electronics

MDPI

Review

Deep-Learning-Based Approaches for Semantic Segmentation of Natural Scene Images: A Review

Busra Emek Soylu [1], Mehmet Serdar Guzel [1], Gazi Erkan Bostanci [1], Fatih Ekinci [1], Tunc Asuroglu [2,*] and Koray Acici [3]

[1] Department of Computer Engineering, Faculty of Engineering, Ankara University, 06830 Ankara, Turkey; bsoylu@ankara.edu.tr (B.E.S.); mguzel@ankara.edu.tr (M.S.G.); ebostanci@ankara.edu.tr (G.E.B.); fatih.ekinci@gsb.gov.tr (F.E.)
[2] Faculty of Medicine and Health Technology, Tampere University, 33720 Tampere, Finland
[3] Department of Artificial Intelligence and Data Engineering, Faculty of Engineering, Ankara University, 06830 Ankara, Turkey; kacici@ankara.edu.tr
* Correspondence: tunc.asuroglu@tuni.fi

Abstract: The task of semantic segmentation holds a fundamental position in the field of computer vision. Assigning a semantic label to each pixel in an image is a challenging task. In recent times, significant advancements have been achieved in the field of semantic segmentation through the application of Convolutional Neural Networks (CNN) techniques based on deep learning. This paper presents a comprehensive and structured analysis of approximately 150 methods of semantic segmentation based on CNN within the last decade. Moreover, it examines 15 well-known datasets in the semantic segmentation field. These datasets consist of 2D and 3D image and video frames, including general, indoor, outdoor, and street scenes. Furthermore, this paper mentions several recent techniques, such as SAM, UDA, and common post-processing algorithms, such as CRF and MRF. Additionally, this paper analyzes the performance evaluation of reviewed state-of-the-art methods, pioneering methods, common backbone networks, and popular datasets. These have been compared according to the results of Mean Intersection over Union (MIoU), the most popular evaluation metric of semantic segmentation. Finally, it discusses the main challenges and possible solutions and underlines some future research directions in the semantic segmentation task. We hope that our survey article will be useful to provide a foreknowledge to the readers who will work in this field.

Keywords: semantic segmentation; computer vision; deep learning; CNN; general; indoor; outdoor; street scenes; SAM; UDA; CRF; MRF

Citation: Emek Soylu, B.; Guzel, M.S.; Bostanci, G.E.; Ekinci, F.; Asuroglu, T.; Acici, K. Deep-Learning-Based Approaches for Semantic Segmentation of Natural Scene Images: A Review. *Electronics* **2023**, *12*, 2730. https://doi.org/10.3390/electronics12122730

Academic Editor: Hüseyin Kusetogullari

Received: 19 April 2023
Revised: 13 June 2023
Accepted: 15 June 2023
Published: 19 June 2023

1. Introduction

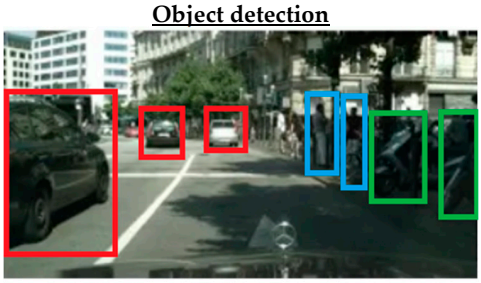
Semantic segmentation, which we sometimes encounter as visual scene understanding, assigns each pixel in an image to predefined semantic labels. After this process, the input image yields an output that turns into a raster map. In other words, it is used to semantically group pixels and analyze data such as 2D, 3D and video. This paper [1] has demonstrated how the semantic representation can be used as an input.

Semantic segmentation is closely related to image classification, object detection, instance segmentation and panoptic segmentation tasks that are very popular in computer vision. Each enables the identification of entities, objects, etc. within the input data. However, each approaches the problem differently and provides different levels of detail in the resulting output. Figure 1 illustrates aspects of these tasks that differ from semantic segmentation.

Image classification

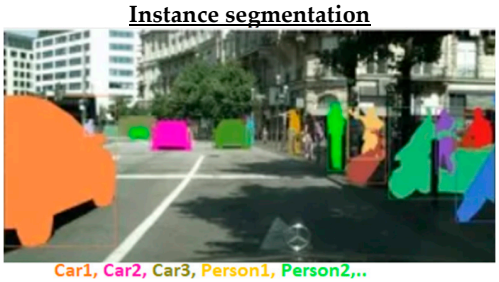
- Detects which objects are existing in the image.
- Does not look at the location of the objects.

Object detection



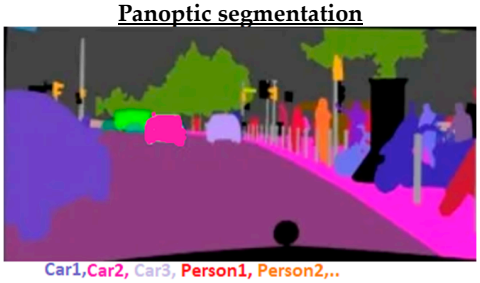
- Not only perceives which objects are existing in the image, but also where those objects are in the image.
- Recognizes objects and positions them with bounding box.

Instance segmentation



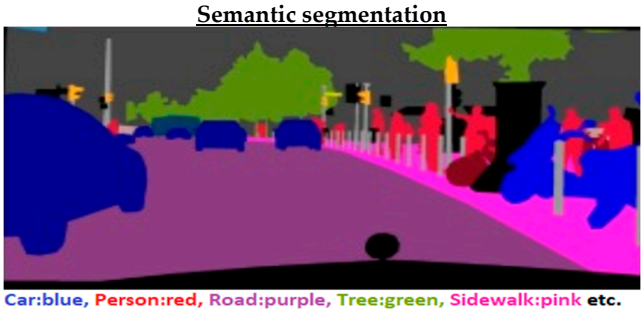
- The pixels for each object are individually labeled.
- In this example, one car pixel is colored in orange, another car in pink, and another car in green.

Panoptic segmentation



- Combines semantic segmentation and instance segmentation.
- Gives a comprehensive understanding of the scene at both the semantic level (car, building, road, etc.) and the instance level (car 1, car 2, etc.).

Semantic segmentation



- Assigns an object category/class label to each pixel in the image.
- In this example, car pixels are colored in blue, person pixels in red and background pixels in gray etc.
- Semantic segmentation handles objects belonging to the same class in the image as a single label, while instance segmentation handles objects belonging to the same class as separate instances. For example, if there are several cars in the example image, the semantic segmentation algorithm assigns all pixels that belong to cars with the 'car' label without distinguishing between different cars. However, the instance segmentation algorithm assigns related pixels to separate labels such as car 1, car 2 and car 3.
- While object detection recognizes more specific and foreground objects (e.g., cars, people, etc. that are obvious in the example image), semantic segmentation not only classifies specific and foreground objects but also classifies unclear and background objects. Since, in contrast to object detection, semantic segmentation can provide pixel-level classification. So, it gives a more detailed output map.

Figure 1. Comparison of image classification, object detection, instance segmentation, panoptic segmentation, and semantic segmentation.

Deep learning is perhaps the most popular and fastest growing area of machine learning in recent times. It covers the development of new methods and the improvement of existing methods for semantic segmentation. Most of the methods investigated in this study have been created to solve pixel-based labeling problems based on "Convolutional Neural Networks (CNN)" [2], one of the most commonly used techniques of deep learning.

The goal of deep-learning-based semantic segmentation is to estimate a class label for each image pixel; this is an important but difficult task to understand the image. Recent approaches have applied CNNs, the most popular model in deep learning, to the pixel-level labeling task and had remarkable success. However, the classical CNNs used in image classification are not effective in semantic segmentation. The main reason for this is that CNN applies pooling and subsampling operations to the input data to classify the image. These processes cause a loss of resolution and local and global information in the input image. These losses are not a big problem in the classification task for most situations that do not require very fine details because it is enough to give a global label as output in the classification. However, for the semantic segmentation task, which requires a prediction of a class label for each pixel and more detail, these losses negatively affect accuracy. To solve these problems, different CNN architectures have been developed that recover the loss of spatial, global and local information. Studies in this area have progressed by incorporating local information obtained from CNN and global information obtained from deeper parts of the network. It is explained in detail in Section 2.

The several important surveys on semantic segmentation can be summarized as follows: This paper [3] has categorized the architectures for semantic segmentation in deep learning into ten distinct classes. These methods are based on: Feature encoder, Regional proposal, Recurrent neural network, Up sampling / Deconvolution, Increase resolution of feature, Enhancement of features, Semi and weakly supervised, Spatio-temporal, Methods using CRF / MRF, Alternative to CRF. They have summarized approximately 100 models and 33 publicly available datasets. The deep learning methods for semantic segmentation have been classified by [4] based on varying degrees of supervision during the training process. Furthermore, the authors have provided a concise overview of the techniques that are specifically geared towards real-time segmentation, a topic that has received comparatively less attention in previous surveys. The paper [5] has presented the comprehensive knowledge on deep learning required for semantic segmentation tasks. The survey, comprising of 28 datasets and 29 methods, has been presented. In addition to the models using RGBD and 3D data, models that perform instance segmentation are also mentioned in the study. This paper [6] has provided a summary of the advancements made in the field of semantic segmentation, specifically in the areas of weakly supervised learning, domain adaptation, multi-modal data fusion, and real-time processing. This paper [7] has divided semantic segmentation methods into three categories: Region-based, FCN-based, and Weakly supervised. Moreover, an overview of the strengths, weaknesses and significant challenges associated with these approaches has been provided. This paper [8] has provided an overview of segmentation models that utilize semi-supervised and weakly supervised learning techniques. The emphasis has been placed on the fundamental aspects of the model's structure, operational mechanism, and primary functionalities. This paper [9] has focused on the decade-long progression observed in this domain, which can be classified into three distinct chronological phases: the pre- and early-deep-learning era, the fully convolutional era, and the post-FCN era. This paper [10] has focused on studies that performed semantic segmentation using deep learning for autonomous driving. The study includes a comparative analysis of 14 frameworks, 12 datasets, various data augmentation and domain adaptation techniques, and the benefits of these techniques. According to the chronological progression of image segmentation technology, Ref. [11] have sorted the classic segmentation algorithms (e.g., Edge Detection, Clustering Method, Random Walks), Co-Segmentation Methods (e.g., MRF-based, Object-Based) and the presently popular deep learning algorithms.

The goals and main contribution of this paper can be mentioned as follows: Our review has involved a comprehensive and structured analysis of approximately 150 methods of semantic segmentation based on CNN. We have classified and categorized these methods as shown in Figure 2. Nevertheless, these categories ought not to be viewed in isolation from one another, as there are no clear demarcations between them. There exist mutually advantageous commonalities between the categories, and an approach may potentially belong to multiple categories. Then, we have created a table for each category, in chronological order. These tables have summarized to include the network structures, the backbone networks of these models, the datasets used and their accuracies. Then, we have compared and analyzed the performance evaluation of the most popular backbone networks, the pioneering methods, and the state-of-the-art methods for each category. Moreover, the study scrutinizes 15 widely recognized datasets in semantic segmentation. Furthermore, the paper references various modern methodologies, including SAM, UDA and conventional post-processing algorithms such as CRF, MRF and Random Walker. The article ultimately addresses the primary obstacles, potential remedies, and prospective avenues for further investigation in the realm of semantic segmentation.

