Artificial Intelligence Review (2024) 57:149
https://doi.org/10.1007/s10462-024-10775-6

Deep learning implementation of image segmentation in agricultural applications: a comprehensive review

Lian Lei[1] · Qiliang Yang[2] · Ling Yang[1,3] · Tao Shen[1,3] · Ruoxi Wang[2] · Chengbiao Fu[1]

Accepted: 24 April 2024 / Published online: 22 May 2024

Abstract

Image segmentation is a crucial task in computer vision, which divides a digital image into multiple segments and objects. In agriculture, image segmentation is extensively used for crop and soil monitoring, predicting the best times to sow, fertilize, and harvest, estimating crop yield, and detecting plant diseases. However, image segmentation faces difficulties in agriculture, such as the challenges of disease staging recognition, labeling inconsistency, and changes in plant morphology with the environment. Consequently, we have conducted a comprehensive review of image segmentation techniques based on deep learning, exploring the development and prospects of image segmentation in agriculture. Deep learning-based image segmentation solutions widely used in agriculture are categorized into eight main groups: encoder-decoder structures, multi-scale and pyramid-based methods, dilated convolutional networks, visual attention models, generative adversarial networks, graph neural networks, instance segmentation networks, and transformer-based models. In addition, the applications of image segmentation methods in agriculture are presented, such as plant disease detection, weed identification, crop growth monitoring, crop yield estimation, and counting. Furthermore, a collection of publicly available plant image segmentation datasets has been reviewed, and the evaluation and comparison of performance for image segmentation algorithms have been conducted on benchmark datasets. Finally, there is a discussion of the challenges and future prospects of image segmentation in agriculture.

Keywords Image segmentation · Deep learning · Agricultural image processing · Plant diseases identification · Crop growth monitoring · Plant image datasets · Crop yield estimation and counting · Weed identification

1 Introduction

Image segmentation is a fundamental task in computer vision, which divides images (or video frames) into multiple segments and objects (Minaee et al. 2021). It plays a pivotal role in many crucial applications, such as scene understanding, image analysis, robotic perception, disease diagnosis (Pradhan et al. 2023), medical images (Kaur et al. 2021), video monitoring, augmented reality, and image compression (Yu et al. 2023). Image segmentation task contains three sub-tasks: semantic segmentation, instance segmentation, and

Extended author information available on the last page of the article

panoptic segmentation. Semantic segmentation, also called scene labeling, refers to the process of assigning a semantic label (e.g. car, people, and road) to each pixel of an image (Yu et al. 2018a, b). Instance segmentation is designed to identify and segment pixels that belong to each object instance (Gu et al. 2022). Going further, panoptic segmentation includes semantic segmentation and instance segmentation (Chuang et al. 2023). Recently, with the rapid development of modern agriculture based on information technology, image segmentation is widely used to monitor crop and soil health, predict the best time to sow, fertilize and harvest, estimate crop yields, and detect plant diseases (Elbasi et al. 2023; Wang et al. 2022a, b). However, the practical application of image segmentation in agriculture is difficult due to the complexity of the agricultural environment. Disease stage identification and labeling inconsistency are the two main challenges of image segmentation applied to plant diseases. Early, mid, and late stage diseases may appear similar in images, which makes it difficult to utilize image data for disease stage identification (Yao et al. 2022). Simultaneously, variable criteria for determining disease type and severity may result in inconsistent labeling of training datasets. Such inconsistencies can adversely affect model training and evaluation, consequently diminishing prediction accuracy (Chen et al. 2021a, b). In addition, leaf and root system environments often undergo fluctuations, leading to corresponding changes in morphological characteristics such as texture, size, and shape (Kang et al. 2021; Yan et al. 2023). In this study, the Web of Science database was used to search for strings designed to capture a wide range of articles with plant image segmentation. The following search strings are used to search for articles on database: ("Crop" OR "Plant") AND ("Image segmentation" OR "Semantic segmentation" OR "Instance segmentation") AND ("Deep learning"). Time limited to the past five years. We searched 2205 relevant articles and selected the top 1500. After deleting irrelevant articles such as machine learning and system design, 300 highly relevant articles were retained.

In early agriculture image segmentation, numerous traditional methods were used to address the above problems. Such methods include thresholding (Otsu 1979), edge-based segmentation (Rosenfeld 1981), region growing (Ikonomatakis et al. 1997), k-means clustering (Dhanachandra et al. 2015), watershed algorithms (Longzhe and Enchen 2011), and graph-based methods (Boykov and Jolly 2001). However, the above methods have higher requirements for image quality, and the recognition result may be significantly degraded or even invalidated if the environmental conditions change during the image acquisition process. Therefore, the generality and robustness of those methods are unsatisfactory, and the accuracy in practical application is not guaranteed. In contrast, deep learning methods, which require less preprocessing and manual selection of potential features, have improved both accuracy and robustness. To address the segmentation issues arising from diverse growth stages and overlapping plant objects, an Encoder-Decoder deep network was utilized, taking 14 various vegetation indices as input for weed/crop/background segmentation and achieving the highest Mean Intersection over Union (mIoU) value of 88.91% (Wang et al. 2020a, b). To mitigate the effects of soil disturbance and minor color differences on plant root segmentation, an attention mechanism was incorporated into the DeepLabv3 + semantic segmentation model. By training the model on a mature cotton root dataset, the model scored 98.75% Intersection over Union (IoU) (Kang et al. 2021). In addition to CNN architectures, graph convolutional networks also perform well in image segmentation tasks. Pei et al. (2023) proposed a multiscale global graph convolutional neural network (MSG-GCN) by embedding a multi-scale graph convolutional neural block into the last layer of the U-Net + + encoder module. The MSG-GCN outperforms U-Net and U-Net + + on the University of Tokyo Chiba Forest aerial dataset. In particular, advanced Transformers and generative adversarial networks have been widely used in plant disease detection (Douarre etl a. 2019; Wu et al. 2022a, b, c) and weed identification (Espejo-Garcia et al. 2021).

The recent surge in image segmentation development underscores its growing significance in agriculture. However, a comprehensive understanding of image segmentation in agriculture requires a thorough examination of existing literature and related studies. To this end, numerous reviews have been conducted with varying focus areas. Yu et al. (2018a, b) conducted a comprehensive review of publicly available scene annotation datasets and semantic segmentation methods based on hand-crafted features, learned features, and weakly supervised learning. Hu et al. (2018) reviewed common RGB-D datasets used for semantic segmentation, conventional machine learning approaches, and deep learning-based image segmentation technologies relevant to RGB-D segmentation. Guo et al. (2018) provided an overview of deep learning-based image semantic segmentation, categorizing it into region-based, fully convolutional network based, and weakly supervised segmentation. Asgari Taghanaki et al. (2021) organized medical and non-medical image segmentation techniques into six different modalities: deep architecture-based, data synthesis-based, loss function-based, ranking models, weak supervision, and multi-tasking, while elucidating the limitations associated with current methods and suggesting prospective research directions in semantic image segmentation. Minaee et al. (2021) offered a comprehensive review of deep learning-enabled image segmentation methods, summarizing seminal work in both semantic and instance segmentation. Additionally, they scrutinized widely utilized datasets, and performance comparisons, and deliberated on the progression and challenges faced by image segmentation technology. Gu et al. (2022) synthesized existing fully supervised, weakly supervised, and semi-supervised instance segmentation techniques, dividing strongly supervised approaches into three subcategories based on the number of stages, and delineating datasets and metrics relevant to instance segmentation.

Although numerous studies have comprehensively reviewed various methods for image segmentation (see Table 1), only four articles focus specifically on plant image segmentation. In plant image segmentation, Hamuda et al. (2016) provided a review of segmentation methods for plants in the field, including color index-based segmentation, threshold-based segmentation, and learning-based segmentation. Maheswari et al. (2021) delve into deep learning semantic segmentation for smart orchard yield estimation. This paper also discusses the challenging issues that arise in the process of intelligent fruit yield estimation, such as sampling, collection, annotation and data enhancement, fruit detection, and counting. Buckner et al. (2021) showed how the segmentation and classification methods differ due to the diversity of physical features discovered at these different scales. Luo et al. (2023) reviewed two traditional segmentation methods and three deep learning methods including U-Net, SegNet and Deeplab. In addition to the above review, numerous studies have focused on a wide range of applications of deep learning in agriculture, e.g., Hasan et al.(2020), on recent deep learning advances in plant disease detection. These studies provide a broad overview, of which image segmentation is a small part. In summary, current research is mainly focused on orchard and plant organ etc. and there is no comprehensive review in agriculture. The image segmentation methods mainly focus on traditional image segmentation methods and early methods based on deep learning, with no mention of the latest models such as Transformer and GANs. Also, there are no related articles summarizing public plant image segmentation datasets. Therefore, we comprehensively summarize the commonly used network architectures in plant image segmentation, as well as other studies in these architectures. The public plant image segmentation datasets are summarized, and four applications, plant disease identification, weed identification, plant growth

Table 1 A summary of the review studies on image segmentation

Time	Reference	Advantages	Disadvantages
2018	Methods and datasets on semantic segmentation: A review (Yu et al. 2018a, b)	Review public datasets comparing manual features, learning-based and weakly supervised learning methods, and scene labeling	The applications and datasets of these studies are mainly focused on autonomous driving, medicine, etc., and there is no comprehensive review for the agricultural
2018	RGB-D Semantic Segmentation: A Review (Hu et al. 2018)	Summarizes the commonly used RGB-D datasets for semantic segmentation, traditional RGB-D segmentation methods and DL-based methods	
2018	A review of semantic segmentation using deep neural networks (Guo et al. 2018)	An overview of DL-based semantic segmentation is presented, divided into regions, full convolutional networks and weakly supervised segmentation	
2020	Deep Semantic Segmentation of Natural and Medical images: A Review (Asgari Taghanaki et al. 2021)	Natural and medical image segmentation techniques are reviewed. Six methods are summarized and potential research directions for image segmentation are suggested	
2021	Image Segmentation Using Deep Learning: A Survey (Minaee et al. 2021)	DL-based image segmentation techniques are reviewed in detail and their network structures are classified. Common datasets and challenges for image segmentation are summarized	
2022	A review of 2D instance segmentation based on deep neural networks (Gu et al. 2022)	The strong supervision method for instance partitioning is divided into one-step, two-step, and multi-step methods. Examples of partitioning datasets and metrics are presented, as well as future challenges	
2016	A survey of image processing techniques for plant extraction and segmentation in the field	Attention is given to the color index based approach and the segmentation performance of the color index based approach is discussed in detail	Recent advances in agriculture, such as the application of deep learning algorithms, have been less reviewed
2021	High-throughput image segmentation and machine learning approaches in the plant sciences across multiple scales (Buckner et al. (2021)	Explored the latest technologies in plant image segmentation and machine learning at the agricultural, organ, and cellular scales of plants	Lack of comprehensive summary of agricultural applications
2021	Intelligent Fruit Yield Estimation for Orchards Using Deep Learning Based Semantic Segmentation Techniques (Maheswari et al. 2021)	Methods for fruit yield estimation using deep learning-based semantic segmentation are explored, along with associated challenges	The main focus is on yield estimation and counting
2023	Semantic segmentation of agricultural images: A survey (Luo et al. 2023)	Traditional and DL methods for semantic segmentation of agricultural images are presented and the obstacles to training and evaluating small datasets are summarized	Only summarizing relevant algorithms, lacking a summary of agricultural applications

monitoring, and yield estimation, are reviewed in this paper. The specific contributions are as follows:

- Based on network structure, we categorized the widely used agricultural segmentation models into eight different categories.
- A number of common publicly available datasets for plant image segmentation are reviewed.
- We evaluate and compare the performance of image segmentation algorithms widely used on benchmark datasets.
- The applications of image segmentation in agriculture are presented including disease detection, weed identification, crop growth monitoring, and yield estimation.
- Finally, for future directions and challenges of image segmentation in agriculture are outlined.

2 Deep learning network architecture

With the rapid development in Computer Vision and Deep Learning theories and models, the process of image segmentation and classification becomes easier when compared with the traditional approaches (Solanki et al. 2023a, b). This chapter reviews the widely used theological neural network architecture in the field of computer vision, including convolution, generation of antagonistic, graph, and Transformer networks. The abbreviations for the full text are shown in Table 2.

2.1 Convolutional neural networks (CNNs)

CNNs are a class of artificial neural networks that are widely used in a variety of computer vision tasks. They have achieved substantial success in areas such as image recognition, object detection, and image segmentation, solidifying their position as one of the most important components within the field of deep learning. Compared to machine learning, CNN also improved in terms of speed and accuracy (Solanki et al. 2023a, b). The CNN structure is shown in (Fig. 1). The core of CNNs includes a convolution layer, pooling layer, and fully connected layer Convolutional layers are responsible for extracting features via convolution operations, whereas pooling layers endeavor to reduce the dimensionality of feature maps whilst bolstering their robustness. Nonlinear layers serve to introduce nonlinearity, thereby endowing neural networks with the ability to learn increasingly complex patterns and relationships. The merits of CNNs include capitalizing on spatial correlations between adjacent pixels, parameter sharing, invariance, large data processing, and adaptive feature learning. Some notable CNN architectures include VGGNet (Simonyan and Zisserman 2014), GoogLeNet (Szegedy et al. 2015), and ResNet (He et al. 2016).

2.2 Generative adversarial networks (GANs)

GAN (Goodfellow et al. 2020) (Fig. 2) framework consists of two interconnected components: a generator and a discriminator. In traditional GAN architectures, the generator network G is responsible for learning the mapping from a latent noise vector Z to the target distribution y, approximating "real" samples in the process. Concurrently, the discriminator network D determines whether the artificially generated sample goal of the properties of genuine samples. Notable variants of GANs include, Conditional GAN (CGAN) (Mirza and Osindero 2014), Deep Convolutional GAN (DCGAN) (Radford et al. 2015), CycleGAN (Zhu et al. 2017), Wasserstein-GAN (WGAN) (Arjovsky et al. 2017), DualGAN (Yi et al. 2018), and Semi-Supervised GAN (SGAN) (Trinh and O'Brien 2020), which have been extensively applied and studied within the domain of computer vision.

2.3 Graph neural networks (GNNs)

Deep learning has demonstrated remarkable success in a variety of domains. However, researchers have recognized its limitations in addressing and solving all situations and problems. In particular, when processing graph-structured data within non-Euclidean spaces, there exists an inherent challenge in leveraging both structural and semantic information effectively. GNNs (Scarselli et al. 2009) have emerged as a prominent research focus in the realm of graph data structures due to their capacity to concurrently learn topological information and preserve structural attributes (Fig. 3). Among various GNN models, Graph Convolutional Networks (GCN) (Kipf and Welling 2017) constitute a distinctive convolutional neural network architecture that can directly operate on graphs while exploiting their inherent structural information. In addition, popular GNN techniques encompass advanced methodologies such as Graph Attention Networks (GAT) (Veličković et al. 2017) and Graph Generative Adversarial Networks (Graphical GAN) (Li et al. 2018), further enriching the landscape of graph-based learning algorithms.

2.4 Transformer

In 2017, Google introduced the groundbreaking Transformer model (Vaswani et al. 2017), which caused a sensation in the natural language processing field. In recent years, a number of innovative research papers have successfully applied the Transformer technology to cross-disciplinary computer vision tasks, as shown in Fig. 4, ushering in a new era in the visual domain. Dosovitskiy et al. (2020) introduced a model named ViT (Vision Transformer), which is a fully self-attention-based image classification approach. ViT's approach is to divide an image into patches of fixed size, perform linear transformations and position coding on each patch, and then feed the patches into the transform encoder to perform feature extraction and classification on the entire image. Compared to traditional CNNs, the ViT model fully relies on the self-attention mechanism to capture relevant information within images, offering higher interpretability and transferability.

3 Deep learning-based image segmentation models

Minaee et al. (2021) provided a comprehensive categorization of image segmentation algorithms into 11 classes, based on their network structure. This classification serves as the most detailed demarcation in existing literature. Within the context of agriculture, however, in the field of agriculture, recurrent neural network applications teach less and are gradually being replaced. Subsequently, recent trends suggest an inclination towards Transformer and GNN based image segmentation methods. In light of this, we've tailored our

Table 2 A summary of abbreviations

Abbreviation	Full Name	Abbreviation	Full Name
DL	Deep Learning	DCGAN	Deep Convolutional Generative Adversarial Network
CNN	Convolutional Neural Network	FCN	Fully Convolutional Network
GAN	Generative Adversarial Network	SDN	Stack Deconvolutional Network
SGAN	Semi-Supervised Generative Adversarial Network	CGAN	Conditional Generative Adversarial Network
GNN	Graph Neural Network	DPN	Deep Parsing Network
GCN	Graph Convolutional Networks	FPN	Feature Pyramid Network
VGGNet	Visual Geometry Group Network	PSPNet	Pyramid Scene Parsing Network
GAT	Graph Attention Networks	APCNet	Adaptive Pyramid Context Network
ViT	Vision Transformer	ACM	Adaptive Context Module
UPerNet	Unified Perceptual Parsing Network	CCN	Context Contrasted Network
DMNet	Dynamic Multiscale Filters Network	ASPP	Atrous Spatial Pyramid Pooling
CRF	Conditional Random Field	LS-DeconvNet	Locality-Sensitive Deconvolution Network
DUC-HDC	Dense Upsampling Convolution and Hybrid Dilated Convolution	DenseASPP	Densely Connected Atrous Spatial Pyramid Pooling
SC-loss	Semantic Context Loss	GI Unit	Graph Interaction Unit
EVDR	Exploit Visual Dependency Relation	GMNet	Graph Matching Network
Graph-BAS3Net	Boundary-Aware Semi-Supervised Segmentation Network	RoI	Region of Interest
SDS	Simultaneous Detection and Segmentation	RPN	Region Proposal Network
SVM	Support Vector Machine	MNC	Multi-task Network Cascades
NMS	overlapping regions are further refined with Non-Maximum Suppression	PANet	Path Aggregation Network
R-CNN	Region-based Convolutional Neural Network	MS R-CNN	Mask Scoring Region-based Convolutional Neural Network
FCIS	Fully Convolutional Instance-Aware Semantic Segmentation	CXR	chest X-ray
PSANet	Point-wise Spatial Attention Network	DFANet	Deep Feature Aggregation Network
DANet	Dual Attention Network	BiSeNet	real-time semantic segmentation Bilateral Segmentation Network
CA	Recurrent Attention	ANNNet	Asymmetric Non-Local Neural Networks
CCA	Criss-Cross Attention	SETR	Segmentation Transformer
SANet	Squeeze-and-Attention Network	MLP	Multi-Layer Perceptron

Table 2 (continued)

Abbreviation	Full Name	Abbreviation	Full Name
SENet	Squeeze-Excitation Network	VisTR	Video Instance Segmentation Transformer
OCNet	Object Context-aware Network	LSC	Leaf Segmentation Challenge
EMANet	Expectation–Maximization Attention Network	MSU-PID	Multi-modality Plant Imagery Database
mIoU	Mean intersection over Union	FrCNnet	Fully resolution Convolutional Network
DS-DETR	Disease Segmentation Detection Transformer	PDCD	Disease Classification Dataset
SMCA	Spatial Modulation Common Attention	TDSD	Tomato Leaf Disease Segmentation Dataset
RLDCP	Rice Leaf Disease Copy Paste	CWFID	Crop Weed Field Image Dataset
UAV	Unmanned aerial vehicle	MLC	Maximum Likelihood Classification
ST	Segmentation Time	AdaIN	Adaptive Instance Normalization
CROP	Central Round Object Painting	SSL	Semi-supervised learning
DDPM	Denoising Diffusion Probability Model	ReTree	Retina Tree
SCTNet	Single-Branch CNN with Transformer	FPS	Frame Per Second

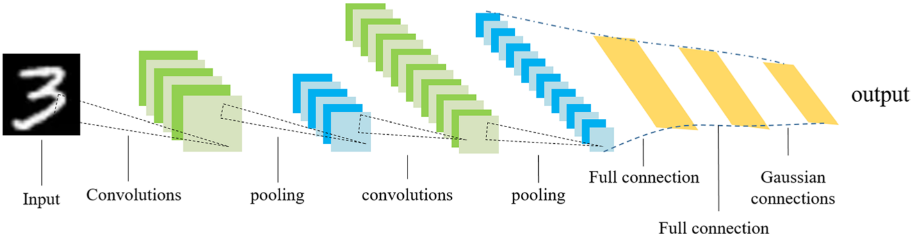

Fig. 1 The architecture of CNNs (Lecun et al. 1998)

classifications to fit the agricultural domain, thus presenting the image segmentation algorithms based on network structures into 8 distinct classes.

3.1 Encoder-decoder network

The encoder-decoder network, a prominent deep learning architecture, has demonstrated success in various image processing tasks, including image classification, object detection, and semantic segmentation. Semantic segmentation requires assigning semantic labels to individual pixels in an image while carefully balancing granular pixel-level details and global contextual information. The architecture consists of two essential components: firstly, the encoder extracts salient features from the input image, condensing them into a compact, low-dimensional representation; secondly, the decoder expands this representation to the original dimensions, allocating meaningful labels to each pixel.

Shelhamer et al. (2017) proposed a streamlined, yet efficacious Encoder-Decoder Network architecture (Fig. 5), denoted as Fully Convolutional Network (FCN). Harnessing the VGG-16 network for feature extraction, this innovative design supplants the final fully connected layer with a convolutional layer to preserve crucial spatial information. The decoder upsamples low-resolution feature maps from the encoder output to the original image resolution and then fuses the feature maps of the final layer of the model with those of the previous layers via skip connections (Fig. 6). This sophisticated fusion facilitates the integration of semantic information derived from deeper, coarser layers with appearance information gleaned from shallower, finer layers, ultimately yielding highly precise and intricate segmentations.

