future internet

MDPI

Review

2D Semantic Segmentation: Recent Developments and Future Directions

Yu Guo [1], Guigen Nie [1,2,*], Wenliang Gao [1] and Mi Liao [1]

[1] Wuhan University GNSS Research Center, Wuhan University, Wuhan 430079, China; yguowh@whu.edu.cn (Y.G.)
[2] Hubei Luojia Laboratory, Wuhan 430079, China
* Correspondence: ggnie@whu.edu.cn

Abstract: Semantic segmentation is a critical task in computer vision that aims to assign each pixel in an image a corresponding label on the basis of its semantic content. This task is commonly referred to as dense labeling because it requires pixel-level classification of the image. The research area of semantic segmentation is vast and has achieved critical advances in recent years. Deep learning architectures in particular have shown remarkable performance in generating high-level, hierarchical, and semantic features from images. Among these architectures, convolutional neural networks have been widely used to address semantic segmentation problems. This work aims to review and analyze recent technological developments in image semantic segmentation. It provides an overview of traditional and deep-learning-based approaches and analyzes their structural characteristics, strengths, and limitations. Specifically, it focuses on technical developments in deep-learning-based 2D semantic segmentation methods proposed over the past decade and discusses current challenges in semantic segmentation. The future development direction of semantic segmentation and the potential research areas that need further exploration are also examined.

Keywords: overview; semantic segmentation; image segmentation; technology development; deep learning

Citation: Guo, Y.; Nie, G.; Gao, W.; Liao, M. 2D Semantic Segmentation: Recent Developments and Future Directions. *Future Internet* **2023**, *15*, 205. https://doi.org/10.3390/fi15060205

Academic Editor: Paolo Bellavista

Received: 5 May 2023
Revised: 23 May 2023
Accepted: 29 May 2023
Published: 1 June 2023

1. Introduction

Semantic segmentation is one of the three major research topics in computer vision and the basis of various complex visual tasks. From a classification perspective, all pixels in semantic segmentation should be assigned label information. Pixel-level semantic segmentation in region-level classification is more challenging than image classification and region-level object detection. The task of semantic segmentation is to assign each pixel of an image to a corresponding class label [1] from a predefined set of classes; that is, semantic segmentation is based on pixel-to-pixel classification of labeled data [2,3]. The massive success of deep learning has had a great impact on semantic segmentation methods [4–6] because it has improved their accuracy, and it has attracted the interest of many people in technical and research fields that require computer vision capabilities. Semantic segmentation has been widely used in various fields because it can provide scene understanding at the pixel level. For example, effective extraction of traffic road networks from satellite images is essential in car navigation and road planning [7]. In robotics and surveillance, temporary localization and classification of action clips are particularly important [8]. Meanwhile, in unmanned driving, semantic segmentation is one of the key technologies for road-scene understanding [9–12], and it guarantees the safe driving of autonomous vehicles. In addition, semantic segmentation has certain applications in pedestrian detection [13,14], remote sensing [15,16], pose estimation [17,18], and the medical field [19–22].

However, despite important progress in recent years, pixel-level semantic segmentation still faces many challenges, especially when dealing with complex scenarios and varying environmental conditions.

We chose this specific subject because we recognize the importance of having a deep understanding of the current techniques and challenges in image semantic segmentation and the direction of future development. Furthermore, despite the multitude of studies on image semantic segmentation, a comprehensive, systematic overview of recent advancements and emerging technologies remains lacking. This situation makes understanding the research progress in this field difficult, especially for those new to this area.

Centering on the key technical methods that emerged during the development of image semantic segmentation, this study expounds upon the evolution from traditional methods to deep learning and discusses the strengths and weaknesses of each method and the directions that may be explored further in the future. An appropriate analysis is conducted on some commonly used and challenging data sets along with their evaluation metrics. We hope that this review will provide readers a clear global perspective toward understanding existing challenges and potential future solutions.

The remainder of this paper is organized as follows. Section 2 discusses some traditional methods that preceded deep learning. Section 3 provides a qualitative analysis of existing deep learning approaches. Section 4 describes the key technological developments in deep learning for semantic segmentation and the challenges at this stage. It also provides a detailed analysis of current issues and identifies future research directions. Section 5 briefly introduces common data sets and evaluation metrics used in the field, and Section 6 discusses the current challenges in semantic segmentation and the potential research directions.

2. Traditional Methods before Deep Learning

Before the application of deep learning techniques in image processing, traditional machine learning methods were widely used in various image semantic segmentation tasks. These methods typically relied on manually designed feature extractors and classifiers such as support vector machines (SVMs), decision trees, K-means, and conditional random fields (CRFs). These algorithms usually depend on manual feature extraction methods such as SIFT and HOG to achieve image segmentation through feature matching or classification. These traditional methods still have advantages in certain cases such as when processing small-sample data or images with minimal noise. However, with the development of deep learning techniques, deep learning-based image semantic segmentation methods have gradually become mainstream. Deep learning methods automatically learn feature representations by utilizing multilayer convolutional neural networks (CNNs), thus avoiding the tedious process of manually designing feature extractors. Overall, traditional machine learning methods have laid the foundation for the development of deep learning methods in image semantic segmentation tasks.

2.1. SVMs

The support vector machine (SVM) is a classic machine learning method that is widely used in computer vision fields such as image classification, object recognition, and object tracking. The main idea of an SVM is to map data to a high-dimensional space and find the optimal hyperplane in that space for classification. This method can handle non-linear and high-dimensional data, making it widely applicable for solving practical problems. SVMs have also been widely used in image semantic segmentation tasks. Generally, an SVM is employed as a pixel classifier; each pixel in the image is regarded as a sample, and an SVM is used to classify each pixel to obtain the semantic segmentation result of the image. SVMs have good generalization performance and can achieve good results for some small image semantic segmentation problems. However, SVMs can be used only for binary classification problems and are ineffective for multiclass problems. In image semantic segmentation tasks, multiclass problems such as segmenting multiple objects in an image

are commonly encountered. In addition, SVMs have a high training complexity and require much time and computational resources, so their application in some large-scale image segmentation problems is limited. In comparison with SVMs, deep learning methods can better handle multiclass classification problems and have more advantages in training time and computational resources; thus, they are gradually becoming the mainstream methods in image semantic segmentation.

2.2. Decision Trees

The decision tree is common machine learning method that can be used in image segmentation tasks. In traditional image segmentation, decision trees are usually utilized for image object classification; that is, to label and classify different objects or regions in an image. Decision trees classify data sets by gradually dividing them on the basis of different features from the root node to a leaf node, achieving the classification purpose. This method is simple and easy to understand and interpret, and it can handle multiple output problems. However, decision trees perform poorly when dealing with high-dimensional data because the feature space of the data set is often huge, and the classification results of decision trees are easily affected by noise and data uncertainty, resulting in poor generalization ability of the model. In addition, the performance of decision trees is limited by the size and complexity of the data set. Decision trees often become too complex and difficult to maintain with increasingly complex and large data sets. By contrast, deep learning methods can handle high-dimensional data through deep neural networks, thus performing well in image segmentation tasks.

2.3. K-Means

In traditional image segmentation, the K-means algorithm is commonly used to cluster image pixels into several categories and facilitate subsequent segmentation. In image segmentation, the K-means algorithm needs to divide image pixels into K clusters, each of which has similar color or texture features. It is a simple and effective algorithm that can cluster a large amount of unlabeled data, so it has been widely used in image segmentation. However, the K-means algorithm also has disadvantages. First, the K-means algorithm requires the manual setting of the clustering number K, and different clustering numbers result in different segmentation results. Second, the K-means algorithm is sensitive to noisy data and outliers, which may lead to inaccurate segmentation results. Lastly, the K-means algorithm can only generate block segmentation results and cannot handle images with complex textures and edges. By contrast, deep learning uses deep CNNs to extract features from images. Through the analysis and learning of these features, the method can automatically obtain image segmentation results and handle images with complex textures and edges, thus obtaining accurate segmentation results.

2.4. CRFs

A CRF is mainly applied at the post-processing stage of images to improve the accuracy and continuity of segmentation results. It is a probabilistic graphical model that models the relationships between pixels as an undirected graph and obtains the best segmentation result by minimizing the energy in the graph. In image segmentation, CRFs are usually used to consider the contextual information between pixels; namely, the influence of surrounding pixels on the current pixel and the interaction between pixels. They can also eliminate noise by smoothing adjacent pixels, refine the segmentation results, and improve segmentation quality. Although CRFs have been widely used in image segmentation before the rise of deep learning, they still have some shortcomings such as the need for a large amount of manual feature engineering and parameter tuning as well as a high computational complexity, leading to low efficiency. By comparison, deep learning methods can obtain the best segmentation results through end-to-end training, thus avoiding the tedious process of manual parameter adjustment. In addition, deep learning models have good performance in terms of calculation speed and accuracy and can obtain high segmentation accuracy in a

short time. Although a CRF can still be used as a post-processing layer for deep learning models, it is rarely utilized as the main method of image segmentation at present due to the superiority of deep learning.

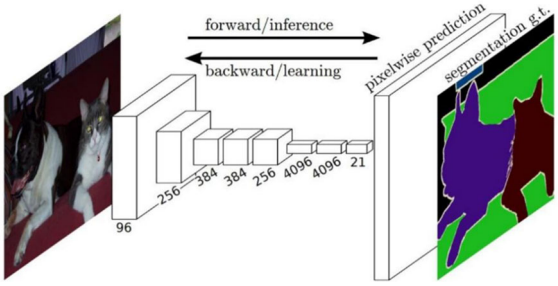
3. Classical Networks Based on Deep Learning

With the rapid development of deep learning, many deep neural networks have emerged and achieved considerable progress in various fields, especially computer vision. Among them, CNNs have elicited considerable attention and are regarded highly because of their superior performance in numerous tasks. LeCun et al. [23] proposed Lenet-5, which has served as a groundbreaking model that only contains five layers of neurons and has paved the way for future CNN development. Subsequently, a series of CNN models were introduced; these include VGG [24], AlexNet [25], GoogLeNet [26], and ResNet [27]. The introduction of these networks has revolutionized the field of computer vision and provided new solutions for semantic segmentation, allowing for enhanced accuracy and efficiency in various applications.

3.1. Fully Convolutional Networks

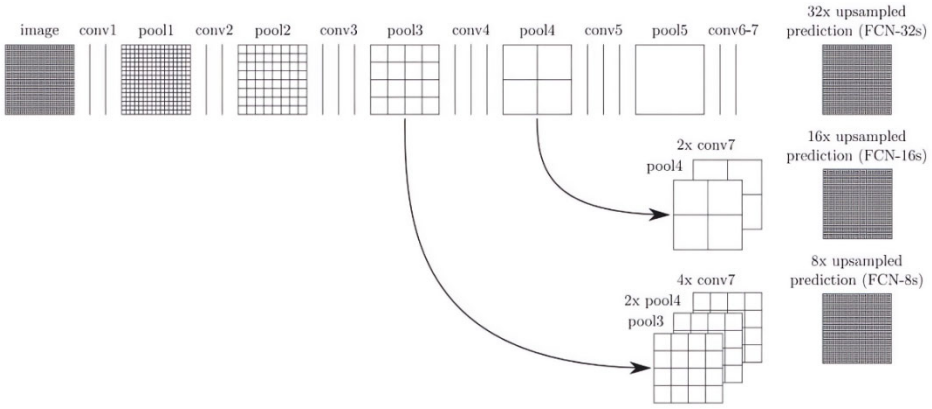
Compared with traditional statistical methods, methods based on CNNs have natural advantages. They can automatically extract semantic features, so they can replace manual feature extraction in original methods. CNNs have end-to-end processing structures [28,29], which take images as an input and generate pixel-level labels through multiple layers of convolution and deconvolution. The fully convolutional networks (FCNs) proposed by Long et al. [30] were the first application of CNN in semantic segmentation. FCNs perform pixel-level classification of images. CNNs [31] have achieved great success in semantic segmentation tasks [32]. They use fully connected layers in convolutional layers to obtain fixed-length feature vectors. Meanwhile, FCNs accept images of any size and use deconvolution layers to upsample the feature maps of the last convolutional layer and restore them to the same size as the input image. Subsequently, predictions are generated for each pixel while retaining the spatial information of the original input image. Pixel classification is then performed on the upsampled feature maps. FCN network structure as shown in Figure 1.

In the past few years, FCNs have achieved great success in segmentation [33–37]. Nearly all semantic segmentation methods adopt the idea of the FCN method, indicating that FCNs are a breakthrough in the field of semantic segmentation. FCNs have also created new opportunities for further improving the architecture of deep semantic segmentation. The main disadvantages of FCNs are their low efficiency in label localization within the feature hierarchy, inability to handle global contextual information, and lack of a multiscale processing mechanism. The majority of subsequent research has attempted to address these issues by proposing various architectures or techniques.



(**a**)

Figure 1. *Cont.*



(**b**)

Figure 1. FCN network structure (**a**,**b**) [30].

3.2. U-Net

U-Net [38] is a neural network designed for medical image segmentation to address challenges in the segmentation of cell-level medical images, as shown in Figure 2. It uses a U-shaped network structure to obtain context and location information for improved segmentation efficiency. U-Net generally has an encoder–decoder structure in which the first half is for feature extraction and the second half is for upsampling. However, the semantic differences between feature maps in U-Net increase the difficulty of network learning. Thus, a series of improved networks such as UNet++ [39] have been proposed. Although U-Net has certain advantages, its shortcomings (such as large semantic differences between feature maps and high network learning difficulty) are apparent. Therefore, further research and improvement of the U-Net network structure are necessary to enhance its applicability in medical image segmentation.

