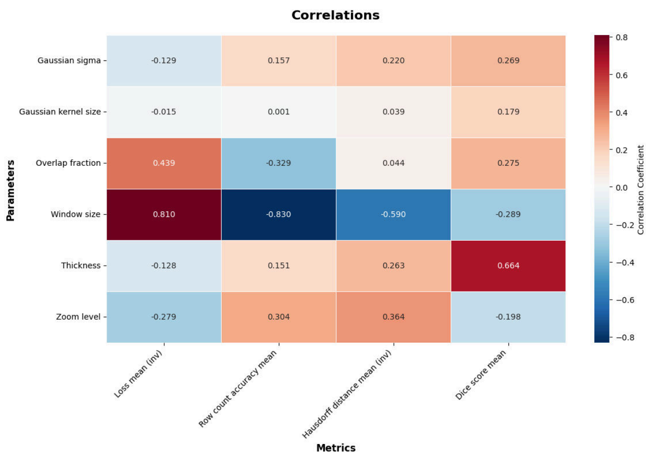


Figure 4.5: Comparison of different target representations and their impact on model performance. Loss and Hausdorff distance metrics are inverted for easier comparison, since higher is not better for these metrics when not inverted.

Quantitative analysis reveals several significant parameter relationships. Window size demonstrates a positive correlation with loss reduction, potentially attributable to the increased proportion of easily classified background pixels in larger patches. Overlap fraction exhibits a positive impact on model performance, likely due to the resulting increase in training dataset size. However, the Hausdorff distance metric exhibits apparent worse results with larger window sizes, which requires careful interpretation since this metric is not normalized and larger windows permit greater absolute distances, potentially introducing measurement bias. A consistent trend observed across multiple experimental configurations indicates that smaller window sizes (256×256 pixels) combined with WMTS zoom level 20 (14.93 cm per pixel) produce better detection performance. These configurations, utilizing reduced patch dimensions with fewer rows per sample and increased line thickness, facilitate more accurate row localization. The thicker line representations with Gaussian blur demonstrate improved Dice coefficient performance through easier spatial matching between predicted and ground truth annotations, while simultaneously enabling faster convergence during early training phases. This representation also somewhat mitigates the class imbalance problem, since the positive class is more prevalent in the training dataset, which is a common problem in semantic segmentation tasks. The final configuration that was chosen to be used for the rest of the experiments is a zoom level of 20, a window size of 256 pixels, a line thickness of 9 pixels, a gaussian kernel size of 5 pixels with a sigma of 3 pixels, and a overlap fraction of 0.9 for large training dataset size. This configuration provides a good balance between performance and computational efficiency, while also allowing for better generalization across different vineyard types.

4.2.2 Loss Function Component Analysis

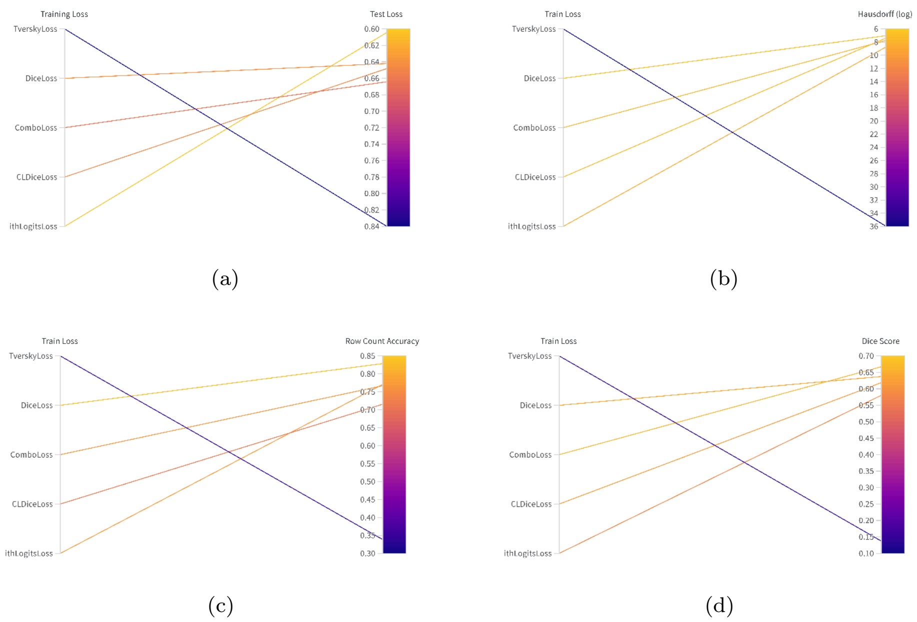


Figure 4.6: Comparison of different loss functions and their impact on model performance.

Figure 4.6 presents the comparative evaluation of multiple loss functions assessed using a fixed model architecture to determine their impact on target performance metrics. Unified Focal Loss and Focal Loss were discarded due to issues with their implementations and failure to converge. For Combo Loss an α of 0.7 was used, for ClDice a α of 0.5. The experimental results indicate that Centerline Dice (clDice) does not demonstrate positive impact on Row Count Accuracy and exhibits performance equivalent to standard Dice Loss. Furthermore the Combo Loss formulation outperforms Dice Loss on the Dice Score itself, while Dice Loss provided the highest Row Count Accuracy. This indicates that Dice or a modified combination of it are both promising loss functions. Due to the focus on Row Count Accuracy as primary indicator for good model performance, further experiments were conducted with Dice Loss.