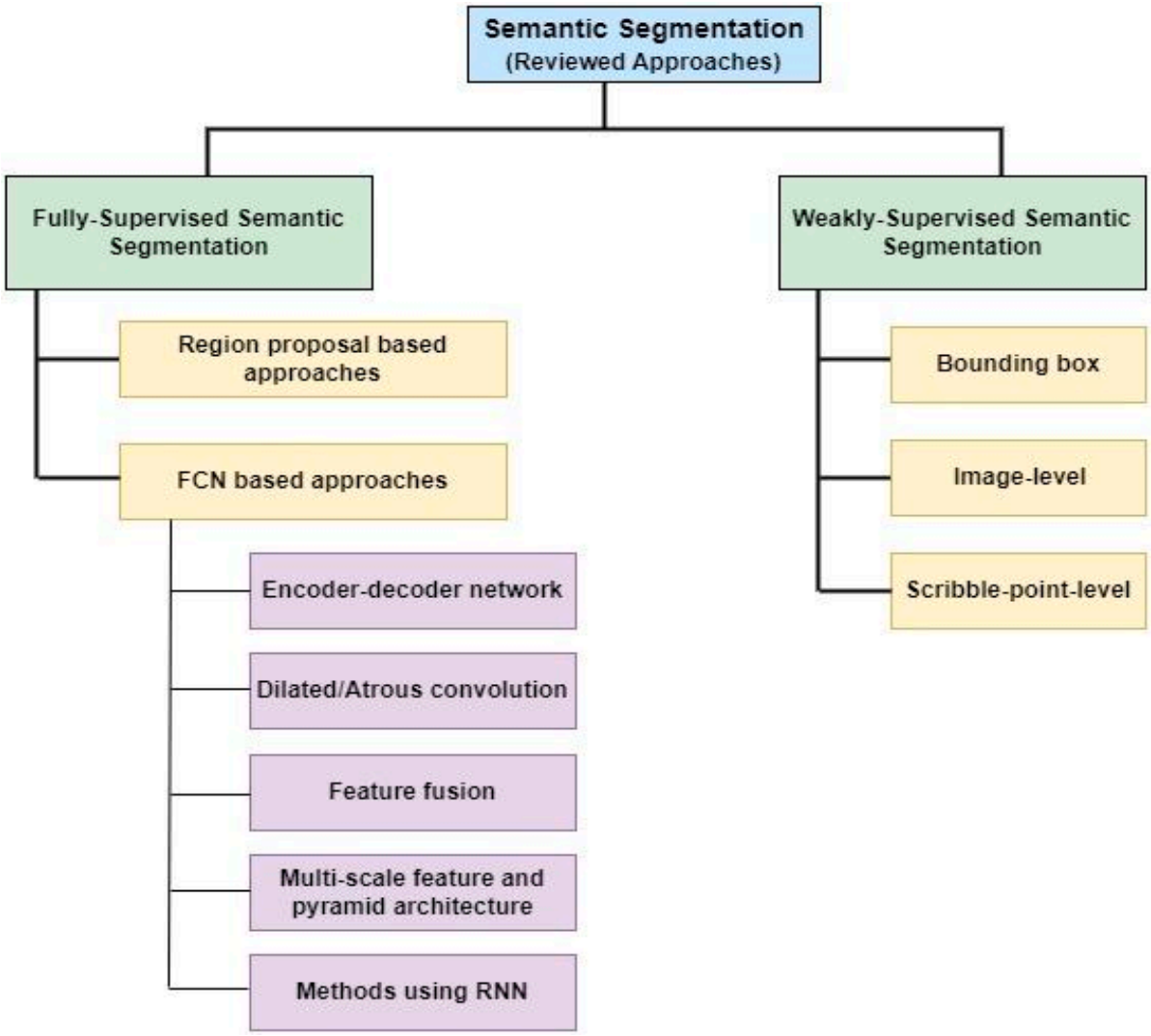


Figure 2. CNN-based architectures for semantic segmentation.

We have aimed at comprehensively addressing deep neural networks related to semantic segmentation. During the process, we tried to relate all the reviewed methods with respect to the architectural design, and we handled them in chronological order. Additionally, we have attempted to indicate the advantages and disadvantages of the methods. We hoped that this article would provide a general understanding of semantic segmentation for researchers who intends to conduct work in this field.

The remainder of this article is organized as follows: In Section 2, an overview is provided of deep network semantic segmentation techniques that rely on fully supervised learning. In Section 3, an overview is provided of deep network semantic segmentation techniques that rely on weakly supervised learning. In Section 4, several recent techniques for semantic segmentation are presented. Section 5 overviews the common post-processing algorithms in this area. Section 6 reviews well-known scene parsing datasets that in used semantic segmentation. Section 7 pertains to the comparison of state-of-the-art methods and common backbone networks on the most widely used datasets. Furthermore, an analysis is presented regarding the performance comparison of all datasets referenced in Section 6. Section 8 discusses the common challenges faced by the current methods, possible solutions and underscores some future research directions in the field. The paper concludes in Section 9.

2. Fully Supervised Semantic Segmentation

Fully supervised methods require many original images and corresponding pixel-based semantically annotated images. That is, there must be sufficient labeled training data. These approaches can be divided into two types according to the mode of operation: Region-Proposal-based and Fully Convolutional Network (FCN)-based methods. Region-Proposal-based methods structure is given in Figure 3 and its explanation is given in Section 2.1. The structure of the FCN network model is shown in Figure 4, and its explanation is given in Section 2.2. Region-based networks assign a category label to each proposal after converting images into a set of region proposals. FCN-based methods take the entire image as input and predict labels on a pixel-by-pixel basis (without removing region suggestions) by mapping them directly to the relevant segmentation results with fully convolutional layers. In addition, they can be trained end-to-end, as they consist of convolution, pooling and upsampling layers.

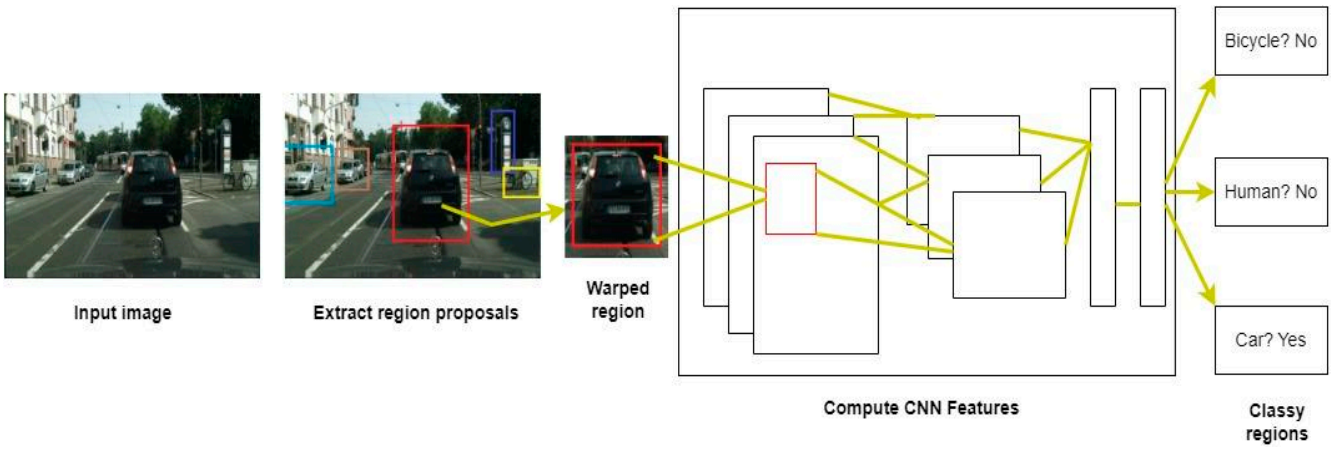


Figure 3. "R-CNN: Regions with CNN features". (Reproduced with permission from authors, Rich feature hierarchies for accurate object detection and semantic segmentation [12]. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2014).

The methods based on FCN semantic segmentation have been organized into the following categories: Section 2.1. Region-Proposal-Based Approaches, Section 2.2. Fully Convolutional Network (FCN)-Based Approaches.

2.1. Region-Proposal-Based Approaches

Region-based methods take an arbitrary-size image as input, extract a series of region proposals from that image, and then transform region-based proposals into pixel estimates by labeling a pixel according to the region with the highest score it contains.

In "R-CNN" [12], semantic segmentation is performed according to object recognition results. This model generates about 2000 category-independent region proposals from the input image. It then uses CNN to extract fixed-size features for each proposal and finally classifies each region using "a linear Support Vector Machine (SVM)" [13]. "R-CNNs" can also be built on top of any CNN structure, such as "VGG" [14], "GoogLeNet" [15], "ResNet" [16] and "AlexNet" [17]. Since the original "R-CNN" is computationally expensive and slow, newer architectures such as "Fast R-CNN" [18] and "Faster R-CNN" [19] have made this approach faster. "Mask R-CNN" [20] has extended "Faster R-CNN" with a branch for forecasting an object mask in parallel to the available branch for bounding box detection. This advanced method is mentioned within the instance segmentation subject that is both semantic and a form of detection. The "Path Aggregation Network (PANet)" [21] is based on the "Mask R-CNN" and improves it in important aspects. Hariharan has argued that the "R-CNN" algorithm is fine-tuned to classify bounding boxes (i.e., to extract features for all regions), but is inadequate to extract foreground features. To address this problem, they used a jointly trained CNN to develop a model based on region proposal classification using features extracted from both bounding boxes and foreground regions. Moreover, based on the proposed "Convolutional Feature Masking (CFM)" layer [22] has explored two possible ways to do this. Aforementioned studies using region-proposal-based approaches are given in Table 1.

Table 1. Region-proposal-based approaches.

Paper, Year	Method	Backbone Network	Dataset	Accuracy mIoU (%)
[12] (2014)	Regional CNN (R-CNN)	AlexNet [17]	Pascal VOC 2010 Pascal VOC 2012	53.7 47.9
[23] (2014)	Simultaneous Detection& Segment. (SDS)	MCG [24]	Pascal VOC 2010 Pascal VOC 2012	52.6 51.6
[18] (2015)	Fast R-CNN	VGG-16 [14]	Pascal VOC 2010 Pascal VOC 2012	66.1 65.7
[22] (2015)	Convolutional feature masking (CFM)	VGG + MCG	Pascal VOC 2012	61.8
[25] (2016)	Multi-scale, overlapping regions	VGG-16	Pascal Context SIFT Flow	49.9 64.0
[26] (2020)	Region Attention Network (RANet)	ResNet-101 [16]	Cityscapes Pascal Context COCO Stuff	81.9 54.9 40.7

2.2. Fully Convolutional Network (FCN)-Based Approaches

A classic CNN consists of two components: Convolutional layers and fully connected layers located at a deeper level of the network. Convolutional layers operate as a floating window, are not bound to a fixed-size image, and can create feature maps of arbitrary-size. Fully linked layers, on the other hand, must have a fixed-size input. This requirement can reduce recognition accuracy for images and sub-images of arbitrary size. In this approach, the fully connected layer is removed and replaced by the fully convolutional layer, thus converting CNN to FCN. Thus, it is ensured that CNN takes images of arbitrary size as input and obtains an output of arbitrary size.

This study [27] is pioneering work in this area. In their work, they have adapted classification networks such as "VGGNet", "GoogleNet" and "AlexNet", which have been very popular in recent years, to fully convolutional networks. The backbone net-

work involves the primary structure of the network, which is produced for the image classification task. These structures, essentially, perform feature extraction for the task of semantic segmentation. These classification networks are called backbone networks within our study.

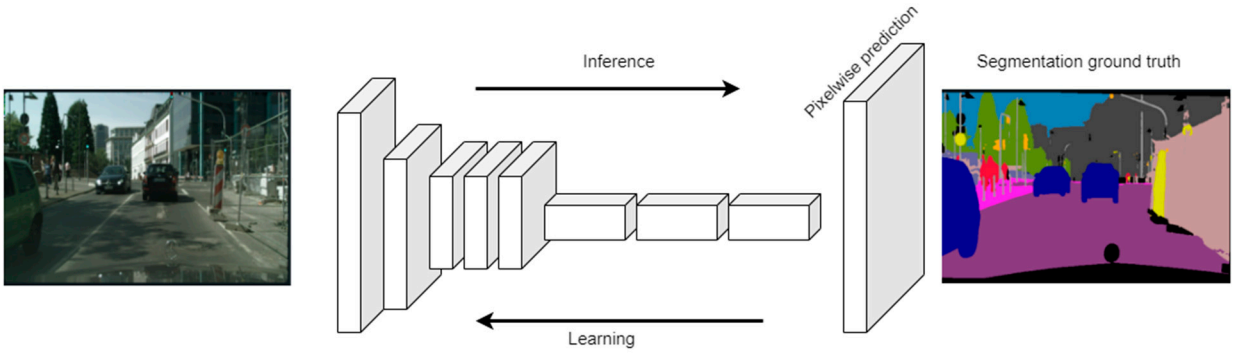


Figure 4. Framework of FCN. The figure is adapted from [27].