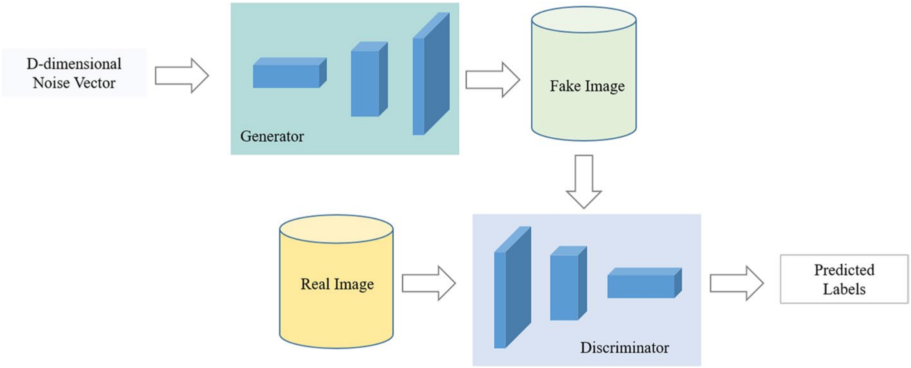

Fig. 2 The architecture of a GAN (Goodfellow et al. 2020)

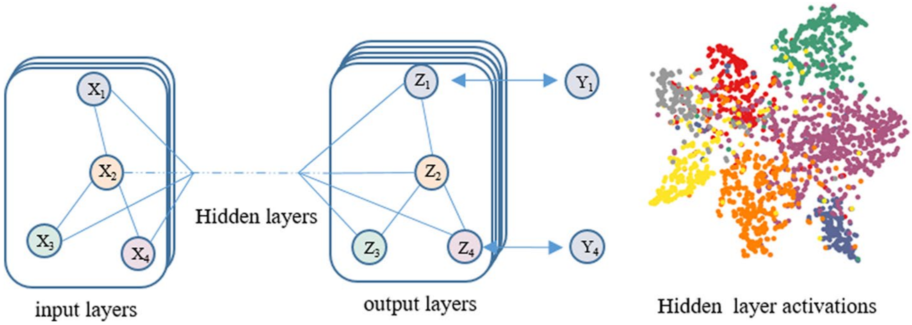

Fig. 3 The architecture of a GCN layer (Kipf and Welling 2017)

Ronneberger et al. (2015) introduced the highly efficient U-Net for segmenting biological microscopy images, building upon the FCN (Fig. 7). The U-Net embodies the Encoder-Decoder paradigm, with the encoder employing pooling layers for systematic down-sampling and the decoder leveraging deconvolution for iterative up-sampling. This approach enables the progressive restoration of spatial information and edge details from the original input image, ultimately mapping low-resolution feature maps to pixel-level segmentation outcomes. To mitigate information loss during the encoding phase's down-sampling, U-Net incorporates skip connections to amalgamate corresponding feature maps from both encoder and decoder segments. This sophisticated fusion allows the decoder to access enhanced high-resolution information during up-sampling, refining the restoration

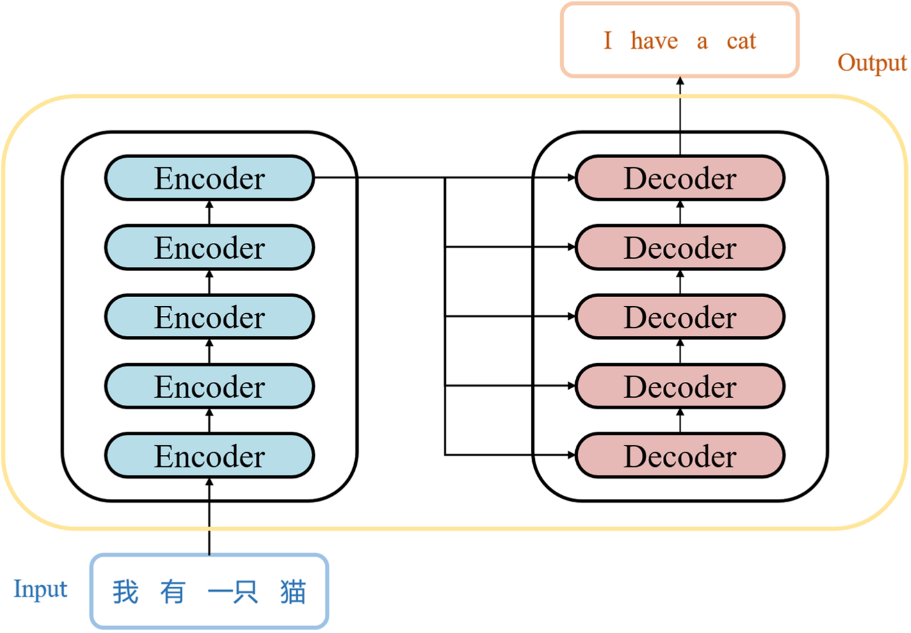

Fig. 4 The architecture of Transformer (Vaswani et al. 2017)

of intricate details in the original image and bolstering segmentation accuracy. Exhibiting prowess in pixel-level segmentation and adeptly handling small datasets, the symmetric U-shaped architecture ensures comprehensive feature integration between the encoder and decoder. To effectively extrapolate from a limited annotated image repository, U-Net harnesses data augmentation as its training strategy cornerstone.

Although U-Net demonstrates impressive performance, network redundancy still poses a substantial challenge. Owing to the patch-based per-pixel training approach, high similarity among adjacent patches results in redundancy and sluggish training processes. Furthermore, information loss remains pervasive within CNN architectures, irrespective of the number of optimizations applied during the down-sampling phase. U-Net + + (Zhou et al. 2018) is an improvement of the original U-Net model, which mainly introduces dense connection and multi-scale feature fusion mechanisms. Dense connectivity enables each decoder layer to directly access feature maps from all corresponding encoder layers to improve feature propagation and information flow efficiency; The multi-scale feature fusion uses the feature maps of different levels and fuses them into the decoder through concatenation and upsampling operations to improve the model's perception of the target and segmentation accuracy. Drawing inspiration from residual and dense connections, Res-U-Net (Xiao et al. 2018a, b) and Dense-U-Net (Guan et al. 2020) individually replace each submodule of U-Net with variants incorporating respective connection types. Despite its inherent limitations, U-Net continues to be the preeminent segmentation model within the medical domain (Liu et al. 2021a, b, c).

Badrinarayanan et al. (2017) proposed SegNet, a bespoke encoder-decoder fully convolutional architecture devised explicitly for image segmentation (Fig. 8). Analogous to deconvolution networks, SegNet's core trainable segmentation engine encompasses an encoder network, topologically congruent with VGG16's 13 convolutional layers, succeeded by a corresponding decoder network and a per-pixel classification stratum. SegNet's key innovation resides in the decoder's strategy for upsampling low-resolution input feature maps, which specifically employs pooling indices computed during the respective encoder's max-pooling operation to execute nonlinear upsampling. By capitalizing on this upsampling technique to recuperate contour and positional information, SegNet adeptly

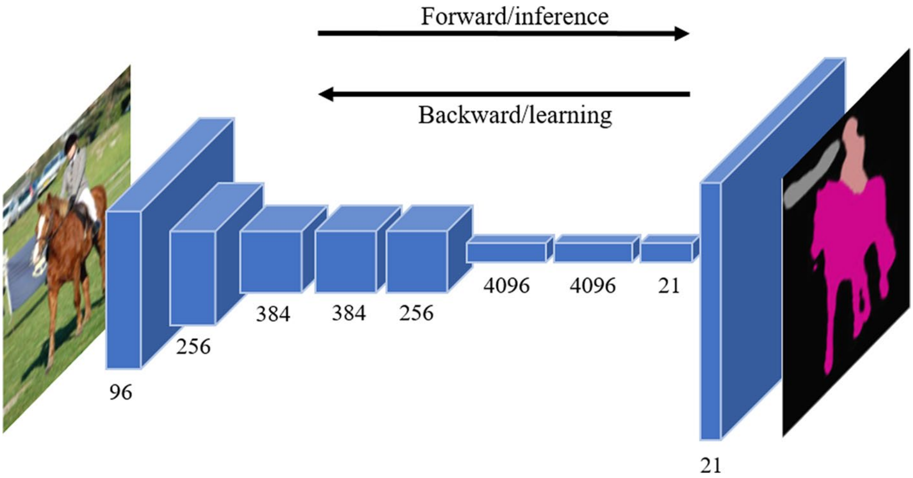

Fig. 5 The FCN learns to make pixel-accurate predictions (Shelhamer et al. 2017)

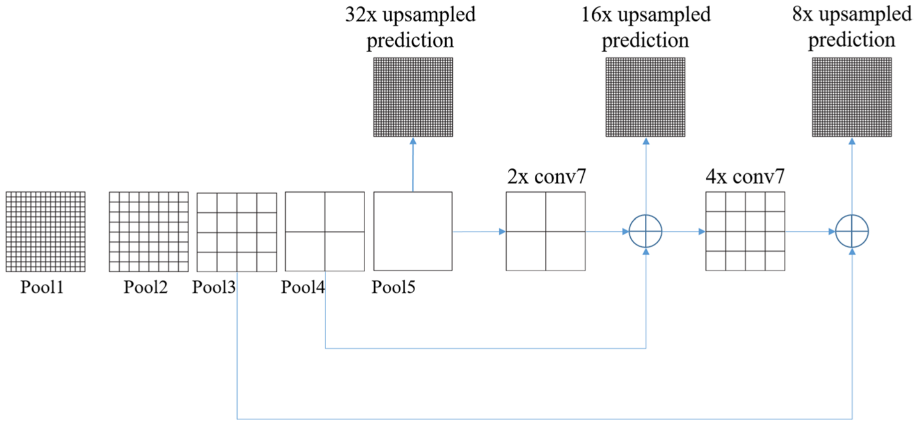

Fig. 6 Skip connections combine coarse and fine information. From (Shelhamer et al. 2017)

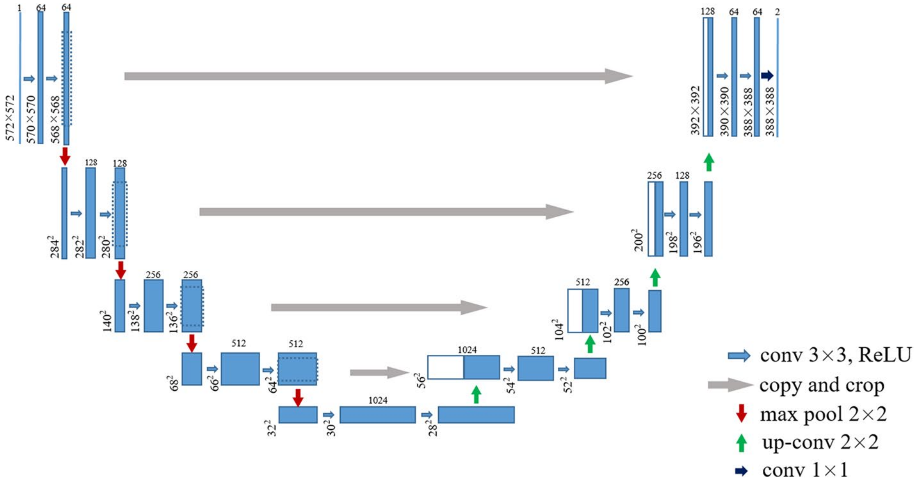

Fig. 7 The U-Net model (Ronneberger et al. 2015)

extracts and retains positional data of target edge features within images restores image dimensions and fine details, and achieves high-precision image segmentation. SegNet excels in regional segmentation, rendering it particularly suitable for segmenting punctiform and block-like objects.

Additional research adopting encoder-decoder frameworks for image segmentation comprises RefineNet (Lin et al. 2017a, b), premised on ResNet-inspired residual connections; global convolutional network (Peng et al. 2017), addressing the precision trade-off between localization and classification; lightweight network LinkNet (Chaurasia and Culurciello 2017); a skin lesion segmentation model implementing Jaccard distance as the loss function (Y. Yuan et al. 2017); Stack Deconvolutional Network (SDN) (Fu et al. 2019a, b); and Semantic Image Segmentation via Deep Parsing Network (DPN) (F. Yuan et al. 2019).

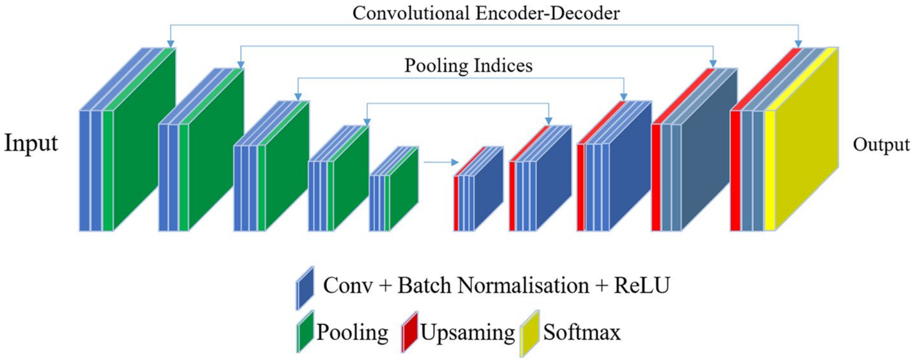
Convolutional Encoder-Decoder
Pooling Indices
Input
Output
Conv + Batch Normalisation + ReLU
Pooling
Upsaming
Softmax

Fig. 8 The SegNet model (Badrinarayanan et al. 2017)

3.2 Multiscale model

Multiscale models represent an advanced category of computer vision architectures that are capable of handling visual information across different levels of granularity. By facilitating feature extraction at multiple scales, these models excel at recognizing the contextual foundations embedded in given images. The emergence of multiscale models has been driven primarily by challenges associated with scale variance and intricate contextual relationships, which traditional computer vision paradigms frequently fail to address due to their limited ability to extract features at fixed scales.

An excellent instance of multiscale models is the Feature Pyramid Network (FPN), as proposed by Lin et al. (2017a, b). FPN was initially created for object detection applications but has now been effectively expanded for segmentation tasks. This architecture exploits the inherent multiscale pyramid hierarchy found in deep CNNs to construct a feature pyramid with minimal supplementary computational burden. FPNs incorporate bottom-up pathways, top-down pathways, and lateral connections to amalgamate low-resolution and high-resolution features. Following this, cascaded feature maps are processed through 3 × 3 convolutional layers to produce outputs for each stage. At each point in the top-down pathway, predictions relevant to object detection are generated. For image segmentation, a pair of multilayer perceptrons (MLPs) is utilized to engender masks.

Zhao et al. (2017) introduced the Pyramid Scene Parsing Network (PSPNet), an advanced multi-scale architecture that is adept at capturing global contextual representations in scenes (Fig. 9). The PSPNet uses a ResNet for feature extraction and enhances its capabilities with atrous convolutions to extract diverse patterns from input images. Feature maps are fed into a pyramid pooling module designed to capture patterns at different scales. Maps are integrated at four unique scales, corresponding to pyramid hierarchy levels, and processed through a 1 × 1 convolutional layer to reduce dimensionality. Outputs from pyramid levels are up-sampled and fused with the initial feature map, encapsulating both local and global contextual information. A convolutional layer then generates per-pixel predictions. PSPNet outperforms state-of-the-art models such as FCN, DeepLab-v2, DPN, and CRF-RNN on multiple datasets, demonstrating superior segmentation performance. However, handling occlusions between objects remains challenging, with edge segmentation being suboptimal in partially occluded areas. A key limitation of FCN-based models is their inability to effectively exploit category cues within global scenes, resulting

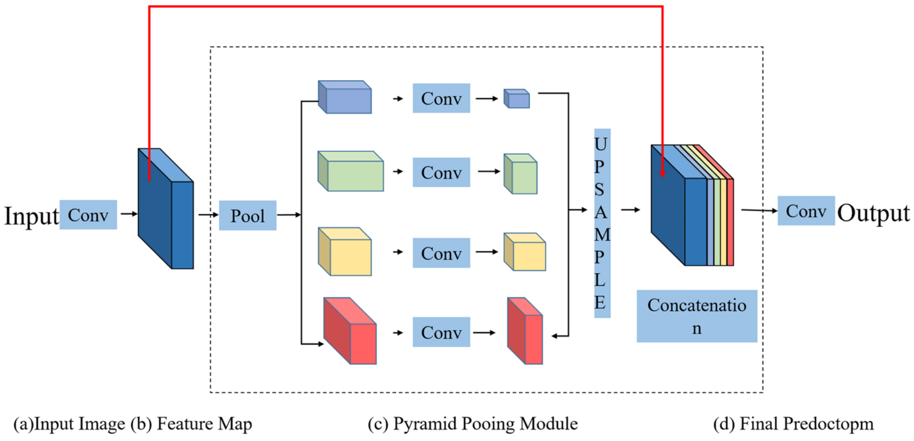

Fig. 9 The architecture of PSPN (Zhao et al. 2017)

in reduced segmentation accuracy and insufficient contextual integration. By implementing the PSPNet, the authors effectively aggregate context from various regions, providing the model with improved global context understanding. PSPNet uses different pool sizes to expand the receptive field, fostering a more comprehensive understanding of global contextual information. The pyramid pooling module gathers hierarchical information more efficiently than global pooling. Computationally, PSPNet imposes minimal overhead on the original atrous convolution FCN network, allowing simultaneous training of global pyramid pooling modules and local FCN features during end-to-end learning.

Ghiasi and Fowlkes (2016) described a multi-resolution reconstruction architecture based on Laplacian pyramids, which refines segmentation boundaries reconstructed from low-resolution maps using skip connections and multiplicative gating from high-resolution feature maps. In recent years, studies have demonstrated that the use of contextual features can significantly improve the performance of deep semantic segmentation networks. Contemporary semantics-based approaches differences notable variations in their methods for constructing semantic structures. He et al. (2019a, b) proposed the Adaptive Pyramid Context Network (APCNet) specifically for semantic segmentation tasks. The APCNet adaptively assembles multi-scale context representations using a series of well-designed Adaptive Context Modules (ACMs). Each ACM estimates local affinity coefficients for individual sub-regions under the guidance of global image information and subsequently computes context vectors based on these affinities. APCNet achieves state-of-the-art results on a variety of semantic segmentation benchmark datasets.

Additional models utilize multi-scale analysis for segmentation, encompassing the likes of Unified Perceptual Parsing Network (UPerNet) (Xiao et al. 2018a, b), Context Contrasted Network with gated multiscale aggregation (CCN) (Ding et al. 2018), Multi-Scale Context Intertwining (MSCI) (Lin et al. 2018), Dynamic Multiscale Filters Network (DMNet) (He et al. 2019a, b), Enhanced Feature Pyramid Network (EFPN) (Wang et al. 2021a, b), and Feature Pyramid Aggregation Network (FPAN) (Wu et al. 2022a, b, c).

3.3 Atrous convolutional models

Atrous convolution, alternatively known as dilated convolution, has a hyper-parameter referred to as the atrous rate. This parameter defines the distance between values as the convolution kernel processes the data. For example, a 3×3 kernel exhibiting an atrous rate of 2 has a receptive field the size of a 5×5 kernel while using only nine parameters. When the atrous rate is set to 3, the kernel achieves a receptive field equivalent in size to an 8×8 kernel. This allows the receptive field to be expanded without incurring additional computational costs. Moreover, by manipulating different atrous rates, a wide range of receptive fields can be obtained, effectively capturing multi-scale information.

Chen et al. (2014, 2016) proposed DeepLab-v1, an innovative approach that first employs atrous convolutions to mitigate information loss stemming from pooling operations, followed by the utilization of Conditional Random Fields (CRFs) to further improve segmentation accuracy In the subsequent iteration, DeepLab-v2 (Chen et al. 2018a, b, c, d), the more powerful and expressive ResNet-101 replaces VGG16 (Fig. 10). Within this refined version, Chen et al. skillfully exploit atrous convolutions and propose the Atrous

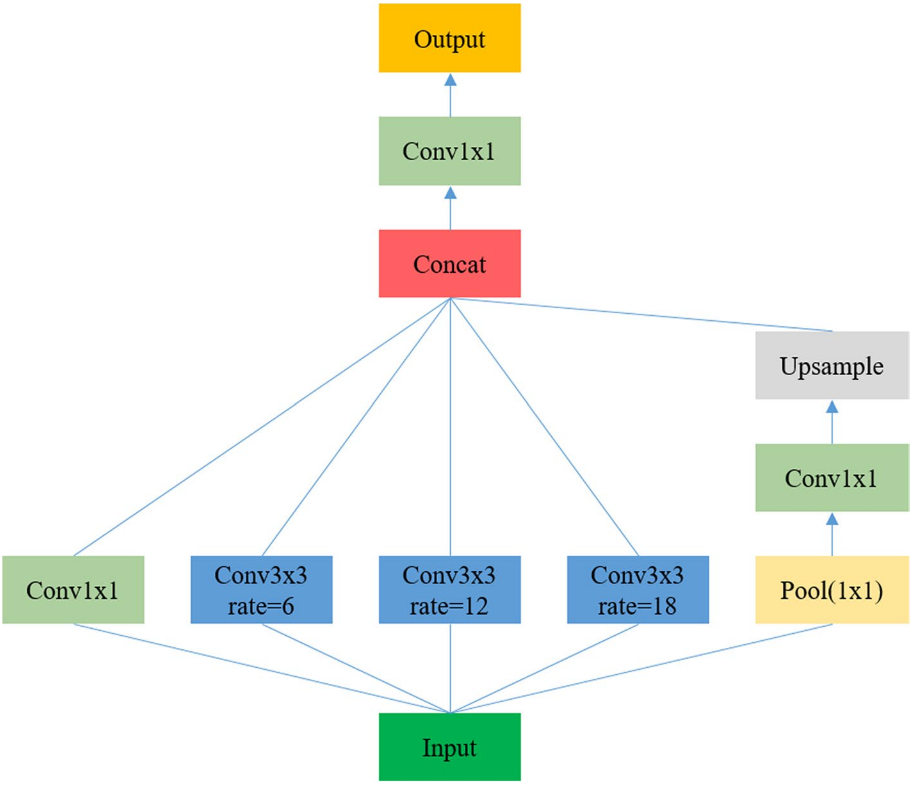

Fig. 10 Atrous Spatial Pyramid Pooling (ASPP). To classify the center pixel (orange), ASPP exploits multi-scale features by employing multiple parallel filters with different rates. From (L.-C. Chen et al. 2018a, b, c, d)

Spatial Pyramid Pooling (ASPP) module, as shown in the accompanying figure, while preserving the fully connected CRF components.