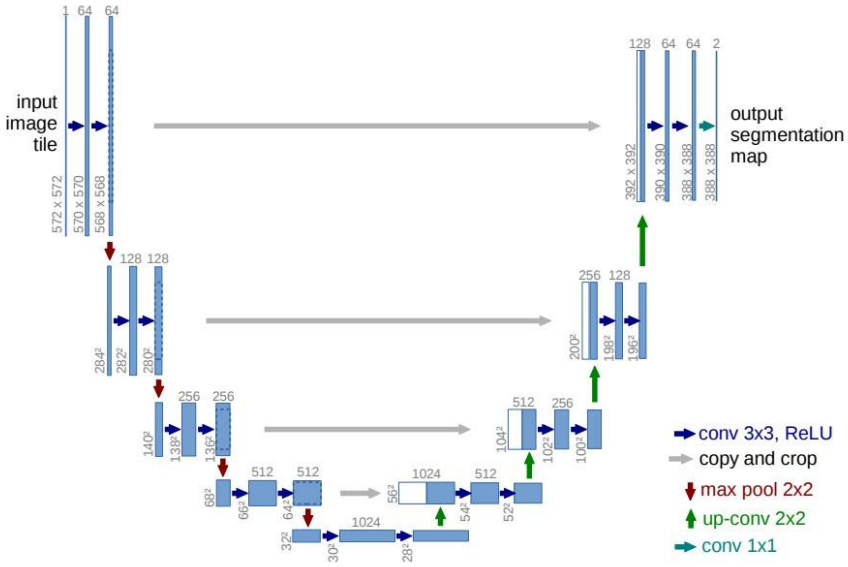


Figure 2. U-net network structure [38].

3.3. DeepLab Family

Image segmentation has witnessed the development of various techniques over the years. For instance, DeepLab V1 [40], which is based on VGG and image segmentation models such as FCN and U-Net, uses different approaches to maintain resolution consistency. Specifically, DeepLab V1 employs dilated convolution, and FCN and U-Net use deconvolution and pooling.

DeepLab V2 [41] is an improvement of the V1 version through the introduction of ASPP, which involves obtaining feature maps of different scales through different rates before making predictions. This version also includes multiscale training and large receptive fields to enhance performance.

DeepLab V3 [42] continued with the development of the previous version by upgrading modules to further enhance performance. Notably, the version includes the addition of a multigrid module, improvement of the ASPP module structure, and removal of the post-processing module of CRFs.

DeepLab V3+ [43] adopts an encoder–decoder architecture in which the encoder architecture employs DeepLab V3 and the decoder part employs a simple and effective module for recovering the details of the target boundary, the network structure is shown in Figure 3. The main limitations of the DeepLab series include considerable loss of detail segmentation, high computation demands, and poor relevance of semantic information between contexts. The disadvantages and improvement directions are as follows:

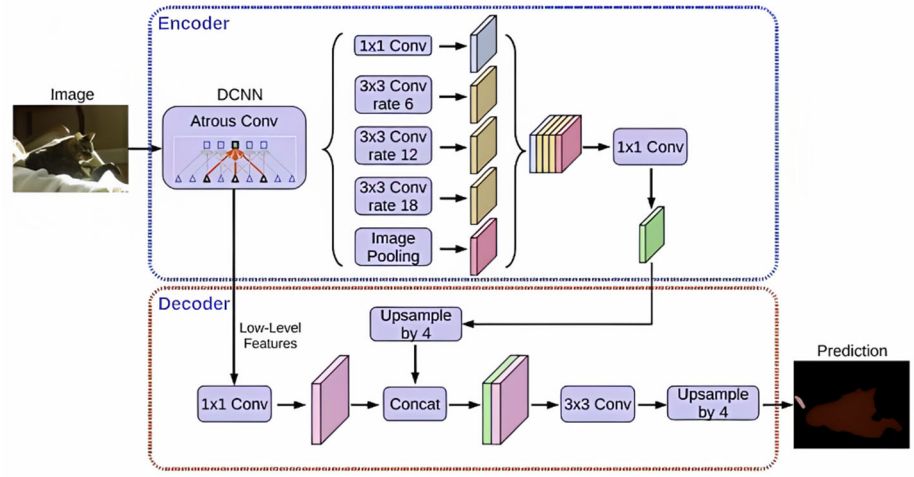


Figure 3. DeepLab V3+ network structure [43].

DeepLab V1. Shortcomings: DeeplabV1 has certain computational bottleneck and overfitting issues due to the use of dilated convolutions and fully connected layers, and the segmentation performance on edge regions is unsatisfactory. Improvement direction: lightweight network structures and attention mechanisms can be explored while considering the introduction of additional contextual and semantic information.

DeepLab V2. Shortcomings: The multiscale pyramid pooling used in DeepLab V2 has a high computational complexity and feature information loss, and objects of different scales have large differences in segmentation performance. Improvement direction: other efficient multiscale feature fusion methods can be considered while introducing fine segmentation heads and accurate pixel classifiers.

DeepLab V3. Shortcomings: DeepLab V3 has a complex ASPP module and multiscale feature fusion mechanism, leading to high computational and training time costs and inaccurate segmentation on image edges. Improvement direction: other efficient attention mechanisms and lightweight network structures can be explored while applying optimization based on edge information.

DeepLab V3+ (as shown in Figure 3). Shortcomings: The encoder–decoder structure used in DeepLab V3+ is complex, leading to computational bottlenecks and high memory consumption. The segmentation performance on details is also unsatisfactory. Improvement direction: lightweight encoder–decoder structures and attention mechanisms can be considered while introducing fine feature extraction and fusion mechanisms for details.

3.4. SegNet

SegNet [44] adopts an encoder–decoder structure to combine shallow and deep information to optimize segmentation results, as shown in Figure 4. Skip connections are used to directly connect the feature maps of the encoder network with those of the corresponding decoder network, thus improving the memory utilization and computational efficiency of SegNet. The innovation of SegNet is that it records the position of max pooling during each downsampling because max pooling and downsampling operations reduce the spatial resolution of feature maps. However, recording the position of max pooling increases memory consumption. Therefore, only the index of max pooling is recorded. In addition to effectively combining transpose convolution and non-cooling, SegNet adopts batch normalization [45] to solve the problem of vanishing gradients.

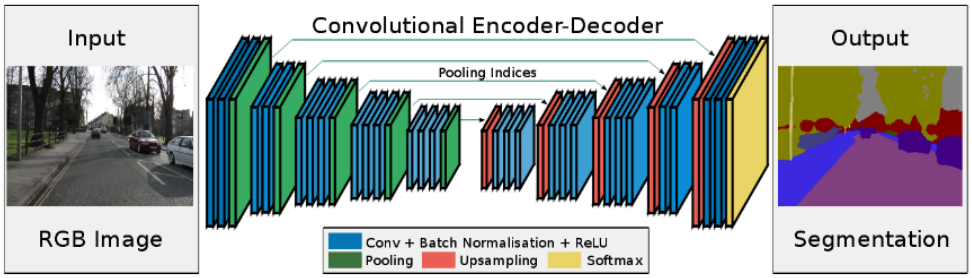


Figure 4. SegNet network structure [44].

However, SegNet also has some limitations. First, SegNet employs an encoder–decoder structure to combine shallow and deep information for optimizing segmentation results. However, segmentation accuracy may decrease due to the lack of multiscale information guidance. Second, although SegNet records the maximum pool position for each downsampling operation to preserve spatial resolution, it increases memory consumption. Therefore, SegNet may be unsuitable for use in computationally and memory-constrained scenarios. Lastly, SegNet employs batch normalization to address vanishing gradients, but in some cases, batch normalization may lead to overfitting. The following directions can be considered to solve these problems:

(1) Introducing multiscale information such as pyramid pooling and multiscale fusion can improve segmentation accuracy. SegNet's model structure can be simplified using a lightweight network structure to reduce memory and computational resource consumption.
(2) Other methods such as residual connections and attention mechanisms can be explored to address the problem of vanishing gradients. Although the SegNet algorithm has certain limitations, it is still a classic semantic segmentation algorithm that can perform well in specific scenarios.

3.5. PSPNet

PSPNet [46], a common semantic segmentation model, uses dilated convolutions to expand the receptive field and retain abundant spatial position information, as shown in Figure 5. The algorithm combines multiscale features by using an image pyramid and its variants to improve segmentation performance. This method is effectively applied to complex scene analysis, in which global pyramid pooling features provide additional contextual information and a deep supervision optimization strategy based on ResNet-FCN is introduced. The algorithm performs better than common semantic segmentation models such as FCN, SegNet, and DeepLab V1.

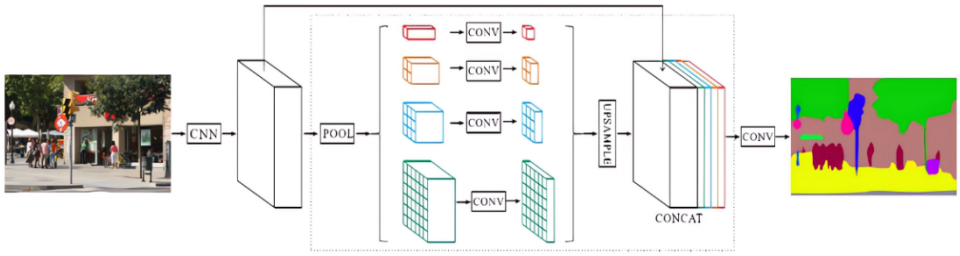


Figure 5. PSPNet network structure [46].

However, the PSPNet algorithm still has shortcomings. First, the dilated convolution method used in this algorithm requires a large amount of computational resources, which may lead to increased training and inference time. Second, the PSPNet model is relatively complex and requires large amounts of computational resources and storage space. Lastly, the algorithm is sensitive to the size and scale of the input image, and preprocessing and scaling are required for segmentation. Therefore, the PSPNet algorithm needs to be improved to address these issues.

The improvement directions are as follows. First, a lightweight network structure can be used to reduce the consumption of computational resources. Second, new receptive field-expansion methods can be explored. Lastly, the changes in the size and scale of the input image need to be handled properly to improve the stability and applicability of the model.

4. Key Technologies to Improve the Effect of Semantic Segmentation

During the continuous development of semantic segmentation technology, many innovative methods have emerged and played a key role in technological breakthroughs. The application of these methods has considerably improved the accuracy of image-level segmentation tasks. These innovative methods include but are not limited to the following: deep-learning-based semantic segmentation algorithms, multiscale image segmentation techniques, and methods that utilize contextual information for segmentation. These methods not only improve segmentation accuracy but also greatly reduce computational complexity. They provide support for the practical application of semantic segmentation technology. In the future, we can further explore these innovative methods and apply them to a wide range of scenarios to achieve accurate and efficient image segmentation.

4.1. Development Trend of Encoder–Decoder Systems

The advent of encoder–decoder architecture has notably influenced semantic segmentation over the past few years. The convolutional neural network model it employs has the capability to extract image feature data and transform it into semantic segmentation outcomes through the decoder. This effectively addresses the challenge of multiclass pixel categorization in semantic segmentation tasks, thereby enhancing both the precision and speed of segmentation. The encoder–decoder structure is composed of two segments(as shown in Figure 6). The encoder's role is to construct an efficient feature extraction network. As the network deepens, the feature map dimension progressively shrinks while the semantic features consistently grow. The decoder's aim, on the other hand, is to develop a feature map restoration model. Throughout the decoding stage, semantic information is articulated and details are continually enriched.

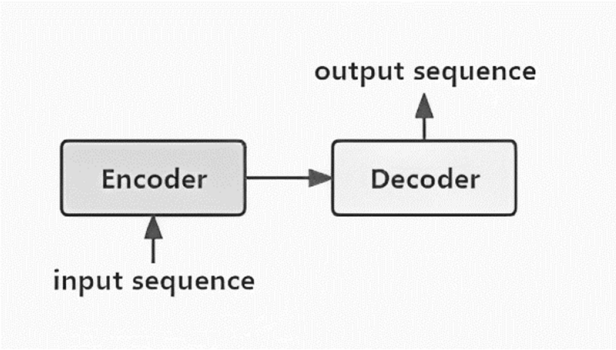


Figure 6. Brief structure diagram of a transformer.

As for FCN, it conserves spatial information by substituting the final fully connected layer of VGG16 with a convolutional one, maintaining the feature map in a two-dimensional arrangement. Further, this approach also remedies the issue of a required fixed input size. Zhou et al. [47] proposed the novel encoder–decoder network CENet to collaboratively explore hierarchical convolutional features for accurate pixel-level semantic segmentation. Primarily, two methodologies are used to strike a balance between the employment of superficial and deep features: encoding and decoding methods and multipath fusion techniques. The former concentrates on refining the structures of both encoding and decoding, hence enhancing the capability for information extraction and restoration. The latter places emphasis on the amalgamation of both types of information.

Another hot direction for encoder–decoder architectures is the emergence of network architectures based on transformers, such as the Visual Transformer (Vit) [48] and its variants [49–52]. Due to their powerful long-range dependency modeling capabilities, which have achieved breakthroughs in various vision tasks, their appearance provides a promising research direction for semantic segmentation [53].

Looking ahead, there remains immense potential for the evolution of semantic segmentation technology grounded in the encoder–decoder architecture. To expedite the progression of this technology, we postulate the following concepts are viable:

(1) Adaptive Feature Fusion: Traditional encoder–decoder architectures often rely on fixed feature fusion strategies. However, these strategies may not fully exploit the multiscale information in images. Therefore, we suggest continuing research on adaptive feature fusion methods that dynamically adjust the feature fusion weights based on the input image characteristics, thereby achieving more precise semantic segmentation.

(2) Task-Oriented Encoder Design: Current encoder–decoder architectures often employ generic encoder structures such as ResNet and VGG. Although these generic encoders exhibit excellent feature extraction capabilities, they may not be optimized for specific semantic segmentation tasks. To address this issue, we propose designing a task-oriented encoder structure that incorporates modules specifically designed for semantic segmentation, thereby improving feature extraction effectiveness.

(3) Dynamic Decoder Optimization: Traditional decoder structures usually adopt fixed upsampling and fusion strategies. However, these strategies may not fully recover the image's detail information. To address this issue, we suggest further investigation of dynamic decoder optimization methods by introducing adaptive upsampling and fusion strategies to dynamically recover detail information based on input image characteristics.

(4) Integration of Vision Transformers (ViTs): Vision Transformers (ViTs) have achieved significant success in the computer vision field. It is worth continuing in-depth research on better integrating ViTs with encoder–decoder architectures, leveraging ViTs' powerful remote dependency modeling capabilities, and achieving more accurate semantic segmentation through end-to-end training.