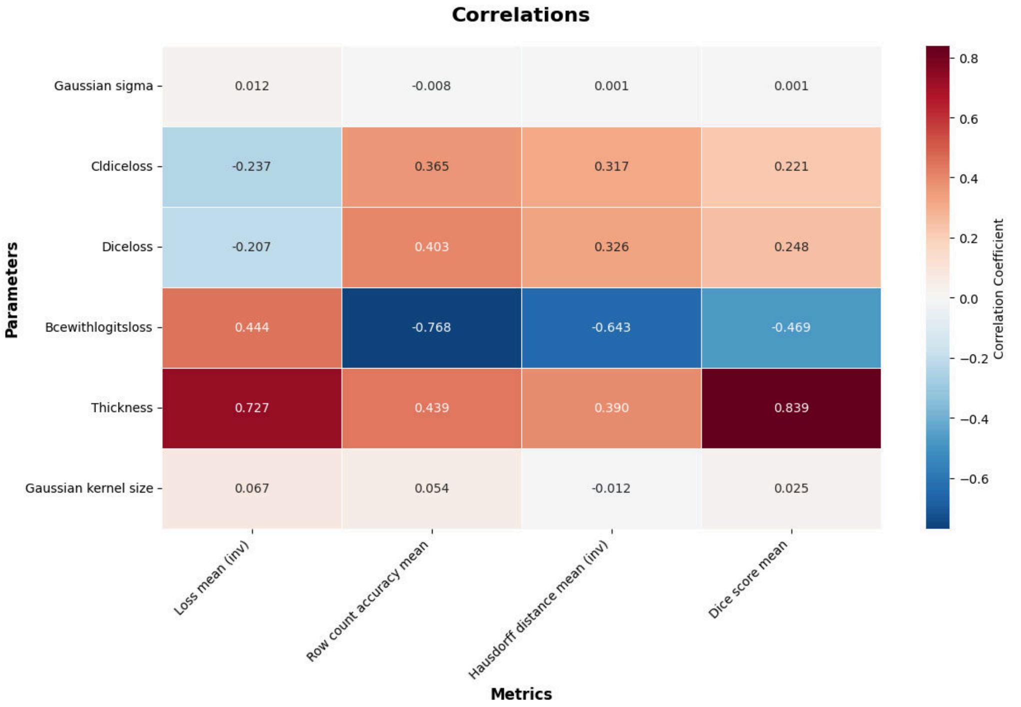


Figure 4.7: Comparison of loss functions and line thickness. Loss and Hausdorff distance metrics are inverted.

Subsequent investigation of line thickness and α parameter effects on combination loss functions demonstrated that clDice more consistently converged to a good solution compared to Combo Loss, which was more sensitive to the line thickness. clDice also showed better Hausdorff distances when using thinner lines, which would make sense, since part of the advantage of clDice is that it is less sensitive to class imbalance, which gets higher with thinner lines. However, comparing clDice to Dice, there does not seem to be a clear advantage, as they perform very similar and Dice outperforms clDice very slightly on all metrics. We can also see that one of the most important correlations is line thickness, as it directly positive impacts almost all metrics, especially when using Dice based losses, which is why the final configuration for the next experiments was set to a line thickness of 9 pixels, combined with DiceLoss.

4.2.3 Model Architecture Variations

For the model architecture experiments, three different experimental settings were evaluated. The first focussed on core architecture variables, such as the number of layers, increase in number of channels per layer, kernel sizes, dilation, different activation functions, different normalization layers, affine parameters in normalization layers, and different upsampling strategies. This experiment was conducted with the basic SegNet architecture, for faster turnaround times in the tuning process, since the effects with this architecture and insights can likely be transferred to the more complex and bigger architectures.

The model architecture investigation encompassed three distinct experimental phases. The initial phase focused on fundamental architectural variables, including layer depth, channel expansion factors, kernel dimensions, dilation parameters, activation functions, normalization strategies, affine parameters in normalization layers, and upsampling methodologies. This preliminary investigation utilized the basic SegNet architecture to enable rapid experimental iteration, with the assumption that architectural insights would transfer to more complex network designs.

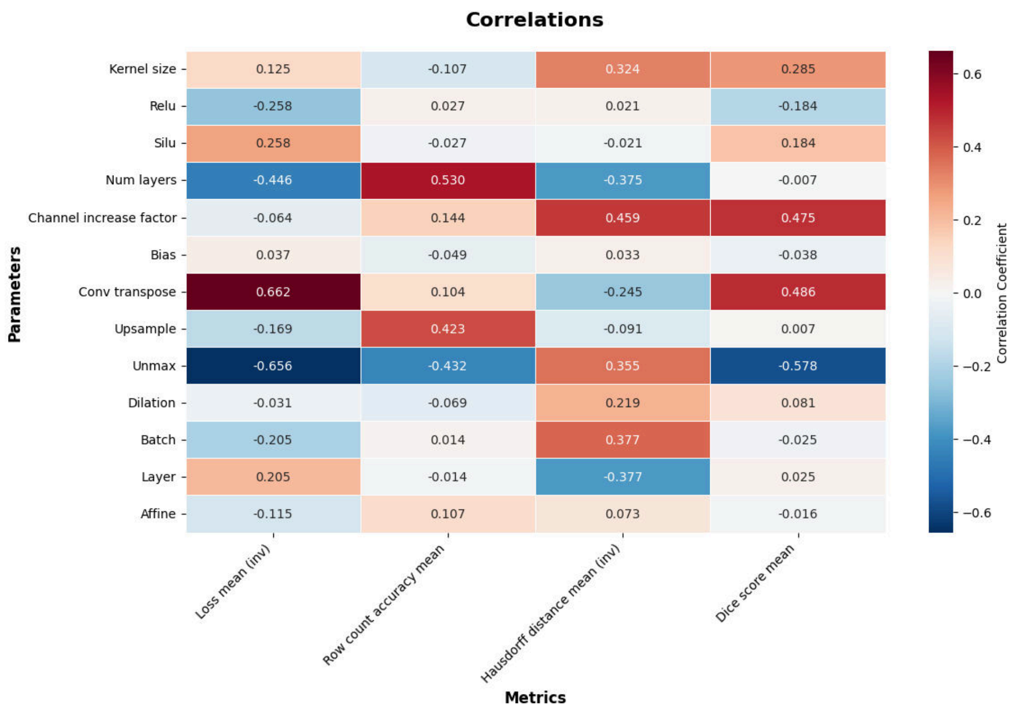


Figure 4.8: Quantitative analysis of architectural parameter effects on model performance metrics. Loss and Hausdorff distance metrics are inverted.

From Figure 4.8 we can see that some settings have a clear positive impact on all important metrics, such as the channel increase factor, as it allows for more complex feature extraction and better representation of the input data. The number of layers presents a tradeoff between row count accuracy and Hausdorff distance, with more layers leading to better row count accuracy but worse Hausdorff distance. As expected, kernel size and dilation have a similar effect, translating to a better Hausdorff distance, but worse row count accuracy, as the bigger kernels tend to produce larger predictions, resulting in smaller distances between predicted skeleton and ground truth. We can also see that althoug Relu is the default activation function for SegNet and UNet like models, Silu performs slightly better on the dice score. We can also observe a difference between transposed Convolutions and bilinear upsampling, with ConvTranspose performing better on the dice score, but worse on Row Count Accuracy, which is flipped for the bilinear upsampling. Since our primary focus is on the row count accuracy, bilinear upsampling was chosen for the final architecture, in this tradeoff. A similar tradeoff is made between batch normalization and layer normalization, where we prioritse the Hausdorff metric and thus choose batch normalization.