Chen et al. (2017) advanced DeepLab-v2 by introducing DeepLab-v3, a notable improvement that eliminates the use of fully connected CRFs. They developed a deeper network architecture by cascading, replicating the last block of the ResNet (i.e., block4) multiple times (i.e., block5-block7), and appending these duplicates in a cascading fashion to the network's backend. Each block encompasses three 3 × 3 convolutional layers, where all but the last block's final convolutional layer adopts a stride of 2. Moreover, they proposed a technique termed Multi-grid, which applies atrous convolutions at different rates in blocks 4 to 7. Within the ASPP module, batch normalization was incorporated, the atrous convolution with rate = 24 was replaced by a 1 × 1 convolution, and image-level features were integrated.

Subsequently, Chen et al. (2018a) presented the DeepLab-v3 +. In contrast to its predecessors, v3 + exhibits significant architectural changes, including the assimilation of a simple but effective decoding module, as shown in Fig. 11. The v3 + model uses the v3 network for encoding purposes, substituting the ResNet101 with the more profound Xception (Chollet 2017) architecture. In addition, depth-separable convolution is introduced in both the ASPP module and decoding model to reduce network parameters. Atrous separable convolutions replace the standard atrous convolutions within the framework.

Dilated convolution expands the comprehension capabilities of CNNs regarding input images, enhancing their performance in image processing tasks and garnering significant attention within the image segmentation domain. Apart from the DeepLab family, numerous related models have been developed, such as multiscale context aggregation

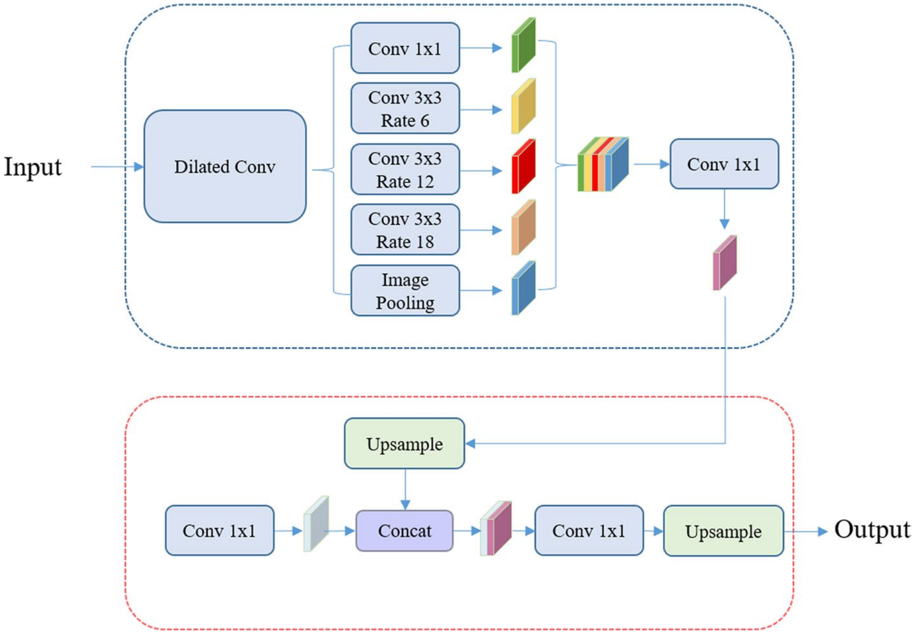

Fig. 11 The architecture of DeepLabv3 + (Chen et al. 2018b)

(Yu and Koltun 2016), Locality-Sensitive Deconvolution Networks (LS-DeconvNet) (Cheng et al. 2017), Dense Upsampling Convolution and Hybrid Dilated Convolution (DUC-HDC) (Wang et al. 2018), and densely connected Atrous Spatial Pyramid Pooling (DenseASPP) (Yang et al. 2018).

3.4 Graph convolutional neural network

Contrasting with traditional grid-structured CNNs, Graph Convolutional Networks (GCNs) can process data of arbitrary geometries, encompassing unordered point clouds, complex surfaces, and intricate polygons. GCNs facilitate local feature extraction for individual nodes within a graph, while simultaneously accounting for interrelations between a node and its neighboring counterparts. This capability enables GCNs to adeptly handle multiscale information in diverse contexts.

In the field of deep learning, advanced feature extraction often overlooks the significance of local positional information, which is crucial for semantic segmentation. To address this limitation, Lu et al. (2019) introduced a graph model initialized by an FCN, aptly named Graph-FCN (Fig. 12), explicitly designed for semantic segmentation tasks. The authors ingeniously constructed a graph network model using the intermediate feature layer derived from a semantic segmentation network, where each pixel location within the feature layer acts as a graph node connected to its neighboring nodes. The graph network model was introduced to the GCN, facilitating classification on individual nodes, thereby transforming the semantic segmentation challenge of classifying discrete pixels into a graph node classification task. This pioneering approach harnessed the prowess of Graph Convolutional Networks to tackle the graph node classification puzzle, marking a groundbreaking application of GCNs to image semantic segmentation.

Demonstrating prodigious potential in scene parsing, and contextual inference using image regions beyond local convolutions has attracted considerable attention. In this pursuit, Wu et al. (2020) incorporated linguistic knowledge to facilitate contextual reasoning within image regions, formulating a Graph Interaction Unit (GI Unit) and Semantic Context Loss (SC-loss). GI Units augment high-level semantic features in convolutional networks and adaptively learn semantic coherence for individual samples. Specifically, dataset-based linguistic knowledge is initially embedded within GI Units to encourage contextual reasoning on visual graphs. This is followed by the mapping of evolved visual graph representations onto each local representation to strengthen discriminative capabilities in scene parsing. SC-loss further refines GI Units, enhancing semantic representation on sample-based semantic graphs.

GCNs have made significant progress in graph-centric tasks, such as image segmentation, resulting in the creation of numerous high-performing models. Notable works include Spatial Pyramid Based Graph Reasoning (Li et al. 2020), Graph Matching Network (GMNet) (Michieli et al. 2020), Boundary-Aware Semi-Supervised Segmentation Network (Graph-BAS3Net) (Huang et al. 2021a, b, c), Boundary-aware Graph Convolution (BGC) (Hu et al. 2021a, b), Exploit Visual Dependency Relations (EVDR) (Liu et al. 2021a, b, c), and weakly supervised image semantic segmentation predicated on image-level class labels (Pan et al. 2021). Despite Graph Convolutional Networks showcasing formidable performance across segmentation domains, they may confront computational and storage constraints when processing large-scale images and necessitate copious training data to attain superior prediction results. Moreover, GCN's susceptibility to graph topology structure may precipitate suboptimal performance under particular circumstances.

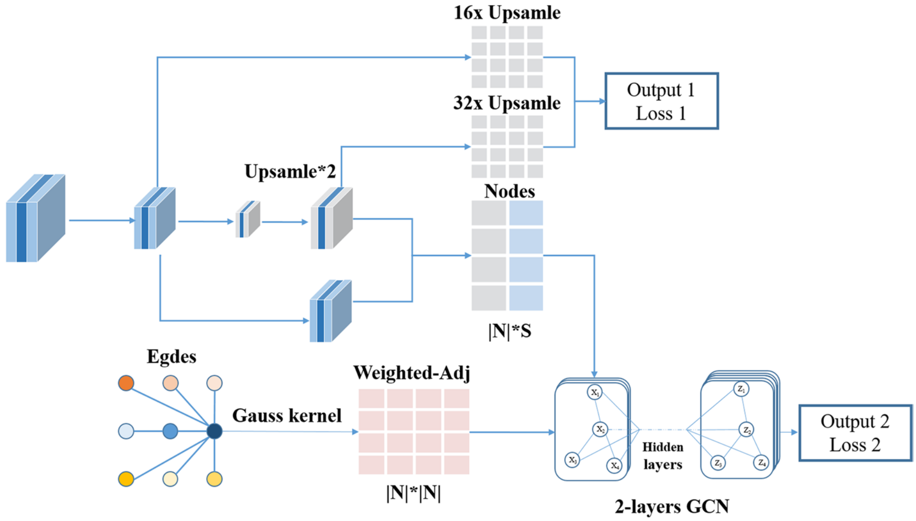

Fig. 12 The structure of the Graph-FCN. There are two outputs of the model, and two losses L1 and L2. They share the weights of the feature extracted by the convolutional layer. L1 is calculated by output1 and L2 is calculated by the output2. By minimizing L1 and L2, the FCN-16s can improve performance (Lu et al. 2019)

3.5 Instance segmentation network

Instance segmentation is intimately intertwined with other computer vision tasks, involving not only the per-pixel classification intricacies of semantic segmentation but also attributes of object detection, such as identifying unique instances within an image and assigning individual masks.

Hariharan et al. (2014) developed the Simultaneous Detection and Segmentation (SDS) model, which revolutionized instance segmentation research. The model's originality and capability expanded the possibilities, for instance, segmentation research. This model leverages the MCG algorithm to extract candidate regions for each image while simultaneously obtaining feature vectors for both detection bounding boxes and regional foreground via dual pathways. The region classification is performed using Support Vector Machine (SVM) (Chang and Lin 2011), and the overlapping regions are further refined with Non-Maximum Suppression (NMS) (Neubeck and Van Gool 2006) further refining the overlapping regions. Ultimately, CNN-generated features are leveraged for mask prediction, culminating in meticulously segmented images.

Distinctively, the Faster R-CNN (Ren et al. 2016) architecture (Fig. 13) includes a region proposal network (RPN) for generating bounding box candidates. The RPN obtains a region of interest (RoI), and the RoIPool layer calculates features from these proposals to infer object bounding box coordinates and categories. Expanding upon Faster R-CNN's foundation, (He et al. 2017a) developed the Mask R-CNN model, establishing it as the benchmark for instance segmentation tasks. This model adds a segmentation subnetwork to the existing object detection framework (Fig. 14). It first uses RPN to isolate RoIs of objects within input images, conducting RoI Align operations on the generated RoIs, and then predicts detection boxes, class labels, and segmentation masks for all RoIs. The

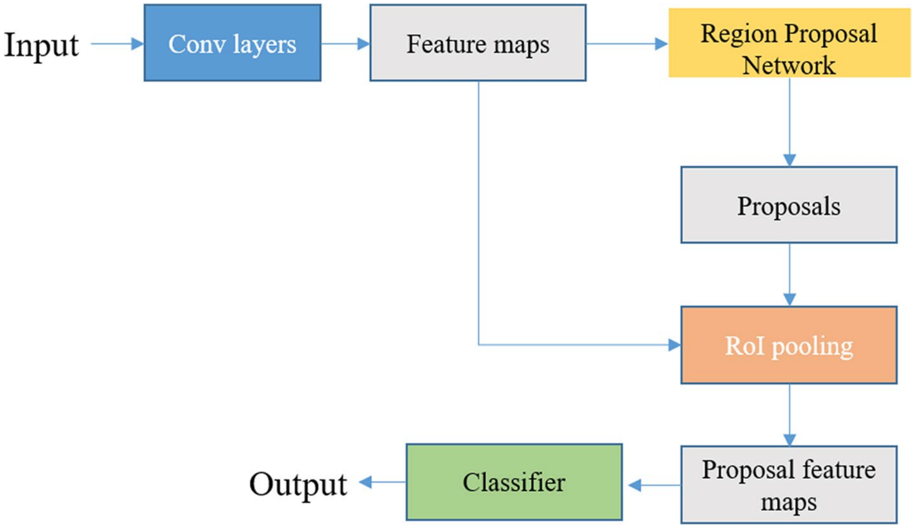

Fig. 13 The architecture of Fast R-CNN (Ren et al. 2016)

language used is objective, value-neutral, and follows standard sentence structure. Technical terms are explained when first used and spelling follows American conventions.

Dai et al. (2016a, b) proposed a more effective approach to instance segmentation by implementing Multi-task Network Cascades (MNC) that feature cascaded structure sharing convolutional characteristics. The MNC model dissects instance segmentation into three distinct tasks. Initially, an RPN is engineered through FCN to predict bounding box positions and object scores. Then, mask estimation is executed, projecting pixel-level masks separately for each instance object. Ultimately, the process of object classification combines convolutional features from the first two stages and assigns appropriate class labels to each mask, resulting in the categorization of objects.

Liu et al. (2018) introduced the Path Aggregation Network (PANet) based on the Mask R-CNN and FPN models (Fig. 15). The network's feature extractor utilizes an

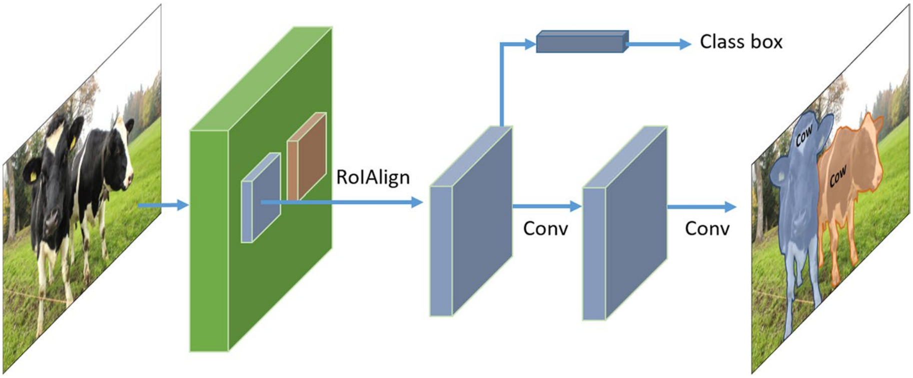

Fig. 14 The architecture of DeepLabv3 + Mask R-CNN (He et al. 2017b)

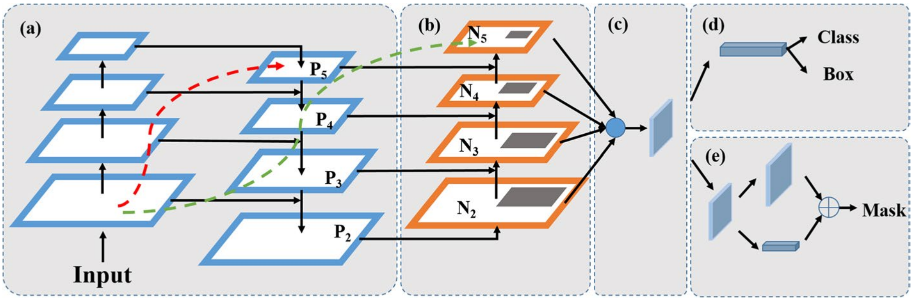

Fig. 15 The architecture of PANet. (**a**) FPN backbone. (**b**) Bottom-up path augmentation. (**c**) Adaptive feature pooling. (**d**) Box branch. (**e**) Fully-connected fusion. (Liu et al. 2018)

FPN backbone and includes a new bottom-up path to enhance the propagation of lower-level features. Each stage in the third path takes the feature map of the preceding stage as input, which is then processed by a 3×3 convolutional layer. A lateral connection supplements the output by directing it to the same-level feature map on the top-down pathway, thus preparing it for the following stage.

Huang et al. (2019a, b) introduced the Mask Scoring R-CNN (MS R-CNN) model, which expands on Mask R-CNN by including a Mask IoU evaluation branch that calculates scores using features produced by RoI Align and predicted masks. Wang et al. (2019a, b, c, 2020) analyzed the relationship between object detection and instance segmentation duties and put forward RDSNet. In this approach, images are divided into object and segmentation branches after moving through an FPN backbone network. The segmentation branch conducts pixel clustering using target embedding, resulting in target masks, while the object branch distinguishes instance object categories and location information. This approach guarantees complete feature information with minimal loss. Chen et al. (2019) proposed the instance segmentation model Mask-Lab to improve object detection by utilizing semantic and directional features predicated on Faster R-CNN. Moreover, direct mask generation instance segmentation models encompass SISDLF (De Brabandere et al. 2017), DeepMask (Chen et al. 2018a, b, c, d), and CenterMask (Lee and Park 2020). PolarMask (Xie et al. 2020) employs polar coordinates to model contour-encoded masks. Instance segmentation models predicated on positional information include InstanceFCN (Dai et al. 2016a, b) FCIS (Li et al. 2017), and SOLO (Wang et al. 2020a, b).

3.6 Generative models and adversarial training

GANs improve the accuracy and robustness of image segmentation while effectively addressing data imbalance and limited samples. By concurrently training a generator and a discriminator, GANs can produce more authentic and accurate segmented images, which can increase diagnostic and decision-making accuracy in various domains such as medical imaging, autonomous driving, and disease detection. Moreover, GANs can augment datasets, providing better assurance of a model's ability to generalize. Consequently, GANs hold substantial significance in the field of image segmentation.

Luc et al. (2016) proposed a method for semantic segmentation utilizing GANs in 2016. Their approach employed an FCN as the generator model (Fig. 16). The training process involved feeding the generated semantic segmentation images and ground truth images into

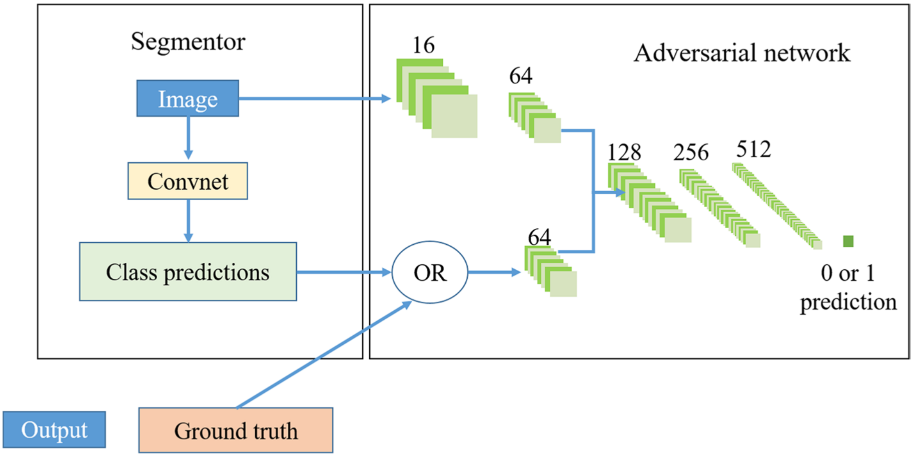

Fig. 16 Overview of the proposed approach. Left: segmentation net takes RGB image as input, and produces per-pixel class predictions. Right: Adversarial net takes label map as input and produces class label (1 = ground truth, or 0 = synthetic). Adversarial optionally also takes RGB images as input (Luc et al. 2016)

the discriminator model to facilitate adversarial training. Using this paradigm, the generator model learns to create precise semantic segmentation images in increments. Meanwhile, the discriminator model improves its ability to discriminate between generated semantic segmentation images and ground truth images. This methodology has been shown to yield high-caliber semantic segmentation results, surpassing traditional unsupervised learning techniques in numerous tasks.

Souly et al. (2017) proposed a semi-weakly supervised semantic segmentation paradigm based on GANs. This study offers a new perspective by regarding the segmentation network as a discriminator and using the GAN generator to augment the training dataset to optimize training efficacy. The model is classified into two categories: semi-supervised and weakly-supervised, based on the absence or presence of classification labels in the supplementary data. In the case of the weakly-supervised model, which employs classification labels, a conditional GAN is used by the GAN generator, with the image's classification label functioning as input.

Xue et al. (2018) presented an innovative medical image segmentation technique that utilizes a distinctive adversarial network structure, called SegAN. This architecture comprises dual components: a generator and a discriminator. The generator utilizes multiscale L1 loss for training, which produces more accurate segmentation results, while the discriminator uses adversarial loss during training to discriminate between real and generated images. Dai et al. (2018) presented the Structure Correcting Adversarial Network (SGAN), which includes a critic network that imposes structural regularities emerging from chest X-ray (CXR) imagery onto convolutional segmentation networks. This approach has undergone testing across multiple datasets and has achieved state-of-the-art outcomes in numerous instances.

Li et al. (2021) introduced a semisupervised learning and robust out-of-domain generalization methodology for semantic segmentation that uses a generative model to determine representations of unlabeled images. The model first extracts representations from labeled images and then applies them to unlabeled counterparts. This method supports better performance on unlabeled data and superior generalization to novel domains.

Alimanov and Islam (2023) proposed a new Denoising Diffusion Probability Model (DDPM), which is superior to GAN in image synthesis. The authors developed a Retina Tree (ReTree) dataset consisting of retinal images and corresponding blood vessel trees. They also trained a DDPM-based segmentation network using images from the ReTree dataset. In the first stage, the Retina Tree is generated using standard normally distributed random numbers. Then, the model is guided to generate fundus images based on a given blood vessel tree and a random distribution.

GANs offer several advantages for image segmentation in the following ways: they generate high-quality images, facilitate data augmentation, enable unsupervised/semi-supervised learning, improve feature learning, suppress overfitting, and enhance contextual understanding. These methods have shown remarkable success in tasks such as medical images (Xun et al. 2022), agriculture (Lu et al. 2022), and autonomous driving (Liu et al. 2021a, b, c) by improving segmentation accuracy and robustness. Despite these advancements, issues related to model generalization capabilities and dataset-specific problems remain significant challenges within this field. Therefore, researchers have focused on techniques such as unsupervised domain adaptation (Ma et al. 2024) and semi-supervised learning (Peláez-Vegas et al. 2023) to explore these areas further.

3.7 Attention-based models

The attention mechanism has become a pivotal and sophisticated technique utilized extensively in various deep learning tasks. It requires thorough exploration and understanding within the realm of advanced technologies. Attention mechanisms were inspired by human visual cognition, where the brain rapidly discerns the focus of attention from incoming visual signals. Consequently, when humans perceive images, they determine where to concentrate their attention in forthcoming instances, allocating less attention to peripheral regions instead of processing every pixel in the entire image instantaneously and adjusting the focal point over time. In 2014, Google Mind's team (Mnih et al. 2014) pioneered the incorporation of attention mechanisms into RNN models for image classification, which subsequently gained traction. In 2015, Bahdanau et al. (2014) were the pioneers of implementing attention mechanisms in NLP. They utilized attention components to enhance the original encoder-decoder architecture, resulting in remarkable outcomes and enhanced performance in English-French translation tasks.