(5) Guiding Semantic Segmentation with Prior Knowledge: In many practical applications, images often exhibit specific structures and prior knowledge. For example, in medical image segmentation, we usually have some understanding of the target structure's shape and location. Therefore, introducing this prior knowledge into encoder–decoder architectures can guide the model toward learning more reasonable semantic segmentation results. This can be achieved by adding prior knowledge constraints to the loss function or designing specific prior knowledge modules.

(6) Adaptive Domain Adaptation: In practical applications, domain differences may cause performance fluctuations in the model across different data sets. To address this issue, we suggest researching an adaptive domain adaptation method based on encoder–decoder architectures to learn the mapping relationship between source and target domains to improve the model's generalization capabilities in the target domain.

(7) Integration of Multimodal Information: In many practical scenarios, besides RGB images, other modalities of data such as depth information and spectral information can be obtained. This multimodal information provides a richer context for semantic segmentation. Therefore, we propose incorporating multimodal information into encoder–decoder architectures to improve segmentation performance. Specific approaches may include designing multimodal fusion modules or using multitask learning to simultaneously learn feature representations of different modalities.

In summary, image semantic segmentation technology based on encoder–decoder architectures has achieved significant success in many visual tasks. However, to further improve its performance in the field of semantic segmentation, more exploration and research are needed.

4.2. Skip Connections

Skip Connections(as shown in Figure 7) are a crucial structure within deep learning networks that are primarily used to combat gradient vanishing and exploding problems found in deep networks. They operate by directly linking the feature maps of preceding layers to subsequent layers. In other words, the output from the earlier layers is passed not only to the next layer but also to several layers after that. This form of connection facilitates more effective backpropagation within the network, thereby accelerating its training.

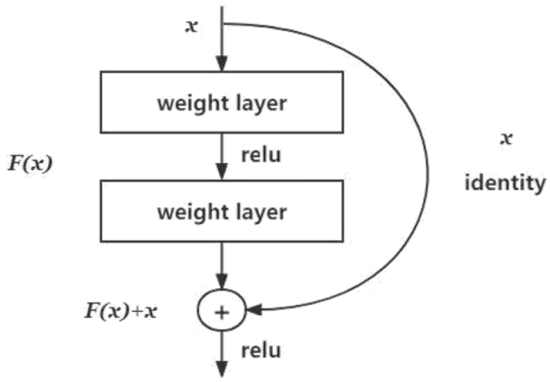


Figure 7. Schematic diagram of a Skip Connection in ResNet.

Skip Connections carry significant relevance in semantic segmentation. In tasks involving semantic segmentation, it is crucial to classify every pixel accurately, necessitating the simultaneous acquisition of global semantic information and local detail information. In this case, convolutional and pooling layers of the deep neural network can extract global semantic information, while Skip Connections aid in preserving local detailed information. Therefore, Skip Connections help us achieve more accurate segmentation results. Jiao et al. [54] used Skip Connections to better utilize the multilevel features of the encoder backbone to enrich and recover more details in the final semantic feature map.

While Skip Connections effectively solve the problem of gradient vanishing and exploding, they also increases the complexity and computational load of the model. The design and quantity of Skip Connections need to be adjusted according to the specific task. An excessive amount may lead to model overfitting, while an insufficient amount might not yield the expected results.

Currently, most research on Skip Connections is empirically based. How to scientifically determine the optimal positions and quantity of Skip Connections is a worthy research topic. Most Skip Connections are static, meaning they are determined during the model construction phase. However, dynamic Skip Connections, which are adjusted during the model's training and prediction phases based on necessity, remain a promising research direction.

4.3. Spatial Pyramid Pooling

In semantic segmentation, traditional CNNs typically use the FCN structure, which adds upsampling layers at the end of the network to upsample low-resolution feature maps to the original input image size for pixel-level classification. However, this approach requires a high computational cost and large amounts of memory, making it unsuitable for real-time semantic segmentation tasks.

Spatial pyramid pooling (SPP) [55] was proposed to address this problem. The core idea of the SPP network is to introduce pyramid pooling layers in CNNs that map different sizes of feature maps to a fixed-size feature vector through pooling operations, thereby avoiding the need for upsampling operations and greatly reducing the computational cost and memory overhead without sacrificing accuracy. Moreover, an SPP network supports inputs of arbitrary sizes and has translation and scale invariance, making it suitable for object detection and semantic segmentation tasks involving various scales and orientations.

With the introduction of the SPP network, semantic segmentation has made considerable progress. Future research directions include optimizing the SPP network structure, exploring multiscale feature fusion methods, and introducing attention mechanisms and context information to further improve the performance and speed of semantic segmentation.

4.4. Dilated Convolutions

The emergence of dilated convolutions has had a remarkable impact on semantic segmentation. The basic idea of dilated convolutions [56] is to continuously use convolutional filters to expand their effective receptive field (as shown in Figure 8). The difference between dilated and ordinary convolutions is that a dilation rate parameter (indicating the size of expansion) is added to the former. The similarity between dilated and ordinary convolutions is that the size of the convolution kernel is the same; that is, the number of parameters is unchanged. Another difference is that the receptive field of dilated convolutions is larger than that of ordinary convolutions. Wu et al. [57] introduced a mixed dilated convolution (MDC) module to improve the model's ability to recognize objects with various scales and irregular shapes. The MDC module not only increases the diversity of receptive fields but also solves the problems in conventional dilated convolutions (the pervasive "mesh" problems). The authors in [58] proposed a grouped dilated convolution module that combined existing grouped convolutions and atrous SPP techniques. The experimental results showed that the mean intersection over union (MIoU) of the method based on the CamVid data set was 73.15%, and the MIoU of the method based on SBD was 72.81%. The performance was excellent. However, compared with other technical methods, the GPU memory usage of dilated convolutions was higher.

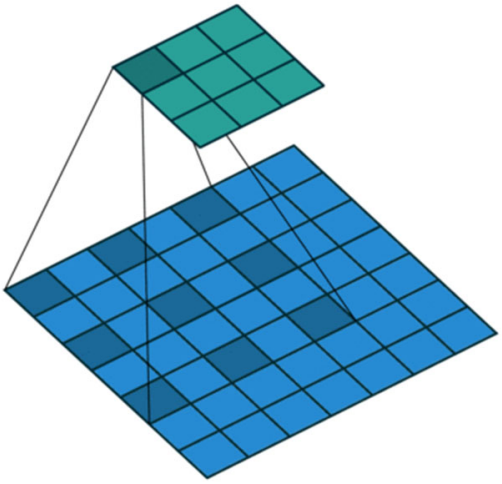


Figure 8. Schematic diagram of dilated convolutions.

Dilated convolutions also have some drawbacks. First, when the dilation rate is too high, a "checkerboard effect" may occur, resulting in false shadows and noise in the segmentation results. Second, the receptive field of dilated convolutions is limited by the size of the kernel and the dilation rate, so the method cannot handle global information.

Future research directions for dilated convolutions can be explored in several aspects. First, new dilation rate selection strategies can be explored to address the checkerboard effect and local information loss. Second, dilated convolutions can be combined with other operations to further improve the accuracy and efficiency of semantic segmentation. Lastly, the application and optimization of deep learning models based on dilated convolutions can be performed in practical scenarios.

4.5. Knowledge Distillation

The concept of knowledge distillation was first introduced by Hinton in the paper titled "Distilling the Knowledge in a Neural Network [59]". This method involved the use of a Teacher Network (typically more complex yet with superior prediction accuracy) to train a Student Network characterized by a leaner structure, reduced complexity, and being more suited for inference deployment. The crux of the technique lies in utilizing a soft target related to the Teacher Network as part of the loss function, thereby facilitating the transfer of knowledge from the teacher to the student.

In the field of semantic segmentation, knowledge distillation(as shown in Figure 9) can assist us in reducing the size and computational complexity of models without sacrificing accuracy. Semantic segmentation tasks typically require predictions at the pixel level, demanding that the model handle a vast amount of detailed information, thereby leading to significant model size and computational burden. Knowledge distillation, through enabling a smaller model to mimic the behavior of a larger one, can aid us in designing smaller, faster, and more efficient semantic segmentation models.

Knowledge distillation can be broadly divided into three categories: Pixel-wise distillation, Pair-wise distillation, and Holistic distillation [60]. While these methods share the common goal of transferring knowledge, each offers a distinct approach to distillation. Pixel-wise distillation focuses on individual pixel values, Pair-wise distillation compares the relationship between pairs of instances, and Holistic distillation aims at capturing more global features. Thus, these distillation methods provide diverse means of enhancing the learning capability of the Student Network.

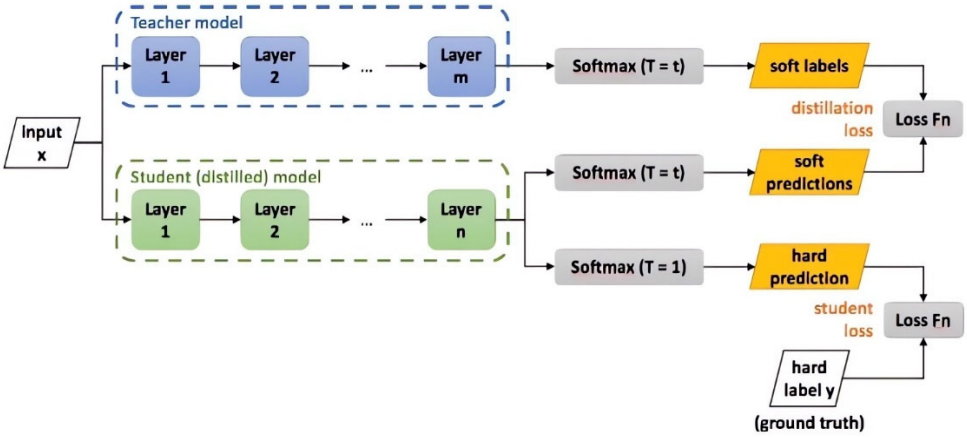


Figure 9. Schematic diagram of knowledge distillation. The first step is to train Net-T; the second step is to distill the knowledge of Net-T to Net-S at high temperature T.

Wu et al. [61] proposed a prediction model based on an encoder–decoder architecture (called SPNet) to further determine additional unlabeled data; they presented a knowledge distillation scheme to distill the structured knowledge from cumbersome to compact networks. Without using any pretrained models, this method demonstrated its state-of-the-art performance on several challenging data sets. On the cityscape test data set, it achieved 75.8% mIoU at 61.2 FPS. Khosravian et al. [62] proposed a multiteacher framework. Experiments validated that the multiteacher framework resulted in an improvement of 9% to 32.18% compared to the single-teacher paradigm. Moreover, it was demonstrated that paradigm surpassed previous supervised real-time studies in the semantic segmentation challenge. Feng et al. [63] proposed a simple yet general and effective knowledge distillation framework called double similarity distillation (DSD) to improve the classification accuracy of all existing compact networks by capturing the similarity knowledge in pixel and category dimensions. Extensive experiments on four challenging data sets demonstrated its effectiveness and generality.

However, knowledge distillation requires some computing resources and time to train the large model and transfer knowledge to the small model, resulting in additional computational costs. Moreover, during the knowledge distillation process, issues such as incomplete knowledge transfer and information loss may arise, thereby affecting the accuracy and performance of the model.

Based on existing research, we believe that the following directions hold significant potential for further investigation:

(1) Adaptive Knowledge Distillation: Traditional knowledge distillation methods often require manually setting fixed loss weights. To reduce human intervention and improve distillation performance, it is worth exploring adaptive methods for adjusting loss weights. This can be achieved by introducing dynamic weight allocation strategies that automatically adjust loss weights based on the model's performance during training.

(2) Task-Driven Knowledge Distillation: To enhance the effectiveness of knowledge distillation in semantic segmentation tasks, prior knowledge of the target task can be incorporated into the distillation process. For instance, task-specific loss functions can be designed to guide the smaller model toward learning more effective feature representations for the target task.

(3) Weakly Supervised Knowledge Distillation: High-quality annotation of semantic segmentation data is often time-consuming and expensive. To reduce annotation costs, it is worth investigating the application of knowledge distillation in weakly supervised semantic segmentation tasks. By utilizing weakly labeled data (e.g., image-level labels or edge labels), the performance of lightweight models can be effectively improved.

(4) Integration of Model Architecture Search and Knowledge Distillation: To select a more suitable lightweight model, knowledge distillation can be combined with model architecture search (NAS). By automatically searching for the optimal lightweight model structure, the performance of the distilled model can be further enhanced.
(5) Online Knowledge Distillation: To reduce the computational cost of the distillation process, online knowledge distillation methods can be explored. Online knowledge distillation performs knowledge distillation of smaller models simultaneously with the training of larger models, thus avoiding additional distillation computation overhead. By updating the smaller model's parameters in real time, knowledge can be effectively transferred, thereby accelerating model convergence.
(6) Cross-Model Knowledge Distillation: In practical applications, it may be necessary to transfer knowledge from multiple large-scale models into a single lightweight model. Investigating cross-model knowledge distillation methods can efficiently integrate the knowledge from multiple large-scale models, further improving the performance of the lightweight model.
(7) Explainable Knowledge Distillation: Although knowledge distillation can enhance the performance of lightweight models, the distillation process may result in a reduction in model interpretability. To improve model interpretability, it is worth exploring the introduction of explainability constraints into the knowledge distillation process. By constraining the feature representations learned by the lightweight model to have similar explainability to those of the larger model, the model can maintain its performance while exhibiting better interpretability.

4.6. Domain Adaptation

Domain adaptation is a critical challenge that involves applying models trained in the source domain to the target domain, and it is a subset of transfer learning. Its goal is to leverage the knowledge learned in a different but consistent source domain to improve the model's effectiveness in the target domain [64]. Domain adaptation semantic segmentation is similar to semi-supervised semantic segmentation; the only difference between the two is whether a domain gap exists between labeled and unlabeled images.

Domain adaptation methods can be divided into three main categories.

a. Instance-based domain adaptation: This method improves the classifier's performance in the target domain by weighting the samples from the source and target domains. Typical methods include maximum mean discrepancy and kernel mean matching.
b. Feature-based domain adaptation: This method improves the model's performance in the target domain by finding a mapping in the feature space that minimizes the distribution difference between the source and target domains. Typical methods include domain-invariant feature extraction and deep adversarial domain adaptation networks.
c. Adversarial domain adaptation: This method improves the model's generalization ability in the target domain by making the feature distributions similar in the source and target domains through adversarial training. Typical methods include generative adversarial networks (GANs) and domain adversarial neural networks.