Figure 4.8 demonstrates the quantitative impact of individual architectural parameters on model performance metrics. Channel increase factor exhibits consistent positive correlation across all evaluated metrics, enabling enhanced feature extraction capabilities and improved input data representation. Layer depth presents a performance trade-off, with increased depth improving Row Count Accuracy while degrading Hausdorff distance measurements. Kernel size and dilation parameters demonstrate similar effects, resulting in improved Hausdorff distance performance but reduced Row Count Accuracy. This relationship occurs because larger kernels generate broader predictions, consequently reducing distances between predicted skeleton structures and ground truth annotations. The activation function analysis reveals that while Rectified Linear Unit (ReLU) represents the standard choice for SegNet and U-Net architectures, Sigmoid Linear Unit (SiLU) achieves marginally superior Dice score performance. Upsampling strategy comparison indicates that transposed convolutions outperform bilinear interpolation on Dice score metrics but exhibit inferior Row Count Accuracy. Given the prioritization of Row Count Accuracy as the primary performance criterion, bilinear upsampling was selected for subsequent experiments. Similarly, batch normalization demonstrated superior Hausdorff distance performance compared to layer normalization, leading to its adoption in the final architecture configuration.

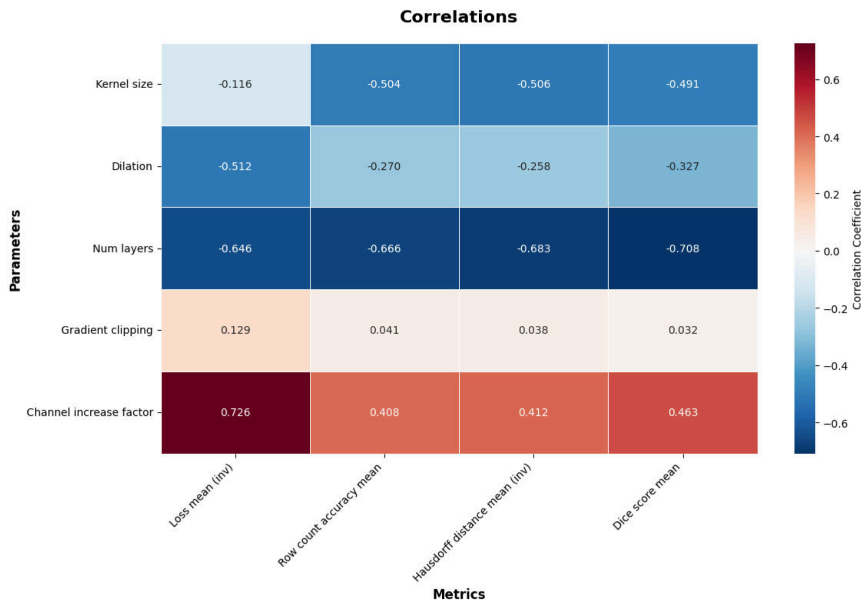
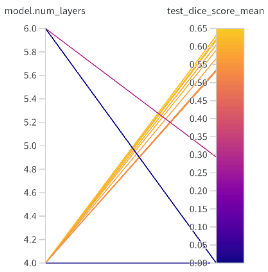


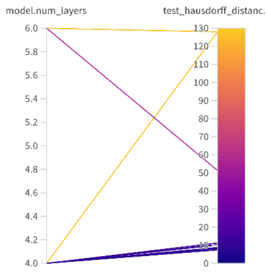
Figure 4.9: Architectural parameter evaluation using U-Net structure with double convolution layers. Loss and Hausdorff distance metrics are inverted.

The second experimental phase employed increased architectural complexity, utilizing U-Net like structure with double convolution layers, same-padding for dimension preservation, and standard convolutions replacing transposed convolutions in the decoder pathway. Channel increase factors were applied at the block level rather than individual convolution layers, consistent with standard U-Net implementation. Figure 4.9 presents the evaluation results for a reduced parameter subset, emphasizing specific parameter performance rather than general correlational trends. The layer depth analysis reveals that four-layer architectures achieve superior overall performance compared to six-layer implementations. Figure 4.10 demonstrates this relationship quantitatively, indicating that deeper structures experience information degradation that compromises detection performance. The visual comparison between four-layer and six-layer predictions confirms that excessive depth reduces prediction quality and spatial accuracy.

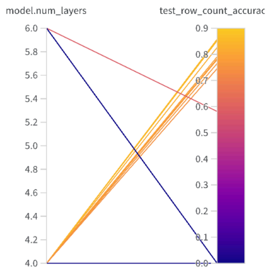
The third experimental phase introduced skip connections and multi-scale convolution layers incorporating mixed kernel sizes and dilation parameters. However, neither architectural enhancement demonstrated performance improvements over the established baseline, suggesting that the existing model complexity approaches the optimal configuration given the available dataset constraints.



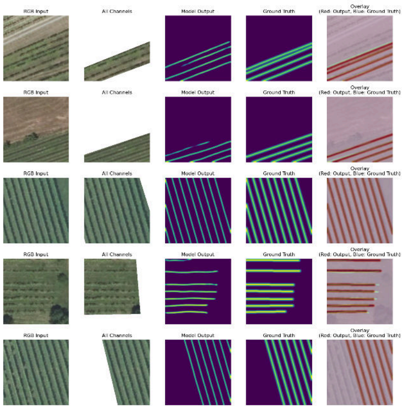
(a)



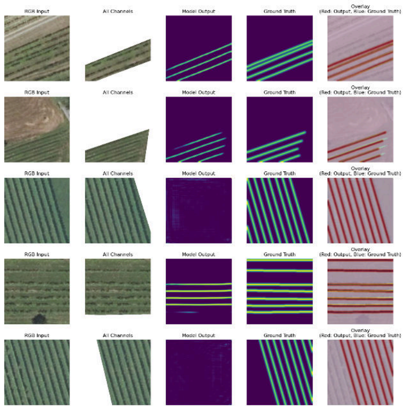
(b)



(c)



(d) Prediction with 4 layers



(e) Prediction with 6 layers

Figure 4.10: Comparison of 4 and 6 layers and their impact on performance metrics and prediction quality.

4.3 Performance Evaluation

4.3.1 Quantitative Results

Table 4.3 presents the quantitative performance metrics for the highest-performing models identified across all experimental configurations. The evaluation encompassed Dice coefficient, Hausdorff distance (pixels), Row Count Accuracy (RCA), training epochs, learning rate, kernel size, layer depth, Gaussian kernel parameters, line thickness, and loss function methodology. Parameters demonstrating consistent optimal performance across all experimental configurations were excluded from the tabular presentation to emphasize variable factors. These standardized parameters include WMTS zoom level 20, window size of 256 pixels, overlap fraction of 0.9, bilinear upsampling methodology, and channel increase factor of 3. These parameter selections achieved superior performance in individual experimental phases and maintained consistency across the complete experimental evaluation. The experimental results demonstrate that despite preliminary investigations indicating superior performance for standard Dice Loss, the highest-performing model configurations utilized Centerline Dice (clDice) loss functions. This finding suggests that clDice optimization provides advantages that become apparent only under optimal architectural and preprocessing conditions.