Chen et al. (2014, 2016) introduced an innovative scale attention module capable of soft-weighting multi-scale features at each pixel location (Fig. 17). This adaptive technique allows for the identification of significant regions of an image, leading to improved model performance. This study stands as one of the earliest endeavors to utilize attention mechanisms in the domain of image segmentation.

The design of convolutional filters in CNNs to local areas hinders a comprehensive understanding of complex scenes. To address this issue, Zhao et al. (2018) devised the Point-wise Spatial Attention Network (PSANet) to alleviate the constraints of the local neighborhood. By using adaptively learned attention masks, each position on the feature map establishes connections with all others. This fosters bidirectional information propagation essential for scene parsing. Gathering information from different locations improves predictions for the current position while disseminating information from the current position bolsters predictions for other locations.

CNNs extract features using local receptive fields, which can result in different representations of identical pixel labels in the final feature map. This can cause class

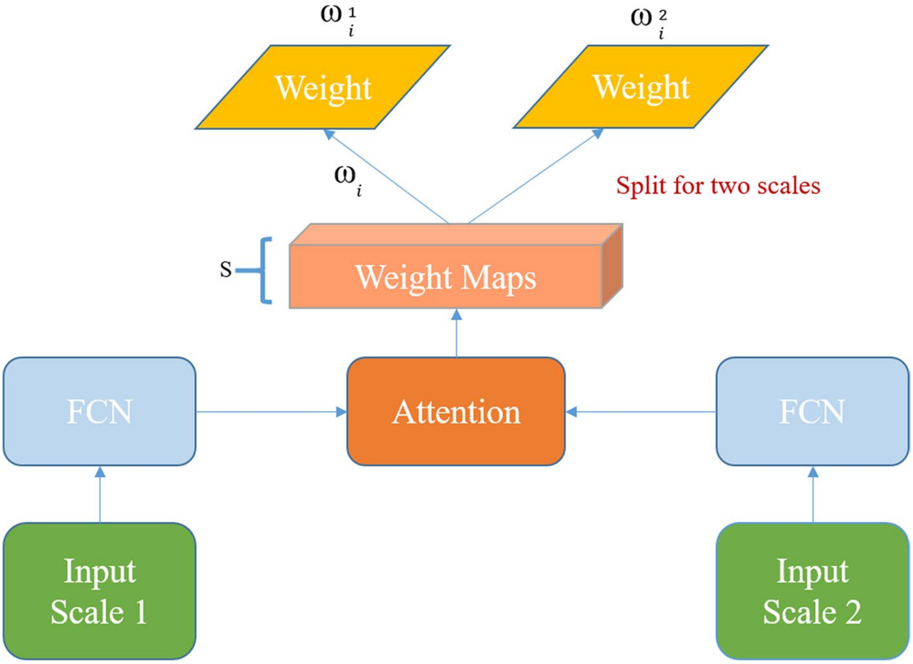

Fig. 17 The attention model makes use of features from FCNs and produces weight maps, reflecting how to do a weighted merge of the FCN-produced score maps at different scales and different positions (Chen et al. 2016)

inconsistency and inaccurate segmentation. To address this issue, RNN-based models (and their LSTM variants) capture long-range dependencies within feature maps, improving model precision. However, the effectiveness of these methods hinges on the outcomes derived from long-term memory learning, which can be rather generic. Fu et al. (2019a, b) presented the Dual Attention Network (DANet) as a solution. DANet utilizes a self-attention mechanism to capture feature dependencies across spatial and channel dimensions independently. Additionally, it adaptively combines local features with their global dependencies.

Huang et al. (2019a, b) proposed the novel CCNet, positing that the capture of long-range dependencies in visual understanding problems could provide informative contextual information. The CCNet integrates two modules: Criss-Cross Attention (CCA) and Recurrent Attention (CA). For every pixel, the ground-breaking CCA module extracts contextual data from neighboring pixels through criss-cross paths. With further iterations, every pixel eventually captures long-range dependencies originating from all other pixels. In comparison to non-local CCA modules, Recurrent CA modules demonstrate superior computational efficiency and reduced GPU memory consumption.

Zhong et al. (2020) introduced the Squeeze-and-Attention Network (SANet), an innovative approach to semantic segmentation tasks. They postulated that semantic segmentation involves two distinct sub-tasks: pixel prediction and pixel grouping. To address the pixel grouping challenge, they developed the SA module, which improves prediction accuracy. Drawing inspiration from Squeeze-Excitation Networks (SENet), SANet expands upon this concept by mitigating local constraints imposed by convolutional kernels and integrating attention-based convolution channels. This approach effectively supplements traditional convolutional layers with attention-focused processing for pixel groupings, thereby accounting for spatial-channel dependencies.

Additional applications of attention mechanisms within the realm of semantic segmentation include the EncNet (Zhang et al. 2018a, b), real-time semantic segmentation Bilateral Segmentation Network (BiSeNet) (Yu et al. 2018a, b), Expectation–Maximization Attention Network (EMANet) (Li et al. 2019a, b), which features the Context Encoding Module, Deep Feature Aggregation Network (DFANet) (Li et al. 2019a, b), Asymmetric Non-Local Neural Networks (ANNNet) (Zhu et al. 2019), Object Context-aware OCNet (Yuan et al. 2021), and SegNeXt (Guo et al. 2022). These models focus on object recognition, feature aggregation, and attention networks to improve the quality of semantic segmentation.

3.8 Transformer-based models

Zheng et al. (2021) utilized Transformers for semantic segmentation by introducing the SETR (Segmentation Transformer) model. The model reformulates semantic segmentation as a sequence-to-sequence prediction task, mitigating the need for the model to acquire local-to-global features by lowering the resolution. SETR exclusively employs a Transformer-based architecture. Initially, the ViT (Vision Transformer) model dissects images into fixed-size patches that undergo linear transformation. Subsequently, pixel vectors and positional encodings for each patch are integrated to allow the encoder. After 24 layers of Transformer learning, global features of the image are extracted, and a decoder is used to restore the original image resolution (Fig. 18).

Strudel et al. (2021) introduced the Segmenter, a customized Transformer model designed for semantic segmentation tasks. Since individual patches in image segmentation often present ambiguity, the inclusion of contextual information is essential to reach a label consensus. During the encoding phase, the Segmenter utilizes a ViT architecture to split images into patches before applying linear mapping and generating an embedded sequence post-encoder processing. During the decoding phase, learned embeddings for classes are

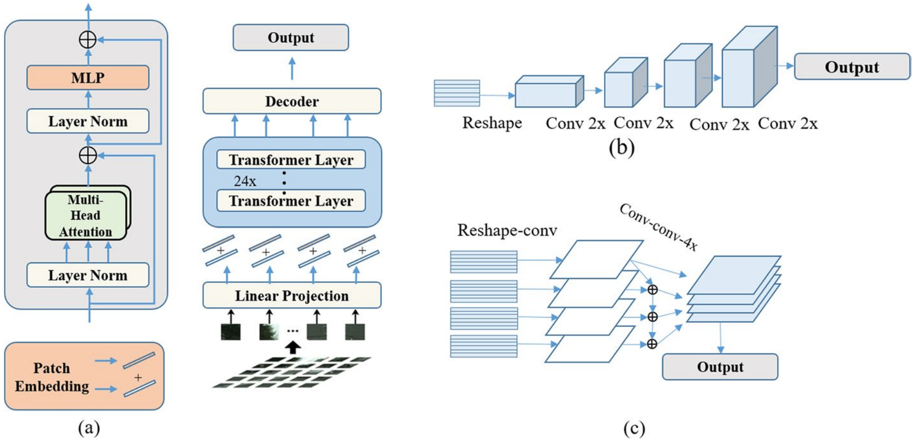

Fig. 18 The architecture of SETR. (**a**) The images are split into fixed-size patches, each patch is linearly embedded, positional embeddings are added, and the resulting vector sequence is fed to a standard Transformer encoder. (**b**) Progressive upsampling. (**c**) Multi-level feature aggregation. (Zheng et al. 2021)

incorporated into the decoder alongside the output of the encoder. Technical terms are explained on their first use. Class labels are procured employing either a linear decoder or a masked transformer decoder, resulting in the final pixel segmentation map after executing operations such as softmax and up-sampling.

Xie et al. (2021) proposed SegFormer, which combines Transformers with lightweight Multi-Layer Perceptron (MLP) decoders. SegFormer embraces hierarchical feature representation, reducing output feature dimensions at each Transformer layer during the encoding process to capture multi-scale feature information. The position embeddings found in ViT are omitted, preventing performance degradation caused by differences between training and testing image dimensions. The proposed MLP decoder utilizes a simple MLP framework to combine features across disparate encoder layer scales, integrating local and global attention mechanisms. This research substantiated that such basic and lightweight designs are crucial for achieving efficient segmentation on Transformers.

SCTNet (Xu et al. 2024) is a new state-of-the-art real-time semantic segmentation network that combines the features of Transformer semantic information with single-branch CNN. The network not only maintains the efficiency of a lightweight single-branch CNN but also possesses rich semantic representation capabilities, enabling it to achieve the best balance between performance and speed on multiple semantic segmentation datasets. On the Cityscapes dataset, SCTNet achieved an excellent performance of 80.5% mIoU and 62.8 FPS.

Transformers exhibit impressive feature learning capabilities, making their efficiency and capacity in segmentation an important area for future research, due to their strong long-range modeling capabilities and dynamic responsiveness. A variety of Transformer-based segmentation approaches have been developed, including the pioneering end-to-end panoptic segmentation model MaX-DeepLab (Wang et al. 2021a, b), TransUNet Chen et al. 2021a, b) incorporating ViT into U-Net for medical image segmentation, Video Instance Segmentation Transformer (VisTR) (Wu, Jiang, et al. 2022a, b, c), the panoptic segmentation benchmark Panoptic SegFormer (Li et al. 2022c, a, b), distortion-aware transformers (Zhang et al. 2022), weakly supervised semantic segmentation (Xu et al. 2022), End-to-End Weakly Supervised Semantic Segmentation with Transformers (Ru et al. 2022), and One-Stage Camouflaged Instance Segmentation with Transformers (OSFormer) (Pei et al. 2022).

4 Dataset and performance comparison

In this section, we review the most prevalent plant image datasets used for training and testing DL-based image segmentation models, provide an overview of typical metrics employed for evaluating segmentation model performance, and present evaluation results of DL-based segmentation models on established benchmark datasets.

4.1 Plant datasets

In contrast to prominent image segmentation benchmarks such as PASCAL VOC 2012 (Everingham et al. 2010), PASCAL-Context (Mottaghi et al. 2014), CamVid, Cityscapes (Cordts et al. 2016), and ADE20K, the domain of plant image segmentation lacks a comprehensive and unified dataset. Collecting plant image segmentation datasets can occur through methods such as custom data collection or online retrieval. We have compiled

Table 3 A summary of parameters related to the plant public dataset collected

Dataset	Resolution	Number of images	Images collection	Download link
LeafSnap (Kumar et al. 2012)	Multiple	30,866	Indoor/Outdoor	https://www.kaggle.com/datasets/xhlulu/leafsnap-dataset
LSC (Minervini et al. 2014, 2016)	A1: 500*530 A2: 530*565 A3: 2248*2048 A4: 441*441	A1: 128 A2: 31 A3: 27 A4: 624	Indoor	https://www.plant-phenotyping.org/CVPPP2014-dataset
GrowliFlowerL (Kierdorf et al. 2022)	368×448	2198	Outdoor	http://rs.ipb.uni-bonn.de/data/growliflower/
Fig dataset (Fuentes-Pacheco et al. 2019)	2000×1500	10	Outdoor	https://github.com/jofuepa/fig-dataset
KOMATSUNA (Uchiyama et al. 2017)	RGB-D: 166×190 Multi-view: 480×480	RGB-D: 300 Multi-view: 600	Indoor	https://limu.ait.kyushu-u.ac.jp/~agri/komatsuna/
MSU-PID (Cruz et al. 2016)	116×119	576	Indoor	https://cvlab.cse.msu.edu/multi-modality-imagery-database-msu-pid.html

various publicly available datasets containing plant images (shown in Table 3). These datasets are commonly utilized for training and evaluating image segmentation algorithms. A detailed analysis of each dataset's unique characteristics is provided, along with the associated parameters.

LeafSnap (Kumar et al. 2012) constitutes a substantial image repository tailored for plant identification, showcasing an extensive variety of leaf images from a multitude of tree species (Fig. 19). This dataset encompasses 30,866 leaf images distributed across 185 distinct species, primarily located along the eastern seaboard of the United States, incorporating abundant urban park, arboreal, and botanical garden specimens. Each tree species is represented by a corpus exceeding two thousand leaf images, with individual images meticulously annotated with pertinent species information and supplementary metadata. Beyond the scope of visual data, the dataset also supplies exhaustive descriptions correlated with each species, encapsulating facets such as foliar characteristics and growth habitats. The LeafSnap dataset has been widely harnessed in research endeavors spanning machine learning, computer vision, and artificial intelligence domains, thereby fostering a deeper comprehension and preservation of plant species within the natural world.

The Leaf Segmentation Challenge (LSC) dataset (Minervini et al. 2014, 2016) is a specialized collection designed for leaf image segmentation (Fig. 20). It includes four separate directories (A1, A2, A3, and A4) and a total of 810 digital images, which consist of 783 top-view Arabidopsis thaliana plant images and 27 high-resolution Nicotiana tabacum (tobacco) plant images. The visual data covers a wide range of growth stages and presents several challenges, including complex backgrounds, varying resolutions, and overlapping plants. Each image in the LSC dataset comes with a binary mask file that marks the foreground region of individual leaves, which is crucial for validating and benchmarking leaf image segmentation algorithms. The dataset is divided into four directories. Directory A1 contains 128 images, each with dimensions of 500×530 pixels, featuring complex

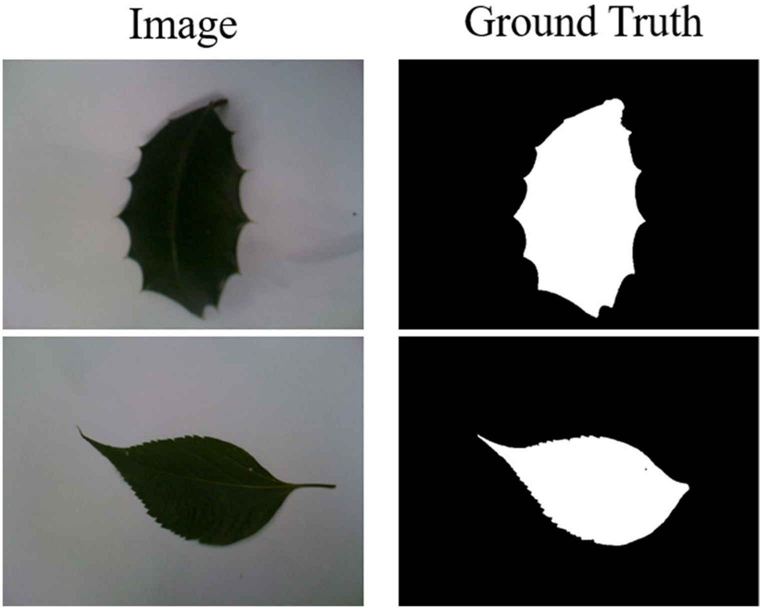

Fig. 19 An example of LeafSnap (Kumar et al. 2012)

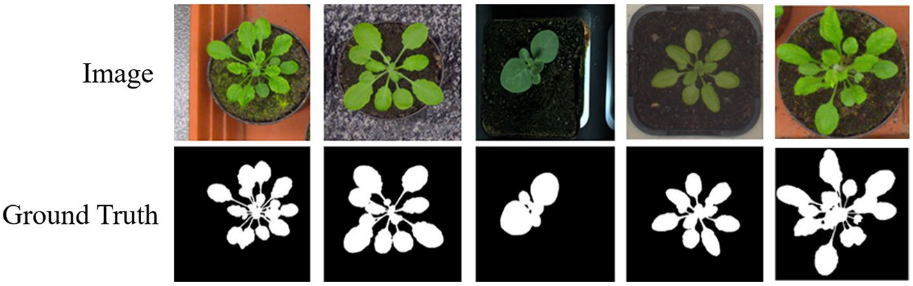

Fig. 20 Sample images from the LSC dataset with their corresponding ground truths (Minervini et al. 2014, 2016)

and varied backgrounds. Directory A2 contains 31 images, with uniform dimensions of 530 × 565 pixels, and featuring homogeneous and simple backgrounds. Directory A3 consists of 27 high-resolution images of the Nicotiana tabacum plant, with dimensions of 2448 × 2048 pixels. Finally, Directory A4 comprises 624 images of the Arabidopsis thaliana plant, each with dimensions of 441 × 441 pixels.

GrowliFlower (Kierdorf et al. 2022) is a georeferenced dataset of time-series captured by UAVs from two cauliflower fields (0.39 and 0.60 hectares) monitored in 2020 and 2021 (Fig. 21). The dataset comprises orthoimages with RGB and multispectral bands and coordinates of about 14,000 plants, allowing for the extraction of complete and partial image blocks. Phenotypic traits of 740 plants, such as plant sizes and developmental stages, are also included. GrowliFlower provides four subsets that can be used for distinct machine learning tasks. The GrowliFlowL subset is composed of pixel-level images that have been manually annotated by humans and possess block dimensions of 368 × 448 pixels. The

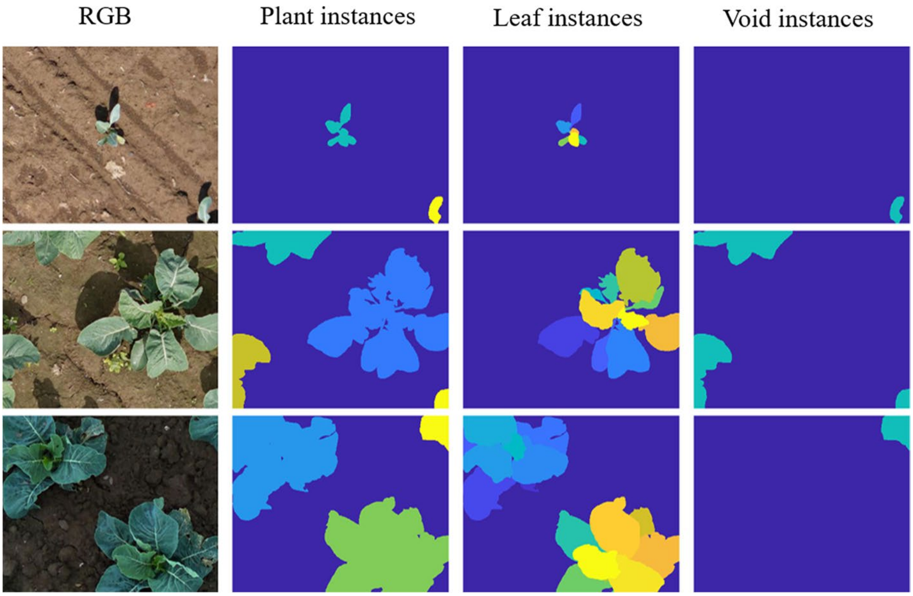

Fig. 21 An example of GrowliFlowerL (Kierdorf et al. 2022)

dataset, which is divided among training, validation, and test sets, contains a total of 2,198 plant images, with 1,972 containing plants and 226 lacking them.

The MSU-PID (Cruz et al. 2016) dataset comprises 576 annotated images of Arabidopsis thaliana and 175 images of broad bean plants. The images were captured indoors, using a Hitachi KP-F145GV CCD camera over 9 days (15 images per day) and 5 days (13 images per day), respectively. The camera integrates fluorescence, infrared, RGB color, and depth sensors, resulting in varying spatial resolutions for each mode. The dimensions of the Arabidopsis images are: 240 × 240 for fluorescence and IR, 120 × 120 for RGB, and 25 × 25 for depth. The fava bean image, on the other hand, has dimensions of 1000 × 640 for fluorescence and IR, 380 × 720 for RGB, and 90 × 190 for depth.

The Fig dataset (Fuentes-Pacheco et al. 2019) comprises aerial RGB images of fig shrubs captured by a modern RGB camera affixed to an unmanned aerial vehicle (UAV). The dataset features 10 precisely labeled RGB images with resolutions up to 2000×1500 pixels (Fig. 22). The corresponding labeled ground truth images were manually generated utilizing advanced image annotation tools. The leaves of most fig shrubs are overlapped due to the dry season and dust accumulation. These images pose various challenges in computer vision, such as variations in lighting and shading, the presence of weeds, diverse shades of soil, camouflaged vegetation, and an array of debris (e.g., rocks, dried branches, and tools used by agricultural workers).

The Komatsuna dataset (Uchiyama et al. 2017) comprises of two segments. The first segment includes 300 images captured using an advanced RGB-D camera (Fig. 23). The second segment consists of a multi-view dataset generated from several RGB cameras. It comprises 600 images with a resolution of 480×480 pixels. These images have a 166×190 pixel resolution. In total, these datasets provide 1200 images. The Komatsuna dataset is divided into two distinct segments. The initial segment involves 300 images captured through a high-tech RGB-D camera, with a resolution of 166 × 190 pixels. The subsequent segment involves a multi-view dataset extracted from numerous RGB cameras, including 600 images with a 480×480 pixel resolution.