Domain adaptation in semantic segmentation is a challenging problem for two reasons. One of the reasons is that annotating labels is an expensive endeavor. Another reason is that the domain gap between the source and target domains limits the performance of semantic segmentation [65]. Previous domain-alignment strategies aim to explore the largest domain-invariant space to expand the knowledge learned in the target domain. Given the lack of guidance of the corresponding classes of the source domain, outlier classes and negative shifts arise. To address this issue, the authors of [66] proposed a partial domain adaptation method for semantic segmentation to guide the target model as it selectively learned category-level knowledge. A module called a partial adaptive map was introduced to incentivize the target model; abundant knowledge was acquired

from non-outliers, and minimal knowledge was obtained from outliers, thus avoiding negative transfer.

However, this approach has some limitations and challenges. Domain adaptation techniques require adequate consideration of the differences between the source and target domains, including data and feature distributions. Additionally, domain adaptation methods have technical restrictions such as the need for sufficient labeled data, appropriate feature selection and transformation methods, and careful consideration of different task characteristics.

Domain adaptation, as an essential method in transfer learning, is important in various practical applications, particularly semantic segmentation. To overcome the challenges posed by existing technology (such as negative transfer, insufficient annotations, and multisource domain adaptation), future researchers should focus on the following aspects:

(1) Domain-specific information transfer methods: Domain-specific information transfer methods should be investigated by analyzing structured information, local and global semantic relationships, and high-order features within the domain. This task can help enhance the model's adaptability and generalization capabilities for domain differences.
(2) Application of adversarial training in domain adaptation: Adversarial training strategies must be utilized to strengthen the model's robustness against the distribution differences between the source and target domains. By introducing domain-adversarial loss functions, the distribution gap between source and target domain features can be reduced, thus improving domain adaptation performance.
(3) Data augmentation and sample generation using generative models: Generative models such as GANs can be explored to generate samples with target-domain characteristics during the training process, thereby enhancing the model's generalization ability in the target domain. Furthermore, generative models can be used for data augmentation to expand the training data for source and target domains and therefore increase the model's robustness.
(4) Incorporation of multitask learning and domain knowledge: Models' generalization ability can be improved by learning multiple related tasks in a single model. Simultaneously, domain knowledge can be integrated to provide additional information about the source and target domains, which can guide the model in domain adaptation.
(5) Enhancement of model interpretability: The interpretability of domain adaptation models can be increased to make the domain transfer process transparent. This task can be achieved through the introduction of interpretability metrics and visualization methods. It can also help researchers understand the model's behavior and influential factors during the transfer process.
(6) Online domain adaptation and incremental learning: Online domain adaptation methods and incremental learning algorithms can be designed to enable a model to adjust in real time as it continuously receives new data and adapt to the changes in the target domain. This task can improve the model's adaptability and practicality in dynamic environments.
(7) Incorporation of unsupervised or weakly supervised learning methods: Considering the scarcity of annotated data, domain adaptation techniques can be optimized by using unsupervised or weakly supervised learning methods. Doing so can effectively reduce annotation costs while enhancing the generalization ability of models in the target domain.
(8) Multimodal data fusion: Multimodal data (such as images, point clouds, and depth information) can be utilized for domain adaptation to fully leverage information from different data sources. Fusing multiple data types can enhance the performance and robustness of domain adaptation models.
(9) Knowledge-graph-based domain adaptation: Knowledge graphs can be employed to provide domain adaptation models with rich background knowledge and semantic

information. Combining knowledge graphs with domain adaptation techniques can improve the model's ability to understand and transfer complex scenarios.

4.7. Few-Shot/Zero-Shot Semantic Segmentation

Weakly supervised learning [67] and unsupervised learning [68] are becoming active research areas. Zero-shot and few-shot learning methods were introduced based on the premise that humans can recognize new concepts in a scene by using only a small amount of visual sample data. In the past, unsupervised segmentation was achieved by applying clustering algorithms such as K-means and graph cut [69] on hand-crafted image features. The few-shot semantic segmentation (FSS) method recognizes objects by leveraging few annotated examples, and it has a remarkable potential to segment new objects with samples annotated with few pixels. FSS places general semantic segmentation in few-shot scenarios in which the model performs dense pixel labeling of new classes with only a small number of support samples. However, existing FSS methods rely heavily on the integrity of visible image information. Therefore, a challenge in FSS is that under insufficient lighting or complex working conditions, visible images often fail to gain enough information, resulting in a sharp decrease in segmentation performance [70]. In this context, researchers have exploited the complementary and similar information of visible and thermal images to improve the performance of FSS [71].

The zero-shot semantic segmentation (ZSS) method uses an embedded vector to generate visual features under the condition of zero training samples [72–76]. Generally, pixels are aligned with semantic texts, and the segmentation model is applied to invisible object categories. This direction is promising because it can break the restriction in the number of segmentation categories. The main drawback of the ZSS method is that its prediction ability is insufficient, and it cannot distinguish the seen class from the unseen class. Notably, a broad ZSS (GZSL) was proposed in Ref. [77]. GZSL can identify seen and unseen classes simultaneously. The training of the feature extractor is achieved without considering semantic characteristics. The GZSL method decreases the prediction performance for the unseen class [78].

Recent work has been conducted on ZSS. To further reduce the burden of annotations, researchers introduced a challenging task called weakly supervised zero-shot semantic segmentation (WZSS). WZSS improved the performance of the HIOU of the model from 25.9 to 31.8. The results for HIOU (31.8) and MIO (22.0) in Ref. [79] were obtained on the PASCAL VOC 2012 data set. The researchers in [80] also proposed a novel context-aware feature generation network. This network can represent pixel-level context information based on category-level semantics and synthesize context-aware pixel visual features for the unseen classes. Experiments on PASCAL VOC, PASCAL Context, and COCO-Stuff showed that this method is much better than the existing zero-lens semantic division method because medical terms are specially set in the professional field and difficult to obtain. Therefore, the researchers in [81] put forward a new ZSL paradigm with three main contributions. First, prior knowledge of the division target was extracted from the previous model, which was called a relationship prototype. Second, a cross-mode adaptation module that could extend the prototype to the zero-shot model was developed. Lastly, a relationship prototype perception module was proposed to allow the zero-shot model to perceive the information contained in the prototype. Numerous experiments showed that the framework was much better than the existing framework.

The advantage of few-shot and zero-shot semantic segmentation methods is their ability to perform semantic segmentation in cases of data scarcity or when no annotated data are available, which makes them highly practical and economical. Furthermore, these methods can support model transferability, enabling their application to new scenarios and tasks.

However, few-shot and zero-shot semantic segmentation methods have certain limitations. First, due to the limited or absent annotated data, the accuracy and robustness of these methods may be reduced. Second, these methods require modeling of the semantic

information and properties of the classes, leading to computational complexity and increased time cost for model learning and inference. Lastly, the generalization ability of these methods needs further improvement.

Future research can focus on several areas. Few-shot and zero-shot semantic segmentation models can be improved to enhance their accuracy and robustness in scenarios involving scarce annotated data. Moreover, other efficient methods of learning semantic information and properties can be explored to reduce the computational complexity and time cost of models. The research on the generalization ability of few-shot and zero-shot semantic segmentation methods can also be deepened to explore effective generalization methods and strategies.

4.8. Attention Mechanism

An attention mechanism [82] is a visual mechanism that mimics humans. Human vision can pay more attention to the target area while suppressing other temporarily useless information. An attention mechanism can be viewed as a dynamic weight adjustment process based on the features of the input image. Attention mechanisms in deep learning are similar to human visual attention mechanisms [83]. Wang et al. [84] proposed a non-local operation to obtain image features. With the introduction of channel attention and spatial attention [85], different attention mechanisms have been applied to networks for semantic segmentation. Currently, attention mechanisms are mainly attached to the encoder–decoder structure.

Li et al. [86] constructed a pyramid attention network for semantic segmentation that exploited global context information for semantic segmentation by bypassing the decoder network and combining an attention mechanism and spatial pyramid for pixel-level extraction of precise dense features. Fu et al. [87] proposed a dual-attention network that could capture contextual dependencies based on a self-attention mechanism. Kang et al. [88] added an attention mechanism to the model based on the DeepLab V3+ semantic segmentation model and designed an image semantic segmentation model based on the attention mechanism. The experimental results showed that the model had a higher segmentation accuracy and running efficiency than the DeepLab V3+ model, U-Net model, and SegNet model. The authors in [89] proposed a hybrid attention semantic segmentation network (HAssNet) based on the FCN model that could extract the object and its surrounding environment through the large receptive field of the multiscale object. A spatial attention mechanism was introduced into the FCN, and a channel attention mechanism was designed to achieve semantic consistency. Experimental results on open remote sensing data sets showed that the MIoU of HAssNet was improved by an average of 6.7% compared with that of state-of-the-art segmentation networks.

The ultimate goal of attention mechanisms is to obtain abundant critical information on the current task among the many pieces of information; however, with the addition of an attention mechanism, the number of parameters of the model increases. Rational use of attention mechanisms can improve semantic segmentation accuracy. The emergence of numerous attention mechanisms can further promote the development of semantic segmentation.

The use of attention mechanisms in semantic segmentation tasks has two main advantages(as shown in Figure 10). On the one hand, it can enhance the model's focus on the target objects, reduce the interference of background noise, and improve the accuracy of segmentation. On the other hand, attention mechanisms can improve the recognition and utilization of feature information at different scales, allowing the model to effectively adapt to objects of different scales.

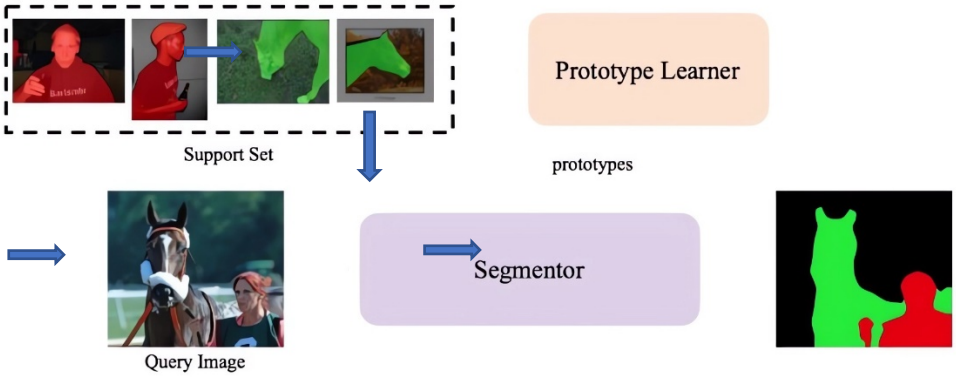


Figure 10. Few-shot semantic segmentation framework. Picture from "Few-Shot Semantic Segmentation with Prototype Learning".

However, attention mechanisms also have some drawbacks such as the need for network structure improvement(as shown in Figure 11), which increases computational complexity and consumes storage space. Additionally, attention mechanisms have a strong dependence on training data, which may lead to overfitting.

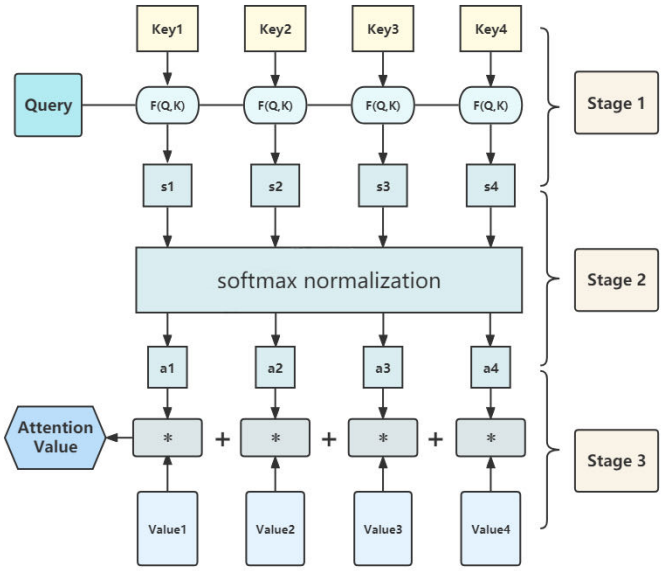


Figure 11. Schematic diagram of attention mechanism, * representing the process of weighted summation.

Future research directions include the establishment of efficient attention mechanisms that can improve segmentation accuracy without increasing computational complexity. Another direction is to study the applicability of attention mechanisms in different scenarios and investigate their effects on different data sets and tasks. Additionally, combining attention mechanisms with other techniques such as depth-separable convolution and graph convolutional networks (GCNs) can further enhance the performance of semantic segmentation.

4.9. Method Based on Multimodal Fusion

With the rapid development of deep learning, its excellent performance in various scene understanding tasks has been well demonstrated. However, in some complex or challenging conditions, multiple modalities need to be used in the same scene to provide complementary information. Multisource data fusion is a key step in technological development, and the rise of artificial intelligence technology has promoted the development of sensor fusion technology. The current research direction is to achieve semantic segmentation of multimodal data fusion by utilizing data acquired from multiple sensors. This

method can effectively improve the performance and robustness of semantic segmentation algorithms [90–92]. Examples include fusing RGB and thermal depth images [93] and fusing RGB images and lidar point clouds [94]. Zou et al. [95] developed a multimodal fusion network as a joint encoding model for verification. With this multimodal fusion assumption, multiple multimodal models were built in accordance with the proposed fusion method and evaluated on the KITTI and A2D2 data sets. The best fusion network achieved an over 85% lane line accuracy and an over 98.7% overall accuracy.

The method based on multimodal fusion has advantages and disadvantages when applied to semantic segmentation. The advantages lie in the method's ability to incorporate complementary information from multiple modalities, which can improve the accuracy and robustness of the segmentation results. However, the fusion of different modalities may also introduce additional noise and uncertainty, which can negatively affect segmentation performance. In addition, the fusion process itself can be computationally intensive and may require specialized hardware or software for efficient implementation.