Quantitative analysis reveals a mean Hausdorff distance of 6.6 ± 0.8 pixels across the highest-performing models. At WMTS zoom level 20 (14.93 cm per pixel), this corresponds to spatial accuracy of $0.99 \pm 0.12 meters$, which satisfies the 1-2 meter precision

Table 4.3: Best performing models across all experiments. V2 is the SegNet Model, V3 adds the DoubleConvs, V4 is a full UNet with skip connections, V5 is a UNet modification with multi-scale convolutions.

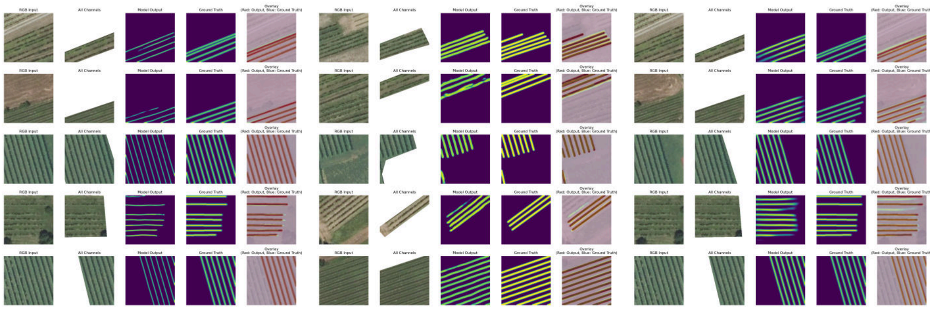
Run No.	**Model**	**Dice Score**	**Hausdorff**	**RCA**	**Epochs**	**LR**	**Kernel**	**Layers**	**GK Size**	**GK Sigma**	**Thickness**	**Criterion**
#22	v3	0.623	7.806	0.883	30	0.0003	5	4	9	5	5	BCE
#13	v3	0.632	7.352	0.880	30	0.0003	3	4	9	5	5	BCE
#509	v4	0.757	6.227	0.879	20	0.0001	3	4	5	3	9	CLDice
#505	v3	0.774	6.603	0.878	20	0.0001	3	4	5	3	9	CLDice
#504	v3	0.761	6.610	0.869	20	0.0001	3	4	5	3	9	CLDice
#508	v4	0.755	6.329	0.854	20	0.0001	3	4	5	3	9	CLDice
#308	v2	0.539	13.378	0.838	30	0.0003	7	6	9	5	5	BCE
#510	v5	0.750	7.315	0.823	20	0.0001	3	4	5	3	9	CLDice

requirement established for practical agricultural applications. The Row Count Accuracy is also very high, with most models achieving above 0.85, which indicates that the models are able to detect the rows with a high degree of accuracy, even in challenging vineyard configurations. Performance limitations must be contextualized within the constraints of manual annotation accuracy and geometric representation challenges. Row segments positioned partially outside vineyard polygon boundaries, while included in ground truth targets, remain predominantly outside the polygon mask, creating systematic evaluation constraints. Additionally, inconsistencies in manual annotation and temporal variations in imagery capture conditions establish practical performance ceilings for the evaluated metrics.

4.3.2 Qualitative Analysis

Figure 4.11 demonstrates representative test set predictions from the highest-performing model configurations. The visualization format presents input imagery, polygon mask channels, predicted segmentation masks, ground truth representations, and overlay comparisons to enable comprehensive qualitative assessment. The model predictions exhibit high-quality row detection capabilities, with notable differences attributable to kernel size selection and loss function methodology. Dice-based loss functions generate predictions characterized by binary mask characteristics, while Binary Cross-Entropy (BCE)-based approaches produce probability distributions with varying confidence scores. Although this distinction does not significantly impact final performance through skeletonization post-processing, the confidence score information may prove valuable for active learning implementations where sample difficulty assessment guides annotation prioritization.

Analysing failure modes reveals consistent challenge patterns across model configurations. Primary difficulties arise from row segments positioned near vineyard polygon boundaries, where ground truth annotations include structures predominantly located outside the polygon mask region. This geometric inconsistency creates systematic evaluation constraints that limit achievable performance metrics. Temporal and maintenance-



(a) #13 (b) #508 (c) #308

Figure 4.11: Example of Test set predictions from some of the top models. First column is the input image, second column is the added 4th channel with the polygon max, 3rd column is the predicted mask, 4th column the ground truth representation, and the last column both the predicted and ground truth mask overlaid on the input image. (a) is a 4-Layer, DoubleConv, BCE trained model, (b) is a 4-Layer, UNet, clDice trained model and (c) is a 6-Layer, Segnet, BCE trained model

related factors constitute additional challenge categories. Seasonal variations in imagery capture timing result in vineyard rows with minimal vegetation coverage or completely bare structures, substantially increasing detection difficulty. Furthermore, vineyard sites experiencing inadequate maintenance exhibit missing plants, significant gaps in row continuity, and structural irregularities. The current model configurations demonstrate limited robustness to these conditions due to insufficient representation of such cases in the training dataset, highlighting the need for expanded data collection encompassing diverse maintenance states and seasonal conditions.

Chapter 5

Discussion

5.1 Technical Achievements

The research demonstrates technical achievements in automated vineyard row detection through the integration of multiple methodological techniques. The comprehensive experimental evaluation validates the effectiveness of domain-specific architectural modifications and data processing strategies designed to address the unique challenges inherent in agricultural remote sensing applications.

5.1.1 Data Structure and Processing Pipeline

The four-channel input representation demonstrates its usefulness in addressing fundamental challenges in agricultural computer vision. The integration of RGB orthophoto information with binary vineyard polygon masks enables the model to focus computational attention on only relevant regions while maintaining spatial context essential for accurate row detection, thus making the segmentation task easier and more efficient, not needing to learn the distinction between vineyard and non-vineyard areas.

The processing pipeline successfully transforms geospatial data formats into standardized training datasets suitable for deep learning applications. The integration of publicly available INVEKOS vineyard boundaries with high-resolution basemap.at orthophoto imagery also demonstrates an approach that can be transferred to other areas and contexts, using data from other regulatory agencies, as well as existing commercially availbale high resolution satellite imagery. The parameterized approach to target mask generation, incorporating adjustable line thickness and Gaussian blur parameters, provides benefits for addressing class imbalance while maintaining spatial precision requirements.

The sliding window patch extraction strategy with configurable overlap demonstrates effectiveness in handling variable vineyard geometries while maintaining consistent spatial resolution. The experimental validation confirms that 256×256 pixel patches with 90% overlap is a setting that can achieve optimal balance between computational efficiency and detection performance, generating sufficient training data density to enable robust model convergence despite limited annotated vineyard availability.