4.2 Metrics for image segmentation models

In image segmentation tasks, intersection over union (IoU) is a crucial evaluation statistic. By calculating the ratio between the intersection and union of the predicted and ground

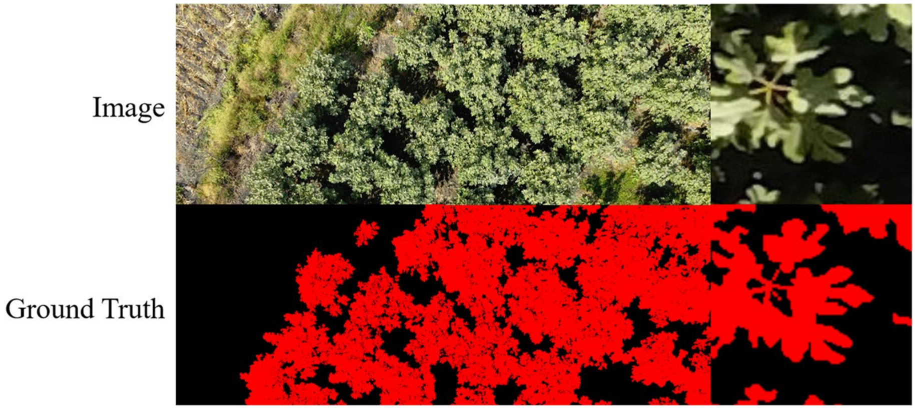

Fig. 22 An example of Fig dataset (Fuentes-Pacheco et al. 2019)

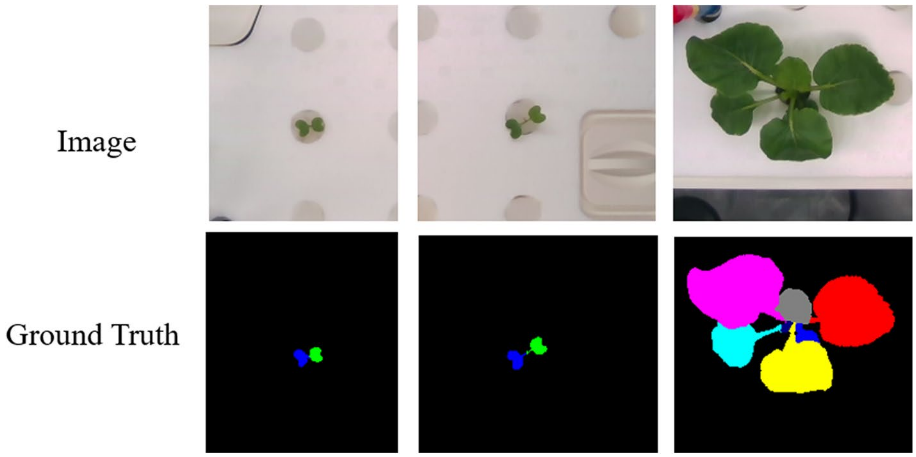

Fig. 23 Sample images from the Komatsuna dataset with their corresponding ground truths (Uchiyama et al. 2017)

truth labels, it can determine the percentage of accurately predicted pixels in relation to the total number of pixels for each class. The formula is:

$$\mathrm{IoU} = \frac{TP}{TP + FN + FP}$$

Within this framework, TP (True Positive) refers to the pixel count that shows the same classification in both the ground truth labels and the predicted results. FN (False Negative) identifies cases where pixels are assigned to a class in the ground truth annotations but are erroneously predicted as a different category. FP (False Positive) is defined as cases where pixels are labeled as the class in the predicted results, but they differ from the corresponding ground truth classification.

Pixel accuracy, a pertinent metric in computer vision, is derived by computing the quotient of the aggregate count of accurately classified pixels and the total pixel count. This particular metric demonstrates efficacy in scenarios featuring balanced class distributions; however, it may exhibit bias when confronting class imbalances. The formula can be articulated as follows

$$\text{Pixel Accuracy} = \frac{\sum_{i=1}^{N} T_i}{\sum_{i=1}^{N} \left(T_i + F_i\right)}$$

In this context, T_i represents the number of correctly predicted pixels for the i-th class, F_i denotes the number of incorrectly predicted pixels for the i-th class, and N signifies the total number of classes.

Mean accuracy is a metric obtained by dividing the number of correctly classified pixels for each class by the total pixel count of that class and then averaging the results across all classes. It is more robust than pixel accuracy in situations with class imbalance. The specific formula is as follows:

$$\text{Mean Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{T_i + F_i}$$

Indeed, the meanings of T_i and F_i in mean accuracy are identical to those in pixel accuracy.

Mean Intersection over Union (mIoU) is obtained by averaging the Intersection over Union (IoU) for each class and can be used to evaluate the overall performance of a model. The specific formula is as follows:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i + FP_i}$$

Among them, the meanings of TP_i, FN_i, and FP_i in mIoU are consistent with their definitions in IoU.

F1 Score is an evaluation metric that combines precision and recall, suitable for assessing binary classification problems (e.g., background/foreground). It provides a comprehensive evaluation of accuracy and completeness. The specific formula is as follows:

$$\text{F}_1\text{Score} = 2 \times \frac{Precision \times Recall}{Precision + +Recall}$$

In this context, Precision represents the proportion of true positive samples among all samples predicted as positive, while Recall denotes the proportion of true positive samples among all samples that are positive.

4.3 Quantitative performance of DL-based models

In this section, we list the performance of some algorithms discussed previously on popular segmentation benchmarks in Table 4. Although most published models report their performance on standard datasets and use standard metrics, some do not, making a comprehensive comparison challenging. Furthermore, only a few publications provide additional information in a reproducible manner, such as execution time and memory consumption, which is important for industrial applications that may run on embedded systems (e.g., drones, autonomous vehicles, robots, etc.) where computational capacity and storage space are limited, necessitating lightweight models.

5 Agricultural application

Presented in this section is a summary review of application studies of image segmentation in the agricultural field. These applications are loosely organized into four areas, including plant disease identification (Table 5), weed identification (Table 6), crop growth monitoring (Table 7), and crop yield estimation and counting (Table 8).

5.1 Plant disease identification

Plant diseases and pests have a significant impact on crop yield and quality, limiting plant growth and ultimately decreasing the quantity and quality of crops, thus reducing

Table 4 A summary of papers for semantic segmentation of natural images applied to Public dataset

Reference	PASCAL VOC 2012 Mean IoU	PASCAL-Context Mean IoU		CamVid	Cityscapes:	ADE20K
FCN-8 s (Shelhamer et al. 2017)	67.2		39.1	NA	NA	NA
DPN (Yuan et al. 2019)	74.1		NA	NA	NA	NA
SegNet (Badrinarayanan et al. 2017)	59.1		NA	60.1	NA	NA
MCANet (Yu and Koltun 2016)	73.9		NA	NA	NA	NA
RefineNet (G. Lin et al. 2017a, b)	84.2		47.1	NA	73.6	40.2
GCN (Peng et al. 2017; Pei et al. 2023)	82.2		NA	NA	76.9	NA
LinkNet (Chaurasia and Culurciello 2017)	NA		NA	68.3	76.4	NA
SDN (Fu et al. 2019a, b)	83.5		NA	71.8	NA	NA
PSPNet (Zhao et al. 2017)	85.4		NA	NA	80.2	NA
DeepLabv3 + (Chen et al. 2018a)	89.0		NA	NA	82.1	NA
DUC-HDC (Wang et al. 2018)	83.1		NA	NA	80.1	NA
DenseASPP (Yang et al. 2018)	NA		NA	NA	80.6	NA
BiSeNet (Yu et al. 2018a, b)	NA		NA	68.7	74.7	NA

Table 4 (continued)

Reference	PASCAL VOC 2012 Mean IoU	PASCAL-Context Mean IoU		CamVid	Cityscapes:	ADE20K
EncNet (Zhang et al. 2018a, b)	85.9		51.7	NA	NA	44.94
APCNet He et al. 2019a, b)	84.2		NA	NA	NA	NA
DMNet He et al. 2019a, b)	84.4		54.4	NA	NA	45.50
DANet (Fu et al. 2019a, b)	NA		52.6	NA	81.5	NA
DFANet (Li et al. 2019a, b)	NA		NA	59.3	70.3	NA
EMANet (Li et al. 2019a, b)	88.2		53.1	NA	NA	NA
ANNNet (Zhu et al. 2019)	NA		52.8	NA	81.3	45.24
Graph-FCN (Lu et al. 2019)	65.91		NA	NA	NA	NA
GINet (Wu et al. 2020)	NA		54.9	NA	NA	45.54
DGCN (Zhang 2019)	NA		53.7	NA	82.0	NA
CCNet (Huang et al. 2019a, b)	NA		NA	79.1	81.9	45.76
PSANet (Zhao et al. 2018)	85.7		NA	NA	81.4	43.77
BGCNet (Hu et al. 2021a, b)	84.2		NA	NA	82.1	NA
EVDR (Liu et al. 2021a, b, c)	NA		NA	NA	81.9	NA

Table 4 (continued)

Reference	PASCAL VOC 2012 Mean IoU	PASCAL-Context Mean IoU		CamVid	Cityscapes:	ADE20K
OCNet (Yuan et al. 2021)	NA		56.2	NA	82.5	45.50
SETR (Zheng et al. 2021)	NA		55.83	NA	81.6	50.28
Segmenter (Strudel et al. 2021)	NA		59	NA	81.3	51.82
SegNeXt (Guo et al. 2022)	90.6		59.2	NA	83.2	51.0

Table 5 Summary of DL approaches in plant disease detection

Reference	Species	Network architecture	Methodology	Performance
Wang and Zhang 2018	Corn	Encoder-decoder network	Add preprocessing and data augmentation	96% accuracy
Saleem et al. 2021	Mango	Encoder-decoder network	Color, texture, and geometric feature extraction, fusion, and PCA-based feature reduction were applied before feature fusion	99.2% accuracy and 0.8% false negative rate
Yao et al. 2022	Peach	Instance segmentation network	Applying Focal Loss to Mask Scoring R-CNN	0.463 accuracy of segm_mAP_50
Li et al. 2022a, b, c	Potato	Instance segmentation and dilated convolutional networks	Designed an integrated framework that combines instance segmentation, classification, and semantic segmentation	94.24% mean pixel accuracy (MPA) and 89.91 mIoU
Douarre et al. 2019	Apples	Based on GAN	Designed two data generation methods based on plant canopy simulation and GANs	64.3% F1 score
Cap et al. 2022	Cucumber	Based on GAN	Add attention mechanism to GAN	83.9% F1 score
Wu et al. 2022a, b, c	Tomato	Based on Transformer	Spatially Modulated Co-Attention (SMCA) was used to assign Gaussian-like spatial weights to the query box of DS-DETR	96.4% accuracy
Li et al. 2022a, b, c	Rice	Based on Transformer	Proposed a data expansion method based on disease copying and pasting	85.38% mIoU

Table 6 Summary of DL approaches in weed identification

Reference	Network architecture	Methodology	Performance
Das and Bais 2021	Multiscale network	A novel deep learning architecture with skip connections, residual blocks, and pyramid scene parsing layers for segmentation of the highly imbalanced and complex dataset	76.79% mIoU and 97.65% accuracy
Kim and Park 2022	Encoder-decoder network	Increase crop, weed, and combined losses, and design a model for intensive training in the target area	The mIoU values of the segmentation for the crops and weeds are 0.9164, 0.8372, and 0.8260
Lan et al. 2021	Encoder-decoder network	Reduce the computational complexity of the original model parameters and improve the segmentation accuracy of the original model	80.28% mIoU and 93.09% accuracy
Nong et al. 2022	Dilated convolutional and multiscale network	By fusing selective kernel attention with encoding features	70% mIoU
Zou et al. 2021a, b	Dilated convolutional	Use Hybrid division convolution to modify the backbone network of U-Net	Green plant segmentation accuracy was 93.5%, and crop segmentation IoU was 93.40%
Zou et al. 2021a, b	Encoder-decoder network	Reduce Encoder-Decoder layers	98.57% accuracy
Ullah et al. 2021	Dilated convolutional	Add three parallel layers with atrous rates of 2, 6, and 9 at the input end to extract features at multiple scales	89.12% mIoU

diminishes farmers' productivity and income (Hasan et al. 2022). Consequently, rapid and accurate detection of plant diseases is critical (Chouhan et al. 2019a, b). Image segmentation technology has emerged as an efficient detection tool, surpassing traditional manual inspection methods by extracting valuable information from digital images. Despite the continued relevance of traditional image segmentation techniques (Abdu et al. 2019; Zhang et al. 2018a, b), deep learning has made remarkable advancements in digital image processing, significantly surpassing traditional approaches (Liu and Wang 2021; Wang et al. 2022a, b). Plant disease and pest identification utilizing deep learning methodologies has become a central area of research. Segmentation networks address plant disease and pest detection by performing semantic or instance segmentation on affected and healthy regions. This approach enables the fine-grained delineation of diseased areas while capturing location, category, and associated geometric attribute, which include length, width, area, contour, and centroid. However, detecting plant diseases and pests in complex natural environments presents numerous challenges, including subtle distinctions between diseased regions and backgrounds, low contrast, considerable variation in the size and type of affected areas, and extensive image noise. Traditional classical methods often fail to provide accurate detection results in these situations.

CNNs demonstrate exceptional image feature extraction capabilities, making them suitable for task in plant disease recognition and segmentation. By training binary or multi-class segmentation models, CNNs can accurately differentiate between healthy and infected plant tissues in input images, allowing for effective disease localization. Wang et al. (2018) proposed an innovative method for segmenting corn leaf diseases utilizing Fully Convolutional Neural Networks (FCNNs). To addresses the limitations of traditional computer vision, methods that are susceptible to variable illumination and complex backgrounds. The proposed method achieved a remarkable segmentation accuracy of 96.26%. Chouhan et al. (2019a, b) proposed an optimization method of bacterial foraging optimization algorithm. The method is used to initialize the weights of the artificial neural network for image segmentation, and a dice similarity coefficient of 86.79% is obtained. Saleem et al. (2021) presented a full resolution convolutional network (FrCNnet) model based on CNNs for accurately segmenting mango leaf damage in computer-aided systems. The FrCNnet learns the features of each pixel in the input data after applying specific preprocessing techniques. Nagaraju et al. (2022) proposed two learning algorithms, the image preprocessing and transformation algorithm and the image masking and REC-based hybrid segmentation algorithm (IMHSA), to solve the problem of limited data sets and overfitting of convolutional neural network models in the classification process. Yao et al. (2022) implemented Mask R-CNN and Mask Scoring R-CNN for the segmentation and identification of peach diseases. Utilizing instance segmentation models enables the extraction of disease names, locations, and segmentations, with the foreground area serving as the fundamental feature for subsequent segmentation. Focal Loss addresses challenges posed by difficult and imbalanced samples and is employed in this dataset to improve segmentation precision.

By utilizing the image generation capabilities of the generator and the discernment abilities of the discriminator, GANs can generate high-quality maps for segmenting plant diseases. Douarre et al. (2019) presented a study employing Deep Convolutional Generative Adversarial Networks (DCGANs) to generate images for the segmentation of apple rust disease, a condition characterized by spots on leaves and fruits. Infrared (IR) images of vegetation were partitioned into 64 × 64 pixel sub-images and then fed into a SegNet model for pixel-level segmentation. CycleGAN was evaluated for its capability to synthesize and augment image data for tasks related to plant health detection. Nerkar and Talbar (2021) implemented CycleGAN together with U-Net as an image synthesis generator to rebalance

Table 7 Summary of DL approaches in crop growth monitoring

Reference	Species	Network architecture	Methodology	Performance
Huang et al. 2021a, b, c	Rice, beans, cotton	Encoder-decoder network	Integrating semantic segmentation and style transfer	62.54% mIoU
Akiva et al. 2021	Cranberry	Multiscale network	The model includes a cloud motion estimation branch, built on an unsupervised Siamese-style recurrent spatial transformer network, and a fully-convolutional cloud segmentation branch	62.54% mIoU and 13.46% Mean Absolute Error
Zhang et al. 2020	Purple rapeseed	Encoder-decoder network	Regression analysis was performed between the purple rapeseed leaf ratios and the measured N content	82.41% IoU
Fukuda et al. 2021	Pear	Encoder-decoder network	Replacing U-Net Convolutional Kernel Size and Increasing Convolutional Times	97.5% IoU
Weyler et al. 2022	Beet	Encoder-decoder network	The upper decoder predicts normalized confidences and the lower decoder predicts offsets and vote weights	98.57% accuracy
Yu et al. 2022	Maize	Encoder-decoder network	U-Net model with a lightweight network MobileNet as the feature extraction network	79% IoU

Table 8 Summary of DL approaches in crop yield estimation and counting

Reference	Species	Network architecture	Methodology	Performance
Wang et al. 2019a, b, c, 2020	Wheat	Encoder-decoder network	Using the Otsu algorithm to binarize initial segmentation	97.4% accuracy
Alkhudaydi and De La Iglesia 2022	Wheat	Encoder-decoder network	Applying repeated convolution and subsampling	NA
Shao et al. 2021	Rice	Encoder-decoder network	Combining localization-based FCN with watershed algorithm	The mean absolute error (MAE) of the model on the 300-size test set is 2.99
Tan et al. 2019	Rice	Encoder-decoder network	Replacing U-Net Convolutional Kernel Size and Increasing Convolutional Times	97.5% IoU
Wang et al. 2022a, b	Wheat	Instance segmentation network	Construct a Regression Convolutional Neural Network (RCNN) to count wheat ears based on the segmentation results of FCN	R2 and root mean square error (RMSE) are 0.980, 0.996, and 9.437
Malambo et al. 2019	Sorghum	Encoder-decoder network	Post-process the segmented output to remove small objects and separate merged panicles	Overall detection accuracy of 94%
Liu et al. 2023	NA	Instance segmentation network	Integrated Unsupervised Style Transfer Network (STNet) and Dual Task Based Split Counting Network (SCNet)	88.2% mIoU

a dataset containing nine classes of tomato diseases. The synergy between CycleGAN and U-Net exhibited superior performance in perceiving image quality metrics by capturing low-level details and realistic textures. However, the study did not investigate the effectiveness of synthetic images in plant disease identification tasks. Aiming to enhance the diversity of image generation, Cap et al. (2022) integrated a leaf segmentation module, comprising a weakly supervised segmentation network, into CycleGAN, resulting in a novel model called LeafGAN. This model is designed to transform regions of interest within plant disease images. Upon testing on a dataset containing five classes of cucumber leaf diseases, LeafGAN contributed to a 7.4% increase in diagnostic performance compared to the unmodified CycleGAN, which only resulted in a 0.7% improvement.

In plant disease segmentation tasks, Transformers demonstrate proficiency in extracting both local and global features, capitalizing on self-attention mechanisms to accomplish precise disease localization and segmentation. Wu et al. (2022a, b, c) proposed a Disease Segmentation Detection Transformer (DS-DETR) based on the DETR network for segmenting early blight and late blight of tomatoes. DS-DETR initially uses the Plant Disease Classification Dataset (PDCD) for unsupervised pre-training, which effectively resolves the problem of lengthy training cycles and slow convergence in DETR. Through pre-training the Transformer structure, leaf disease features can be acquired beforehand, and these pre-trained model weights are employed to hasten the convergence rate in DS-DETR. Next, Spatial Modulation Common Attention (SMCA) was utilized to assign Gaussian-like spatial weights to the DS-DETR query frames. This method permits the usage of query frames with varying weightings to train different areas of the image, enhancing the model's accuracy. In addition, the Transformer structure of DS-DETR introduces improved relative position coding to enhance recognition of the sequence order of input markers, further strengthening the spatial position features. Finally, testing the DS-DETR model was conducted on the self-constructed Tomato Leaf Disease Segmentation Dataset (TDSD). The experimental results indicate that DS-DETR outperforms all other models on APmask with improvements of 12.87%, 8.25%, 3.67%, 1.95%, 10.27%, and 9.52% compared to Mask RCNN, BlendMask, CondInst, SOLOv2, ISTR (Hu et al. 2021a, b) and DETR models. Furthermore, it achieved a disease classification accuracy of 0.9640. However, the method still needs to be improved for the segmentation of small light spots.

If one is able to identify a disease, but is unable to quantify its severity, then one is still far from meeting the requirements of precision agriculture. Disease symptoms look symmetric at different stages of infection, with the possibility of overlapping symptoms appearing on the same leaf. So, the wide variety of symptom characteristics in qualitative and quantitative terms makes it very challenging to collect disease samples (Hasan et al. 2023). Li et al. (2022a, b, c) proposed a lightweight network based on copy-paste and semantic segmentation for accurate disease region segmentation and severity assessment. The RSegformer model was trained using a lightweight Segformer semantic segmentation network that features an attention mechanism and an up-sampling operator. This modification enables the model to balance local and global information, accelerate the training process, and reduce overfitting. The results of this model show that RLDCP can effectively improve the accuracy and generalization performance of the semantic segmentation model compared to traditional data augmentation methods, and can improve the mIoU of the semantic segmentation model by about 5% with only twice the size of the dataset. RSegformer can achieve 85.38% mIoU with a model size of 14.36 M.

5.2 Weed identification

Weeds are one of the principal factors undermining crop yields. They sporadically appear in fields and compete with crops for water, nutrients, and light, resulting in a negative impact on crop productivity and quality. This ultimately leads to a significant decline in global crop production (Wang et al. 2019a, b, c, 2020). At present, herbicide application represents the most prevalent method of weed control worldwide (Christensen et al. 2009). Conventional weed eradication involves indiscriminately spraying herbicides across entire fields, regardless of weed density, leading to excessive herbicide use in areas without weeds. This approach results in herbicide waste and contamination of agricultural ecosystems (Zou et al. 2021a, b). Image segmentation methods offer an efficient way to attain accurate weed detection and density assessment. The primary aim for image segmentation in weed detection is to distinguish plants from the background, including soil and residue. The key task for successful weed detection is efficient vegetation segmentation.

Kamal et al. (2022) evaluated deep machine learning algorithms for differentiating weeds from crop plants, utilizing an open carrot field image database. Das and Bais (2021) introduced DeepVeg, which focuses on the smallest (damage) class without impacting other classes to address the problem of class imbalance. Mishra et al. (2022) proposed an Inception V4 architecture approach based on deep convolutional neural networks, using RGB weed and crop images. It provides data cleaning to eliminate the background and uses segmentation masks to eliminate foreground vegetation. The early rapeseed field image dataset was used to train and test the proposed model. The evaluation results show that the DeepVeg model performs better than the union score with an average intersection greater than 0.76 and an accuracy greater than 0.97 in the four categories of segmentation. The model also shows robustness in detecting unlabeled, newly grown weeds and rapeseed, and can train the model to distinguish between rapeseed and weeds with similar circular structures with a small amount of data, which is suitable for early damage and weed segmentation. However, there is room for further improvement in the segmentation of such complex and highly imbalanced datasets. To address the question of the correlation between crop and weed classes, Kim and Park (2022) proposed the multi-task semantic segmentation-convolutional neural network for detecting crops and weeds (MTS-CNN) using one-stage training. This method incorporates crop, weed, and combined (crop and weed) losses to increase the associations between crop and weed classes, and trains the object (crop and weed) region intensively. In experiments performed with three different open databases—the BoniRob dataset, a crop/weed field image dataset (CWFID), and a rice seedling and weed dataset—the mIoU values of the segmentation for the crops and weeds in the MTS-CNN are 0.9164, 0.8372, and 0.8260, respectively.