The basic FCN-based method [27] has major limitations for semantic segmentation. Low-level features from shallow layers of the network contain more detailed information at higher resolution, i.e., they have richer spatial details. High-level features from the deeper part of the network have higher semantic information, but due to the pooling layer, the feature map resolution is lowered; that is, spatially detailed information is lost. For this reason, an encoder-decoder network has been developed to extract the features, which reduces the spatial size and then gradually recovers the spatial size of the features obtained through upsampling. Another restriction is that the FCN has a predefined fixed-size receptive field because of the convolution operation. This ignores the global information in the image when it encounters an object larger or smaller than the receiving field. To use this global information, that is, to include the semantic context, methods based on generating features with larger receptive fields without sacrificing spatial resolution have been developed. Dilated convolution methods use dilated/atrous convolutions in FCN to expand the receptive field of convolutions and enable dense predictions, feature fusion methods fuse high-level low-resolution and low-level high-resolution features, thereby visibly improving performance, multi-scale methods combine multi-scale/stage features by modeling local and global information from different layers and pyramid methods significantly increase performance by expanding the receptive field by multi-resolution pyramid-based representations and methods using "Recurrent Neural Networks (RNN)" [28] and "Long Short-Term Memory (LSTM)" [29] capture long-range semantic dependencies in images. A graphical model, the "Conditional Random Field (CRF)" [30] has also been used to introduce global context into an FCN and improve output accuracy. In these studies, segmentation performance is often improved by applying the CRF to the CNN as a post-processing step or by fully integrating the CRF into the CNN to train the entire network end-to-end [31–33].

CRFs can model contextual relationships between different pixels to maximize label conformity. The studies using CRF have been indicated in the tables.

2.2.1. Encoder-Decoder Network

This network has two parts: an encoder and a decoder. On the part of the encoder, features with different receptive fields are extracted from each convolutional layer of the image, while on the part of the decoder, segmentation is made of the features generated by the encoder. The encoder module is a typical CNN that has layers such as convolution, pooling, and nonlinear activation. The pooling layer in this part causes a smaller feature map than the original image. Next, the pooling layer is removed in the decoder part, and then the feature map is expanded (spatial dimension is recovered) using up-sample layers to obtain high-resolution prediction. In addition, skip connections between the encoder

and decoder have provided more accurate results by fusing low-level information with high-level information. An example architecture is shown on Figure 5.

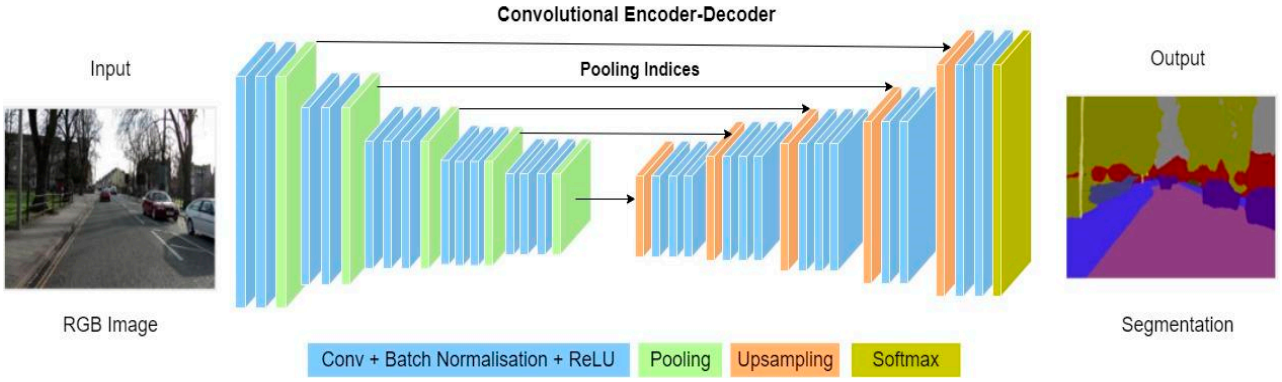


Figure 5. The architecture of SegNet. The figure is reproduced from [34]. (Licensed under CC BY 4.0).

"DeconvNet" [35], one of the most important studies in this area, is an approach with components that complement the simple FCN-based approach, which is good at extracting a general form of the object. On the other hand, "DeconvNet" collects the proposals in descending order of size and effectively renders multi-scale objects by identifying finer object details. The innovation of "SegNet" [34,36] is in the way the decoder upsamples feature maps with low spatial dimensions. Additionally, "SegNet", can store the max-pooling indexes of the encoder feature maps and use them in the decoder network, so its performance is quite good. Ref. [37] have introduced a probability-based pixel-based framework, which they named "Bayesian SegNet", by modifying the "SegNet" architecture. The technique they have used to construct a probabilistic encoder-decoder architecture is dropout [38], which is utilized as approximate inference by "Bayesian CNN" [39]. This paper [40] has proposed an architecture consisting of an encoder such as "SqueezeNet" and then a decoder with enhancement modules such as "SharpMask". The design of the "RefineNet" architecture as presented by [41] enables gradient propagation that is useful for efficient training between long-range connections. In their encoder-decoder structure, all operations, including downsampling, are applied as a single stream. Ref. [42] have presented "GridNet" to solve the loss of resolution problem. "GridNet" has followed a grid pattern that allows multiple interconnected streams to run at different resolutions. Ref. [43] has presented "IIE-SegNet" which enhanced boundaries based on image information entropy. Ref. [44] has proposed a novel "SFANet" to alleviate the misalignment problem between two adjacent levels of feature maps. Ref. [45] has constructed a new "Context Aggregation Network (CANet)" employing shallow encoder-decoders to compensate for local ambiguities while capturing sufficient global context and maintaining computational efficiency. Aforementioned studies using encoder-decoder network-based approaches are given in Table 2.

Table 2. Encoder-decoder network-based approaches.

Author, Year	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
[35] (2015)	DeconvNet	yes	VGG-16	Pascal VOC 2012	70.5
[36] (2015)	SegNet	yes	VGG-16	CamVid NYUDv2 KITTI	62.5 41.0 58.4

Table 2. *Cont.*

Author, Year	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
[37] (2015)	Bayesian SegNet	no	VGG-16	Pascal VOC 2012 CamVid	60.5 63.1
[46] (2016)	Efficient neural network (ENet)	no	ResNet	CamVid Cityscapes SUN RGBD	55.6 58.3 19.7
[40] (2016)	SqueezeNet + SharpMask	no	VGG-16	Cityscapes	59.8
[41] (2017)	RefineNet	yes (just Pascal dataset)	ResNet-101	Pascal VOC 2012 Cityscapes SUN RGBD ADE20K	83.4 73.6 45.7 40.2
[42] (2017)	Residual Conv-Deconv Grid Network	no	ResNet-101	Cityscapes	69.4
[47] (2017)	Label refinement network (LRN)	no	VGG-16	Pascal VOC 2012 CamVid SUN RGBD	62.8 61.7 33.1
[48] (2018)	DeepLabV3+	no	ResNet-101	Pascal VOC 2012 Cityscapes	87.8 82.1
[49] (2018)	Gated Feedback Refinement Network (G-FRNet)	yes (just Pascal dataset)	VGG-16 ResNet-101	Pascal VOC 2012	70.4_{VGG16} $79.3_{\text{ResNet101}}$
				CamVid	68.0_{VGG16}
[50] (2018)	Dense Decoder Shortcut Connections	no	ResNeXt [51]	Pascal VOC 2012 CamVid NYUDv2 Pascal Context	81.2 70.9 48.1 47.8
[52] (2019)	Stacked Deconvolutional Network (SDN)	no	DenseNet161 [53]	Pascal VOC 2012 CamVid GATECH RGBD COCO Stuff	83.5 69.6 53.5 35.9
[54] (2019)	Hierarchical adjacency dependent network (HadNet)	no	Xception [55] + ASPP [56]	Pascal VOC 2012	87.9
[43] (2021)	IIE-SegNet	no	Deeplab-v3 [48]	Pascal VOC 2012	89.6
[57] (2021)	HRNet	yes	ResNet + ASPP	Pascal VOC 2012	79.5
[44] (2021)	Stage-aware Feature AlignmentNetwork (SFANet)	no	ResNet-18 [16]	Cityscapes CamVid	78.1 74.7
[58] (2021)	Segmenter	no	ViT-L/16 [59]	ADE20K Pascal Context Cityscapes	53.6 59.0 81.3
[60] (2021)	Multi-level graph conv.RNN (MGCRNN)	no	VGG-16	Pascal VOC 2012 Cityscapes	74.2 73.6
[45] (2022)	Context Aggregation Network (CANet)	no	ResNet-101	Cityscapes CamVid BDD100K	81.8 78.6 66.5

2.2.2. Dilated/Atrous Convolution

This approach aims to remove the limitation of a fixed field of view in simple FCN-based approaches. To this end, they have presented dilated convolutions to obtain feature maps with a larger field of view without reducing the spatial size, i.e., resolution. This has been achieved by placing holes between pixels in the standard convolution cores. Since there is no need to increase the number of parameters during this process, the computation time does not increase either. Compared with standard convolution, dilated convolution increases the hyperparameter of the dilation rate. This hyperparameter represents the number of intervals between cores. Thus, dense feature extraction has been achieved, and significant progress has been made in improving spatial resolution.

DeepLab architecture is one of the pioneering studies in this field. "DeepLab-v1" [31] has improved object boundary localization by integrating CRF and the responses from the last layer of CNN. "DeepLab-v2" [56] has proposed "Sharp Spatial Pyramid Pooling (ASPP)", which parallelizes multiple atrous convolutions to obtain richer multi-scale contextual information. "DeepLab-v3" [61] has further enhanced DeepLab by strengthening the "ASPP" algorithm with image-level features which encode global context. Based on "DeepLab-v3", Ref. [48] has proposed "DeepLab-v3+" by intensely connecting the decoder component to the encoder. Ref. [62] have developed the "DeepLab-v3+" based "Cascade Waterfall ASPP (CWASPP)" module to reduce the parameters and increase the segmentation performance. Ref. [63] have proposed "DilatedNet", which uses dilated convolutions with a "dilated rate" unlike the standard convolution operator. The dilated convolutions have a larger receptive field without downsampling the feature maps much. However, the performance of the network is adversely affected by the grinding artifacts it has. Therefore, Ref. [64] have developed "Dilated Residual Networks (DRN)" to remove the grinding artifacts. Ref. [65] have proposed an optimized algorithm by combining "ASPP" method and CRF. Ref. [66] have proposed "Dense Upsampling Convolution (DUC)" to create pixel-level prediction that can seize and decode the elaborate information lost during upsampling and "Hybrid Dilated Convolution (HDC)" to solve the gridding problem by enlarging the receptive fields of the network. Dilated convolution process can be seen in Figure 6. Aforementioned studies using dilated/atrous convolution-based approaches are given in Table 3.