5.1.2 Architecture Choices and Domain-Specific Modifications

The systematic evaluation of encoder-decoder architectures reveals that domain-specific modifications significantly enhance vineyard row detection performance compared to standard semantic segmentation approaches. The experimental comparison demonstrates that U-Net-based architectures with double convolution layers achieve superior Row

Count Accuracy metrics compared to simpler SegNet implementations, indicating that increased feature extraction capacity provides measurable benefits for linear structure detection tasks.

The investigation of activation functions also reveals that Sigmoid Linear Unit (SiLU) outperforms traditional Rectified Linear Unit (ReLU) implementations on Dice coefficient metrics, suggesting that smooth activation functions facilitate more effective gradient flow during training of linear structure segmentation networks. Similarly, the comparative evaluation of normalization strategies indicates that Batch Normalization demonstrates superior performance compared to Layer Normalization for the specific requirements of vineyard row detection, likely due to improved feature distribution characteristics across batch dimensions.

The upsampling methodology comparison provides insights for deployment considerations, with bilinear interpolation achieving superior Row Count Accuracy performance compared to transposed convolutions despite slightly reduced Dice coefficient scores. This finding suggests that learnable upsampling parameters may introduce artifacts that compromise topological accuracy, making deterministic interpolation methods more suitable for applications requiring precise row enumeration.

Notably, the experimental results indicate that sophisticated architectural enhancements, including skip connections and multi-scale convolution kernels, do not provide measurable performance improvements over optimally configured baseline architectures. This finding suggests that the current dataset size represents a limiting factor for model complexity, and that additional architectural sophistication probably requires correspondingly larger training datasets to realize potential benefits.

5.1.3 Loss Function Optimization and Training Dynamics

The comprehensive evaluation of loss function formulations reveals that topology-preserving approaches provide subtle but meaningful advantages for vineyard row detection applications. The experimental comparison demonstrates that Centerline Dice (clDice) loss achieves marginally superior performance on the highest-performing model configurations, despite preliminary investigations suggesting advantages for standard Dice Loss. This apparent contradiction indicates that clDice benefits become apparent only under optimal preprocessing and architectural conditions, highlighting the importance of hyperparameter tuning the entire system rather than isolated component selection.

The investigation of target representation parameters confirms that line thickness exerts the most significant influence on model performance across multiple loss function categories. The optimal configuration utilizing 9-pixel line thickness with Gaussian blur demonstrates effectiveness in addressing class imbalance while maintaining sufficient spatial precision for accurate row localization. This finding establishes quantitative guidelines for target mask generation that balance training stability with geometric accuracy requirements.

The experimental analysis of combination loss functions, including Combo Loss and Unified Focal Loss implementations, reveals that simple, well-tuned loss functions often outperform complex combinations. This result suggests that careful hyperparameter optimization of established loss functions may provide more reliable performance improvements than sophisticated mathematical formulations that introduce additional tuning complexity and implementation difficulties.

5.2 Limitations and Challenges

Despite the demonstrated technical achievements, the developed vineyard row detection system exhibits several fundamental limitations that constrain its immediate applicability and identify critical areas requiring further research and development.

5.2.1 Resolution Constraints and Spatial Accuracy Trade-offs

The 30 cm per pixel spatial resolution limitation imposed by publicly available aerial imagery creates fundamental constraints on feature detection precision that directly impact system performance. Vineyard rows spanning 80-250 cm create target features measuring only 6-18 pixels in width, representing a challenging detection scenario that approaches the theoretical limits of pixel-based segmentation approaches. This resolution constraint necessitates the use of artificially thickened target representations that may not accurately reflect the geometric characteristics of actual vineyard structures.

The experimental evaluation demonstrates that the achieved spatial accuracy of 0.99 ± 0.12 meters satisfies the established 1-2 meter precision requirement for practical agricultural management applications. The trade-off between spatial resolution and computational efficiency becomes evident in the patch size optimization results. While larger patches theoretically provide increased spatial context for row detection, the experimental findings indicate that smaller patches achieve superior performance metrics, likely due to the increased training dataset density rather than improved contextual information. This relationship suggests that resolution limitations can only really be addressed through larger training datasets or higher-resolution imagery sources, rather than architectural or preprocessing modifications alone.

5.2.2 Dataset Size and Generalization Constraints

The patch size optimization results reveal that smaller patches achieve superior detection performance compared to larger patches, despite the theoretical advantage of increased spatial context in larger windows. This counterintuitive finding appears attributable to increased training dataset density rather than improved contextual information, suggesting that resolution limitations require mitigation through expanded training datasets or higher-resolution imagery sources rather than algorithmic modifications alone.

The dataset comprising 79 manually annotated vineyards presents substantial constraints on model generalization and statistical confidence. This dataset size falls below the thousands of examples typically required for robust deep learning performance. Manual annotation introduces systematic performance limitations through inherent subjectivity in identifying row boundaries from aerial imagery, particularly in cases involving missing plants or structural irregularities. These ground truth inconsistencies directly impact evaluation metric reliability and establish practical performance ceilings. Additionally, temporal variations in imagery capture conditions create dataset heterogeneity that are hard to adress without more diverse training data.

5.2.3 Vineyard Diversity and Environmental Robustness

The experimental evaluation reveals systematic performance degradation when encountering vineyard configurations that deviate from well-maintained, regularly spaced row patterns typical of modern commercial operations. The qualitative analysis identifies consistent failure modes in vineyards exhibiting missing plants, significant gaps in row continuity, or structural irregularities resulting from differing maintenance practices. These challenging conditions represent a substantial portion of real-world agricultural sites but remain underrepresented in the current training dataset.

Temporal variations in imagery capture timing create additional robustness challenges that directly impact detection reliability. Vineyard rows captured during dormant seasons, with minimal vegetation coverage or completely bare structures, demonstrate substantially reduced detection accuracy compared to images acquired during active growing periods. This seasonal sensitivity limits the practical deployment flexibility of the current system and necessitates temporal considerations in imagery acquisition planning, or more data collection across different seasonal conditions to ensure robust performance.

Environmental factors, including shadow patterns, varying soil types, and diverse management practices, contribute to appearance variations that challenge model robustness. The current training approach relies primarily on data augmentation through geometric transformations, which may not adequately address the substantial appearance variations resulting from environmental and temporal factors present in real-world agricultural imagery.

5.3 Real-World Applicability

The assessment of real-world deployment potential requires evaluation of the developed system against practical agricultural management requirements, including accuracy specifications, operational constraints, and integration considerations with existing digital agriculture platforms.

5.3.1 Accuracy Requirements and Performance Validation

The achieved spatial accuracy of 0.99 ± 0.12 meters successfully meets the established 1-2 meter precision requirement derived from the operational constraints of consumer-grade GPS systems utilized in mobile agricultural management applications. This performance level enables meaningful subdivision of vineyard areas for operational planning purposes while supporting GPS-based worker coordination within the accuracy limitations of standard mobile devices.