Unmanned aerial vehicles (UAVs) are frequently used for crop monitoring and weed mapping on farmland. They are preferred over ground vehicles due to their flexibility, cost-effectiveness, ease of operation, and absence of soil compaction in fields (Nong et al. 2022). Aiming at the real-time identification of rice weeds by UAV low-altitude remote sensing. Lan et al. (2021) present two refined recognition models, MobileNetV2-UNet and FFB-BiSeNetV2, for real-time detection of rice weeds via low-altitude remote sensing from UAVs. The models are built on U-Net and BiSeNetV2 semantic segmentation models. The MobileNetV2-UNet model minimizes the computation of the original model parameters, whereas the FFB-BiSeNetV2 model enhances segmentation accuracy. The real-time segmentation effect of the two improved models on rice weeds was confirmed through the collection of low-altitude remote sensing video data. The results indicate that, in comparison to the U-Net model, the MobileNetV2-UNet model reduces network parameters and model size, minimizes floating point calculations, and enhances inference speed by almost three times. The FFB-BiSeNetV2 model enhances segmentation accuracy when compared to the BiSeNetV2 model and achieves maximum pixel accuracy and average cross-over ratio. The optimized models meet the performance requirements for real-time recognition on the embedded hardware platform. This study serves as a point of reference for real-time recognition of rice weeds and precise spraying operation of plant protection UAVs. Nong et al. (2022) proposed SemiWeedNet, a semi-supervised segmentation method for weeds and crops in drone imagery that accurately identifies varying sizes of weeds in complex environments while minimizing the necessity for extensive labeled data. SemiWeedNet integrates labeled and unlabeled images when constructing a unified semi-supervised architecture based on semantic segmentation models. By amalgamating encoded features with selective kernel attention, the researchers created a multi-scale enhancement module that emphasizes salient weed and crop traits while mitigating interference from intricate backgrounds. To address challenges arising from crop-weed similarities and overlaps, they introduced Online Hard Example Mining (OHEM) for optimizing labeled data training. Results demonstrated that SemiWeedNet outperforms extant methods, showcasing the potential of its components in boosting segmentation performance. Zou et al. (2021a, b) proposed a new approach to determine field weed density and create maps. They captured field images using UAVs and applied the excess green minus excess red index combined with the minimum-error threshold segmentation method to differentiate green plants from bare ground. Employing an enhanced U-Net, they conducted crop segmentation, subsequently obtaining weed images by eliminating bare ground and crops from the field. This innovative approach effectively evaluates field weed density from drone imagery, thus offering crucial information for precision weeding initiatives.

Zou et al. (2021a, b) tackled the issue of image annotation complexity in weed semantic segmentation by introducing an image enhancement technique. The semantic segmentation network underwent a two-stage training process, encompassing pre-training and fine-tuning phases. This approach yielded an intersection over union (IoU) value of 92.91% and an average segmentation time (ST) of 51.71 ms per image. Results indicated that the refined U-Net effectively distinguished weeds from images containing a multitude of other plants. The proposed weed target image segmentation method exhibited remarkable accuracy in segmenting weeds within intricate field environments, demonstrating wide-ranging applicability. U-Net, renowned for its robust training capability with limited samples and simplistic architecture, is extensively employed in weed identification tasks. Nasiri et al. (2022) utilized the U-Net architecture, a deep encoder-decoder convolutional neural network (CNN), to achieve pixel-level semantic segmentation of sugar beets, weeds, and soil. For 1385 RGB images collected under various conditions and heights, researchers trained a U-Net architecture with ResNet50 as the encoder module. To address data imbalance and small-region segmentation challenges, they employed a combination of Dice loss and focal loss to create a custom linear loss function. Ullah et al. (2021) used Maximum Likelihood Classification (MLC) and image processing techniques to label field images into three categories: background, crops, and weeds. Sodjinou et al. (2022) resented a segmentation approach that integrated semantic segmentation and the K-means algorithm to tackle crop and weed segmentation in color images. By employing thresholding techniques, all elements besides plants were eliminated from the images.

5.3 Crop growth monitoring

Plant growth monitoring plays a pivotal role in modern agricultural production, significantly contributing to the precise evaluation of crop growth conditions, enhancement of crop yield and quality, and prevention and control of pests and diseases. Conventionally, crop surveillance predominantly entailed manual visual inspections on a field scale. However, with technological advancements, the adoption of sensor technology and automation systems has become increasingly prevalent (Krishnaswamy Rangarajan and Purushothaman 2020). These sophisticated automated systems employ computer vision techniques to monitor plant growth dynamics, overall health, and performance metrics. To evaluate plant behavior under experimental field conditions and gauge their physical responses and symptoms to external stimuli, quantitative methodologies, and algorithms are essential. The integration of these methods establishes relationships between physical plant attributes and sensor-derived data. Image segmentation techniques, pivotal in the realm of computer vision, facilitate the periodic acquisition and analysis of plant imagery. This process yields vital information about crop growth status, encompassing variables such as growth rate and leaf coloration. Consequently, these insights empower automated systems to make informed management decisions and optimize agricultural strategies.

In references Zheng et al. (2009) and Zheng et al. (2010), the mean shift algorithm was utilized for the segmentation and extraction of soybeans and non-green vegetation. Krishnaswamy Rangarajan and Purushothaman (2020) investigated leaf count estimation and chromatic attributes for nine lab-cultivated eggplants (Solanum melgena) saplings. Leaf segmentation was facilitated using a combination of particle swarm optimization and contour growth methodology, achieving an accuracy rate of 89%. The saplings were ranked based on their defect percentages. Automated equipment and Foldscope (an innovative paper-based microscope) were utilized to obtain images for conducting linear regression analysis on the estimated Normalized Green Red Difference Index (NGRDI), ultimately validating regions of health and defects. The achieved R-squared value and least mean square error (LMSE) amounted to 0.86 and 0.1 respectively.

Unmanned aerial vehicles (UAVs), as efficient data collection devices, are widely used in crop monitoring. Akiva et al. (2021) developed a UAV-based field data and ground-based sky data collection system for capturing video images at multiple time points for crop health analysis. The system was designed for crop health analysis, and dataset evaluation showed an impressive level of accuracy in predicting the internal temperature of exposed fruits evaluation showed an impressive level of accuracy. Solar irradiance prediction errors ranged from 8.41–20.36% MAPE within 5–20 min intervals. The system achieved a segmentation accuracy of 62.54% mIoU and an exposed fruit recognition count accuracy of 13.46 MAE, providing informed feedback for growers. Due to temperature and illuminance variations during UAV flights, color bias in UAV images is inevitable. Temperature and illuminance changes during UAV flights cause color cast in imagery, which can mislead crop monitoring assessments. Color calibration is essential to mitigate these effects. Current methods use semantic correspondences for color transfer but often overlook the integration of semantic segmentation and style transfer, leading to issues with semantic mismatch. To address this problem, Huang et al. (2021a, b, c) propose a multi-decoder architecture that integrates semantic segmentation and style transfer for end-to-end color transfer. Additionally, an adaptive instance normalization (AdaIN) method tailored for crops is introduced to estimate color bias in crop regions, using this information to calibrate colors across the entire image through a local-to-global attention mechanism. The objective of this study is to establish a general framework for removing the color cast in UAV imagery for crop monitoring. This framework will provide a solid foundation for subsequent data interpretation.

Crop growth monitoring enables early estimation of final yields and prediction of optimal harvesting times to improve production efficiency. Zhang et al. (2020) employed a U-Net model for pixel-level segmentation of purple rapeseed leaves at the seedling stage using UAV RGB images. Considering the limited spatial resolution and small target size of UAV-acquired rapeseed images, a careful selection of input image block sizes was performed. Experiments demonstrated that the U-Net model with a block size of 256 × 256 pixels achieved better and more stable results, with an F-measure of 90.29% and IoU of 82.41%. In order to solve the problem of changing lighting conditions of RGB images throughout the day, Fukuda et al. (2021) introduced CROP (Central Round Object Painting). This technique employs deep learning for image segmentation, utilizing a neural network architecture based on an enhanced version of U-Net. CROP identifies various central round fruit types in RGB images under diverse illumination conditions and generates corresponding masks. By quantifying mask pixels, the fruits' relative two-dimensional dimensions can be acquired, offering a non-contact approach for automatically tracking fruit growth in time-series imagery. Weyler et al. (2022) proposed a vision-based method for the concurrent instance segmentation of crop plants and leaves within breeding plots. A specialized convolutional neural network was devised to pinpoint plant-specific key points and the location of pixel groups, enabling the detection of individual leaf and plant instances. This supports vision-based systems in delivering extensive automated and regular assessments of plant growth status. Yu et al. (2022) evaluates the effectiveness of the U-Net model, particularly with Vgg16 as the feature extraction network, in accurately segmenting maize tassels from near-ground RGB and UAV images. This method addresses the current labor-intensive and error-prone manual monitoring approach. The results indicate that the U-Net model (with Vgg16) performs better than when using MobileNet, demonstrating good segmentation accuracy across different tasseling stages, maize varieties, and image resolutions. The segmented area changes from the images align well with manual measurements. Even at a resolution of 3.06 mm, the UAV RGB images showed satisfactory segmentation accuracy. Thus, the U-Net model proves to be efficient for the accurate segmentation of maize tassels in various complex scenarios, signaling potential for future use in crop phenotypic experiments.

5.4 Crop yield estimation and counting

Since Seguí et al. (2015) initiated research on CNN-based counting, there has been an increasing interest in deep learning-driven counting methodologies for the automated enumeration of agricultural products. These techniques offer valuable support to agribusinesses, enhancing the optimization and streamlining of harvest yields. Furthermore, they provide essential yield assessments and production efficiency estimates for agricultural producers. Such data is ultimately employed to forecast storage requirements, profitability, and production capability. In dense agricultural contexts, counting and yield estimation share similarities, particularly as yield estimation predominantly relies upon deep-learning-based counting techniques. Image segmentation technologies isolate crops from background elements within field images, yielding distinct crop boundaries to facilitate accurate counting and yield computation.

Wheat, rice, and corn constitute the world's three primary staple crops. Wang et al. (2019a, b, c, 2020) proposed a new method to accurately count wheat ears in field conditions using an FCN and Harris corner detection. The process includes constructing a dataset of wheat-ear images, training an FCN as the segmentation model, testing the model, processing results using the Otsu algorithm for binarization, and applying Harris corner detection. This technique offers high accuracy (0.984 on average), quick computation time (0.033 s for 256 × 256-pixel image), and improved performance under different conditions like wheat-ear occlusion and soil disturbance. The counting accuracy is also commendable with an average score of 0.974, R2 of 0.983, and RMSE of 14.043, showing an improvement of 10% compared to previous methods. This proves crucial for efficient wheat phenotyping studies. Alkhudaydi et al. (2022) introduced Spike Count, a density estimation method related to human crowd counting, specifically tailored for enumerating wheat spikelets. This counting methodology is rooted in a deep learning architecture due to its capacity for automatic feature recognition, employing transfer learning for both segmentation and counting tasks. Experimental outcomes revealed that segmentation is advantageous, as concentrating solely on regions of interest improves counting precision in the majority of scenarios. Notably, transfer learning from analogous images yielded favorable results for counting tasks throughout most stages of wheat development. Shao et al. (2021) generated a dataset comprising 3,300 rice panicle samples, representing an array of intricate situations including diverse lighting conditions, complex backgrounds, overlapping rice plants, and overlapping leaves. They also proposed a hybrid approach incorporating the location-based counting fully convolutional neural network (LC-FCN) model, founded on transfer learning, and the watershed algorithm for identifying dense rice images. This method furnishes reliable baseline data for rice yield estimation and provides researchers with a valuable dataset. Tan et al. (2019) proposed an adhesion rice separation counting algorithm for under-segmented regions, suitable for counting rice grains. This algorithm consists of a watershed algorithm, an improved corner detection algorithm, and a neural network classification algorithm. To quickly and accurately obtain the number of wheat ears in a field, Wang et al. (2022a, b) introduced a method for accurately counting wheat ears in field conditions using FCN and Harris corner detection. The process involves building a dataset with RGB images of wheat ears, training an FCN for segmentation, testing the model, binarizing the results with the Otsu algorithm, and applying Harris corner detection to count the wheat ears. The proposed model achieves high segmentation accuracy (average 0.984), is efficient (requires only 0.033 s for a 256 × 256-pixel image), and performs well under challenging conditions like wheat-ear occlusion and soil disturbance. The counting method also shows excellent results, with an average accuracy of 0.974, R2 of 0.983, and RMSE of 14.043—all metrics represent a 10% improvement over previous methods. Hence, this method provides a reliable technique for wheat phenotyping studies.

Leveraging UAVs or satellite remote sensing imagery for large-scale crop enumeration and yield estimation presents an expeditious method for acquiring comprehensive data, effectively capturing crop growth conditions and production levels across various regions. Malambo et al. (2019) presents an image analysis method using the SegNet deep learning semantic segmentation model to estimate sorghum panicle counts from unmanned aerial system (UAS) images. These counts are crucial for sorghum crop improvement. The model was trained with 462 labeled images (250 × 250) to segment UAS images into sorghum panicles, foliage, and exposed ground, which was then applied to field orthomosaics to generate field-level semantic segmentation. Individual panicle locations were identified by post-processing the segmentation output. Comparisons of model estimates with manually digitized panicle locations in 60 selected plots revealed a detection accuracy of 94%. Spearman correlation between the estimated and reference panicle counts was high, scoring 0.88, while the mean bias reached 0.65. The primary sources of panicle detection errors stemmed from misclassifications during semantic segmentation and mosaicking inaccuracies in the field orthomosaic. Despite these, the method demonstrated promising potential, which could be further enhanced through the collection of more data and comprehensive hyper-parameter tuning.

6 Challenges and prospects

6.1 Agricultural datasets are scarce

Current plant image segmentation datasets are predominantly from single or limited plant varieties, resulting in a lack of diversity. Consequently, the models derived from these datasets struggle to adapt to images of unknown plant species. Compared to datasets of images from other domains, plant image segmentation datasets are smaller in scale, producing models that may be less robust and susceptible to overfitting issues. Agricultural image data present challenges, such as complex backgrounds, varying lighting conditions, and diverse plant postures, which complicate accurate labeling. Crops often grow in intricate environments where other objects or crop types may be present. Agricultural image segmentation technology needs to adapt to different domain applications, imposing higher requirements on the associated datasets. The limitations and challenges of agricultural semantic segmentation datasets remain pronounced, necessitating further in-depth research and exploration to fulfill practical application demands. The limitations inherent in current agricultural semantic segmentation data sets are obvious. Solutions to these challenges may include the deployment of advanced deep learning techniques such as transfer learning, GAN, model architecture, physical information neural networks, and deep synthetic minority oversampling techniques (Alzubaidi et al. 2023). In addition, unsupervised and semi-supervised learning (Sect. 6.2) techniques can improve the performance of small-scale agricultural datasets. Using the above innovative techniques is a good way to overcome the current limitations and meet the practical application needs of agricultural image segmentation.

6.2 Semi-supervised and unsupervised learning

In the case of semantic segmentation, the annotations should be pixel-wise, which is costly to obtain. An alternative to supervised learning is unsupervised learning using the large amount of available unlabeled visual data. Semi-supervised learning (SSL) lies between supervised and unsupervised learning and gives some supervision in addition to unlabeled data, e.g. labeling certain samples (Souly et al. 2017). Unsupervised learning involves constructing models devoid of labeled data, capitalizing on inherent data structures and features. Collecting large-scale data sets in agriculture is difficult and expensive, especially for rare diseases. Data labeling requires a significant investment of time by experts. Semi-supervised and unsupervised learning provides an effective way to solve this problem (Li and Chao 2021). The transfer learning approach is to train a general image segmentation model on a large number of labeled samples, possibly from a common benchmark, and then fine-tune the model on a few samples from some specific target application. Transfer learning can also improve the generalization ability of the model, thus making it more suitable for different types of agricultural images.

6.3 Lightweight network

Lightweight neural networks are characterized by reducing the number of parameters and computational complexity while maintaining strong performance, helping to integrate deep learning techniques into resource-constrained or computationally constrained agricultural equipment. Examples include agricultural robotics, UAV terminals, mobile devices, and embedded systems, which contribute to the proliferation and expansion of intelligent technologies. With the growing ubiquity and demand for smart terminals in the future, lightweight networks will gain increasing significance. In the context of mobile devices, lightweight networks expedite model inference and optimize energy efficiency, thereby enhancing user experiences. For embedded systems, such networks underpin intelligent perception, control, and decision-making capabilities, promoting advancements in the Internet of Things (IoT) and Industrial IoT sectors. While primarily serving smart agriculture applications, lightweight networks possess the versatility to accommodate broader domains, encompassing healthcare, financial services, and urban transportation, ultimately bolstering the development of smart cities and fostering digital transformation. Currently, there are five methods for network lightweight: network pruning, knowledge distillation (Gou et al. 2021), tensor decomposition (Kim et al. 2015), quantization (Wang et al. 2019a, b, c, 2020), and compact convolutional filters (Chen et al. 2023). These approaches focus on reducing computational complexity and memory requirements without significantly impacting performance. They are useful for deploying image segmentation tasks on devices with limited computing resources. This helps reduce inference time and increase efficiency in real-time applications.

6.4 Pretrained foundation model

Pretrained Foundation Models (PFMs) (Zhou et al. 2023) are considered the foundation for various downstream tasks with different data modalities. A PFM is trained on large amounts of data, which provides a reasonable parameter initialization for a wide range of downstream applications. In contrast to previous approaches that use convolutional and recurrent modules to extract features, BERT learns bidirectional encoder representations from transformers trained on large datasets as contextual language models. Similarly, the Generative Pretrained Transformer (GPT) (Bagal et al. 2022) method uses transformers as feature extractors and is trained on large datasets using an autoregressive paradigm. Recently, ChatGPT has shown promising success on large language models, using an autoregressive language model with zero or few shots. The remarkable achievements of PFM have led to significant breakthroughs in various fields of AI in recent years. In the field of segmentation, the Segment Anything model (Kirillov et al. 2023) has been a huge success, even said to be the GPT of CV. Currently, pretrained foundation models have not been applied in agriculture. Pretrained foundation models are suitable for complex patterns and data distribution in agriculture due to their huge network structure and massive parameters. The application of a pretrained foundation model in agriculture image is the future development trend.

6.5 Multimodal image fusion

Most agricultural image segmentation works utilize visible (red–green–blue, RGB) images to perceive scene content. Visible light cameras cannot handle changes in scene lighting and lack the ability to penetrate complex environments. The imaging mechanism of visible light cameras makes it challenging to capture sufficient and effective scene information in poor lighting conditions and adverse weather. Additionally, visible light cameras cannot effectively deal with complex scenes featuring similar target appearances, multiple scene areas, and significant changes. Depth cameras provide the physical distance of objects from the camera's photo center in the imaging scene, while thermal infrared imagers reflect the thermal radiation characteristics of objects with temperatures above absolute zero in various lighting and weather conditions, providing accurate target contour and semantic information. Multi-modal image semantic segmentation leverages the complementary characteristics between different modal images through fusion, enhancing the segmentation model's learning and reasoning abilities in complex scenes (Zhang et al. 2021). This method has been successfully applied in remote sensing (Jin et al. 2023), automatic driving (Huang et al. 2021a, b, c), medical imaging (Zhu et al. 2023), and other fields. In agriculture, identifying plant diseases, weed infestation, and yield estimation can be done more effectively through multimodal image fusion. By combining different types of images, models can better understand complex agricultural scenarios, thus improving decision-making in crop management, pest control, and yield forecasting.

7 Conclusion

Image segmentation is a crucial technology for practitioners to comprehensively analyze and understand vegetation and diseases in farmland. This technology helps to improve the quality of crop growth, increase yields, reduce costs, and ultimately achieve sustainable agricultural production. The applications of image segmentation technology in agriculture are comprehensively reviewed in this paper. We categorize image segmentation solutions based on deep learning into eight categories and discuss their specific applications in agriculture, including disease detection, weed identification, crop growth monitoring, and yield estimation. Despite the wide development of image segmentation, the latest advances in this field have found limited application in agriculture. Most research continues to focus on earlier achievements such as U-Net, Deeplab, and other established methods. This study aims to assist researchers in selecting appropriate approaches for specific agricultural applications by offering insights and applications regarding the effectiveness of various models. Additionally, this paper presents the first comprehensive summary and analysis of the most widely used plant image segmentation datasets, providing direction and reference for related fields. By analyzing public plant image segmentation datasets, researchers can better understand the commonalities and disparities between different datasets, thereby facilitating the sharing of resources and experience among researchers. However, the current public datasets have limitations, such as a lack of diversity and a limited number of images, only containing plant images of a few species. Therefore, there is an urgent need for a larger dataset with a wider range of species in plant image segmentation. Finally, this paper discusses some open challenges and promising research directions for deep learning-based plant image segmentation in the forthcoming years.

Acknowledgements This work is supported by the Applied Basic Research Key Project of Yunnan (No. 202201AS070034), Yunnan Science and Technology Talent and Platform Program (No. 202305AM070006), Yunnan Fundamental Research Projects (NO. 202401AU070159), Yunnan Fundamental Research Projects (NO. 202301AV070003), Natural Science Foundation of China (NO. 42067029). We thank all research participants for their assistance and participation in this study.

Author contributions L.Y and T.S conceived and designed the framework of review; L.L wrote the main manuscript text; R.W and C.F wrote and modified the paper. All authors reviewed the manuscript.

Declarations

Competing interests The authors declare no competing interests.