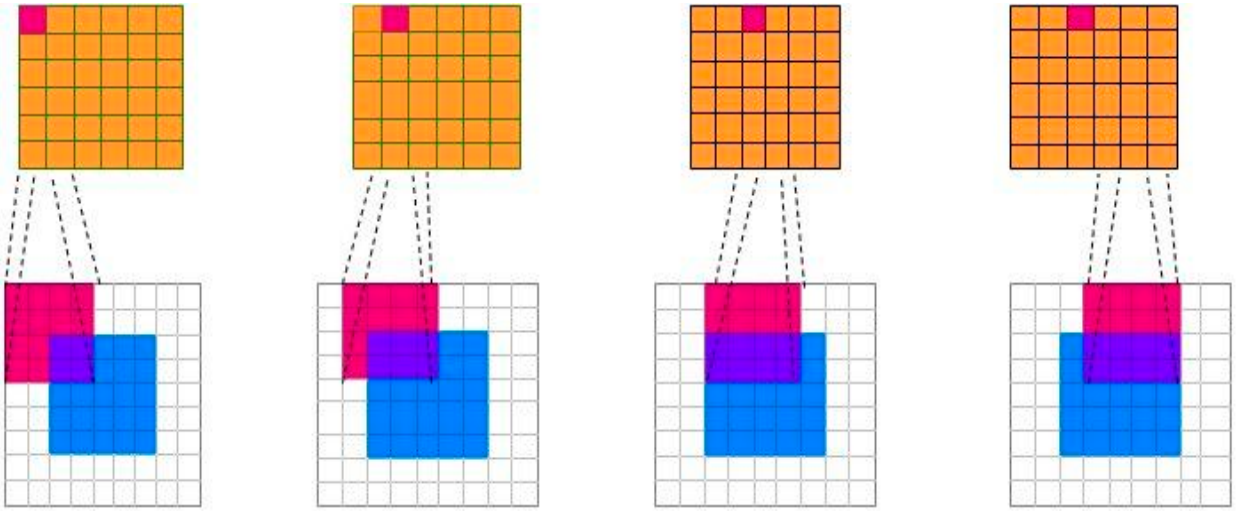


Figure 6. Dilated convolution (**right**) and standard convolution (**left**). The figure is reproduced from [67]. (Licensed under CC BY 4.0).

Table 3. Dilated/Atrous convolution-based approaches.

Author, Year	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
[31] (2014)	DeepLab-v1 (LargeFOV)	yes	VGG-16	Pascal VOC 2012	67.6
[63] (2015)	DilatedNet	yes	VGG-16	Pascal VOC 2012	67.6
[46] (2016)	Efficient neural network (ENet)	no	ResNet	CamVid Cityscapes SUN RGBD	55.6 58.3 19.7
[64] (2016)	Dilated Residual Network (DRN)	no	ResNet-101	Cityscapes	66.6
[56] (2017)	DeepLab-v2 (ASPP)	yes	ResNet-101	Pascal VOC 2012 Cityscapes	79.7 70.4
[61] (2017)	DeepLab-v3	no	ResNet	Pascal VOC 2012 Cityscapes	85.7 81.3
[68] (2017)	Depth fully-connected CRF (DFCN-DCRF)	yes	VGG-16	SUN RGBD	39.3
[48] (2018)	DeepLab-v3+	no	ResNet-101	Pascal VOC 2012 Cityscapes	87.8 82.1
[66] (2018)	Dense upsampling convolution (DUC) + Hybrid Dilated Convolution (HDC)	yes	DeepLab-v2 ResNet-101	Pascal VOC 2012 Cityscapes	83.1 77.6
[69] (2018)	Context Encoding Network (EncNet)	no	ResNet	Pascal VOC 2012 Pascal Context ADE20K	82.9 51.7 44.6
[65] (2019)	Atrous Conv. + fully connected CRFs	yes	ResNet-101	Pascal VOC 2012	77.6
[70] (2020)	Multi-Receptive Atrous Convolutional Network (MRACN)	no	ResNet-101	Pascal VOC 2012 DTMR-DVR	80.2 60.4
[71] (2021)	Multi-source fusion generative adver.net.(SCAGAN)	no	DeepLab-v2	Pascal VOC 2012	70.1
[72] (2021)	SEgmentation TRansformer (SETR)	no	T-Large [72]	ADE20K Pascal Context Cityscapes	50.2 55.8 82.1
[67] (2021)	Efficient Spatial Pyramid of Dilated Conv.(ESPNet)	yes	DeepLab-v2	Cityscapes	60.3
[62] (2022)	Cascade Waterfall ASPP Module (CWASPP)	no	MobileNetv2 [73]	Pascal VOC 2012	73.3

2.2.3. Feature Fusion

Dilated/Atrous convolution strategy has made significant progress to overcome the spatial resolution loss problem. Still, however, FCN's largest receptive field is not sufficient to directly capture and model the global context as needed. Another way to add global context is feature fusion. This technique aims to fuse the features extracted from the previous layer of the network with the localized feature map extracted from the next layer. There are two cases, early and late fusion, for combining the global context feature with a local feature map. Refs. [27,31] have used skip connections to achieve a late fusion by combining the two predictions into a single classification result. Another study, "Enhancing Feature Fusion (ExFuse)" [74], has shown that incorporating semantic information with low-level features and high-resolution details with high-level features is useful in late fusion. Ref. [75] has proposed "ParseNet", which adds global context directly to FCNs. They have spatially separated the global feature into the same dimension as the local feature map, combined them, and finally used the combined feature to learn the classifier for the early fusion. Ref. [76] have proposed "RGB-D Fusion Network (RDFNet)" that effectively extracts and fuses multi-level RGB-D features in very deep networks by extending the core idea of residual learning to RGB-D semantic segmentation. Ref. [77] have used the feature fusion method to improve the feature information extracted by the model. Aforementioned studies using feature fusion-based approaches are given in Table 4.

Table 4. Feature fusion-based approaches.

Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
[75] (2015)	ParseNet	yes	DeepLab-v1	Pascal VOC 2012 Pascal Context	65.8 36.6
[76] (2017)	RGB-D fusion network (RDFNet)	no	ResNet-101	NYUDv2 SUN RGBD	50.1 47.7
[74] (2018)	ExFuse	no	ResNet-101	Pascal VOC 2012	86.2
[77] (2021)	Self-attention feature fusion network (SA-FFNet)	no	ResNet-18	Cityscapes CamVid	75.0 69.5
[67] (2021)	Efficient Spatial Pyramid of Dilated Conv.(ESPNet)	yes	DeepLab-v2	Cityscapes	60.3

2.2.4. Multi-Scale Feature and Pyramid Architecture

These approaches involve extracting features from multiscale or a set of nested regions. Combining multi-scale features with FCN has outperformed single-scale features. In [78], "a multiscale convolutional network" has been developed to extract intensive feature vectors that encode multidimensional regions clustered around each pixel. "A multi-scale network" proposed by [79] has predicted a coarse global output from the entire input image, and then improved it using finer scale local networks. This model does not use superpixels or contours while capturing image details. Ref. [80] have adapted "DeepLab-MSc" to a share-network and proposed an attention mechanism that learns to softly weight multi-scale features at each pixel location. According to [78,81], background information can be effectively captured by combining features extracted by a multi-scale network, thus improving performance for semantic segmentation. Ref. [82] has used "multi-scale CNNs" [78] and the "floating pyramid pool" [83] to encode this rich background information. The floating pyramid pool in the feature map can obtain information from background regions of different sizes. Ref. [84] has suggested "a gated summation scheme" to collect multi-scale features for each spatial location. The gates in this scheme check

the flow of information about the various scaling features. Ref. [52] have produced a "Stacked Deconvolutional Network (SDN)" in which the connections within and between units have been designed to advance the flow of information and propagation of gradients throughout the network. Inter-unit connections have made it efficient to reuse multi-scale information between different units. Ref. [85] has proposed a method that entails the creation of a multi-scale meta-relational network (MSNN). This network has utilized an optimized initialization representation to augment the generalization capacity of learned measurements. The "Multi-scale Relational Network (MSRN)" [86] algorithm involves the removal of the fully connected layer from a four-layer CNN model, followed by the stitching of thirty-four-layer feature maps in the depth direction to generate multi-scale features [87]. The method also includes the integration of multi-scale features from the target set of samples, followed by the computation of relational features through the subtraction of elements and subsequent calculation of absolute values.

The multi-scale pyramid architecture has a multi-scale and pyramid structure that detects objects of different scales. These studies have combined the pyramid strategy with CNN. There are two common image pyramids, named Gaussian and Laplacian [88]. The main problem with current FCN-based models is their inability to capture sufficiently good features at the global image level. To address this shortcoming, previous studies have developed "global pooling" [75], "floating pyramid pooling" [83] and "spatial pyramid pooling" [89]. Ref. [90] have proposed a "Pyramid Scene Parsing Network (PSPNet) (Figure 7)" that incorporates appropriate global features by region-based aggregation. Those in [56] have developed the "Sharp Spatial Pyramid Pooling (ASPP)" method, which performs multi-scale segmentation, inspired by the image pyramid strategy. Ref. [91] have introduced "CiSS-Net" that have "Context Net (CNet)" and "Segment Net (S-Net)" named subnets. The "C-Net" learns high-level semantic context information from p-maps, and the "S-Net" incorporates the learned context into FCN-based semantic segmentation. Ref. [92] has applied graphical convolution to solve a fixed receptive field problem because of the convolution operation and has proposed an improved Laplacian. Graphical reasoning [93] has been performed directly in the original feature space, which has been organized as a spatial pyramid. Aforementioned studies using multi-scale feature and pyramid-architecture-based approaches are given in Table 5.

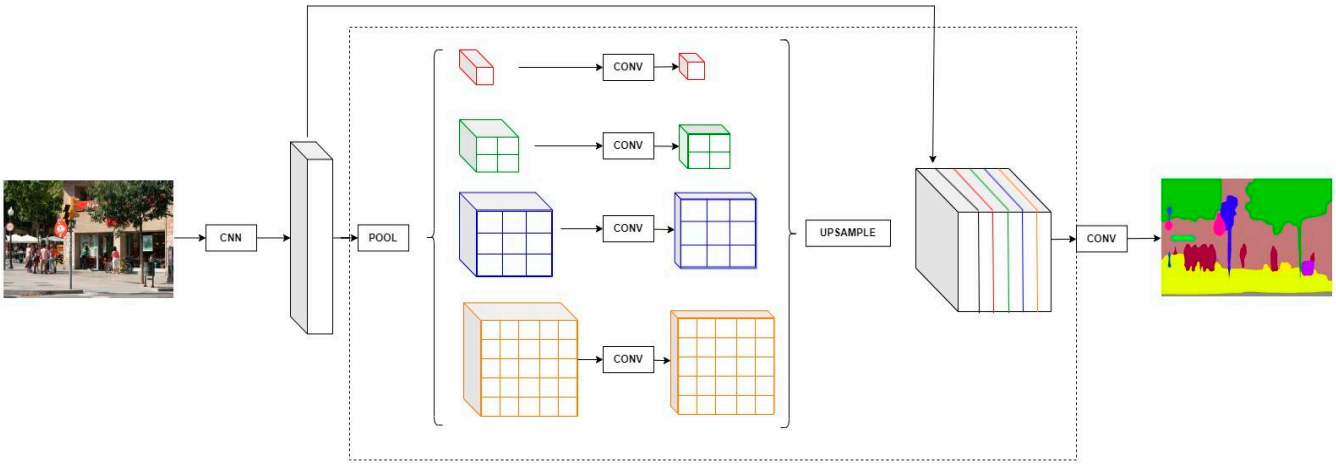


Figure 7. Overview of PSPNet. (Reproduced with permission from authors, Pyramid scene parsing network [90]. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017).

Table 5. Multi-scale feature and pyramid-architecture-based approaches.

	Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
Multi-scale pyramid architecture	[78] (2012)	Multiscale ConvNet	yes	ConvNet [94]	SIFT Flow Stanford Background	50.8 (MA_{CC}) 76.0 (MA_{CC})
	[82] (2016)	FeatMap-Net	yes	VGG-16	Pascal VOC 2012 Pascal Context SIFT Flow NYUDv2 (40 class)	75.3 43.3 44.9 40.6
	[95] (2016)	Laplacian Pyramid Reconst.&Refine. (LRR)	yes	VGG-16 ResNet-101	Pascal VOC 2012 Cityscapes	$74.7_{ResNet101}$ 69.7_{VGG16}
	[90] (2017)	Pyramid scene parsing network (PSPNet)	no	ResNet-101	Pascal VOC 2012 Cityscapes ADE20K	82.6 78.4 41.9
	[91] (2019)	Context-reinforced Network (CiSS-Net)	no	ResNet-50	Cityscapes ADE20K Pascal Context	79.2 42.5 48.7
	[92] (2020)	Spatial Pyramid Based Graph Reasoning (SpyGR)	no	ResNet-101	Cityscapes COCO Stuff Pascal Context	81.6 39.9 52.8
	[96] (2014)	Recursive Context Propagation Network (RCPN)	yes	Multiscale ConvNet [78]	Stanford Background. SIFT Flow	78.8 (MA_{CC}) 48.0 (MA_{CC})
	[97] (2015)	pure-node (PN) RCPN tree-MRF (TM) RCPN	yes	RCPN [96]	Stanford Background	$64.0_{PN\text{-}RCPN}$ $64.5_{TM\text{-}RCPN}$
					SIFT Flow	$30.2_{PN\text{-}RCPN}$ $31.4_{TM\text{-}RCPN}$
	[31] (2014)	DeepLab-MSc	yes	VGG-16	Pascal VOC 2012	71.6
	[98] (2016)	DeepLab-CRF-Attention	yes	DeepLab-v1	Pascal VOC 2012 COCO Stuff	75.1 35.7
	[56] (2017)	DeepLab-v2 (ASPP)	yes	ResNet-101	Pascal VOC 2012 Cityscapes	79.7 70.4
	[63] (2015)	DilatedNet	yes	VGG-16	Pascal VOC 2012	67.6
	[79] (2015)	Multiscale Convolutional Network	no	AlexNet VGG-16	Pascal VOC 2012	72.4_{VGG} (MA_{CC})
					SIFT Flow	$48.2_{AlexNet}$ 55.7_{VGG}
					NYUDv2	$41.3_{AlexNet}$ 45.1_{VGG}
					NYUDv2 (4 class)	$79.1_{AlexNet}$ 82.0_{VGG}

Table 5. *Cont.*

	Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
Multi-scale pyramid architecture	[81] (2015)	Zoom-out	yes	VGG-16	Pascal VOC 2012	69.6
	[99] (2015)	Multi-scale deep ConvNet VGGM	no	VGG-16	NYUDv2 (4 class)	70.4 (PA_{CC})
	[100] (2016)	A network composed by four multi-scale CNNs	no	VGG-16	NYUDv2	49.5
	[101] (2016)	Quadratic Optimization (QO)	yes	Deeplab-v1	Pascal VOC 2012	75.4
	[61] (2017)	DeepLab-v3	no	ResNet	Pascal VOC 2012 Cityscapes	85.7 81.3
	[102] (2017)	Contextual deep structured model	yes	VGG-16	Pascal VOC 2012 Pascal Context SIFT Flow Cityscapes SUN RGBD KITTI NYUDv2	75.3 43.3 44.9 71.6 42.3 70.3 40.6
	[103] (2017)	Deep layer cascade (LC)	no	IRNet [104]	Pascal VOC 2012 Cityscapes	80.3 (PA_{CC}) 71.1
	[84] (2018)	Context Contrasted Local (CCL)	yes	ResNet-101	Pascal Context SUN RGBD COCO Stuff	51.6 47.1 35.7
	[52] (2019)	Stacked Deconvolutional Network (SDN)	no	DenseNet [49]	Pascal VOC 2012 CamVid GATECH RGBD COCO Stuff	83.5 69.6 53.5 35.9
	[105] (2020)	Parallel fully convolutional neural network	no	FCN + HED [96]	Pascal VOC 2012 Pascal Context Cityscapes	66.7 43.6 67.1
	[57] (2021)	HRNet	yes	ResNet + ASPP	Pascal VOC 2012	79.5
	[106] (2017)	Structured patch prediction (SegModel)	yes	ResNet-101	Pascal VOC 2012 Cityscapes ADE20K	82.5 79.2 54.5

2.2.5. Methods Using Recurrent Neural Networks (RNN)

FCNs are limited to small, fixed-size filters that limit their ability to learn long-range dependencies. Recurrent Neural Networks (RNN) are not affected by this restriction. Thanks to its iterations, it spreads the activity, allowing them to model long-range dependence. RNNs are artificial neural networks with cyclical connections. Thanks to these loops, RNN networks can learn complex dynamics. Thus, sequential data can be processed. For example, time series, video frames, etc. However, RNNs are difficult to train due to vanishing gradients and overshooting. To overcome this issue, the Long Short-Term Memory (LSTM) method, which is a type of RNN architecture, has been proposed. The LSTM network is more successful than RNNs in solving the vanishing gradient problem and processing data over a long period of time. Another difference between LSTMs and RNNs is memory cells. In this way, they store status information for short or long periods of time. Thus, for the semantic segmentation problem, RNN and LSTM models have be-

come increasingly popular to model long-short-distance semantic dependencies (semantic connections of local features) in the image using recurrent links.

Some of the most important studies in this field [107] have used a repetitive convolution network that combines a coarser sampled image input with local prediction from the previous iteration, where each iteration contains more and more context. This aspect contrasts with the [79] approach, which first makes a global estimate and then iteratively improves it. Ref. [108] has investigated a "2D LSTM RNN" architecture to efficiently capture local (pixel by pixel) and global (label by label) dependencies (contextual information) within a single model. Ref. [33] have presented "CRFasRNN", which merges the strengths of CNNs and CRFs into a single framework. More specifically, they have formulated the mean field inference of dense CRF with Gaussian binary potentials as an RNN. RNNs can improve the coarse output of a conventional CNN in the forward pass while feeding error differences during training to the CNN. Ref. [109] have proposed to extend the "ReNet" architecture [110], originally designed for image classification, to extract contextual information from images using RNN. Ref. [111] has adopted "Undirected Cyclic Graphs (UCGs)" to decompose pixel connectivity in images. Because of the cyclic nature of UCGs, RNNs cannot be directly applied to UCG-structured images. Therefore, they have decomposed the UCG into several "Directed Acyclic Graphs (DAGs)". Next, they have developed DAG-RNNs, a generalization of RNNs, to process DAG structured images. Aforementioned studies using RNN-based approaches are given in Table 6.

Table 6. Methods using RNN-based approaches.

Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
[107] (2014)	Recurrent CNN ($_{R}$CNN)	no	RCNN	Stanford Background SIFT Flow	69.5 (MA_{CC}) 30.0 (MA_{CC})
[108] (2015)	Two-dimensional LSTM Network (2D LSTM)	no	Multidimensional RNNs	Stanford Background SIFT Flow	68.2 (MA_{CC}) 22.5 (MA_{CC})
[33] (2015)	CRFasRNN	yes	VGG-16	Pascal VOC 2012 Pascal Context	72.0 39.2
[112] (2016)	Higher order CRF-RNN	yes	VGG-16	Pascal VOC 2012 Pascal Context	77.9 41.3
[109] (2016)	ReSeg	no	ReNet [110]	CamVid	58.8
[113] (2016)	Directed acyclic graph RNN (DAG-RNN)	no	VGG-16	SiftFlow CamVid	55.7 (MA_{CC}) 78.1 (MA_{CC})
[111] (2017)	DAG-RNN + CRF	yes	VGG-16	SIFT Flow Pascal Context COCO Stuff	44.8 43.7 31.2
[114] (2018)	Dense RNN (DD-RNN)	yes	VGG-16	Pascal Context ADE20K SIFT Flow	45.3 36.3 46.3

Table 6. *Cont.*

Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
[115] (2018)	Multi-Level Contextual RNN (ML-CRNN)	no	VGG-16	CamVid KITTI SIFT Flow Stanford Background Cityscapes	66.8 60.1 44.7 65.7 71.2
[116] (2020)	Recursive conv. with residual unit	no	VGG-16	Pascal VOC 2012 Cityscapes	55.1 44.0
[117] (2020)	CGBNet	yes	ResNet-101	Pascal Context SUN RGBD SIFT Flow COCO Stuff ADE20K Cityscapes	53.4 48.2 46.8 36.9 44.9 81.2
[60] (2021)	Multi-level graph conv.RNN (MGCRNN)	no	VGG-16	Pascal VOC 2012 Cityscapes	74.2 73.6

3. Weakly Supervised Semantic Segmentation

The fully supervised deep learning models that have been examined so far are approaches that include pixel-level annotated segmentation masks and have achieved significant success in this area. However, these methods require very large numbers of training images, which are very laborious to obtain. To overcome this drawback, weakly supervised techniques have been developed that are faster and less costly than pixel-level labeling. This approach includes such techniques as bounding box, image level, point level and scribble level for each class. A bounding box indicates the location of the object in the image. An image level label indicates the presence or absence of semantic classes. A point level puts a point at the object's position. A scribble level scratches each semantic category in the image. It then fine-tunes using the segmentation losses identified based on these poor descriptions. The difficulty here is how exactly these annotations are mapped to their corresponding pixels. To put it another way, the key task is how to directly relate high-level semantics to a low-level view. An example of weak supervision is given in Figure 8.

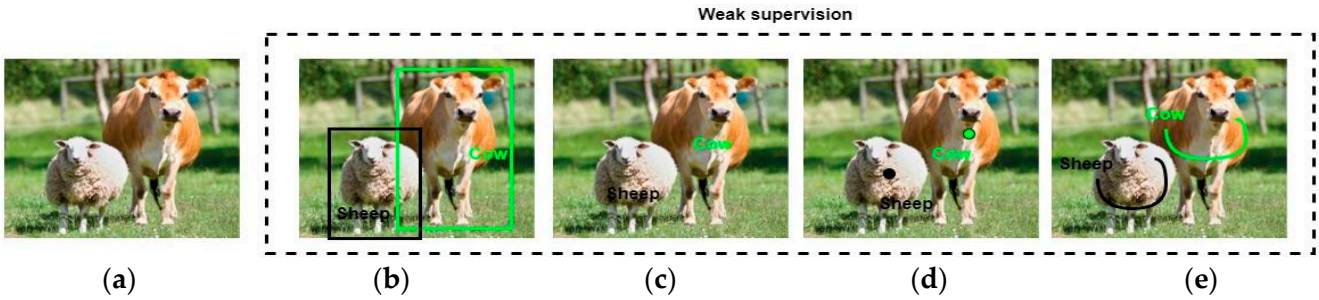


Figure 8. An example of "weakly supervised semantic segmentation". (**a**) Original image; (**b**) Bounding box; (**c**) Image-level; (**d**) Point-level; (**e**) Scribble-level.

3.1. Bounding Box

A bounding box-level label provides the rough localization and size of an object in the image but does not provide detailed shape information about the object. Each object in the image is enclosed in the frame of the video with a rectangle to extract a segmentation mask from each bounding box. With this rectangle, object locations are detected, and weak annotations are extracted.

Ref. [118] has introduced a simple voting scheme to estimate shape guidance for each bounding box. The derived shape guidance is used in the graph-cut-based formulation. "Bounding Boxes to Supervise (BoxSup)" [119] has been proposed to create a set of segments by iterating among automatically composing region proposals and training convolutional networks. Through these two steps, segmentation masks are progressively rescued to evolve the networks. According to [120], input label noise is a problem for weak supervision. To eliminate this noise problem, they have proposed recursive training, where the convnet predictions of the previous training round are used as supervision for the next round. Ref. [121] have developed "Box-driven Class-wise Masking (BCM)" model that they implement the BCM via segmentation-guided learning with box-like supervision. The proposed BCM can help softly remove the irrelevant regions of each class. It also provides an obvious hint of the foreground region, which could greatly contribute to segmentation learning. Aforementioned studies using bounding box level-based approaches are given in Table 7.

Table 7. Bounding box level-based approaches.

Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
[118] (2013)	DET3	no	CPMC [122]	Pascal VOC 2012_{test}	48.0
[119] (2015)	Bounding Boxes to Supervise CNN (BoxSup)	yes	DeepLab-v1	Pascal VOC 2012_{val} Pascal $Context_{val}$	62.0 40.5
[120] (2017)	SimpleDoselt (SDI)	no	DeepLab-v1	Pascal VOC 2012_{val}	65.7
[121] (2019)	Box-driven class-wise masking (BCM + FR-loss)	yes	DeepLab-v1 ResNet-101	Pascal VOC 2012_{val}	$66.8_{DeepLabv1}$ $70.2_{ResNet101}$
[123] (2020)	FCN+Cartesian/ Polar Coordinate System (CCS) + (PCS)	yes	VGG-16	Pascal VOC 2012_{val}	68.7
[124] (2021)	Background-Aware Pooling (BAP) and Noise-Aware Loss (NAL)	yes	DeepLab-v1	Pascal VOC 2012_{val}	68.1
[125] (2022)	Pixel-as-Instance Prior (PIP)	yes	DeepLab-v1	Pascal VOC 2012_{val}	67.9

3.2. Image-Level

- Multi-Instance Learning (MIL)

The image-level label only determines which classes are present without specifying the location of the objects in the image. The first few studies in this section used the "Multi-Instance Learning (MIL)" [126] framework, which diminished the level of supervision required, reducing the need for expensive annotations in tasks such as semantic segmentation. Ref. [127] has represented each image as a bag of pixel-level-instances and defined a pixel-wise multi-class matching of MIL for loss to learn the segmentation model from image labels. Another example based on MIL, "Log-Sum-Exp (LSE)" [128], has not been trained with pixel-label or annotations such as bounding boxes or scribbles. In place of that, it just obtained a single object class tag for a given image and limited it to giving

more weight to pixels significant for classification. Ref. [129] have developed the "Globally Weighted Ranking Pool (GWRP)" that is used by dilation loss to widen the object seeds to regions of acceptable size. Since these MIL-based methods are just image labels, they rely on classification networks to locate objects. On the other hand, since there is no pixel-wise annotation, classification networks produce faulty and coarse object regions, reducing semantic segmentation performance. Additional methodologies, such as CAM and pseudo-labeling, have been devised to address this issue.

- Class Activation Maps (CAM)

The utilization of Class Activation Maps (CAM) enables the visualization of the specific areas within an image that are prioritized by a CNN model in the process of classification. The utilization of maps can be considered valuable in understanding the significant portions of an image that a model finds essential for its predictive capabilities.

The utilization of CAM is feasible in a weakly supervised scenario within the domain of semantic segmentation. Rather than utilizing a dataset that is completely annotated with pixel-level labels for segmentation, it is possible to initially train a classification model using image-level labels. The present model has the capability to produce CAMs that effectively emphasize the most distinctive areas within the image that are utilized for the purpose of classification.

Ref. [130] explained how to create CAM with CNNs' global average pooling (GAP). The discriminative image areas that the CNN utilized to identify a given category are shown on a class activation map for that category. CNNs that have been trained for classification can be taught to perform object localization without the need for bounding box annotations. Using CAM, they can see the projected class scores for any given image and see where the CNN found discriminative object features.

Another study about CAM, Erased CAM Supervision Net (ECSNet) method proposed in [131]. They utilized connections between CAMs to suggest a unique weakly supervised technique, which was motivated by the fact that removing differentiating features forces networks to accumulate new ones from non-discriminative object areas. In this work, they used segmentation supervision, driving networks, and the characteristics learnt from deleted pictures to examine resilient representation. CAM-derived object sections are initially removed from pictures. Erased CAM Supervision Net (ECSNet) creates pixel-level labels by anticipating the segmentation outcomes of those processed pictures to give segmentation supervision to other areas. Additionally, they developed the rule of minimizing noise to choose trustworthy labels. Except for ground truth image-level labels, their trials on the Pascal VOC 2012 dataset demonstrate that their ECS-Net outperforms earlier state-of-the-art techniques, achieving 67.6% mIoU on the test set and 66.6% mIoU on the validation set.

Most existing approaches to image-level labelling consist of two components. The first component is the FCN, the second is the one that aims to provide an efficient pseudo-mask at the label level. This mask is utilized to control the pixel level required by the training process. Because of this, it is very important to acquire an effective mask for this part. The CAM [130] has often been used to generate pseudo-masks and then train segmentation models. CAM creates an attention map to localize the most distinctive regions of the object. The approaches mentioned in the table as the CAM-based method have adopted the CAM method to select the most distinctive regions. However, it has been observed that CAM is successful for small objects but can localize only the small distinctive region of the target object when faced with large-scale objects. This will reduce performance as every undetected object will be labeled as background for the semantic segmentation task. To address this problem, most studies in this group first utilized CAMs to locate objects in each category and used salience detection techniques to select background regions. For example, Ref. [66] has suggested "Saliency-Guided Refinement Method" that considers both extended object regions and salience maps under a Bayesian frame. Refs. [132–134] have focused on discovering invisible semantic objects with the "erasing strategy" they followed in their studies. Unlike the others, Ref. [134] have used "attention maps" created by

"SeeNet", which they developed, not CAM. To solve the same problem, another study [135] has proposed to transfer discriminant information from sparsely highlighted regions to adjacent object regions, thereby creating dense object localization, which can essentially lift segmentation model learning favorably. Refs. [136,137] have created "localization maps" of the training images for classification. The localization maps are then used as pseudo-labels to train a segmentation network.

As a weakly supervised learning method, pseudo-labeling uses the model's own predictions on unlabeled data as ground truth labels for further training of the model. This method can be utilized in semantic segmentation tasks to make use of a significant amount of unlabeled data, which is particularly useful in situations where completely annotated data is insufficient. When using pseudo-labels in a semantic segmentation task, the model is trained on labeled data, and predictions are made on unlabeled data using the trained model. Pseudo-labels are assigned to the unlabeled data based on the model's predictions. Pseudo-labels can be improved via post-processing with CRFs and other methods depending on the accuracy of the original model's predictions. Adding pseudo-labels to a true-label dataset makes for a more robust training set. Finally, the model is retrained with both true and pseudo labels. This allows the model to make use of more data, which may improve its overall performance. This procedure can be repeated. Following each iteration of training, the model should be better able to predict the unlabeled data, leading to more precise pseudo-labels and an overall boost in performance.

The web-based methods mentioned in the table have retrieved relevant videos automatically from the web and generated fairly accurate object masks of the classes from the videos to simulate supervision for semantic segmentation. Aforementioned studies using image-level-based approaches are given in Table 8.

Table 8. Image-level-based approaches.

	Content	Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
MIL based method		[127] (2014)	Multiple instance learning (MIL-FCN)	no	VGG-16	Pascal VOC 2012_{test}	25.6
		[138] (2015)	Constrained CNN (CCNN)	yes	VGG-16	Pascal VOC 2012_{val}	45.1
		[128] (2015)	Log-Sum-Exp (LSE)	no	OverFeat [139]	Pascal VOC 2012_{val}	42.0
		[140] (2016)	Built-in Fore/Backgr. Prior for WSS	yes	VGG-16	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	46.6 52.9
		[129] (2016)	Seed, Expand and Constrain (SEC):	yes	DeepLab-v1	Pascal VOC 2012_{val}	51.7
	CAM used method	[141] (2017)	(multi-class masks) +CRF	yes	VGG-16	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	50.9 52.6
	EM Alg. used method	[142] (2015)	Weakly Semi-Supervised Learning (WSSL)	yes	DeepLab-v1	Pascal VOC 2012_{val}	$60.6_{Bound.box}$ $38.2_{Image-level}$

Table 8. *Cont.*

Content		Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
		[143] (2016)	Augmented Feedback	yes	DeepLab-v1	Pascal VOC 2012_{val}	52.6_{SS} 54.3_{MCG}
Pseudo mask-based method		[144] (2016)	(HCP) [145]-(MCG) [24]	no	VGG-16	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	41.9 43.2
	CAM used method	[132] (2017)	Adversarial erasing (AE)-Prohibitive seg. learn. (PSL)	yes	DeepLab-v1	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	55.0 55.7
		[133] (2018)	Guided attention inference Netw. (GAIN)	yes	VGG-16	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	60.5 62.1
		[134] (2018)	Self-Erasing Network (SeeNet)	yes	VGG-16 ResNet-101	Pascal VOC 2012_{val}	61.1_{VGG16} $63.1_{ResNet101}$
						Pascal VOC 2012_{test}	60.7_{VGG16} $62.8_{ResNet101}$
		[135] (2018)	Multi-dilated convolu-tional (MDC)	yes	VGG-16	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	60.4 60.8
		[146] (2018)	AffinityNet	yes	DeepLab ResNet-38 [147]	Pascal VOC 2012_{val}	$58.4_{DeepLab}$ $61.7_{ResNet38}$
						Pascal VOC 2012_{test}	$60.5_{DeepLab}$ $63.7_{ResNet38}$
		[148] (2018)	Deep seeded region growing (DSRG)	yes	VGG-16 ResNet-101	Pascal VOC 2012_{val}	59.0_{VGG16} $61.4_{ResNet101}$
						Pascal VOC 2012_{test}	60.4_{VGG16} $63.2_{ResNet101}$
		[136] (2019)	Ficklenet	no	VGG-16 ResNet-101	Pascal VOC 2012_{val}	61.2_{VGG16} $64.9_{ResNet101}$
						Pascal VOC 2012_{test}	61.9_{VGG16} $65.3_{ResNet101}$
		[66] (2018)	Mining Common Object Features (MCOF)	no	VGG-16 ResNet-101	Pascal VOC 2012_{val}	56.2_{VGG16} $60.3_{ResNet101}$
						Pascal VOC 2012_{test}	57.6_{VGG16} $61.2_{ResNet101}$
		[149] (2019)	Online attention accumulation (OAA)	no	VGG-16 ResNet-101	Pascal VOC 2012_{val}	63.1_{VGG16} $65.2_{ResNet101}$
						Pascal VOC 2012_{test}	62.8_{VGG16} $66.4_{ResNet101}$

Table 8. *Cont.*

	Content	Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
Pseudo mask-based method		[150] (2021)	PuzzleCAM	yes	ResNet-101	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	66.9 67.7
		[151] (2022)	Suppression Module (SUPM) + Saliency Map Guidance Module (SMGM)	yes	ResNet-101	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	73.3 73.5
		[152] (2020)	Intra-Class Discriminator (ICD)	yes	VGG-16 ResNet-101	Pascal VOC 2012_{val}	64.0_{VGG16} $67.8_{ResNet101}$
						Pascal VOC 2012_{test}	63.9_{VGG16} $68.0_{ResNet101}$
		[137] (2019)	Saliency& segm. network (SSNet)	yes	VGG-16 Densenet	Pascal VOC 2012_{val}	63.3_{VGG16} $57.1_{Densenet}$
						Pascal VOC 2012_{test}	64.3_{VGG16} $58.6_{Densenet}$
		[153] (2016)	Distinct Class Saliency Maps (DCSM)+CRF	yes	VGG-16	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	44.1 45.1
Web-based methods	CAM used method	[154] (2017)	Web-Crawled Videos	no	VGG-16	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	58.1 58.7
		[155] (2016)	Simple to complex (STC)	yes	VGG-16	Pascal VOC 2012_{val} Pascal VOC 2012_{test}	49.8 51.2
		[156] (2017)	WebS-i2	yes	VGG-16	Pascal VOC 2012_{val}	53.4
		[157] (2017)	Dual image segmentation (DIS)	no	ResNet-101	Pascal VOC 2012_{test}	86.8
		[158] (2017)	Weakly supervised Two-stream Network	yes	VGG-16	$Camvid_{test}$ $Cityscapes_{test}$	29.7 47.2

3.3. Scribble-Point Level

A scribble annotation is a set of pixels with a category label. Scribbles are provided infrequently, and unannotated pixels are considered unknown. Compared to a box annotation, it can provide certain boundaries for objects, but scribbles are mostly labeled on the insides of objects. Additionally, box annotations mean that all pixels outside the boxes are not from the corresponding categories. This state does not exist for scribbles, and the information from scribbles must propagate to all other unknown pixels. Compared to image-level annotations, scribbles provide location information at a few pixels, which should lead to better results.

Ref. [159] has addressed "Scribble Supervised Training (ScribbleSup)" by optimizing a graphical model. The graphic model spreads information from scribbles to unmarked pixels according to semantic content, appearance and spatial constraints. Meanwhile, an FCN is trained, which is controlled by the emitted labels and provides semantic predictions for the graphical model. They have formulated this model as a composite loss function and developed an alternative method to optimize it. Ref. [160] have produced a "Random Walk-

based Label Propagation Mechanism (RAWKS)" that has been shown to be differentiable and usable in deep neural network architectures. Ref. [161] has introduced "point-level supervision", where every category is just related to one or many pixels. They have extended CNN to include this point of supervision in its training loss function. Aforementioned studies using scribble-point-level-based approaches are given in Table 9.