The Row Count Accuracy of 88 % across the highest-performing model configuration demonstrates sufficient reliability for practical vineyard management applications. Accurate row enumeration enables critical functionalities including progress tracking across vineyard sections, resource allocation optimization, and treatment verification workflows essential for digital agricultural platforms. The consistency of row counting performance across diverse vineyard configurations suggests robust applicability despite the architectural and preprocessing optimizations required to achieve these results.

However, the performance validation reveals systematic limitations that may impact practical deployment reliability. The identified failure modes, particularly for poorly maintained vineyards and seasonal imagery variations, represent significant operational constraints that require consideration in deployment planning. Agricultural operators utilizing the system must account for potential accuracy degradation in challenging conditions and implement appropriate quality assurance procedures to ensure reliable results.

The geometric accuracy assessment confirms that the system achieves performance levels compatible with the analytical and operational functions required by Software-as-a-Service agricultural platforms. The Hausdorff distance measurements indicate that spatial errors remain within acceptable tolerances for supporting digital management tool functionalities while avoiding the prohibitive costs associated with professional surveying services.

5.3.2 Deployment Considerations and Integration Challenges

Some deployment challenges require consideration for successful real-world implementation. The dependency on high-resolution aerial imagery necessitates reliable access to imagery services with consistent 30 cm per pixel resolution, which may not be universally available across all agricultural regions. Alternative imagery sources, including commercial satellite services such as the Airbus Pleiades constellation, provide comparable resolution but introduce cost considerations that may impact economic viability for smaller agricultural operations.

The requirement for vineyard boundary information, currently sourced from publicly available INVEKOS datasets, presents integration challenges in regions lacking comparable regulatory infrastructure. Successful deployment in international markets may require alternative approaches to vineyard delineation or manual boundary specification that could increase operational complexity and user adoption barriers, although feasible solutions using computer vision and remote sensing techniques exist, as demonstrated in chapter 2.

Quality assurance considerations represent critical deployment requirements given the identified performance limitations and failure modes. Agricultural operators must implement appropriate validation procedures to verify detection results, particularly for challenging vineyard configurations or suboptimal imagery conditions. The development of confidence scoring mechanisms and automated quality assessment tools represents essential future work for supporting reliable operational deployment.

5.3.3 Scalability Potential and Geographic Extension

The methodology demonstrates strong scalability potential for extension to different geographic areas, with the processing pipeline designed to accommodate diverse imagery sources and boundary datasets.

The economic model utilizing publicly available imagery sources and regulatory datasets aligns with the cost-effective deployment requirements characteristic of digital agriculture platforms. This approach enables broader access to detailed vineyard mapping capabilities without the prohibitive costs associated with professional surveying services, potentially democratizing access to advanced agricultural management tools for smaller operations.

However, scalability faces several practical constraints that require addressing for successful geographic extension. The manual annotation requirements for training data collection represent a significant bottleneck for expanding coverage to new regions with different vineyard characteristics or cultivation practices, that could be mitigate through cooperation of local agricultural surveying services, which likely already have the required data. The development of transfer learning approaches and active learning strategies also represents potential future improvements to reduce annotation requirements while maintaining detection performance across diverse agricultural contexts.

Chapter 6

Conclusion and Future Work

This chapter summarizes the research contributions and experimental findings presented throughout this thesis, evaluating the developed vineyard row detection system against the established objectives and identifying promising directions for future research and development.

6.1 Summary of Contributions

This research addresses the challenge of automated vineyard row detection from aerial imagery by developing and evaluating machine learning approaches that utilize publicly available geospatial data sources. The investigation demonstrates the feasibility of achieving practical detection accuracy using cost-effective imagery sources while establishing methodological foundations for broader agricultural remote sensing applications.

Comprehensive experimental evaluation establishes quantitative performance benchmarks for vineyard row detection using 30 cm per pixel aerial imagery across 79 manually annotated Austrian vineyards. The optimal model configuration achieves spatial accuracy of 0.99 ± 0.12 meters, which satisfies the established 1-2 meter precision requirement for practical agricultural management applications, while the Row Count Accuracy of 88% demonstrates sufficient reliability for supporting critical digital agriculture functionalities. The dataset created encompasses substantial variation in vineyard characteristics, with row densities ranging from 1 to 196 rows per site, with a total of 1422 rows annotated.

A reproducible methodology for transforming heterogeneous geospatial data sources into standardized training datasets was established, suitable for deep learning applications. The processing pipeline integrates publicly available INVEKOS vineyard boundaries with high-resolution basemap.at orthophoto imagery, demonstrating transferability to other regions utilizing comparable regulatory datasets and imagery sources. The target representation strategy utilizing 9-pixel line thickness with Gaussian blur ($\sigma = 3$) provides effective mitigation of class imbalance challenges inherent in linear structure detection tasks, while the patch-based extraction methodology using 256×256 pixel windows with 90% overlap generates sufficient training data density despite limited annotated vineyard availability.

The systematic evaluation of encoder-decoder architectures reveals that double convolution layers constitute the primary performance driver in U-Net-based implementations for this task, providing superior Row Count Accuracy compared to simpler SegNet architectures. The investigation identifies channel increase factor as the most influential architectural parameter, with systematic expansion of feature map dimensions providing consistent performance improvements across multiple evaluation metrics. The experimental analysis demonstrates practical limitations on architectural complexity, with layer depth exhibiting diminishing returns beyond four encoder-decoder stages and sophisticated enhancements such as skip connections and multi-scale convolution kernels failing to provide measurable improvements over optimally configured baseline architectures, when provided a limited dataset.

The loss function evaluation establishes that standard Dice Loss and Centerline Dice (clDice) achieve comparable performance when coupled with appropriate target representation strategies, confirming critical interplay between target line thickness and loss function effectiveness.

6.2 Future Research Directions

The experimental findings and identified limitations establish several promising research directions that could substantially enhance vineyard row detection capabilities and expand applicability to broader agricultural contexts.

The development of active learning approaches represents a critical research priority for addressing current dataset size limitations while reducing manual annotation requirements. Implementation of confidence-based sample selection strategies could enable automated identification of challenging vineyard configurations that provide maximum learning value when added to training datasets. Such approaches could leverage prediction uncertainty measures or ensemble model disagreement to prioritize annotation efforts on samples most likely to improve generalization performance. Geographic expansion to include diverse wine-producing regions across European, American, Australian, and South African vineyard sites would enable assessment of model transferability across different climatic conditions, cultivation practices, and varietal characteristics. Temporal dataset expansion incorporating imagery captured across multiple seasons and growth stages would address current seasonal sensitivity limitations while enabling year-round detection reliability.