References

Abdu AM, Mohd Mokji M, Sheikh UU, Khalil K (2019) Automatic disease symptoms segmentation optimized for dissimilarity feature extraction in digital photographs of plant leaves. 2019 IEEE 15th international colloquium on signal processing & its applications (CSPA). pp 60–64. https://doi.org/10.1109/CSPA.2019.8696049

Akiva P, Planche B, Roy A, Dana K, Oudemans P, Mars M (2021) Ai on the bog: Monitoring and evaluating cranberry crop risk. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2493–2502

Alimanov A, Islam MB (2023) Denoising diffusion probabilistic model for retinal image generation and segmentation. 2023 IEEE international conference on computational photography (ICCP). pp 1–12.https://doi.org/10.1109/ICCP56744.2023.10233841

Alkhudaydi T, De La lglesia B (2022) Counting spikelets from infield wheat crop images using fully convolutional networks. Neural Comput Appl 34(20):17539–17560. https://doi.org/10.1007/s00521-022-07392-1

Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-dabbagh BSN, Fadhel MA, Manoufali M, Zhang J, Al-Timemy AH, Duan Y, Abdullah A, Farhan L, Lu Y, Gupta A, Albu F, Abbosh A, Gu Y (2023) A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. J Big Data 10(1):46. https://doi.org/10.1186/s40537-023-00727-2

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. Proc Mach Learn Res 70:214–223

Asgari Taghanaki S, Abhishek K, Cohen JP, Cohen-Adad J, Hamarneh G (2021) Deep semantic segmentation of natural and medical images: a review. Artif Intell Rev 54:137–178

Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Bagal V, Aggarwal R, Vinod PK, Priyakumar UD (2022) MolGPT: molecular generation using a transformer-decoder model. J Chem Inf Model 62(9):2064–2076. https://doi.org/10.1021/acs.jcim.1c00600

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

Boykov YY, Jolly M-P (2001) Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images.Proc Eighth IEEE Int Conf Comput Vis ICCV 2001 1:105–112. https://doi.org/10.1109/ICCV.2001.937505

Buckner E, Tong H, Ottley C, Williams C (2021) High-throughput image segmentation and machine learning approaches in the plant sciences across multiple scales. Emerg Topics Life Sci 5(2):239–248. https://doi.org/10.1042/ETLS20200273

Cap QH, Uga H, Kagiwada S, Iyatomi H (2022) LeafGAN: an effective data augmentation method for practical plant disease diagnosis. IEEE Trans Autom Sci Eng 19(2):1258–1267. https://doi.org/10.1109/TASE.2020.3041499

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27. https://doi.org/10.1145/1961189.1961199

Chaurasia A, Culurciello E (2017) LinkNet: exploiting encoder representations for efficient semantic segmentation. 2017 IEEE visual communications and image processing (VCIP). pp. 1–4.https://doi.org/10.1109/VCIP.2017.8305148

Chen F, Li S, Han J, Ren F, Yang Z (2023) Review of lightweight deep convolutional neural networks. Arch Comput Methods Eng. https://doi.org/10.1007/s11831-023-10032-z

Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y (2021) Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. Comput Sci (4):357–361

Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018a) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen LC, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587

Chen LC, Yang Y, Wang J, Xu W, Yuille AL (2016) Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3640–3649

Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018b) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision – ECCV 2018, vol 11211. Springer International Publishing, pp 833–851. https://doi.org/10.1007/978-3-030-01234-2_49

Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018c) Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), pp 801–818

Chen S, Zhang K, Zhao Y, Sun Y, Ban W, Chen Y, Zhuang H, Zhang X, Liu J, Yang T (2021b) An approach for rice bacterial leaf streak disease segmentation and disease severity estimation. Agriculture 11(5):420. https://doi.org/10.3390/agriculture11050420

Chen T, Lin L, Wu X, Xiao N, Luo X (2018d) Learning to segment object candidates via recursive neural networks. IEEE Trans Image Process 27(12):5827–5839. https://doi.org/10.1109/TIP.2018.2859025

Chen X, Girshick R, He K, Dollar P (2019) TensorMask: A foundation for dense object segmentation. 2019 IEEE/CVF international conference on computer vision (ICCV). pp 2061–2069.https://doi.org/10.1109/ICCV.2019.00215

Cheng Y, Cai R, Li Z, Zhao X, Huang K (2017) Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation. 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 1475–1483.https://doi.org/10.1109/CVPR.2017.161

Chollet F (2017) Xception: deep learning with depthwise separable convolutions. 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 1800–1807.https://doi.org/10.1109/CVPR.2017.195

Chouhan SS, Kaul A, Singh UP (2019) Radial basis function neural network for the segmentation of plant leaf disease. 2019 4th international conference on information systems and computer networks (ISCON). pp 713–716. https://doi.org/10.1109/ISCON47742.2019.9036299

Chouhan SS, Kaul A, Sinzlr UP (2019) Plants leaf segmentation using bacterial foraging optimization algorithm. 2019 international conference on communication and electronics systems (ICCES). pp 1500–1505.https://doi.org/10.1109/ICCES45898.2019.9002039

Christensen S, Søgaard HT, Kudsk P, Nørremark M, Lund I, Nadimi ES, Jørgensen R (2009) Site-specific weed control technologies. Weed Res 49(3):233–241. https://doi.org/10.1111/j.1365-3180.2009.00696.x

Chuang Y, Zhang S, Zhao X (2023) Deep learning-based panoptic segmentation: recent advances and perspectives. IET Image Proc 17(10):2807–2828. https://doi.org/10.1049/ipr2.12853

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3223

Cruz JA, Yin X, Liu X, Imran SM, Morris DD, Kramer DM, Chen J (2016) Multi-modality imagery database for plant phenotyping. Mach vis Appl 27(5):735–749. https://doi.org/10.1007/s00138-015-0734-6

Dai J, He K, Ren S, Sun, jian. (2016) Instance-sensitive fully convolutional networks, vol 9910. Springer International Publishing. https://doi.org/10.1007/978-3-319-46466-4

Dai J, He K, Sun J (2016b) Instance-aware semantic segmentation via multi-task network cascades. 2016 IEEE conference on computer vision and pattern recognition (CVPR). pp 3150–3158.https://doi.org/10.1109/CVPR.2016.343

Dai W, Dong N, Wang Z, Liang X, Zhang H, Xing EP (2018) SCAN: Structure Correcting Adversarial Network for Organ Segmentation in Chest X-Rays. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, Tavares JMRS, Bradley A, Papa JP, Belagiannis V, Nascimento JC, Lu Z, Conjeti S, Moradi M, Greenspan H, Madabhushi A (eds) Deep learning in medical image analysis and multimodal learning for clinical decision support, vol 11045. Springer International Publishing, pp 263–273. https://doi.org/10.1007/978-3-030-00889-5_30

Das M, Bais A (2021) DeepVeg: deep learning model for segmentation of weed, canola, and canola flea beetle damage. IEEE Access 9:119367–119380. https://doi.org/10.1109/ACCESS.2021.3108003

De Brabandere B, Neven D, Van Gool L (2017) Semantic instance segmentation with a discriminative loss function. arXiv preprint arXiv:1708.02551

Dhanachandra N, Manglem K, Chanu YJ (2015) Image segmentation using K -means clustering algorithm and subtractive clustering algorithm. Procedia Comput Sci 54:764–771. https://doi.org/10.1016/j.procs.2015.06.090

Ding H, Jiang X, Shuai B, Liu AQ, Wang G (2018) Context contrasted feature and gated multi-scale aggregation for scene segmentation. 2018 IEEE/CVF conference on computer vision and pattern recognition. pp 2393–2402. https://doi.org/10.1109/CVPR.2018.00254

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T., Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Douarre C, Crispim-Junior CF, Gelibert A, Tougne L, Rousseau D (2019) Novel data augmentation strategies to boost supervised segmentation of plant disease. Comput Electron Agric 165:104967

Elbasi E, Mostafa N, AlArnaout Z, Zreikat AI, Cina E, Varghese G, Shdefat A, Topcu AE, Abdelbaki W, Mathew S, Zaki C (2023) Artificial intelligence technology in the agricultural sector: a systematic literature review. IEEE Access 11:171–202. https://doi.org/10.1109/ACCESS.2022.3232485

Espejo-Garcia B, Mylonas N, Athanasakos L, Vali E, Fountas S (2021) Combining generative adversarial networks and agricultural transfer learning for weeds identification. Biosys Eng 204:79–89. https://doi.org/10.1016/j.biosystemseng.2021.01.014

Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (VOC) challenge. Int J Comput Vision 88(2):303–338. https://doi.org/10.1007/s11263-009-0275-4

Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 3141–3149.https://doi.org/10.1109/CVPR.2019.00326

Fu J, Liu J, Wang Y, Zhou J, Wang C, Lu H (2019) Stacked deconvolutional network for semantic segmentation. IEEE Trans Image Process 1–1. https://doi.org/10.1109/TIP.2019.2895460

Fuentes-Pacheco J, Torres-Olivares J, Roman-Rangel E, Cervantes S, Juarez-Lopez P, Hermosillo-Valadez J, Rendón-Mancha JM (2019) Fig plant segmentation from aerial images using a deep convolutional encoder-decoder network. Remote Sens 11(10):1157. https://doi.org/10.3390/rs11101157

Fukuda M, Okuno T, Yuki S (2021) Central object segmentation by deep learning to continuously monitor fruit growth through RGB images. Sensors 21(21):6999. https://doi.org/10.3390/s21216999

Ghiasi G, Fowlkes CC (2016) Laplacian pyramid reconstruction and refinement for semantic segmentation. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision – ECCV 2016, vol 9907. Springer International Publishing, pp 519–534. https://doi.org/10.1007/978-3-319-46487-9_32

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144. https://doi.org/10.1145/3422622

Gou J, Yu B, Maybank SJ, Tao D (2021) Knowledge distillation: a survey. Int J Comput Vision 129(6):1789–1819. https://doi.org/10.1007/s11263-021-01453-z

Gu W, Bai S, Kong L (2022) A review on 2D instance segmentation based on deep neural networks. Image Vis Comput 120:104401. https://doi.org/10.1016/j.imavis.2022.104401

Guan S, Khan AA, Sikdar S, Chitnis PV (2020) Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal. IEEE J Biomed Health Inform 24(2):568–576. https://doi.org/10.1109/JBHI.2019.2912935

Guo MH, Lu CZ, Hou Q, Liu Z, Cheng MM, Hu SM (2022) Segnext: Rethinking convolutional attention design for semantic segmentation. Adv Neural Inf Process Syst 35:1140–1156

Guo Y, Liu Y, Georgiou T, Lew MS (2018) A review of semantic segmentation using deep neural networks. Int J Multimed Inform Retr 7(2):87–93. https://doi.org/10.1007/s13735-017-0141-z

Hamuda E, Glavin M, Jones E (2016) A survey of image processing techniques for plant extraction and segmentation in the field. Comput Electron Agric 125:184–199. https://doi.org/10.1016/j.compag.2016.04.024

Hariharan B, Arbeláez P, Girshick R, Malik J (2014) Simultaneous detection and segmentation. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer vision – ECCV 2014, vol 8695. Springer International Publishing, pp 297–312. https://doi.org/10.1007/978-3-319-10584-0_20

Hasan RI, Yusuf SM, Alzubaidi L (2020) Review of the state of the art of deep learning for plant diseases: a broad analysis and discussion. Plants 9(10):1302. https://doi.org/10.3390/plants9101302

Hasan RI, Yusuf SM, Mohd Rahim MS, Alzubaidi L (2022) Automated masks generation for coffee and apple leaf infected with single or multiple diseases-based color analysis approaches. Inform Med Unlocked 28:100837. https://doi.org/10.1016/j.imu.2021.100837

Hasan RI, Yusuf SM, Mohd Rahim MS, Alzubaidi L (2023) Automatic clustering and classification of coffee leaf diseases based on an extended kernel density estimation approach. Plants 12(8):1603. https://doi.org/10.3390/plants12081603

He J, Deng Z, Qiao Y (2019a) Dynamic multi-scale filters for semantic segmentation. 2019 IEEE/CVF international conference on computer vision (ICCV). pp 3561–3571.https://doi.org/10.1109/ICCV.2019.00366

He J, Deng Z, Zhou L, Wang Y, Qiao Y (2019b) Adaptive pyramid context network for semantic segmentation. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 7511–7520.https://doi.org/10.1109/CVPR.2019.00770

He K, Gkioxari G, Dollar P, Girshick R (2017a) Mask R-CNN. 2017 IEEE international conference on computer vision (ICCV). pp 2980–2988.https://doi.org/10.1109/ICCV.2017.322

He K, Gkioxari G, Dollár P, Girshick R (2017b) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. 2016 IEEE conference on computer vision and pattern recognition (CVPR). pp 770–778.https://doi.org/10.1109/CVPR.2016.90

Hu H, Cui J, Zha H (2021a) Boundary-aware graph convolution for semantic segmentation. 2020 25th international conference on pattern recognition (ICPR). pp 1828–1835. https://doi.org/10.1109/ICPR48806.2021.9412034

Hu J, Cao L, Lu Y, Zhang S, Wang Y, Li K, Huang F, Shao L, Ji R (2021b) Istr: End-to-end instance segmentation with transformers. arXiv preprint arXiv:2105.00637

Hu Y, Chen Z, Lin W (2018) RGB-D semantic segmentation: a review. 2018 IEEE international conference on multimedia & expo workshops (ICMEW). pp 1–6.https://doi.org/10.1109/ICMEW.2018.8551554

Huang H, Lin L, Zhang Y, Xu Y, Zheng J, Mao X, Qian X, Peng Z, Zhou J, Chen Y-W, Tong R (2021a) Graph-BAS ^{3} Net: boundary-aware semi-supervised segmentation network with bilateral graph convolution. 2021 IEEE/CVF international conference on computer vision (ICCV). pp 7366–7375.https://doi.org/10.1109/ICCV48922.2021.00729

Huang H, Yang A, Tang Y, Zhuang J, Hou C, Tan Z, Dananjayan S, He Y, Guo Q, Luo S (2021b) Deep color calibration for UAV imagery in crop monitoring using semantic style transfer with local to global attention. Int J Appl Earth Obs Geoinf 104:102590. https://doi.org/10.1016/j.jag.2021.102590

Huang Z, Huang L, Gong Y, Huang C, Wang X (2019) Mask scoring R-CNN. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 6402–6411.https://doi.org/10.1109/CVPR.2019.00657

Huang Z, Lv C, Xing Y, Wu J (2021c) Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. IEEE Sens J 21(10):11781–11790. https://doi.org/10.1109/JSEN.2020.3003121

Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W (2019) CCNet: criss-cross attention for semantic segmentation. 2019 IEEE/CVF international conference on computer vision (ICCV). pp 603–612.https://doi.org/10.1109/ICCV.2019.00069

Ikonomatakis N, Plataniotis KN, Zervakis M, Venetsanopoulos AN (1997)Region growing and region merging image segmentation. Proc 13th Int Conf Digit Sig Process 1:299–302. https://doi.org/10.1109/ICDSP.1997.628077

Jin J, Zhou W, Yang R, Ye L, Yu L (2023) Edge detection guide network for semantic segmentation of remote-sensing images. IEEE Geosci Remote Sens Lett 20:1–5. https://doi.org/10.1109/LGRS.2023.3234257

Kamal S, Shende VG, Swaroopa K, Bindhu Madhavi P, Akram PS, Pant K, Patil SD, Sahile K (2022) FCN network-based weed and crop segmentation for IoT-aided agriculture applications. Wirel Commun Mob Comput 2022:1–10. https://doi.org/10.1155/2022/2770706

Kang J, Liu L, Zhang F, Shen C, Wang N, Shao L (2021) Semantic segmentation model of cotton roots in-situ image based on attention mechanism. Comput Electron Agric 189:106370. https://doi.org/10.1016/j.compag.2021.106370

Kaur P, Harnal S, Tiwari R, Alharithi FS, Almulihi AH, Noya ID, Goyal N (2021) A hybrid convolutional neural network model for diagnosis of COVID-19 using chest X-ray images. Int J Environ Res Public Health 18(22):12191. https://doi.org/10.3390/ijerph182212191

Kierdorf J, Junker-Frohn LV, Delaney M, Olave MD, Burkart A, Jaenicke H, Muller O, Rascher U, Roscher R (2022) GrowliFlower: an image time-series dataset for growth analysis of cauliflower. J Field Robot rob.22122. https://doi.org/10.1002/rob.22122

Kim YH, Park KR (2022) MTS-CNN: multi-task semantic segmentation-convolutional neural network for detecting crops and weeds. Comput Electron Agric 199:107146. https://doi.org/10.1016/j.compag.2022.107146

Kim YD, Park E, Yoo S, Choi T, Yang L, Shin D (2015) Compression of deep convolutional neural networks for fast and low power mobile applications. arXiv preprint arXiv:1511.06530

Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR)

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollár P, Girshick R (2023) Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4015–4026

Krishnaswamy Rangarajan A, Purushothaman R (2020) A vision based crop monitoring system using segmentation techniques. Adv Electr Comput Eng 20(2):89–100. https://doi.org/10.4316/AECE.2020.02011

Kumar N, Belhumeur PN, Biswas A, Jacobs DW, Kress WJ, Lopez IC, Soares JVB (2012) Leafsnap: a computer vision system for automatic plant species identification. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (eds) Computer vision – ECCV 2012, vol 7573. Springer Berlin Heidelberg, pp 502–516. https://doi.org/10.1007/978-3-642-33709-3_36

Lan Y, Huang K, Yang C, Lei L, Ye J, Zhang J, Zeng W, Zhang Y, Deng J (2021) Real-time identification of rice weeds by UAV low-altitude remote sensing based on improved semantic segmentation model. Remote Sens 13(21):4370. https://doi.org/10.3390/rs13214370

Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791

Lee Y, Park J (2020) CenterMask: real-time anchor-free instance segmentation. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 13903–13912.https://doi.org/10.1109/CVPR42600.2020.01392

Li C, Welling M, Zhu J, Zhang B (2018) Graphical generative adversarial networks. Advances in neural information processing systems 31

Li D, Yang J, Kreis K, Torralba A, Fidler S (2021) Semantic segmentation with generative models: Semisupervised learning and strong out-of-domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8300–8311

Li H, Xiong P, Fan H, Sun J (2019a) DFANet: deep feature aggregation for real-time semantic segmentation. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 9514–9523.https://doi.org/10.1109/CVPR.2019.00975

Li X, Yang Y, Zhao Q, Shen T, Lin Z, Liu H (2020) Spatial Pyramid Based Graph Reasoning for Semantic Segmentation. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 8947–8956.https://doi.org/10.1109/CVPR42600.2020.00897

Li X, Zhong Z, Wu J, Yang Y, Lin Z, Liu H (2019b). Expectation-maximization attention networks for semantic segmentation. 2019 IEEE/CVF international conference on computer vision (ICCV). pp 9166–9175.https://doi.org/10.1109/ICCV.2019.00926

Li X, Zhou Y, Liu J, Wang L, Zhang J, Fan X (2022a) The detection method of potato foliage diseases in complex background based on instance segmentation and semantic segmentation. Front Plant Sci 13:899754. https://doi.org/10.3389/fpls.2022.899754

Li Y, Chao X (2021) Semi-supervised few-shot learning approach for plant diseases recognition. Plant Methods 17(1):68. https://doi.org/10.1186/s13007-021-00770-1

Li Y, Qi H, Dai J, Ji X, Wei Y (2017) Fully convolutional instance-aware semantic segmentation. 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 4438–4446.https://doi.org/10.1109/CVPR.2017.472

Li Z, Chen P, Shuai L, Wang M, Zhang L, Wang Y, Mu J (2022b) A copy paste and semantic segmentation-based approach for the classification and assessment of significant rice diseases. Plants 11(22):3174. https://doi.org/10.3390/plants11223174

Li Z, Wang W, Xie E, Yu Z, Anandkumar A, Alvarez JM, Luo P, Lu T (2022c) Panoptic SegFormer: delving deeper into panoptic segmentation with transformers. 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 1270–1279.https://doi.org/10.1109/CVPR52688.2022.00134

Lin D, Ji Y, Lischinski D, Cohen-Or D, Huang H (2018) Multi-scale context intertwining for semantic segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision – ECCV 2018, vol 11207. Springer International Publishing, pp 622–638. https://doi.org/10.1007/978-3-030-01219-9_37

Lin G, Milan A, Shen C, Reid I (2017a) RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 5168–5177.https://doi.org/10.1109/CVPR.2017.549

Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S (2017b) Feature pyramid networks for object detection. 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 936–944. https://doi.org/10.1109/CVPR.2017.106

Liu J, Wang X (2021) Plant diseases and pests detection based on deep learning: a review. Plant Methods 17(1):22. https://doi.org/10.1186/s13007-021-00722-9

Liu K, Ye Z, Guo H, Cao D, Chen L, Wang F-Y (2021a) FISS GAN: a generative adversarial network for foggy image semantic segmentation. IEEE/CAA J Autom Sin 8(8):1428–1439. https://doi.org/10.1109/JAS.2021.1004057

Liu M, Schonfeld D, Tang W (2021b) Exploit visual dependency relations for semantic segmentation. 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 9721–9730.https://doi.org/10.1109/CVPR46437.2021.00960

Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. 2018 IEEE/CVF conference on computer vision and pattern recognition. pp 8759–8768.https://doi.org/10.1109/CVPR.2018.00913

Liu X, He W, Zhang H (2023) Cross-region plastic greenhouse segmentation and counting using the style transfer and dual-task networks. Comput Electron Agric 207:107766. https://doi.org/10.1016/j.compag.2023.107766

Liu X, Song L, Liu S, Zhang Y (2021c) A review of deep-learning-based medical image segmentation methods. Sustainability 13(3):1224. https://doi.org/10.3390/su13031224

Lu Y, Chen D, Olaniyi E, Huang Y (2022) Generative adversarial networks (GANs) for image augmentation in agriculture: a systematic review. Comput Electron Agric 200:107208. https://doi.org/10.1016/j.compag.2022.107208

Lu Y, Chen Y, Zhao D, Chen J (2019) Graph-FCN for image semantic segmentation. In International symposium on neural networks. Cham: Springer International Publishing, pp 97–105