Table 9. Scribble-point-level-based approaches.

Content	Author	Method	CRF Used?	Backbone Network	Dataset	Accuracy mIoU (%)
point	[161] (2016)	Sem.Seg.with Point Supervision	yes	VGG-16	Pascal VOC 2012_{val}	42.9
scribble	[159] (2016)	Scribble-Supervised CNN (ScribbleSup)	yes	VGG-16	Pascal VOC 2012_{val} Pascal $Context_{val}$	63.1 39.3
	[160] (2017)	Random-walk based label propagation mech. (RAWKS)	yes	ResNet-101	Pascal VOC 2012_{val} Pascal $Context_{val}$	60.0 37.4
	[162] (2018)	GraphNet	yes	VGG-16	Pascal VOC 2012_{val} Pascal $Context_{val}$	63.3 39.7
	[163] (2018)	NormalCut	yes	ResNet-101	Pascal VOC 2012_{val}	74.5
	[164] (2018)	KernelCut	yes	ResNet-101	Pascal VOC 2012_{val}	75.0
	[165] (2019)	Boundary Perception Guidance (BPG)	yes	ResNet-101	Pascal VOC 2012_{val}	76.0
	[166] (2021)	Progressive segmentation inference (PSI)	no	ResNet-101	Pascal VOC 2012_{val} Pascal $Context_{val}$	74.9 43.1

4. Recent Approaches in Semantic Segmentation

4.1. Segment Anything Model (SAM)

The Segment Anything Model (SAM) [167] was developed by the Meta AI Research team as an automated image segmentation model. It is based on foundation models and operates with a high level of automation and requires minimal human assistance. Foundation models refer to pre-trained models that have been trained on extensive amounts of data and possess the capacity to generalize to novel tasks and data distributions through the utilization of prompt engineering. Several deep learning methodologies require retraining of the model in response to dataset modifications. SAM provides an adaptable image segmentation model that is more comprehensive and effective.

The process of object segmentation in SAM can be achieved by choosing specific points for inclusion or exclusion from the object through selection or clicking. Segmentations can be produced by utilizing bounding boxes or polygon tools, which will align automatically with the object. SAM exhibits the capability to produce numerous valid masks in instances where there is uncertainty in the identification of the object that requires segmentation.

The System for Automated Masking (SAM) exhibits the ability to autonomously identify and generate masks for all entities encompassed within an image.

By precomputing the image embeddings, the Segmentation Attention Module (SAM) can efficiently produce a segmentation mask for a given prompt, enabling seamless interaction with the model in real-time.

The authors in [168] gathered a total of 52 open-source datasets and utilized them to construct a comprehensive medical segmentation dataset. This dataset comprises 16 modalities, 68 objects and a total of 553K slices. A thorough examination of various strategies for SAM testing was carried out. Experimental findings confirm that the utilization of manual cues such as points and boxes enhance the efficacy of SAM in object recognition within medical imagery, resulting in superior performance in the prompt mode as opposed to the

everything mode. Furthermore, SAM exhibits exceptional proficiency in certain objects and modalities, however, it demonstrates inadequacy or complete ineffectiveness in other contexts. Eventually, an analysis was conducted to evaluate the impact of various factors on the segmentation performance of SAM. Experimental studies confirm that the zero-shot segmentation ability of SAM is insufficient for its immediate implementation in medical image segmentation.

Ref. [169] have conducted a comprehensive assessment of the segmentation performance of SAM on a diverse set of 19 medical imaging datasets that include different modalities and anatomical regions. The performance of SAM, as reported, shows substantial variation based on the dataset and task when evaluated on individual prompts. Furthermore, it has been observed that the performance of SAM is significantly enhanced when utilizing box prompts as opposed to point prompts. SAM's performance tends to improve when iterative multiple-point prompts are given. In addition, several illustrations were presented to demonstrate SAM's efficacy across all evaluated datasets, its iterative segmentation capabilities, and its response to prompt ambiguity. The researchers arrived at the conclusion that the zero-shot segmentation performance of SAM is remarkable for specific medical imaging datasets, while it exhibits moderate to inadequate performance for other datasets. The utilization of SAM in medical imaging has the potential to yield substantial advancements in automated medical image segmentation.

4.2. Unsupervised Domain Adaptive in Semantic Segmentation

The fully supervised methods require a significant amount of pixel-level annotations, as previously stated. This is quite expensive and time-consuming. Weakly supervised learning is a technique that leverages a restricted or imprecise set of labels, whereas unsupervised learning operates without any labels whatsoever. However, as per the existing literature, it is widely acknowledged that weak and unsupervised methods tend to exhibit lower performance compared to their supervised counterparts. The "Unsupervised Domain Adaptation (UDA)" approach has garnered significant attention in the semantic segmentation area in recent times. The UDA semantic segmentation approach involves adapting a model that has been trained on a source domain containing labeled data, to effectively perform on a target domain that contains unlabeled data. This is achieved by utilizing shared features between the two domains. Ref. [170] have inferred that utilizing an UDA approach is a highly recommended method for training semantic segmentation models that are intended to operate dependably and effectively in real-world scenarios, utilizing both labeled and unlabeled data. This is particularly noteworthy given that the acquisition of unlabeled data is considerably simpler and more cost-effective than that of labeled data. The UDA process is generally like this:

- The process of source domain training involves the utilization of conventional supervised learning methods to train a model on labeled data obtained from the source domain. The model that has undergone training is capable of accurately segmenting images within the source domain.
- The process of feature alignment involves utilizing the model to extract features from both the source and target domain data. The objective is to achieve a high degree of similarity in the distribution of features across both domains. Typically, this stage entails a form of adversarial training [171]. The approaches based on adversarial training, such as [172–174] have made remarkable progress for UDA semantic segmentation.
- Many UDA semantic segmentation techniques employ self-training or self-supervision, whereby the model's predictions on the target domain are utilized as pseudo-labels for subsequent training. Typically, this process is executed meticulously and incrementally, whereby the model's highly assured predictions are employed to progressively enhance its capacity to manage the intended field. Approaches based on self-training or self-supervision such as [175–177] have demonstrated significant advancements in UDA semantic segmentation.

- The model is subjected to evaluation in the target domain, employing conventional segmentation metrics such as pixel accuracy or 'IoU', followed by fine-tuning. If deemed essential, the second and third steps are reiterated to achieve additional adaptation.

"DecoupleNet" has been proposed [178] as a means of mitigating the reliance on data and prioritizing "UDA". Two challenges have been identified in current domain-invariant learning approaches, namely, task entanglement and source domain overfitting. Additionally, the self-discrimination (SD) technique has been proposed, utilizing pseudo-labels to enhance the acquisition of more discriminative features for the target domain.

Ref. [179] has introduced "HRDA" which is the first work to learn a multi-resolution input fusion for UDA semantic segmentation. Because of the adaptability of small objects and segmentation details is facilitated by high-resolution (HR) inputs, whereas the adaptability of large regions is facilitated by low-resolution (LR) inputs, HRDA is designed to be applicable to most UDA methods. The results have indicated that HRDA consistently enhances performance by a minimum of +2.4 mIoU, suggesting that the HRDA pseudo-labels serve as a positive reinforcement for the UDA process.

Table 10 presents the comparison with state-of-the-art methods for UDA on the Cityscapes dataset. Even though these methods have proposed new adaptation strategies, they have mostly used ResNet-101 as the backbone network architecture. These outdated networks have not provided a UDA performance gain. Ref. [180] has proposed a novel "DAFormer" method and identified the transformer-based SegFormer as a powerful backbone architecture for UDA. DAFormer's network architecture comprises a transformer encoder and a decoder that fuse context-aware features at multiple levels. The stabilization of training and prevention of overfitting to the source domain are facilitated by three essential training strategies. The quality of pseudo-labels can be enhanced by rare class sampling in the source domain, which helps reduce the confirmation bias of self-training towards common classes. Additionally, feature transfer from ImageNet pretraining can be promoted using a Thing-Class ImageNet Feature Distance and a learning rate warmup. The DAFormer model constitutes a significant breakthrough in the field of UDA.

Table 10. Comparison with state-of-the-art methods for UDA on the GTA5 → Cityscapes benchmark.

Author	Method	CRF Used?	Backbone Network	Accuracy mIoU (%)
[176] (2018)	Class-balanced self-training (CBST)	no	ResNet-38	48.4
[181] (2021)	Domain Adaptation Cross Sampling (DACS)	no	ResNet-101	52.1
[182] (2021)	Correlation-Aware Domain Adaptation (CorDA)	no	ResNet-101	56.6
[183] (2021)	Prototypical pseudo label denoising (ProDA)	no	ResNet-101	57.5
[184] (2022)	Continual test-time adaptation approach (CoTTA)	no	ResNeXt-29 [51] SegFormer [185]	$32.5_{\text{ResNeXt-29}}$ $58.6_{\text{SegFormer}}$
[180] (2022)	DAFormer	no	SegFormer	68.3

5. Post-Processing Algorithms in Semantic Segmentation

Post-processing is an essential part of machine learning domains, including semantic segmentation, which helps improve the output of the model, enhancing the quality of the results and removing noise. In this section, the most common post-processing algorithms used in semantic segmentation are briefly introduced.

5.1. Conditional Random Fields (CRF)

"Conditional Random Fields (CRF)" is a type of probabilistic graphical model that can be effectively employed to represent the interdependencies among the output variables in a structured prediction task. Within the realm of semantic segmentation, the resultant variables are representative of the categorical labels assigned to individual pixels, with the interdependence among these variables reflecting the spatial associations between said pixels.

CRFs can represent complex interconnections and interdependencies among various labels within a structured output. The mentioned attribute is an essential aspect that makes them appropriate for applications such as semantic segmentation within the realm of computer vision.

When assigning a class label to a pixel in semantic segmentation, CRFs are utilized to consider the neighborhood context of that pixel. This is executed to enforce local consistency in the labeling of the image: it is likely that two adjacent pixels that are part of the same object will share the same class label.

Ref. [32] have examined the utilization of fully connected CRF models that are defined on the entire set of pixels present in an image. The graphs that result from the process exhibit a vast number of edges, rendering conventional inference algorithms unfeasible. The primary contribution of our study is a notably efficient method for approximate inference in fully connected CRF models. This method is specifically designed for models in which the pairwise edge potentials are established through a linear combination of Gaussian kernels. Ref. [186] have stated that CRF post-processing is no longer commonly used in newer publications. They argue that this is because knowing the underlying CRF parameters is challenging and that CRFs are slow during both training and inference. They have proposed enhancing the fully connected CRF framework with the premise of conditional independence to address both problems. They can then rewrite the inference so that it uses convolutions, an operation that runs very quickly on GPUs. Inference and training times are improved by an order of magnitude. Backpropagation can be used to easily fine-tune all the convolutional CRFs' parameters. The authors in [187] have suggested the utilization of a Gaussian CRF model for semantic segmentation, as opposed to the current methods that employ discrete CRF models. The authors introduce a new deep neural network architecture, denoted as Gaussian Mean Field (GMF) network, in which the individual layers execute mean field inference on a Gaussian CRF. The GMF network under consideration possesses the desirable characteristic whereby every layer of the network generates an output that is in closer proximity to the maximum a posteriori solution of the Gaussian CRF in comparison to its respective input. The authors suggest the integration of the proposed GMF network with deep CNNs to introduce a novel GCRF network that can be trained end-to-end. Upon being trained end-to-end in a discriminative manner and subsequently evaluated on the demanding Pascal VOC 2012 segmentation dataset, the Gaussian CRF network proposed in this study surpasses several recent semantic segmentation methodologies that integrate CNNs with discrete CRF models.

5.2. Markov Random Field (MRF)

Markov Random Fields (MRFs) are a type of probabilistic graphical model that has been widely employed in the field of semantic segmentation. This is related to their ability to capture the spatial associations and interdependencies among adjacent pixels or regions within an image. MRFs approach the task of semantic segmentation as a labeling problem, in which the objective is to assign the most probable label to each pixel, considering the observed image and the spatial context.