The investigation of transfer learning approaches utilizing larger datasets from related domains could provide substantial performance improvements while addressing current data scarcity limitations. Pretraining on road detection datasets, medical vessel segmentation datasets, or general linear structure detection tasks could provide beneficial feature representations that transfer effectively to vineyard row detection applications. The development of self-supervised pretraining strategies specifically designed for agricultural remote sensing applications represents another promising direction, with techniques such as masked autoencoding, contrastive learning, or temporal consistency modeling leveraging unlabeled agricultural imagery to learn useful representations before fine-tuning on vineyard-specific detection tasks.

The development of specialized loss functions incorporating geometric constraints and topological properties specific to linear agricultural structures could provide substantial improvements over current Dice-based approaches. Advanced skeletonization-based loss functions that explicitly optimize for connectivity preservation, geometric accuracy, and row counting performance warrant investigation, alongside differentiable image processing techniques that incorporate domain-specific knowledge into loss function design. Multi-task learning approaches that simultaneously optimize for segmentation accuracy, row counting performance, and geometric properties could provide more robust optimization compared to single-objective approaches.

The adaptation of developed methodologies to other crop types and agricultural structures represents a natural extension of this research, with applications to orchard detection, field crop row identification, or greenhouse structure mapping demonstrating broader applicability. The development of unified frameworks capable of handling multiple crop types and agricultural structures within single models could provide substantial practical advantages for agricultural service providers. Investigation of scalability to regional or national-level agricultural monitoring could enable large-scale applications for policy support, yield forecasting, or environmental monitoring through cloud-based processing architectures and distributed computing approaches.

Bibliography

[1] Nabila Abraham and Naimul Mefraz Khan. 2019. A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. ISSN: 1945-8452. (April 2019), pp. 683–687. DOI: 10.1109/ISBI.2019.8759329. Retrieved 08/30/2025 from https://ieeexplore.ieee.org/document/8759329.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 12, (December 2017), 2481–2495. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2016.2644615. Retrieved 08/30/2025 from https://ieeexplore.ieee.org/document/7803544.

[3] T. Barros, P. Conde, G. Gonçalves, C. Premebida, M. Monteiro, C. S. S. Ferreira, and U. J. Nunes. 2022. Multispectral vineyard segmentation: A deep learning comparison study. *Computers and Electronics in Agriculture*, 195, (April 2022), 106782. ISSN: 0168-1699. DOI: 10.1016/j.compag.2022.106782. Retrieved 08/30/2025 from https://www.sciencedirect.com/science/article/pii/S0168169922000990.

[4] John Canny. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8, 6, 679–698. DOI: 10.1109/TPAMI.1986.4767851. https://ieeexplore.ieee.org/document/4767851.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40, 4, (April 2018), 834–848. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2017.2699184. Retrieved 08/30/2025 from https://ieeexplore.ieee.org/document/7913730.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. en. In *Computer Vision – ECCV 2018*. Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, (Eds.) Springer International Publishing, Cham, pp. 833–851. ISBN: 978-3-030-01234-2. DOI: 10.1007/978-3-030-01234-2_49.

[7] Pengfei Chen, Xiao Ma, Fangyong Wang, and Jing Li. 2021. A New Method for Crop Row Detection Using Unmanned Aerial Vehicle Images. *Remote Sensing*, 13, 17, 3526. DOI: 10.3390/rs13173526. https://www.mdpi.com/2072-4292/13/17/3526.

[8] James R. Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A. Zimmer, Julia A. Schnabel, and Andrew P. King. 2022. A Topological Loss Function for Deep-Learning Based Image Segmentation Using Persistent Homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 12, (December 2022), 8766–8778. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2020.3013679. Retrieved 08/30/2025 from https://ieeexplore.ieee.org/document/9186664.

[9] Carole Delenne, Sylvie Durrieu, Gilles Rabatel, and Michel Deshayes. 2010. From pixel to vine parcel: A complete methodology for vineyard delineation and characterization using remote-sensing data. *Computers and Electronics in Agriculture*, 70, 1, 78–83. DOI: 10.1016/j.compag.2009.09.012. https://www.sciencedirect.com/science/article/pii/S0168169909002038.

[10] Chengyang Du, Jie Zhuang, and Xinglu Huang. 2024. Deep learning technology in vascular image segmentation and disease diagnosis. *Journal of Intelligent Medicine*, 1, 1, 6–41. DOI: 10.1002/jim4.15. https://onlinelibrary.wiley.com/doi/abs/10.1002/jim4.15.

[11] Richard O. Duda and Peter E. Hart. 1972. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15, 1, 11–15. DOI: 10.1145/361237.361242. https://dl.acm.org/doi/10.1145/361237.361242.

[12] Adrian Galdran, André Anjos, José Dolz, Hadi Chakor, Hervé Lombaert, and Ismail Ben Ayed. 2022. State-of-the-art retinal vessel segmentation with minimalistic models. *Scientific Reports*, 12, 1, 6174. DOI: 10.1038/s41598-022-09675-y. https://www.nature.com/articles/s41598-022-09675-y.

[13] B. R. Z. GmbH. [n. d.] data.gv.at - offene Daten Österreichs – lesbar für Mensch und Maschine. en. (). Retrieved 07/25/2025 from https://www.data.gv.at.

[14] Paul V. C. Hough. 1962. Method and means for recognizing complex patterns. Patent. (1962). https://patents.google.com/patent/US3069654A/en.

[15] [n. d.] INVEKOS Feldstücke Österreich 2024-1. (). Retrieved 07/25/2025 from https://geoportal.inspire.gv.at/metadatensuche/c6dnRBxn/api/records/33078dd6-05e0-4263-a464-f35bd1607ecd.

[16] Zhexin Jiang, Hao Zhang, Yi Wang, and Seok-Bum Ko. 2018. Retinal blood vessel segmentation using fully convolutional network with transfer learning. *Computerized Medical Imaging and Graphics*, 68, 1–15. DOI: 10.1016/j.compmedimag.2018.04.005. https://www.sciencedirect.com/science/article/pii/S0895611118302313.

[17] Davood Karimi and Septimiu E. Salcudean. 2020. Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks. *IEEE Transactions on Medical Imaging*, 39, 2, (February 2020), 499–513. ISSN: 1558-254X. DOI: 10.1109/TMI.2019.2930068. Retrieved 08/30/2025 from https://ieeexplore.ieee.org/document/8767031.

[18] Mohd Jawed Khan and Pankaj Pratap Singh. 2023. Advanced road extraction using CNN-based U-Net model and satellite imagery. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 5, (September 2023), 100244. ISSN: 2772-6711. DOI: 10.1016/j.prime.2023.100244. Retrieved 07/24/2025 from https://www.sciencedirect.com/science/article/pii/S2772671123001390.