To effectively address the aforementioned challenges, future researchers should focus on the following aspects:

(1) Introducing adaptive fusion strategies: Previous multimodal fusion methods mainly relied on fixed fusion strategies. However, in different scenarios and applications, the importance of information from different modalities may vary. Therefore, introducing adaptive fusion strategies that dynamically adjust the weights of different modalities in accordance with the scene context or application requirements can improve fusion results. This goal can be achieved using attention mechanisms, which allow the network to automatically determine the contributions of different modalities.

(2) Utilizing GCNs: Given that GCNs can effectively process irregular structured data, they can be considered for multimodal fusion. By representing multimodal data as graph structures, GCNs can successfully capture the relationships between different modalities, further improving fusion performance.

(3) Cross-modal self-supervised learning: A major challenge in multimodal fusion is effectively transferring information between different modalities. By introducing cross-modal self-supervised learning, models can automatically learn how to share information between modalities. This approach can be realized through alignment and generation tasks such as reconstructing one modality's data by generating another modality's data.

(4) Adopting multiscale fusion strategies: Information from different modalities may be complementary on different scales. To fully exploit this feature, multiscale fusion strategies can be adopted. By fusing on different spatial scales, the local and global relationships between modalities can be captured, thus enhancing fusion performance.

(5) Combining domain adaptation with multimodal fusion: To further improve the robustness of multimodal fusion, domain adaptation techniques can be combined with multimodal fusion. By adopting multimodal data from source and target domains, the distribution discrepancy between domains can be effectively reduced, thus enhancing the model's generalization ability in new domains.

(6) Incorporating knowledge distillation: In multimodal fusion, knowledge distillation can be considered to improve model efficiency and scalability. By allowing small models to learn the relationships between different modalities, computational and storage requirements can be reduced while maintaining high performance.

(7) Applying end-to-end multimodal training methods: Traditional multimodal fusion methods typically require pretraining of single-modality models before fusion, which may lead to computational resource wastage and information loss. By developing end-to-end multimodal training methods, the optimal means to fuse different modalities can be directly learned, thereby improving overall performance.

(8) Utilizing ensemble learning methods: In multimodal fusion, ensemble learning methods can be considered to enhance performance. By combining multiple models with

different fusion strategies, the accuracy and robustness of semantic segmentation can be further improved. Ensemble methods may include voting, bagging, and boosting.

5. Common Data Sets and Evaluation Indicators

Many data sets are available for semantic segmentation tasks. This paper introduces nine representative and general-purpose image segmentation data sets; namely, CamVid (Figure 12), COCO (Figure 13), BDD (Figure 14), Cityscapes (Figure 15), PASCAL VOC 2012 (Figure 16), SBD (Figure 17), KITTI (Figure 18), Mapillary Vistas 3.0 (Figure 19) and VSPW. We have conducted statistics from seven aspects, including the creation time of the dataset, its application scope, classification categories, dataset size, training set, validation set, and test set, as shown in Table 1. These data sets cover different scenes, lighting conditions, and object categories and can be used to evaluate the performance of different algorithms in various complex situations. Moreover, these data sets have extensive application value and can be used to train and test different types of semantic segmentation models.

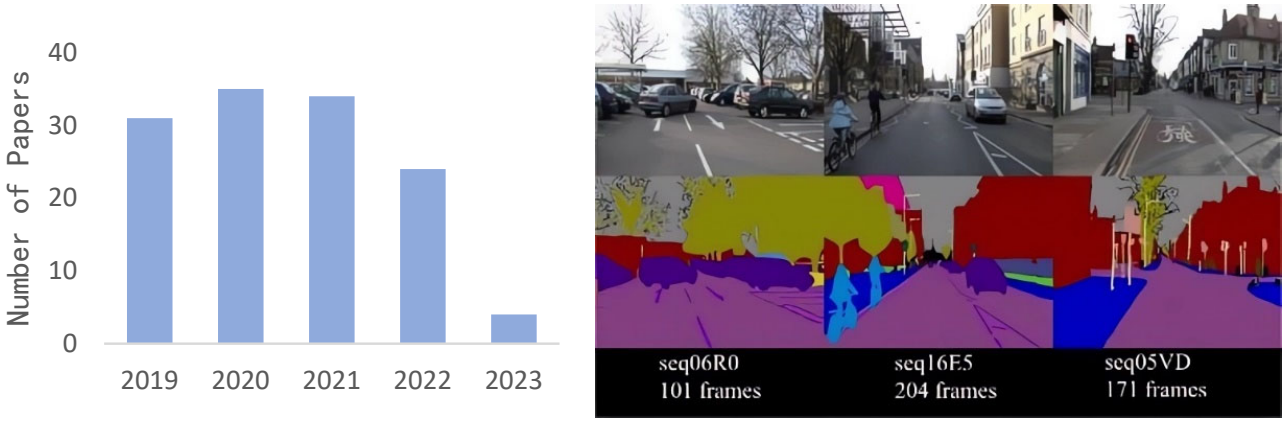


Figure 12. CamVid data set usage and examples.

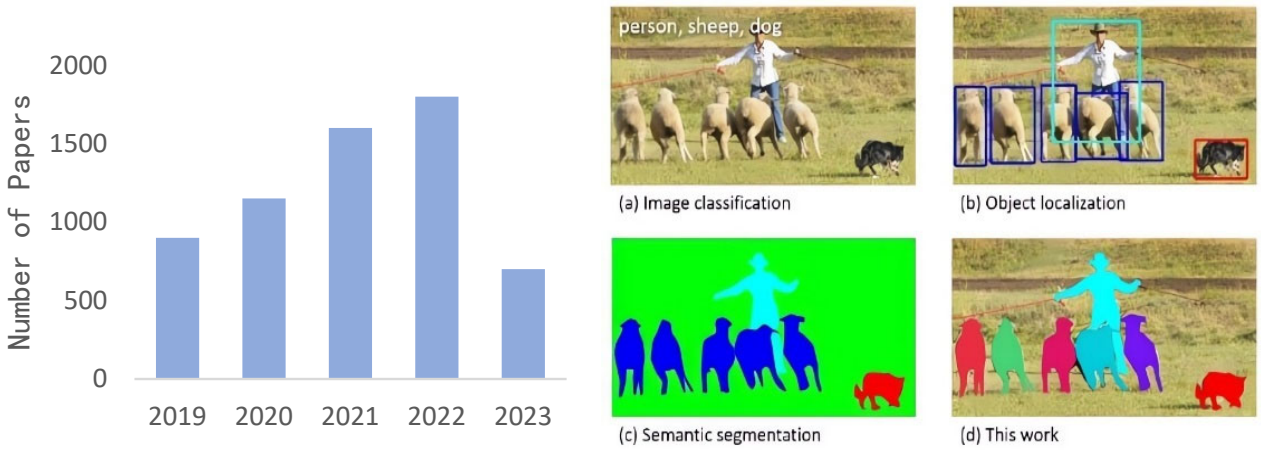


Figure 13. COCO data set usage and examples.

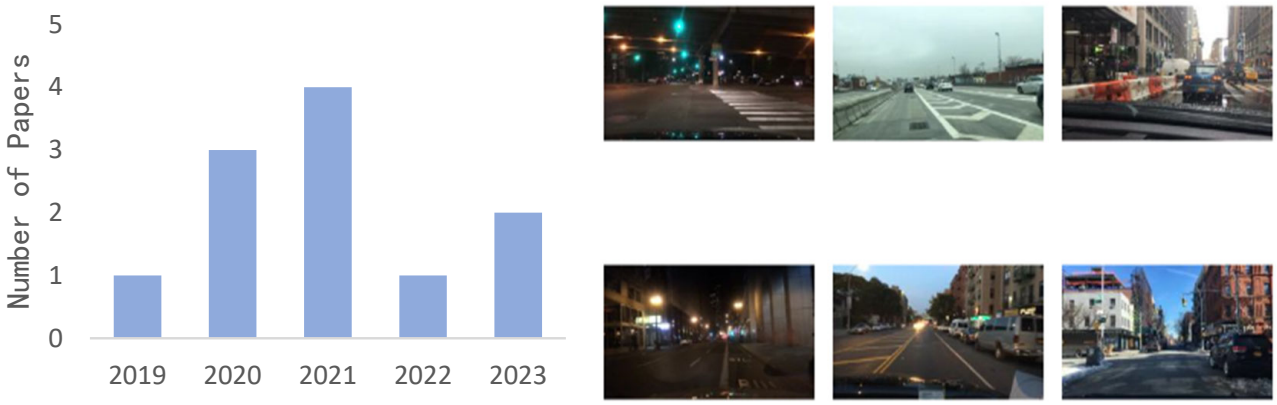


Figure 14. Berkeley Deep Drive data set usage and examples.

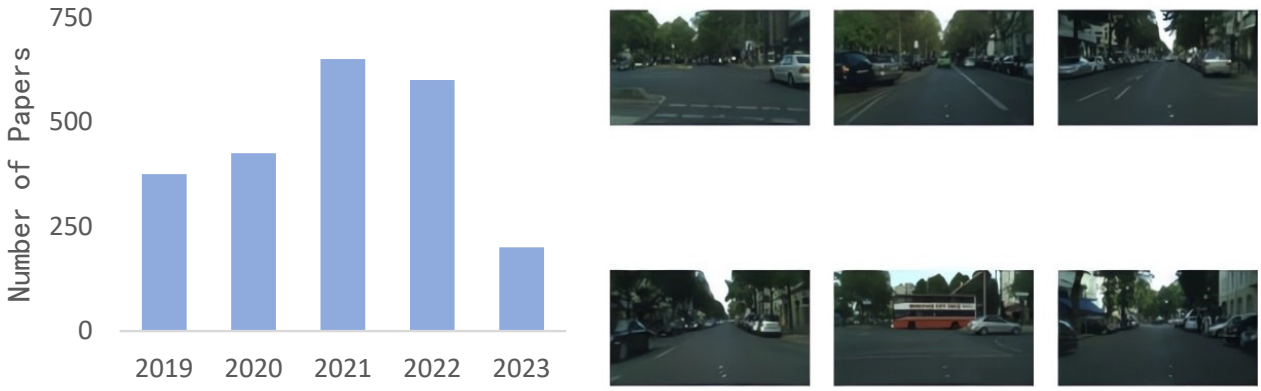


Figure 15. Cityscapes data set usage and examples.

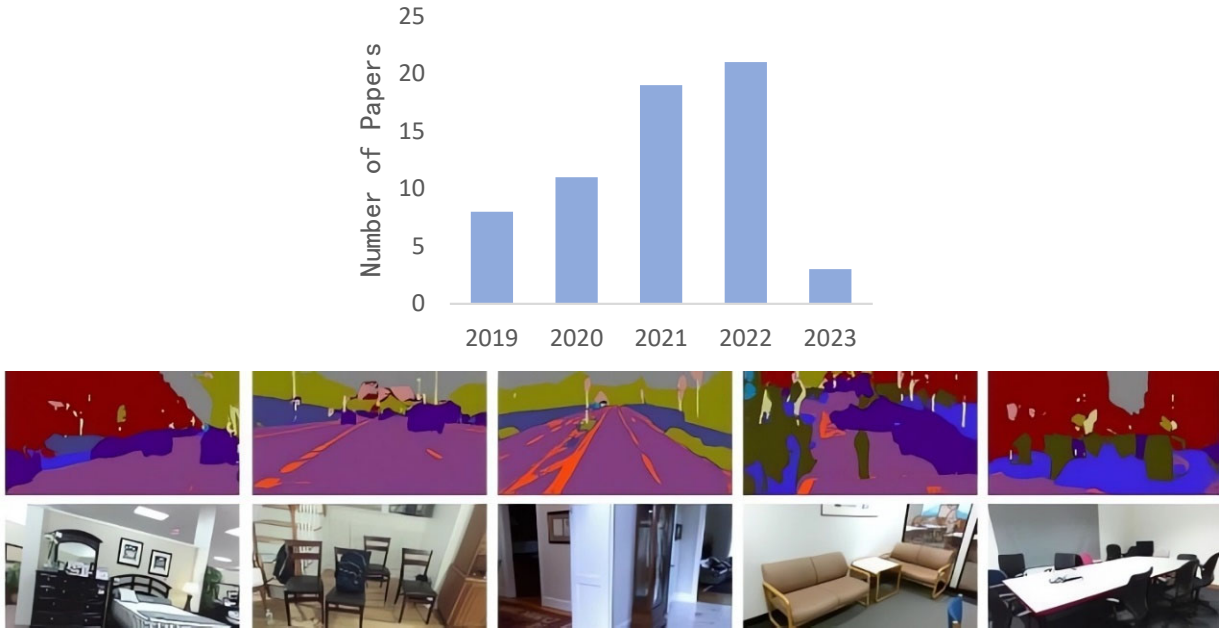


Figure 16. PASCAL VOC 2012 data set usage and examples.

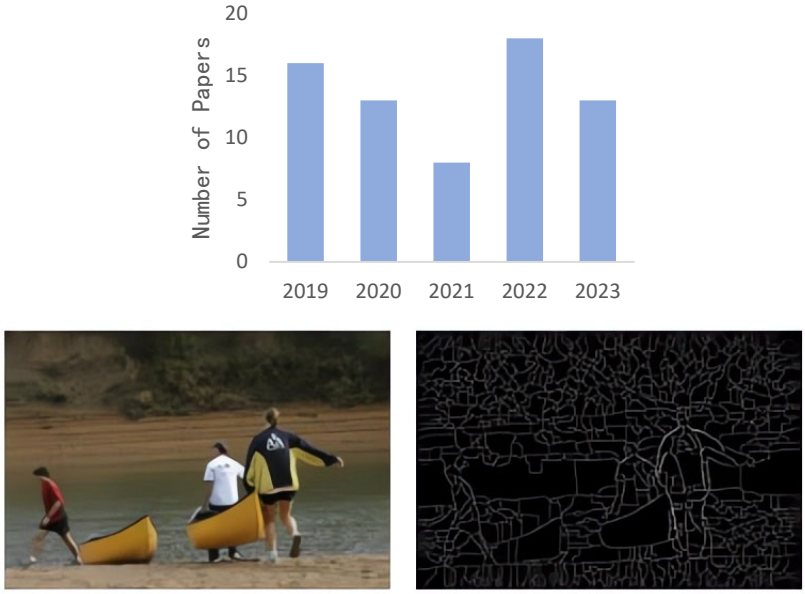


Figure 17. Semantic Boundaries data set usage and examples.