Luc P, Couprie C, Chintala S, Verbeek J (2016) Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408

Luo Z, Yang W, Yuan Y, Gou R, Li X (2023) Semantic segmentation of agricultural images: a survey. Inform Process Agric S2214317323000112. https://doi.org/10.1016/j.inpa.2023.02.001

Ma H, Lin X, Yu Y (2024) I2F: a unified image-to-feature approach for domain adaptive semantic segmentation. IEEE Trans Pattern Anal Mach Intell 46(3):1695–1710. https://doi.org/10.1109/TPAMI.2022.3229207

Maheswari P, Raja P, Apolo-Apolo OE, Pérez-Ruiz M (2021) Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques—a review. Front Plant Sci 12:684328. https://doi.org/10.3389/fpls.2021.684328

Malambo L, Popescu S, Ku N-W, Rooney W, Zhou T, Moore S (2019) A deep learning semantic segmentation-based approach for field-level sorghum panicle counting. Remote Sens 11(24):2939. https://doi.org/10.3390/rs11242939

Michieli U, Borsato E, Rossi L, Zanuttigh P (2020) GMNet: graph matching network for large scale part semantic segmentation in the wild. In: Vedaldi A, Bischof H, Brox T, Frahm J-M (eds) Computer vision – ECCV 2020, vol 12353. Springer International Publishing, pp 397–414. https://doi.org/10.1007/978-3-030-58598-3_24

Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D (2021) Image segmentation using deep learning: a survey. IEEE transactions on pattern analysis and machine intelligence. pp 1–1. https://doi.org/10.1109/TPAMI.2021.3059968

Minervini M, Abdelsamea MM, Tsaftaris SA (2014) Image-based plant phenotyping with incremental learning and active contours. Eco Inform 23:35–48. https://doi.org/10.1016/j.ecoinf.2013.07.004

Minervini M, Fischbach A, Scharr H, Tsaftaris SA (2016) Finely-grained annotated datasets for image-based plant phenotyping. Pattern Recogn Lett 81:80–89. https://doi.org/10.1016/j.patrec.2015.10.013

Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784

Mishra AM, Harnal S, Gautam V, Tiwari R, Upadhyay S (2022) Weed density estimation in soya bean crop using deep convolutional neural networks in smart agriculture. J Plant Dis Prot 129(3):593–604. https://doi.org/10.1007/s41348-022-00595-7

Mnih V, Heess N, Graves A (2014) Recurrent models of visual attention. Adv Neural Inf Process Syst 27

Mottaghi R, Chen X, Liu X, Cho N-G, Lee S-W, Fidler S, Urtasun R, Yuille A (2014) The role of context for object detection and semantic segmentation in the wild. 2014 IEEE conference on computer vision and pattern recognition. pp 891–898.https://doi.org/10.1109/CVPR.2014.119

Nagaraju M, Chawla P, Upadhyay S, Tiwari R (2022) Convolution network model based leaf disease detection using augmentation techniques. Expert Syst 39(4):e12885. https://doi.org/10.1111/exsy.12885

Nasiri A, Omid M, Taheri-Garavand A, Jafari A (2022) Deep learning-based precision agriculture through weed recognition in sugar beet fields. Sustain Comput: Inform Syst 35:100759. https://doi.org/10.1016/j.suscom.2022.100759

Nerkar B, Talbar S (2021) Cross-dataset learning for performance improvement of leaf disease detection using reinforced generative adversarial networks. Int J Inf Technol 13(6):2305–2312. https://doi.org/10.1007/s41870-021-00772-1

Neubeck A, Van Gool L (2006) Efficient non-maximum suppression. 18th international conference on pattern recognition (ICPR'06). pp 850–855. https://doi.org/10.1109/ICPR.2006.479

Nong C, Fan X, Wang J (2022) Semi-supervised learning for weed and crop segmentation using UAV imagery. Front Plant Sci 13:927368. https://doi.org/10.3389/fpls.2022.927368

Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66. https://doi.org/10.1109/TSMC.1979.4310076

Pan S-Y, Lu C-Y, Lee S-P, Peng W-H (2021) Weakly-supervised image semantic segmentation using graph convolutional networks. 2021 IEEE international conference on multimedia and expo (ICME). pp 1–6.https://doi.org/10.1109/ICME51207.2021.9428116

Pei H, Owari T, Tsuyuki S, Zhong Y (2023) Application of a novel multiscale global graph convolutional neural network to improve the accuracy of forest type classification using aerial photographs. Remote Sens 15(4):1001. https://doi.org/10.3390/rs15041001

Pei J, Cheng T, Fan DP, Tang H, Chen C, Van Gool L (2022) Osformer: One-stage camouflaged instance segmentation with transformers. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland, pp 19–37

Peláez-Vegas A, Mesejo P, Luengo J (2023) A survey on semi-supervised semantic segmentation. arXiv preprint arXiv:2302.09899

Peng C, Zhang X, Yu G, Luo G, Sun J (2017) Large kernel matters—improve semantic segmentation by global convolutional network. 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 1743–1751.https://doi.org/10.1109/CVPR.2017.189

Pradhan KS, Chawla P, Tiwari R (2023) HRDEL: high ranking deep ensemble learning-based lung cancer diagnosis model. Expert Syst Appl 213:118956. https://doi.org/10.1016/j.eswa.2022.118956

Longzhe Q, Enchen J (2011) Automatic segmentation method of touching corn kernels in digital image based on improved watershed algorithm. Int Conf New Technol Agric 2011:34–37. https://doi.org/10.1109/ICAE.2011.5943743

Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434

Ren S, He K, Girshick R, Sun J (2016) Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing. pp, 234–241

Rosenfeld A (1981) The max Roberts operator is a Hueckel-type edge detector. IEEE Trans Pattern Anal Mach Intell PAMI 3(1):101–103. https://doi.org/10.1109/TPAMI.1981.4767056

Ru L, Zhan Y, Yu B, Du B (2022) Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 16825–16834. https://doi.org/10.1109/CVPR52688.2022.01634

Saleem R, Hussain Shah J, Sharif M, Jillani Ansari G (2021) Mango leaf disease identification using fully resolution convolutional network. Comput Mater Continua 69(3):3581–3601. https://doi.org/10.32604/cmc.2021.017700

Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G (2009) The graph neural network model. IEEE Trans Neural Netw 20(1):61–80. https://doi.org/10.1109/TNN.2008.2005605

Seguí S, Pujol O, Vitria J (2015) Learning to count with deep object features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 90–96

Shao H, Tang R, Lei Y, Mu J, Guan Y, Xiang Y (2021) Rice ear counting based on image segmentation and establishment of a dataset. Plants 10(8):1625. https://doi.org/10.3390/plants10081625

Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651. https://doi.org/10.1109/TPAMI.2016.2572683

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Sodjinou SG, Mohammadi V, Sanda Mahama AT, Gouton P (2022) A deep semantic segmentation-based algorithm to segment crops and weeds in agronomic color images. Inform Process Agric 9(3):355–364. https://doi.org/10.1016/j.inpa.2021.08.003

Solanki S, Singh UP, Chouhan SS (2023a) Brain tumor classification using ML and DL approaches. 2023 IEEE 5th international conference on cybernetics, cognition and machine learning applications (ICCCMLA). pp 204–208. https://doi.org/10.1109/ICCCMLA58983.2023.10346854

Solanki S, Singh UP, Chouhan SS, Jain S (2023b) A systematic analysis of magnetic resonance images and deep learning methods used for diagnosis of brain tumor. Multimed Tools Appl 83(8):23929–23966. https://doi.org/10.1007/s11042-023-16430-6

Souly N, Spampinato C, Shah M (2017) Semi supervised semantic segmentation using generative adversarial network. 2017 IEEE International conference on computer vision (ICCV). pp 5689–5697.https://doi.org/10.1109/ICCV.2017.606

Strudel R, Garcia R, Laptev I, Schmid C (2021) Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7262–7272

Szegedy C, Wei L, Yangqing J, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. 2015 IEEE conference on computer vision and pattern recognition (CVPR). pp 1–9.https://doi.org/10.1109/CVPR.2015.7298594

Tan S, Ma X, Mai Z, Qi L, Wang Y (2019) Segmentation and counting algorithm for touching hybrid rice grains. Comput Electron Agric 162:493–504. https://doi.org/10.1016/j.compag.2019.04.030

Trinh NH, O'Brien D (2020) Semi-supervised learning with generative adversarial networks for pathological speech classification. 2020 31st Irish signals and systems conference (ISSC). pp 1–5. https://doi.org/10.1109/ISSC49989.2020.9180211

Uchiyama H, Sakurai S, Mishima M, Arita D, Okayasu T, Shimada A, Taniguchi R (2017) An easy-to-setup 3D phenotyping platform for KOMATSUNA dataset. 2017 IEEE international conference on computer vision workshops (ICCVW). pp 2038–2045.https://doi.org/10.1109/ICCVW.2017.239

Ullah HS, Asad MH, Bais A (2021) End to end segmentation of canola field images using dilated u-net. IEEE Access 9:59741–59753. https://doi.org/10.1109/ACCESS.2021.3073715

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30

Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y (2017) Graph attention networks. arXiv preprint arXiv:1710.10903

Wang A, Xu Y, Wei X, Cui B (2020a) Semantic segmentation of crop and weed using an encoder-decoder network and image enhancement method under uncontrolled outdoor illumination. IEEE Access 8:81724–81734. https://doi.org/10.1109/ACCESS.2020.2991354

Wang A, Zhang W, Wei X (2019a) A review on weed detection using ground-based machine vision and image processing techniques. Comput Electron Agric 158:226–240. https://doi.org/10.1016/j.compag.2019.02.005

Wang D, Cao W, Zhang F, Li Z, Xu S, Wu X (2022a) A review of deep learning in multiscale agricultural sensing. Remote Sens 14(3):559. https://doi.org/10.3390/rs14030559

Wang D, Fu Y, Yang G, Yang X, Liang D, Zhou C, Zhang N, Wu H, Zhang D (2019b) Combined use of FCN and Harris corner detection for counting wheat ears in field conditions. IEEE Access 7:178930–178941. https://doi.org/10.1109/ACCESS.2019.2958831

Wang D, Zhang D, Yang G, Xu B, Luo Y, Yang X (2022b) SSRNet: in-field counting wheat ears using multi-stage convolutional neural network. IEEE Trans Geosci Remote Sens 60:1–11. https://doi.org/10.1109/TGRS.2021.3093041

Wang H, Zhu Y, Adam H, Yuille A, Chen LC (2021a) Max-deeplab: End-to-end panoptic segmentation with mask transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5463–5474

Wang K, Liu Z, Lin Y, Lin J, Han S (2019c) HAQ: hardware-aware automated quantization with mixed precision. 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 8604–8612.https://doi.org/10.1109/CVPR.2019.00881

Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Cottrell G (2018) Understanding convolution for semantic segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV). Ieee, pp 1451–1460

Wang S, Gong Y, Xing J, Huang L, Huang C, Hu W (2020) Rdsnet: A new deep architecture forreciprocal object detection and instance segmentation. In Proceedings of the AAAI conference on artificial intelligence 34(07):12208–12215

Wang X, Kong T, Shen C, Jiang Y, Li L (2020) SOLO: segmenting objects by locations. arXivhttp://arxiv.org/abs/1912.04488

Wang X, Wang S, Ning C, Zhou H (2021b) Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification. IEEE Trans Geosci Remote Sens 59(9):7918–7932. https://doi.org/10.1109/TGRS.2020.3044655

Wang Z, Zhang S (2018) Segmentation of corn leaf disease based on fully convolution neural network. Acad J Comput Inform Sci 1(1). https://doi.org/10.25236/AJCIS.010002

Weyler J, Quakernack J, Lottes P, Behley J, Stachniss C (2022) Joint plant and leaf instance segmentation on field-scale UAV imagery. IEEE Robot Autom Lett 7(2):3787–3794. https://doi.org/10.1109/LRA.2022.3147462

Wu J, Jiang Y, Bai S, Zhang W, Bai X (2022a) SeqFormer: sequential transformer for video instance segmentation. arXiv http://arxiv.org/abs/2112.08275

Wu J, Wen C, Chen H, Ma Z, Zhang T, Su H, Yang C (2022b) DS-DETR: a model for tomato leaf disease segmentation and damage evaluation. Agronomy 12(9):2023. https://doi.org/10.3390/agronomy12092023

Wu T, Lu Y, Zhu Y, Zhang C, Wu M, Ma Z, Guo G (2020) GINet: Graph interaction network for scene parsing. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. Springer International Publishing, pp 34–51

Wu Y, Jiang J, Huang Z, Tian Y (2022c) FPANet: feature pyramid aggregation network for real-time semantic segmentation. Appl Intell 52(3):3319–3336. https://doi.org/10.1007/s10489-021-02603-z

Xiao T, Liu Y, Zhou B, Jiang Y, Sun J (2018) Unified perceptual parsing for scene understanding. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision – ECCV 2018, vol 11209. Springer International Publishing, pp 432–448. https://doi.org/10.1007/978-3-030-01228-1_26

Xiao X, Lian S, Luo Z, Li S (2018b) Weighted res-UNet for high-quality retina vessel segmentation. 2018 9th international conference on information technology in medicine and education (ITME). pp 327–331. https://doi.org/10.1109/ITME.2018.00080

Xie E, Sun P, Song X, Wang W, Liu X, Liang D, Shen C, Luo P (2020) PolarMask: single shot instance segmentation with polar representation. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 12190–12199.https://doi.org/10.1109/CVPR42600.2020.01221

Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P (2021) SegFormer: simple and efficient design for semantic segmentation with transformers. arXiv http://arxiv.org/abs/2105.15203

Xu L, Ouyang W, Bennamoun M, Boussaid F, Xu D (2022) Multi-class token transformer for weakly supervised semantic segmentation. arXiv http://arxiv.org/abs/2203.02891

Xu Z, Wu D, Yu C, Chu X, Sang N, Gao C (2024) SCTNet: single-branch CNN with transformer semantic information for real-time segmentation. arXiv http://arxiv.org/abs/2312.17071

Xue Y, Xu T, Zhang H, Long R, Huang X (2018) SegAN: adversarial network with multi-scale L_1 loss for medical image segmentation. Neuroinformatics 16(3–4):383–392. https://doi.org/10.1007/s12021-018-9377-x

Xun S, Li D, Zhu H, Chen M, Wang J, Li J, Chen M, Wu B, Zhang H, Chai X, Jiang Z, Zhang Y, Huang P (2022) Generative adversarial networks in medical image segmentation: a review. Comput Biol Med 140:105063. https://doi.org/10.1016/j.compbiomed.2021.105063

Yan J, Yan T, Ye W, Lv X, Gao P, Xu W (2023) Cotton leaf segmentation with composite backbone architecture combining convolution and attention. Front Plant Sci 14:1111175. https://doi.org/10.3389/fpls.2023.1111175

Yang M, Yu K, Zhang C, Li Z, Yang K (2018) DenseASPP for semantic segmentation in street scenes. 2018 IEEE/CVF conference on computer vision and pattern recognition. pp 3684–3692.https://doi.org/10.1109/CVPR.2018.00388

Yao N, Ni F, Wu M, Wang H, Li G, Sung W-K (2022) Deep learning-based segmentation of peach diseases using convolutional neural network. Front Plant Sci 13:876357. https://doi.org/10.3389/fpls.2022.876357

Yi Z, Zhang H, Tan P, Gong M (2018) DualGAN: unsupervised dual learning for image-to-image translation. arXiv http://arxiv.org/abs/1704.02510

Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) BiSeNet: bilateral segmentation network for real-time semantic segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision – ECCV 2018, vol 11217. Springer International Publishing, pp 334–349. https://doi.org/10.1007/978-3-030-01261-8_20

Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions. arXiv http://arxiv.org/abs/1511.07122

Yu H, Yang Z, Tan L, Wang Y, Sun W, Sun M, Tang Y (2018b) Methods and datasets on semantic segmentation: a review. Neurocomputing 304:82–103. https://doi.org/10.1016/j.neucom.2018.03.037

Yu X, Yin D, Nie C, Ming B, Xu H, Liu Y, Bai Y, Shao M, Cheng M, Liu Y, Liu S, Wang Z, Wang S, Shi L, Jin X (2022) Maize tassel area dynamic monitoring based on near-ground and UAV RGB images by U-Net model. Comput Electron Agric 203:107477. https://doi.org/10.1016/j.compag.2022.107477

Yu Y, Wang C, Fu Q, Kou R, Huang F, Yang B, Yang T, Gao M (2023) Techniques and challenges of image segmentation: a review. Electronics 12(5):1199. https://doi.org/10.3390/electronics12051199

Yuan F, Zhang L, Xia X, Wan B, Huang Q, Li X (2019) Deep smoke segmentation. Neurocomputing 357:248–260. https://doi.org/10.1016/j.neucom.2019.05.011

Yuan Y, Chao M, Lo Y-C (2017) Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance. IEEE Trans Med Imaging 36(9):1876–1886. https://doi.org/10.1109/TMI.2017.2695227

Yuan Y, Huang L, Guo J, Zhang C, Chen X, Wang J (2021) OCNet: object context network for scene parsing. arXiv http://arxiv.org/abs/1809.00916

Zhang H, Dana K, Shi J, Zhang Z, Wang X, Tyagi A, Agrawal A (2018a) Context encoding for semantic segmentation. 2018 IEEE/CVF conference on computer vision and pattern recognition. pp 7151–7160.https://doi.org/10.1109/CVPR.2018.00747

Zhang J, Xie T, Yang C, Song H, Jiang Z, Zhou G, Zhang D, Feng H, Xie J (2020) Segmenting purple rapeseed leaves in the field from UAV RGB Imagery using deep learning as an auxiliary means for nitrogen stress detection. Remote Sens 12(9):1403. https://doi.org/10.3390/rs12091403

Zhang J, Yang K, Ma C, Reiss S, Peng K, Stiefelhagen R 2022) Bending reality: distortion-aware transformers for adapting to panoramic semantic segmentation. 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 16896–16906.https://doi.org/10.1109/CVPR52688.2022.01641

Zhang L, Li X, Arnab A, Yang K, Tong Y, Torr PH (2019) Dual graph convolutional network for semantic segmentation. arXiv preprint arXiv:1909.06121

Zhang S, Wang H, Huang W, You Z (2018b) Plant diseased leaf segmentation and recognition by fusion of superpixel, K-means and PHOG. Optik 157:866–872. https://doi.org/10.1016/j.ijleo.2017.11.190

Zhang Y, Sidibé D, Morel O, Mériaudeau F (2021) Deep multimodal fusion for semantic image segmentation: a survey. Image Vis Comput 105:104042. https://doi.org/10.1016/j.imavis.2020.104042

Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. 2017 IEEE conference on computer vision and pattern recognition (CVPR). pp 6230–6239.https://doi.org/10.1109/CVPR.2017.660

Zhao H, Zhang Y, Liu S, Shi J, Loy CC, Lin D, Jia J (2018) PSANet: point-wise spatial attention network for scene parsing. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018, vol 11213. Springer International Publishing, pp 270–286. https://doi.org/10.1007/978-3-030-01240-3_17

Zheng L, Shi D, Zhang J (2010) Segmentation of green vegetation of crop canopy images based on mean shift and Fisher linear discriminant. Pattern Recogn Lett 31(9):920–925. https://doi.org/10.1016/j.patrec.2010.01.016

Zheng L, Zhang J, Wang Q (2009) Mean-shift-based color segmentation of images containing green vegetation. Comput Electron Agric 65(1):93–98. https://doi.org/10.1016/j.compag.2008.08.002

Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, Fu Y, Feng J, Xiang T, Torr PHS, Zhang L (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. arXiv http://arxiv.org/abs/2012.15840

Zhong Z, Lin ZQ, Bidart R, Hu X, Daya IB, Li Z, Zheng W-S, Li J, Wong A (2020) Squeeze-and-attention networks for semantic segmentation. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 13062–13071.https://doi.org/10.1109/CVPR42600.2020.01308

Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, Zhang K, Ji C, Yan Q, He L, Peng H, Li J, Wu J, Liu Z, Xie P, Xiong C, Pei J, Yu PS, Sun L (2023) A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. arXiv http://arxiv.org/abs/2302.09419

Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2018) UNet++: A nested U-net architecture for medical image segmentation. arXiv http://arxiv.org/abs/1807.10165

Zhu J-Y, Park T, Isola P, Efros AA (2017) unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE international conference on computer vision (ICCV). pp 2242–2251.https://doi.org/10.1109/ICCV.2017.244

Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y (2023) Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. Inform Fusion 91:376–387. https://doi.org/10.1016/j.inffus.2022.10.022

Zhu Z, Xu M, Bai S, Huang T, Bai X (2019) Asymmetric non-local neural networks for semantic segmentation. arXiv http://arxiv.org/abs/1908.07678

Zou K, Chen X, Wang Y, Zhang C, Zhang F (2021a) A modified U-Net with a specific data argumentation method for semantic segmentation of weed images in the field. Comput Electron Agric 187:106242. https://doi.org/10.1016/j.compag.2021.106242

Zou K, Chen X, Zhang F, Zhou H, Zhang C (2021b) A field weed density evaluation method based on UAV imaging and modified U-net. Remote Sens 13(2):310. https://doi.org/10.3390/rs13020310

Authors and Affiliations

Lian Lei[1] · Qiliang Yang[2] · Ling Yang[1,3] · Tao Shen[1,3] · Ruoxi Wang[2] · Chengbiao Fu[1]

✉ Ling Yang
yangling@kust.edu.cn

Lian Lei
19922942130@163.com

Qiliang Yang
yangqilianglovena@163.com

Tao Shen
shentao@kust.edu.cn

Ruoxi Wang
wrx2102022@163.com

Chengbiao Fu
fcb@kust.edu.cn

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Jingmingnanlu 727, Kunming 650500, Yunnan, China

[2] Faculty of Modern Agricultural Engineering, Kunming University of Science and Technology, Jingmingnanlu 727, Kunming 650500, Yunnan, China

[3] Yunnan Key Laboratory of Computer Technologies Application, Kunming University of Science and Technology, Kunming 650500, China