[19] M. Kise, Q. Zhang, and F. Rovira Más. 2005. A Stereovision-based Crop Row Detection Method for Tractor-automated Guidance. *Biosystems Engineering*, 90, 4, 357–367. DOI: 10.1016/j.biosystemseng.2004.12.008. https://www.sciencedirect.com/science/article/pii/S1537511004002260.

[20] 2025. Lanczos resampling. en. Page Version ID: 1301103050. (July 2025). Retrieved 07/25/2025 from https://en.wikipedia.org/w/index.php?title=Lanczos_resampling&oldid=1301103050.

[21] Bo Li, Cheng Chen, Shiwen Dong, and Junfeng Qiao. 2021. Transmission line detection in aerial images: An instance segmentation approach based on multitask neural networks. *Signal Processing: Image Communication*, 96, (August 2021), 116278. ISSN: 0923-5965. DOI: 10.1016/j.image.2021.116278. Retrieved 07/24/2025 from https://www.sciencedirect.com/science/article/pii/S092359652100117X.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42, 2, (February 2020), 318–327. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2858826. Retrieved 08/30/2025 from https://ieeexplore.ieee.org/document/8417976.

[23] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L. Martel. 2021. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71, 102035. DOI: 10.1016/j.media.2021.102035. https://www.sciencedirect.com/science/article/pii/S1361841521000815.

[24] Arpan Mahara, Md Rezaul Karim Khan, Liangdong Deng, Naphtali Rishe, Wenjia Wang, and Seyed Masoud Sadjadi. 2025. Automated Road Extraction from Satellite Imagery Integrating Dense Depthwise Dilated Separable Spatial Pyramid Pooling with DeepLabV3+. *Applied Sciences*, 15, 3, 1027. DOI: 10.3390/app15031027. https://www.mdpi.com/2076-3417/15/3/1027.

[25] [n. d.] ORTHOFOTO – basemap.at. (). Retrieved 07/25/2025 from https://basemap.at/orthofoto/.

[26] Diego Inácio Patrício and Rafael Rieder. 2018. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and Electronics in Agriculture*, 153, 69–81. DOI: 10.1016/j.compag.2018.08.001. https://www.sciencedirect.com/science/article/pii/S0168169918305829.

[27] [n. d.] Pléiades | Very High-Resolution (50cm) Satellites Imagery. en. (). Retrieved 07/29/2025 from https://space-solutions.airbus.com/imagery/our-optical-and-radar-satellite-imagery/pleiades/.

[28] J. Romeo, G. Pajares, M. Montalvo, J. M. Guerrero, M. Guijarro, and A. Ribeiro. 2012. Crop row detection in maize fields inspired on the human visual perception. *TheScientificWorldJournal*, 2012, 484390. DOI: 10.1100/2012/484390.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. en. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, (Eds.) Springer International Publishing, Cham, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

[30] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. 2017. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. en. In *Machine Learning in Medical Imaging*. Qian Wang, Yinghuan Shi, Heung-Il Suk, and Kenji Suzuki, (Eds.) Springer International Publishing, Cham, pp. 379–387. ISBN: 978-3-319-67389-9. DOI: 10.1007/978-3-319-67389-9_44.

[31] Luís Santos, Filipe Neves Dos Santos, Vítor Filipe, and Pranjali Shinde. 2019. Vineyard Segmentation from Satellite Imagery Using Machine Learning. In pp. 109–120.

[32] [n. d.] shapely.MultiLineString — Shapely 2.1.1 documentation. (). Retrieved 07/25/2025 from https://shapely.readthedocs.io/en/stable/reference/shapely.MultiLineString.html.

[33] Pengcheng Shi, Jiesi Hu, Yanwu Yang, Zilve Gao, Wei Liu, and Ting Ma. 2024. Centerline Boundary Dice Loss for Vascular Segmentation. en. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, (Eds.) Springer Nature Switzerland, Cham, pp. 46–56. ISBN: 978-3-031-72111-3. DOI: 10.1007/978-3-031-72111-3_5.

[34] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. 2021. clDice - A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. en. In pp. 16560–16569. Retrieved 08/30/2025 from https://openaccess.thecvf.com/content/CVPR2021/html/Shit_clDice_-_A_Novel_Topology-Preserving_Loss_Function_for_Tubular_Structure_CVPR_2021_paper.html.

[35] [n. d.] Spatial without Compromise · QGIS Web Site. en. (). Retrieved 07/25/2025 from https://qgis.org/.

[36] Saeid Asgari Taghanaki, Yefeng Zheng, S. Kevin Zhou, Bogdan Georgescu, Puneet Sharma, Daguang Xu, Dorin Comaniciu, and Ghassan Hamarneh. 2019. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75, (July 2019), 24–33. ISSN: 0895-6111. DOI: 10.1016/j.compmedimag.2019.04.005. Retrieved 08/30/2025 from https://www.sciencedirect.com/science/article/pii/S0895611118305688.

[37] Lev Teplyakov, Kirill Kaymakov, Evgeny Shvets, and Dmitry Nikolaev. 2021. Line detection via a lightweight CNN with a Hough layer. In *Thirteenth International Conference on Machine Vision*. Volume 11605. SPIE, (January 2021), pp. 376–385. DOI: 10.1117/12.2587167. Retrieved 08/30/2025 from https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11605/116051B/Line-detection-via-a-lightweight-CNN-with-a-Hough-layer/10.1117/12.2587167.full.

[38] Ivan Vidović, Robert Cupec, and Željko Hocenski. 2016. Crop row detection by global energy minimization. *Pattern Recognition*, 55, 68–86. DOI: 10.1016/j.patcog.2016.01.013. https://www.sciencedirect.com/science/article/pii/S0168169916000170.

[39] 2020. Web Map Tile Service. de. Page Version ID: 205162037. (November 2020). Retrieved 07/25/2025 from https://de.wikipedia.org/w/index.php?title=Web_Map_Tile_Service&oldid=205162037.

[40] Michael Yeung, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo. 2022. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95, 102026. DOI: 10.1016/j.compmedimag.2021.102026. https://www.sciencedirect.com/science/article/pii/S0895611121001750.

[41] Wei Yuan and Wenbo Xu. 2022. GapLoss: A Loss Function for Semantic Segmentation of Roads in Remote Sensing Images. *Remote Sensing*, 14, 10, 2422. DOI: 10.3390/rs14102422. https://www.mdpi.com/2072-4292/14/10/2422.

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. (July 2017), pp. 6230–6239. DOI: 10.1109/CVPR.2017.660. Retrieved 08/30/2025 from https://ieeexplore.ieee.org/document/8100143.

[43] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. en. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, (Eds.) Springer International Publishing, Cham, pp. 3–11. ISBN: 978-3-030-00889-5. DOI: 10.1007/978-3-030-00889-5_1.