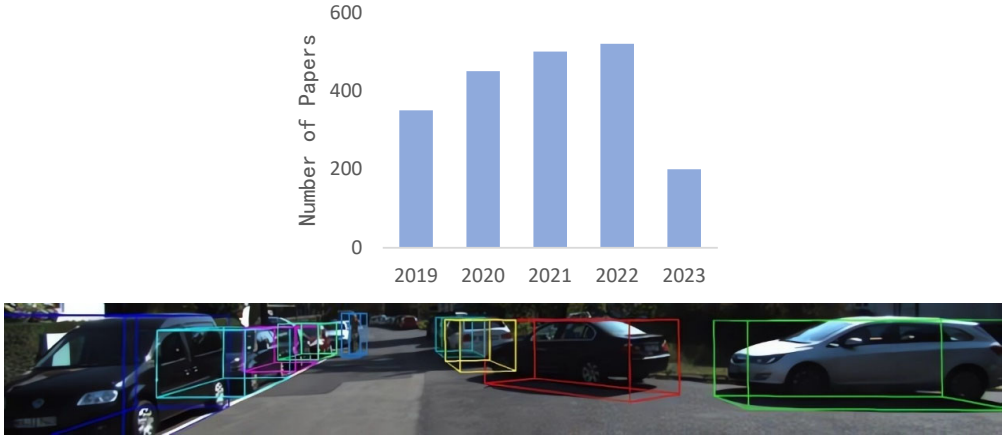


Figure 18. KITTI data set usage and examples.

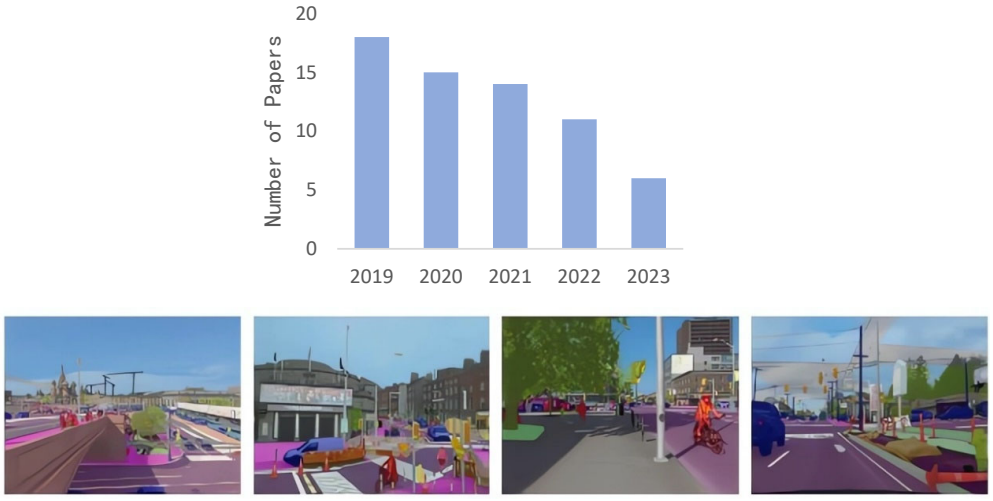


Figure 19. Mapillary Vistas data set usage and examples.

Table 1. Popular semantic segmentation data sets.

Data Set	Time	Application Scenario	Classification	Quantity	Number of Training Sets	Number of Validation Sets	Number of Test Sets
CamVid	2008	City street	32	700+	367	100	233
COCO	2014	Multiple scenarios	81	308,000	82,783	40,504	81,434
BDD	2018	City street	19	100,000	70,000	—	30,000
Cityscapes	2016	City street	30	5000	2975	500	1525
PASCAL-VOC 2012	2015	Multiple scenarios	21	9993	1464	1449	1452
SBD	2014	Multiple scenarios	21	—	8498	2857	—
KITTI	2015	Multiple scenarios	7	7000+	3712	—	3769
Mapillary Vistas 3. 0	2021	City scenarios	66	250 K	180 K	20 K	30 K
VSPW	2021	City scenarios	124	3537 (video)	—	—	—

However, these data sets also have limitations. For example, the annotations of some data sets are inaccurate and not complete enough, which may affect the accuracy and robustness of the algorithms. In addition, some data sets have limitations in image size and quantity, which may restrict the application scope of the algorithms. Therefore, when selecting data sets, their advantages and disadvantages must be considered, and reasonable choices and processing must be made.

We conducted a statistical analysis of the usage of each data set in articles in over the past five years by using Google Scholar data.

1. The CamVid data set [96] is a data set of urban road scenes publicly released by the University of Cambridge. The data set includes more than 700 accurately annotated images that can be used for supervised learning. The images can be divided into training, validation, and test sets. Meanwhile, 11 commonly used categories are employed in the CamVid data set to evaluate segmentation accuracy. CamVid is the first collection of videos with semantic labels of target categories, and in most cases it can adapt to multiple technical approaches [97]. The CamVid data set is one of the most advanced data sets in real-time semantic segmentation [98]. Its image quality is high, and most images were taken under different sunlight, weather, and seasonal conditions. Hence, the CamVid data set is suitable for studying factors that affect semantic segmentation performance. In addition, the annotations of the CamVid data set are detailed, with each pixel accurately labeled with its corresponding category. The data set is thus suitable for supervised learning, especially for deep learning methods that require precise pixel-level labels.

2. The COCO data set [99] is a large-scale image data set and one of the most challenging instance segmentation data sets in the field of image processing. This data set is widely used for various image processing tasks including object detection, key point detection, caption generation, and segmentation. The COCO data set contains over 330,000 images, with 220,000 annotated images that include rich semantic information such as 1.5 million objects and 80 object categories.

The data set's annotations are rich and of high quality, providing important semantic information and object annotations for the training and evaluation of deep learning models. Additionally, the COCO data set exhibits high quality and diversity in terms of data distribution, complexity, and variability, so it can meet the diverse needs of real-world scenarios.

However, the limitations of this data set are evident. First, the data set contains relatively few object categories, so the needs of specific domains or tasks cannot be easily met. Second, the images in the data set are mostly real images from natural scenes and cannot fully satisfy the demands of special scenarios.

3. The Berkeley Deep Drive (BDD) data set [100], a detailed data set of autonomous driving urban scenes, includes tasks such as object detection, multiobject tracking, semantic segmentation, instance segmentation, segmentation tracking, and lane detection. The BDD data set has 19 categories for semantic segmentation, but its application is limited due to its

samples not being practical for semantic segmentation of urban scenes. Specifically, the BDD data set contains approximately 100,000 images, among which about 70,000 are used for training and about 30,000 are used for testing.

The advantage of this data set lies in its detailed urban scene data, which can be applied to multiple autonomous driving scene tasks. In addition, the multitask nature of the data set makes it highly valuable for research and application in the field of autonomous driving.

The application of the BDD data set in semantic segmentation is limited because its samples are not practical for use in the semantic segmentation of urban scenes, which may affect its practicality in certain scenarios. In addition, the data set is large in scale and requires a large amount of computational resources for data processing and model training, which may pose challenges to research and applications.

4. The Cityscapes data set [101] is a popular data set for autonomous driving tasks aimed at object recognition and localization within a scene. The data set comprises a large collection of high-resolution street-level images from 50 different cities, including 5000 annotated images with pixel-level annotation for 19 object categories. To facilitate model training and testing, the data set is divided into three subsets: 2975 training images, 500 validation images, and 1525 test images.

The benefits of using this data set include the large number of high-resolution images with pixel-level annotation that can effectively help models learn object recognition and localization. Additionally, because the data set covers scenes from 50 different cities, it effectively encompasses various scenes and traffic conditions, making the model highly generalizable.

However, the data set's large size requires a large amount of computing resources and storage space, so it may not be suitable for some scenarios with limited resources. Furthermore, the annotations in the data set may contain errors and require certain data cleaning and processing to ensure the reliability of the model's training and testing results.

5. PASCAL VOC 2012 [102] is a widely used data set for scene semantic segmentation that provides labeled data for supervised learning in visual tasks. The data set consists of 21 object categories; the original data set includes 1464, 1449, and 1456 images for training, validation, and testing, respectively. Notably, images in the training and validation sets are pixel-level annotated, which is crucial for scene semantic segmentation. This data set has a high impact and reference value in image segmentation and computer vision because it contains rich scene categories and annotation information, which can effectively support the training and evaluation of scene semantic segmentation and be widely applied in multiple tasks such as image segmentation and object detection in computer vision. Many algorithms and models based on this data set have been proposed, so this data set is of high reference value in academia and industry.

One of the limitations of this data set is the small number of object categories, which cannot cover all types of scenes and may thus affect the generalization performance of the model. In addition, the small image size in the data set may have an adverse effect on certain scene semantic segmentation tasks that require high resolution.

6. The Semantic Boundaries data set (SBD) [103] is commonly used in semantic segmentation tasks to detect boundaries between different objects in an image. The data set was created by researchers at Stanford University and includes images of various indoor and outdoor scenes that can be used to train and evaluate the performance of various semantic segmentation algorithms. The data set is relatively large, with over 20,000 images that are divided into training and testing sets at a ratio of 4:1. The training set contains over 16,000 images, and the testing set contains about 4000 images. The data set can effectively train and evaluate the performance of deep learning models. The limitation of this data set is that it only includes indoor and outdoor scenes and lacks images of some special scenes such as industrial and agricultural production.

The imbalanced categories and diversity of scenarios in the SBD data set provide opportunities for researchers to conduct further in-depth research on related issues such as how to solve the imbalanced category problem through data enhancement or an improved loss

function as well as how to use multitask learning and other technologies to simultaneously improve the performance of the model in semantic segmentation and boundary detection.

7. The KITTI data set [104] was jointly created by the Karlsruhe Institute of Technology in Germany and the Toyota Technological Institute at Chicago. It is currently the largest computer vision algorithm evaluation data set for autonomous driving scenes, including urban, rural, and highway scenes. The data set contains a large number of images, point clouds, semantic segmentation annotations, 2D and 3D bounding box annotations, and other data. The image data include both original RGB and grayscale images, with a total of more than 7000 images. The images in the data set are divided into training and test sets, with 3712 images in the training set and 3769 images in the test set. In addition, the data set provides detailed sensor data such as vehicle trajectories, GPS positioning, and LiDAR.

The advantage of the KITTI data set lies in its rich scenes and diverse data types, which make the data set suitable for evaluating algorithms for various autonomous driving scenes. In addition, the KITTI data set contains detailed sensor data. It serves as a realistic and rich source of data for algorithm research. However, the KITTI data set has some limitations such as a low image resolution and a relatively small scale, which indicates that it may not cover all scenarios in autonomous driving scenes.

8. Mapillary Vistas 3.0 is a large-scale semantic segmentation data set of urban scenes that includes more than 250,000 high-resolution images and pixel-level semantic segmentation annotations (Figure 20). It contains 66 categories (including background categories) with more than 250,000 images (180,000 images in the training set, 20,000 images in the verification set, and 30,000 images in the test set). All images in the training and verification sets provide pixel-level annotation, whereas the images in the test set do not provide annotation information.

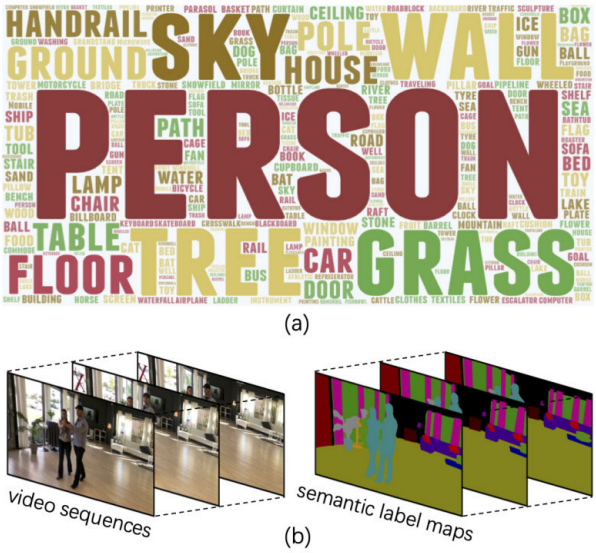


Figure 20. Mapillary Vistas data set examples. (**a**) shows an example of a dataset image, while (**b**) shows the process of segmenting a video sequence and mapping semantic labels.

Although the Mapillary Vistas data set contains up to 66 categories, not all of them have sufficient samples. As a result, the model may demonstrate poor performance in some categories. Given the class imbalance problem in the Mapillary Vistas data set, studying how to handle class imbalance and complex scene problems is an important research direction in the future. For example, this problem can be solved through data enhancement, using an improved loss function, or via multitask learning using the annotations.

9. VSPW [105] is the first multiscene, large-scale, video semantic segmentation data set. It has 3537 labeled videos and 251,632 semantically segmented pictures, and it covers 124 semantic categories. Each video contains a complete shot that lasts for about 5 s on the average. More than 96% of the captured videos have a high spatial resolution from 720 p to

4 K. VSPW was the first attempt to solve the challenging video scene resolution task in the field by considering various scenarios.

Given that the VSPW data set contains video frames, maintaining temporal consistency in semantic segmentation has become an important research direction similar to using other methods such as recurrent neural networks or short-term memory networks to capture temporal information.

In addition to the abovementioned data sets, other data sets are used in the medical field [106,107], remote sensing field [108,109], and others.

The authors of Ref. [110] proposed a novel scene segmentation method that enhanced the accuracy and efficiency of scene segmentation by using a hierarchical feature extraction network. The authors employed dilated convolution and multiscale feature fusion technology at different convolution layers to improve feature extraction accuracy and coverage and enable the model to capture semantic information in the image effectively. An adaptive gating mechanism was introduced to dynamically control the degree of feature fusion between different levels, making the model increasingly flexible and adjustable. Experimental results showed that the proposed method performed well on commonly used scene segmentation data sets, and it had a higher accuracy and speed compared with similar methods. However, the data sets used in the study were common, and the network may not perform optimally in special scenarios involving rain, snow, and fog. Moreover, the complex network structure requires long training and debugging as well as high computing resources.

The authors of Ref. [111] proposed FAPN, which utilizes feature alignment technology to extract information from different resolution feature pyramids to improve dense image prediction accuracy. FAPN has good robustness and is unaffected by changes in input image size and scale, so it has high efficiency and performs well in dense image prediction. However, the network structure of FAPN is complex, and its performance may decrease for images with geometric and other kinds of distortions.

The authors of Ref. [112] introduced a network structure called DDRNet, which effectively utilizes low-resolution and high-resolution feature information via a dual-resolution feature fusion strategy to improve the accuracy of road scene semantic segmentation. The network has a high computing speed and can achieve real-time road scene segmentation. The experimental results demonstrated the robustness and generalization of the network in different scenarios. However, DDRNet was mainly designed for road scene segmentation and may require adjustment and retraining for other scenes. Training and debugging may be difficult and require high-performance computing resources due to the use of a complex feature fusion strategy.

MIFNet (proposed in Ref. [113]) is a lightweight neural network architecture that reduces network parameters and computational complexity while maintaining model accuracy and improving real-time performance on mobile devices. MIFNet adopts multiscale information fusion technology to extract information from feature maps of different scales. It enhances the model's receptive field and improves its recognition and understanding of objects. Experimental verification on multiple publicly available data sets demonstrated the good performance of MIFNet. However, the network structure of MIFNet is simple and may have limited processing capabilities for complex scenes and tasks. The method requires parameter and hyperparameter tuning to achieve optimal performance.

Ref. [114] proposed a new neural network model called ViT-A, which adapts the ViT model to dense prediction tasks and improves the accuracy of image segmentation by using the adapter method. The method has high computational efficiency and flexibility and can be used for transfer learning on different data sets and tasks. Experimental results showed the method's good performance in various complex scenarios. However, the ViT-A method has limitations in the size and scale of input images and requires preprocessing and scaling before segmentation. The adapter structure of ViT-A also requires certain computational resources that may affect the method's real-time performance and deployability.

Ref. [115] proposed a neural network model called DRAN, which uses a dual-relation-aware mechanism to model the relationships between different objects in images and improve the accuracy of scene segmentation. The method incorporates an attention mechanism that can adaptively focus on important regions in images, thus reducing the interference of redundant information and improving the efficiency of image segmentation. Experimental results revealed the model's good performance in various complex scenarios. However, the DRAN model is complex and requires considerable computational resources and training time. Additionally, DRAN is sensitive to the size and scale of input images and requires preprocessing and scaling before segmentation

When comparing Tables 2 and 3, we found that image semantic segmentation technology based on deep learning has achieved great progress in accuracy. Each architecture has its unique advantages. With the common goal of using key technologies to improve segmentation accuracy, the future research direction in this field is the integration of the advantages of multiple architectures.

Table 2. Performances of classical semantic segmentation methods on different data sets.

Method	Time	Backbone	VOC 2012 (MIoU/%)	Cityscapes (MIoU/%)	CamVid (MIoU/%)
FCN	2015	VGG16	62.2	65.3	—
U-Net	2015	VGG16	—	—	—
DeepLab v1	2016	ResNet	71.6	—	—
DeepLab v2	2017	ResNet	79.7	70.4	—
DeepLab v3	2017	ResNet	86.9	81.3	—
DeepLab v3+	2018	ResNet	89	82.1	
SegNet	2017	VGG16	—	—	60.1
PSPNet	2017	ResNet	85.4	80.2	—

Table 3. Performance of recent semantic segmentation methods on different data sets.

Method	Time	Backbone	Cityscapes (MIoU/%)	CamVid (MIoU/%)
HFEN [110]	2021	—	69.5	66.1
FaPN [111]	2021	ResNet-18	75.0	—
DDRNet [112]	2021	DDRNet-39	80.4	—
MIFNet [113]	2021	MobileNetV2	72.2	—
ViT-Adaptor-L [114]	2022	UPerNet	85.2	—
DRANet [115]	2020	ResNet-101	82.9	—

In the process of implementing a deep neural network, accuracy evaluation may vary from problem to problem, and traditional semantic segmentation methods use accuracy or precision as a performance evaluation metric. However, for deep learning, many performance metrics are applicable to classification, object detection, and semantic segmentation.

Typically, the final performance is evaluated in terms of accuracy, including pixel accuracy (PA), mean pixel accuracy (MPA), MIoU, and frequency weighted intersection over union (FWIoU). MIoU is often used to measure the performance of semantic segmentation models. It can be obtained by comparing the ground truth with the output map after passing the image to the export model. PA is the simplest pixel-level evaluation index in semantic segmentation. It only needs to calculate the ratio of correctly classified pixels in the image to the total pixels in the image, as shown in Formula (1):

$$PA = \frac{\sum_{i=0}^{n} p_{ii}}{\sum_{i=0}^{n} \sum_{j=0}^{n} p_{ij}} \quad (1)$$

where p_{ii} represents the number of correctly classified pixels, p_{ij} is the number of pixels that should belong to the *i*-th category but are classified into the *j*-th category, and n is the number of categories.

PA describes the segmentation accuracy of all categories in an image and is commonly used to estimate the overall segmentation performance. The closer PA is to 1, the better the model performs. However, PA contains limited information and may mask poor segmentation results in certain categories, thereby failing to reflect the segmentation accuracy of individual categories. To address this issue, class pixel accuracy (CPA) can be used to evaluate the segmentation accuracy of each category separately. For the segmentation result of the *i*-th category, the CPA calculation formula is as follows (2):

$$CPA = \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{2}$$

MPA represents the average pixel accuracy of all object categories in the image. The main advantages of MPA are its relatively simple computation and its ability to provide an overall assessment of the model's performance across all categories. However, it also has some limitations. In semantic segmentation tasks, the number of pixels per category in the image is often imbalanced, with some categories having considerably more pixels than others. MPA assigns equal weights to all categories during calculation. As a result, categories containing more pixels than others exert a disproportionately large effect on MPA, and the performance of categories with few pixels is disregarded. MPA's calculation considers only the model's correct classification of pixels for each category and does not consider cases in which the model incorrectly classifies pixels as other categories. This situation may lead to an overestimation of the model's performance, as shown in Formula (3):

$$MPA = \frac{1}{n+1} \sum_{i=0}^{n} \frac{p_{ii}}{\sum\limits_{j=0}^{n} p_{ij}} \tag{3}$$

MIoU is the ratio of the intersection of the true value of the segmentation result and its union, which is calculated by class then averaged. MIoU considers the model's correct classification of pixels for each category (true positive; TP), cases in which the model misclassifies pixels as other categories (false positive; FP), and cases in which the model incorrectly classifies category i's pixels as other categories (false negative (FN)). Thus, MIoU provides a comprehensive performance evaluation. Compared with MPA, MIoU is more tolerant of class imbalances in data sets because MIoU accounts for the correct, incorrect, and unclassified pixels of each category during computation, thus better reflecting the model's performance across all categories.

However, MIoU may underestimate the model's performance on small objects with few pixels. To address this issue, researchers have proposed a weighted MIoU in which each category's IoU is weighted based on the number of pixels in that category. The weight of categories with few pixels is increased in the overall score as shown in Formula (4):

$$MIoU = \frac{1}{n+1} \sum_{i=0}^{n} \frac{p_{ii}}{\sum\limits_{j=0}^{n} p_{ij} + \sum\limits_{j=0}^{n} p_{ji} - p_{ii}} \tag{4}$$

FWIoU is an enhancement of MIoU, and it aims to weigh the category of each pixel in accordance with its frequency. FWIoU augments the weight of less frequent categories in the overall score by weighing the IoU of each category, which helps reflect the model's performance across all categories, particularly in data sets with class imbalance. Similar to MIoU, FWIoU comprehensively captures the model's ability to correctly classify pixels of each category (TP), incorrectly classify pixels as other categories (FP), and misclassify pixels of category i as other categories (FN). However, because FWIoU is weighted based

on pixel frequency, it may lead to overemphasis on large targets with a high pixel count. This situation implies that when evaluating model performance, FWIoU might pay too much attention to the performance of large targets and could overlook the performance in small targets as shown in Formula (5):

$$FWIoU = \frac{1}{\sum_{i=0}^{n}\sum_{j=0}^{n} p_{ij}} \sum_{i=0}^{n} \frac{\sum_{j=0}^{n} p_{ij} p_{ii}}{\sum_{j=0}^{n} p_{ij} + \sum_{j=0}^{n} p_{ji} - p_{ii}} \tag{5}$$

Cross-entropy loss plays a pivotal role in the training of deep learning models, especially in classification tasks. The cross-entropy error (CEE) is used to represent the residual between the predicted results and the labels. The fundamental idea is that when the model has high confidence in the correct prediction, the CEE decreases. Conversely, if the model has high confidence in the wrong prediction, the CEE increases. Therefore, CEE optimization is about making the model highly confident in the correct prediction.

Cross-entropy loss is considered the logarithm of the likelihood function for the correct labels to ensure the differentiability of the loss function. Typically, the base of the logarithm can be either two or the natural constant e. Neural network classifiers trained using CEE have distinct performance advantages over mean squared errors on finite data sets. Cross-entropy describes the dissimilarity between two probability distributions; the smaller the difference is, the closer the two probability distributions are. Through gradient descent, cross-entropy loss continually brings the predicted probability distribution close to the label probability distribution in order to accurately reflect the error between the predicted segmentation results and the labels.

However, because the CEE is proportional to the logarithm of the model's predicted probability, it is sensitive to outliers in the model's predictions. This sensitivity could cause the model to overreact to some anomalies, thereby affecting the model's performance. Therefore, when using cross-entropy loss, special attention must be paid to the stability and robustness of model training. The definition of cross-entropy loss is as follows (6):

$$L_{CEE} = -\frac{1}{n}\sum_{n}\sum_{c} t_{n,c} \log_a(y_{n,c}) \tag{6}$$

In addition to the abovementioned metrics, other performance indicators need to be considered when evaluating the performance of a neural network.

These indicators include the frames per second (also known as latency), which refers to the time required for a neural network to process an image; the exported network learning parameters; storage space occupied by the neural network model (that is, network size); the processing capabilities of hardware resources such as the CPU, GPU, and memory; the power consumption of the entire system; and the memory bandwidth utilization rate, which refers to the ratio of the number of bytes transmitted from memory to the total number of transmitted bytes. In addition, other factors such as training indicators and the system environment affect the performance of neural networks. These indicators play an equally important role in optimizing and selecting neural network models.

6. Prospects and Challenges

We investigated the key methods in 2D image semantic segmentation based on deep learning models and showed impressive results for various image semantic segmentation tasks through experiments. We also discussed some challenges and promising research directions for deep-learning-based semantic segmentation.

An important research direction toward improving the generalization ability of models is determining how to obtain data sets that approximate real-world environments (with complex and variable weather conditions). In semantic segmentation, a challenging aspect is 3D semantic segmentation in computer vision, which can be applied to fields such as au-

tonomous driving, medicine, and robotics. The introduction of depth maps enables research to focus on 3D scenes. Although 3D data sets are difficult to obtain and label, they contain more semantic information than 2D data sets, making 3D scene semantic segmentation highly valuable and widely applicable. For example, based on field data collected using 3D scanners, depth cameras, drones, or other 3D capturing devices, data augmentation and enhancement are performed using computer graphics and machine learning techniques to generate additional weather and lighting conditions. This approach leverages both the complexity of the real world and the powerful capabilities of computer technology.

Currently, semantic segmentation for real-time network segmentation tasks still has imperfections. Therefore, balancing the accuracy and efficiency of semantic segmentation remains an important research direction. Recent research trends indicate that an increasing number of scholars are beginning to study alternative pixel-level annotation methods such as unsupervised, semi-supervised, and weakly supervised methods. Among them, weakly supervised semantic segmentation methods typically outperform other methods. In the context of this research background, it becomes crucial to further explore efficient network architectures, optimize computational and storage efficiency, integrate multimodal and multiscale information, incorporate more data, and improve data augmentation techniques.

For instance, in the domain of Domain-Adversarial Neural Networks (DANNs), the training of the model aims to maximize the predictive performance on the target domain while minimizing the distribution discrepancy between the source and target domains (among other factors).

Another important research direction is cross-domain semantic segmentation because data from different domains or scenarios have considerable differences. An important research direction toward improving the generalization ability of models is determining how to apply existing semantic segmentation models to new domains or scenarios. In addition, robust semantic segmentation is a key research direction. For complex real-world scenes such as those containing lighting changes, occlusions, and image noise, researchers can design robust semantic segmentation models to improve these models' performance in complex scenarios.

One approach to address this issue is through data augmentation. By creating and incorporating variations of the training data (such as changing image brightness and contrast, adding random noise, simulating occlusions, etc.), we can enable the model to learn these variations during the training process and improve its performance when faced with such challenges. Another strategy is to utilize adaptive models. These models can dynamically adjust their behavior to adapt to the characteristics of the input data and so on. With further research in the field of artificial intelligence, the concepts of semantic segmentation and image understanding can be integrated to achieve more accurate segmentation by comprehending the scene, objects, and contextual information in images.

These research directions are crucial and can further improve the accuracy, robustness, efficiency, and generalization ability of semantic segmentation methods. The ethical implications of semantic segmentation also need to be considered. As these models become increasingly accurate and widespread, they must be used in a responsible and ethical manner with appropriate safeguards in place to protect individual privacy and prevent misuse.

Author Contributions: Conceptualization, Y.G. and G.N.; methodology, Y.G.; software, Y.G.; validation, W.G., M.L. and G.N.; formal analysis, Y.G.; investigation, Y.G. and W.G.; resources, Y.G.; data curation, Y.G.; writing—original draft preparation, Y.G.; writing—review and editing, G.N and M.L.; visualization, Y.G.; supervision, G.N.; project administration, Y.G.; funding acquisition, G.N. All authors have read and agreed to the published version of the manuscript.

Funding: There is funding support in this article by the National Key Research and Development Project (2018YFE0206500).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors have no conflict of interest to declare that are relevant to the content of this article.

References

1. Janai, J.; Güney, F.; Behl, A.; Geiger, A. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Found. Trends® Comput. Graph. Vis.* **2020**, *12*, 1–308. [CrossRef]
2. Lu, X.; Wang, W.; Ma, C.; Shen, J.; Shao, L.; Porikli, F. See More, Know More: Unsupervised Video Object Segmentation with Co-Attention Siamese Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3618–3627.
3. Lu, X.; Wang, W.; Shen, J.; Crandall, D.; Luo, J. Zero-Shot Video Object Segmentation with Co-Attention Siamese Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2228–2242. [CrossRef] [PubMed]
4. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5168–5177.
5. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 11–18 December 2015.
6. Wei, Z.; Sun, Y.; Wang, J.; Lai, H.; Liu, S. Learning adaptive receptive fields for deep image parsing network. In Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
7. Batra, A.; Singh, S.; Pang, G.; Basu, S.; Jawahar, C.V.; Paluri, M. Improved road connectivity by joint learning of orientation and segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
8. Farha, Y.A.; Gall, J. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
9. Sun, J.; Li, Y. Multi-feature fusion network for road scene semantic segmentation. *Comput. Electr. Eng.* **2021**, *92*, 107155. [CrossRef]
10. Yanc, M. Review on semantic segmentation of road scenes. *Laser Optoelectron. Prog.* **2021**, *58*, 36–58.
11. Li, J.; Jiang, F.; Yang, J.; Kong, B.; Gogate, M.; Dashtipour, K.; Hussain, A. Lane-deeplab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. *Neurocomputing* **2021**, *465*, 15–25. [CrossRef]
12. Ghosh, S.; Pal, A.; Jaiswal, S.; Santosh, K.C.; Das, N.; Nasipuri, M. SegFast-V2: Semantic image segmentation with less parameters in deep learning for autonomous driving. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 3145–3154. [CrossRef]
13. Mao, J.; Xiao, T.; Jiang, Y.; Cao, Z. What can help pedestrian detection? In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
14. Guo, Z.; Liao, W.; Xiao, Y.; Veelaert, P.; Philips, W. Weak segmentation supervised deep neural networks for pedestrian detection. *Pattern Recognit.* **2021**, *119*, 108063. [CrossRef]
15. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, VA, USA, 26 June–1 July 2016.
16. Ouyang, S.; Li, Y. Combining deep semantic segmentation network and graph convolutional neural network for semantic segmentation of remote sensing imagery. *Remote Sens.* **2020**, *13*, 119. [CrossRef]
17. Peng, S.; Liu, Y.; Huang, Q.; Zhou, X.; Bao, H. Pvnet: Pixel-wise voting network for 6dof pose estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
18. Gao, F.; Li, H.; Fei, J.; Huang, Y.; Liu, L. Segmentation-Based Background-Inference and Small-Person Pose Estimation. *IEEE Signal Process. Lett.* **2022**, *29*, 1584–1588. [CrossRef]
19. Cheng, Z.; Qu, A.; He, X. Contour-aware semantic segmentation network with spatial attention mechanism for medical image. *Vis. Comput.* **2022**, *38*, 749–762. [CrossRef]
20. Asgari Taghanaki, S.; Abhishek, K.; Cohen, J.P.; Cohen-Adad, J.; Hamarneh, G. Deep semantic segmentation of natural and medical images: A review. *Artif. Intell. Rev.* **2021**, *54*, 137–178. [CrossRef]
21. Yang, R.; Yu, Y. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front. Oncol.* **2021**, *11*, 638182. [CrossRef]
22. Xia, K.J.; Yin, H.S.; Zhang, Y.D. Deep semantic segmentation of kidney and space-occupying lesion area based on SCNN and ResNet models combined with SIFT-flow algorithm. *J. Med. Syst.* **2019**, *43*, 2. [CrossRef]
23. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
25. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
26. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.

27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 27–30 June 2016.
28. Liu, F.; Shen, C.; Lin, G. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
29. Xu, D.; Ricci, E.; Ouyang, W.; Wang, X.; Sebe, N. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
30. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 11–18 December 2015.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
32. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
33. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
34. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
35. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
36. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. Solo: Segmenting objects by locations. In Proceedings of the 2020 European Conference, Glasgow, UK, 23–28 August 2020.
37. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. OCNet: Object context for semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 2375–2398. [CrossRef]
38. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
39. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; pp. 3–11.
40. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
41. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
42. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
43. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 833–851.
44. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *12*, 2481–2495. [CrossRef]
45. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 2015 International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
46. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
47. Zhou, Q.; Wu, X.; Zhang, S.; Kang, B.; Ge, Z.; Latecki, L.J. Contextual ensemble network for semantic segmentation. *Pattern Recognit.* **2022**, *122*, 108290. [CrossRef]
48. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
49. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale vision transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 6804–6815.
50. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 538–547.
51. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.
52. Yu, Q.; Xia, Y.; Bai, Y.; Lu, Y.; Yuille, A.L.; Shen, W. Glance-and-gaze vision transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12992–13003.
53. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3121–3130.

54. Jiao, J.; Wei, Y.; Jie, Z.; Shi, H.; Lau, R.W.; Huang, T.S. Geometry-aware distillation for indoor semantic segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2864–2873.
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
56. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
57. Wu, H.; Liang, C.; Liu, M.; Wen, Z. Optimized HRNet for image semantic segmentation. *Expert Syst. Appl.* **2021**, *174*, 114532. [CrossRef]
58. Kim, D.S.; Kim, Y.H.; Park, K.R. Semantic segmentation by multi-scale feature extraction based on grouped dilated convolution module. *Mathematics* **2021**, *9*, 947. [CrossRef]
59. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
60. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; Wang, J. Structured knowledge distillation for semantic segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2599–2608.
61. Wu, J.; Ji, R.; Liu, J.; Xu, M.; Zheng, J.; Shao, L.; Tian, Q. Real-time semantic segmentation via sequential knowledge distillation. *Neurocomputing* **2021**, *439*, 134–145. [CrossRef]
62. Amirkhani, A.; Khosravian, A.; Masih-Tehrani, M.; Kashiani, H. Robust Semantic Segmentation with Multi-Teacher Knowledge Distillation. *IEEE Access* **2021**, *9*, 119049–119066. [CrossRef]
63. Feng, Y.; Sun, X.; Diao, W.; Li, J.; Gao, X. Double similarity distillation for semantic image segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 5363–5376. [CrossRef] [PubMed]
64. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
65. Zhang, Y.; Ye, M.; Gan, Y.; Zhang, W. Knowledge based domain adaptation for semantic segmentation. *Knowl.-Based Syst.* **2020**, *193*, 105444. [CrossRef]
66. Tian, Y.; Zhu, S. Partial domain adaptation on semantic segmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3798–3809. [CrossRef]
67. Zhou, Z.H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [CrossRef]
68. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4037–4058. [CrossRef]
69. Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905.
70. Wang, J.; Wang, Y.; Jiang, M.; Yan, X.; Song, M. Moving cast shadow detection using online sub-scene shadow modeling and object inner-edges analysis. *J. Vis. Commun. Image Represent.* **2014**, *25*, 978–993. [CrossRef]
71. Bao, Y.; Song, K.; Wang, J.; Huang, L.; Dong, H.; Yan, Y. Visible and thermal images fusion architecture for few-shot semantic segmentation. *J. Vis. Commun. Image Represent.* **2021**, *80*, 103306. [CrossRef]
72. Bucher, M.; Vu, T.H.; Cord, M.; Pérez, P. Zero-shot semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
73. Gu, Z.; Zhou, S.; Niu, L.; Zhao, Z.; Zhang, L. Context-aware feature generation for zero-shot semantic segmentation. In Proceedings of the 2020 28th ACM International Conference on Multimedia (MM), Seattle, WA, USA, 12–16 October 2020; pp. 1921–1929.
74. Li, B.; Weinberger, K.Q.; Belongie, S.; Koltun, V.; Ranftl, R. Language-driven semantic segmentation. *arXiv* **2022**, arXiv:2201.03546.
75. Zhang, H.; Ding, H. Prototypical matching and open set rejection for zero-shot semantic segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 6954–6963.
76. Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; Wang, X. GroupViT: Semantic Segmentation Emerges from Text Supervision. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 18113–18123.
77. Scheirer, W.J.; de Rezende Rocha, A.; Sapkota, A.; Boult, T.E. Toward open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1757–1772. [CrossRef] [PubMed]
78. Pastore, G.; Cermelli, F.; Xian, Y.; Mancini, M.; Akata, Z.; Caputo, B. A closer look at self-training for zero-label semantic segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 2687–2696.
79. Shen, F.; Lu, Z.M.; Lu, Z.; Wang, Z. Dual semantic-guided model for weakly-supervised zero-shot semantic segmentation. *Multimed. Tools Appl.* **2022**, *81*, 5443–5458. [CrossRef]
80. Gu, Z.; Zhou, S.; Niu, L.; Zhao, Z.; Zhang, L. From pixel to patch: Synthesize context-aware features for zero-shot semantic segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–15. [CrossRef]
81. Bian, C.; Yuan, C.; Ma, K.; Yu, S.; Wei, D.; Zheng, Y. Domain Adaptation Meets Zero-Shot Learning: An Annotation-Efficient Approach to Multi-Modality Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2021**, *41*, 1043–1056. [CrossRef] [PubMed]
82. Kosiorek, A. Attention Mechanism in Neural Networks. *Robot. Ind.* **2017**, *6*, 14.
83. Lambert, J.; Liu, Z.; Sener, O.; Hays, J.; Koltun, V. MSeg: A composite dataset for multi-domain semantic segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 2876–2885.

84. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
85. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, 27 October–2 November 2019; pp. 6687–6696.
86. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
87. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.
88. Kang, J.; Liu, L.; Zhang, F.; Shen, C.; Wang, N.; Shao, L. Semantic segmentation model of cotton roots in-situ image based on attention mechanism. *Comput. Electron. Agric.* **2021**, *189*, 106370. [CrossRef]
89. Lv, N.; Zhang, Z.; Li, C.; Deng, J.; Su, T.; Chen, C.; Zhou, Y. A hybrid-attention semantic segmentation network for remote sensing interpretation in land-use surveillance. *Int. J. Mach. Learn. Cybern.* **2022**, *14*, 395–406. [CrossRef]
90. Wang, K.; He, R.; Wang, L.; Wang, W.; Tan, T. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2010–2023. [CrossRef]
91. Yang, M.; Rosenhahn, B.; Murino, V. *Multimodal Scene Understanding: Algorithms, Applications and Deep Learning*; Academic Press: Cambridge, MA, USA, 2019.
92. Zhang, Y.; Sidibé, D.; Morel, O.; Mériaudeau, F. Deep multimodal fusion for semantic image segmentation: A survey. *Image Vis. Comput.* **2021**, *105*, 104042. [CrossRef]
93. Zhou, W.; Guo, Q.; Lei, J.; Yu, L.; Hwang, J.N. ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1224–1235. [CrossRef]
94. Patel, N.; Choromanska, A.; Krishnamurthy, P.; Khorrami, F. Sensor modality fusion with CNNs for UGV autonomous driving in indoor environments. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 1531–1536.
95. Zou, Z.; Zhang, X.; Liu, H.; Li, Z.; Hussain, A.; Li, J. A novel multimodal fusion network based on a joint coding model for lane line segmentation. *Inf. Fusion* **2022**, *80*, 167–178. [CrossRef]
96. Brostow, G.J.; Fauqueur, J.; Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.* **2009**, *30*, 88–97. [CrossRef]
97. Larsson, M.; Stenborg, E.; Hammarstrand, L.; Pollefeys, M.; Sattler, T.; Kahl, F. A cross-season correspondence dataset for robust semantic segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9524–9534.
98. Orsic, M.; Kreso, I.; Bevandic, P.; Segvic, S. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12599–12608.
99. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the 2014 European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
100. Hu, Y.T.; Chen, H.S.; Hui, K.; Huang, J.B.; Schwing, A.G. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3105–3115.
101. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 3213–3223.
102. Everingham, M.; Eslami, S.M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
103. Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; Malik, J. Semantic contours from inverse detectors. In Proceedings of the 2011 International Conference on Computer Vision (ICCV), Barcelona, Spain, 6-13 November 2011; pp. 991–998.
104. Brostow, G.J.; Shotton, J.; Fauqueur, J.; Cipolla, R. Segmentation and recognition using structure from motion point clouds. In Proceedings of the 2008 European Conference on Computer Vision (ECCV), Berlin, Germany, 12–18 October 2008; pp. 44–57.
105. Miao, J.; Wei, Y.; Wu, Y.; Liang, C.; Li, G.; Yang, Y. Vspw: A large-scale dataset for video scene parsing in the wild. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Virtual, 19–25 June 2021; pp. 4131–4141.
106. Staal, J.; Abràmoff, M.D.; Niemeijer, M.; Viergever, M.A.; Van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **2004**, *23*, 501–509. [CrossRef]
107. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2014**, *34*, 1993–2024. [CrossRef]
108. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Van Den Hengel, A. Semantic labeling of aerial and satellite imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2868–2881. [CrossRef]

109. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
110. Miao, L.; Zhang, Y. A hierarchical feature extraction network for fast scene segmentation. *Sensors* **2021**, *21*, 7730. [CrossRef]
111. Huang, S.; Lu, Z.; Cheng, R.; He, C. Fapn: Feature-aligned pyramid network for dense image prediction. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 10–17 October 2021.
112. Hong, Y.; Pan, H.; Sun, W.; Member, S.; Jia, Y. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. *arXiv* **2021**, arXiv:2101.06085.
113. Cheng, J.; Peng, X.; Tang, X.; Tu, W.; Xu, W. Mifnet: A lightweight multiscale information fusion network. *Int. J. Intell. Syst.* **2021**, *37*, 5617–5642. [CrossRef]
114. Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; Qiao, Y. Vision transformer adapter for dense predictions. *arXiv* **2022**, arXiv:2205.08534.
115. Fu, J.; Liu, J.; Jiang, J.; Li, Y.; Lu, H. Scene segmentation with dual relation-aware attention network. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2547–2560. [CrossRef] [PubMed]