Each pixel within the image is regarded as a node by the MRF. The random variable associated with the node represents the label that has been assigned to the pixel that it corresponds to. The establishment of interconnections among nodes is facilitated by potential functions. The functions represent the cost or likelihood of a set of nodes embracing labels. For instance, a conceivable algorithm may prioritize the similarity of labels between neighboring pixels to enhance spatial coherence. The aim of the MRF is to assign a discrete label to each pixel in a manner that minimizes the overall cost. The MRF inference process is categorized as a problem of combinatorial optimization. Parameterization of potential functions is a common practice in machine learning, which enables their acquisition from a set of annotated training images. To accomplish this task, the parameters are optimized such that the potential functions assign lower costs to the accurate labels. Ref. [188] has handled semantic segmentation by combining high-order relations and label context mixtures into MRF. The authors propose a solution to Markov Random Field (MRF) by introducing a "Deep Parsing Network (DPN)" based on CNNs. This approach facilitates deterministic end-to-end computation within a single forward pass. "DPN" is a technique that enhances a contemporary CNN to effectively represent unary terms. In contrast to prior works that require multiple iterations of MP during back-propagation, DPN can achieve better results by approximating a single iteration of MF. "DPN" incorporates pairwise terms that offer a comprehensive structure for encoding contextual information in high-dimensional data, such as images and videos. "DPN" facilitates the parallelization and acceleration of MF, thereby enabling effective inference. The accuracy of "DPN" is examined on conventional semantic image and video segmentation benchmark datasets. Their results show that "DPN" achieves state-of-the-art performance on Pascal VOC 2012, Cityscapes and CamVid datasets.

5.3. Random Walker

The Random Walker algorithm is a semi-supervised learning methodology that is predominantly employed for the purpose of image segmentation. The process of assigning unlabeled pixels to labeled ones in an image is accomplished through the simulation of a random walk process. In the context of image analysis, individual pixels are regarded as nodes within a graph structure, with interconnections between nodes being established based on their respective similarities. The process involves the emission of walkers or paths from every unlabeled pixel towards its neighboring pixels, and this propagation persists until a labeled pixel is reached. The assignment of a label to an unlabeled pixel is ascertained based on the predominant label among the labeled pixels that have been reached.

In order to employ a random walker algorithm for semantic segmentation, a graph is generated in which each individual pixel within the image is mapped to a corresponding node within the graph. The interconnections linking each vertex are established based on a similarity in pixels, where the magnitude of the edge denotes this resemblance. Following that, certain pixels in the image are assigned labels. The labels mentioned indicate the classification of the respective pixel. Conduct a random walk starting from every unmarked pixel until a marked pixel is reached. The probability of the random walk traversing edges with greater weights is higher. When the random walk arrives at a labeled pixel, the label of the unlabeled pixel is determined. The selection of the label for the random walk can be predicated on either the label that is most frequently encountered or the label that is initially encountered. Iterate the process until all pixels have been labeled or until there is no further change in the labels between iterations.

The acquisition of training data for semantic segmentation poses a significant challenge on a large scale, as it is comparatively costly when compared to other visual tasks. The authors in [160] have suggested an innovative training methodology to tackle this challenge. We utilize sparsely obtained image labeling to generate densely labeled images through label propagation techniques. A segmentation network based on the CNN architecture is trained to replicate the labeling. The process of label propagation is established through the utilization of probabilities of random walk hitting, resulting in a parameterization that is differentiable and includes estimates of uncertainty that are integrated into our loss function. The authors demonstrate that through the joint learning of the label-propagator and segmentation predictor, they successfully acquired knowledge of semantic edges without the provision of explicit edge supervision. The conducted experiments demonstrate that the performance of a segmentation network can be enhanced by training it using the proposed method, as opposed to the conventional approach.

5.4. Domain Transform

The Domain Transform technique is a methodology that makes use of edge-preserving filters to accomplish the objective of semantic segmentation. The utilization of edge-preserving filters serves as a viable means to incorporate local features, thereby enhancing the precision of semantic segmentation, particularly in instances where the segmentation procedure is more arduous, such as object boundaries.

In the context of semantic segmentation, the domain transform process involves an initial pre-processing step of the input image. This pre-processing step typically involves the use of a CNN or another machine learning model to extract initial features and generate a coarse segmentation. The concept of domain transformation involves the conversion of a complex multidimensional image segmentation problem into a more manageable one-dimensional problem by focusing on the edges of the image. This approach is preferred due to the relative simplicity of solving one-dimensional problems. The preservation of image edges is crucial for precise segmentation, a task that the process accomplishes. After the completion of said tasks, an iterative filter that relies on the domain transform is employed to disseminate data while retaining the integrity of the edges. This task is referred to as an edge-preserving filter. This operation can be interpreted as a form of smoothing that tends to integrate information within an object while minimizing the degree of blurring across object boundaries. Finally, a refinement process is typically executed to enhance the precision of the segmentation. This could potentially entail the utilization of supplementary machine learning models or alternative image processing methodologies.

6. Datasets

Semantic segmentation is a topic with a wide variety of applications, so in recent years, many datasets have been created for this task. Semantic segmentation is the primary function for which these datasets are utilized. However, it can also be used for object detection, instance segmentation, and other computer vision tasks. In this section, the common data sets used in the studies examined within the scope of this study have been explained. In Table 11, the datasets have been divided into four groups according to their contents: general, indoor, outdoor and street scenes. Additionally, this table has provided some useful information for these datasets, such as the number of classes, number of images, training/validation/testing split, and image resolution. The images are split into separate sets for training, validation, and testing. This allows researchers and developers to train and evaluate their models effectively. In the parts marked '-' in the validation set, there is no validation set provided, and researchers often use the test set for validation purposes.

- **ADE20K**

ADE20K [189] is a large-scale, diverse dataset that consists of over 20,000 indoor and outdoor scene images with thorough pixel-level annotations. The categories include things such as wall, building, sky, person, road, bed and many more. This dataset is used to construct a scene parsing benchmark using 150 object and stuff classes. All the images have been exhaustively annotated with stuff, objects, and object parts. Additional information about the opacity, cropping, and other properties of each object has been provided. When compared to the training set, the images in the validation set have more comprehensive part annotations.

- **COCO Stuff**

The original COCO dataset [190] has been expanded by adding dense pixel-wise stuff annotations, and the COCO-stuff dataset [191] has created. COCO-Stuff makes it possible to explore deeply the connections between things and their stuff. This dataset contains 172 classes: 80 things, 91 stuff and 1 unlabeled class. The 80 things classes are the same as in classic COCO. An expert annotator has chosen the 91 classes of stuff. If a label does not fit into any of the 171 specified classes or if the annotator is unable to deduce the label of a pixel, the class unlabeled is used. The 'thing' classes include such things as, Bear, Bicycle, Stop sign, Knife, Person, Parking meter, Clock, Traffic light, etc. The 'stuff' classes include a wide variety of materials, regions and other non-object categories. For example, Dirt, Fog, Hair, Dots, Screen, Plastic, Grid, etc. The authors aimed to promote further investigation into contextual relationships between stuff and thing by revealing this dataset.

- **Pascal VOC (Visual Object Classes)**

From 2005 through 2012, the Pascal VOC challenge has been held, and a new version of the dataset has been released annually. It is the most popular dataset in the literature for the semantic segmentation task. The dataset includes 20 different classes for annotation divided into 4 categories: Animals (bird, cat, cow, dog, horse, sheep), Indoor (bottle, chair, dining table, potted plant, sofa, tv), Person and Vehicles (aeroplane, bicycle, boat, bus, car, motorbike, train). Additionally, the difficult object examples have been removed from both training and test sets by masking these objects with the 'Void' label. So, in Pascal VOC 2012 [192], the total number of classes has increased to 21.

- **Pascal Context**

The Pascal Context dataset [193] is an extension of the Pascal VOC 2010 dataset. Pascal Context differs from the PASCAL VOC 2010 dataset in two key respects: the larger number of classes and the greater level of depth in the annotations. There are only about 20 distinct types of objects annotated in Pascal VOC. In contrast, Pascal Context has 540 classes (459 'stuff' classes, 80 'thing' classes and 1 'unlabeled' class) that annotate images in far greater depth. The images are all pixel-by-pixel annotated. The most important 59 of these classes have been selected. These are both indoor and outdoor object classes, such as Bird, Bottle, Cow, Person, Clouds, Floor, Snow, Sea, Wood, Window.

- **NYU-Depth V2 (NYUDv2)**

The NYUDv2 dataset [194] includes 1449 RGBD images, representing 464 indoor scenes, classified into 26 scene types, collected from a wide variety of buildings in three major US locations. Using Amazon Mechanical Turk, per-pixel labeling has been acquired for every single one of the images. The dataset includes 40 different indoor classes such as, Ground, Table, Bed, Television, Wall etc. This dataset's inclusion of color and depth information makes it especially useful for tasks that necessitate a more nuanced comprehension of a scene's 3D geometry.

- **SUN RGBD**

Like the NYUDv2 dataset, SUN RGBD dataset [195] consists of color (RGB) images that also contain depth (D) information for each pixel. The dataset consists of around 10,335 RGB-D video frames that have been captured in 41 different buildings, such as homes, classrooms, stores and offices. The researcher has attached an ASUS Xtion PRO LIVE sensor to a laptop for the capture process. This dataset contains 37 classes such as Bed, Ceiling, Chair, Floor, Furniture, Lamp, Objects, Picture, Sofa, Table, Tv, Toilet, Window and comprehensive pixel-level annotations. Both 2D bounding boxes and 3D point cloud bounding boxes have been used to annotate objects in each image.

- **Berkeley Deep Drive (BDD100K)**

The BDD100K dataset [196] is a large-scale, diverse driving video dataset with extensive annotations that can reveal the difficulties of street-scene understanding. The dataset comprises 100k video sequences with a high resolution of 720p and a high frame rate of 30 fps. These videos have been collected from New York, San Francisco Bay Area, and other regions. The dataset includes scenes (e.g., residential, city street, highway, tunnel), different times of the day (e.g., day, night), and diverse weather conditions (e.g., clear, rainy, snowy, foggy).

- **The Cambridge-driving Labeled Video Database (CamVid)**

The CamVid [197] is a road/driving scene understanding dataset that includes four HD video sequences. Of these, three videos were recorded during daylight hours, while one was recorded at night. The study has employed a Panasonic HVX200 digital camera with 3CCD and high-definition capabilities to capture frames at a resolution of 960×720 pixels and a frame rate of 30 fps. The video resolution is low because it is a very old dataset. The dataset comprises 32 distinct semantic classes that depict diverse objects commonly observed in a road scene. These commonly used classes might include Bicyclist, Building, Car, Fence, Sidewalk, Sky, Pavement, Pole, Road, Tree, Sign Symbol and Void (not labeled and ignored in evaluation).

- **Cityscapes**

Over the course of several months, a vast number of frames were obtained from a mobile platform, capturing the seasons of spring, summer, and fall across 50 urban areas, predominantly in Germany but also in adjacent nations. The decision was made to intentionally refrain from recording during unfavorable weather conditions, such as intense precipitation or snowfall, due to the belief that such conditions necessitate specialized methodologies and datasets. The images were captured utilizing a stereo camera with a 22 cm baseline, which was designed for automotive applications. The camera employed 1/3-inch CMOS sensors with a resolution of 2 megapixels, specifically the On Semi AR0331 model, and utilized rolling shutters. The framerate at which the images were recorded was 17Hz. The dataset [198] contains 30 different classes for annotation divided into 8 categories such as Construction (building, wall, fence, etc.), Flat (road, sidewalk, parking, e.g.), Human (person, rider), Nature (vegetation, terrain), Object (pole, traffic sign, traffic light, etc.), Sky, Vehicle (bus, car, truck) and Void. Among these classes, 19 have been used for evaluation. The remaining 11 classes have been included in the 'Void' class in the evaluation.

- **DTMR-DVR**

The dataset [199] has been made available by the Department of Transport and Main Roads (DTMR) located in Queensland, Australia. Digital Video Recording (DVR) data has been collected through the utilization of cameras that are mounted on vehicles. The team has curated a DTMR-DVR dataset for the purpose of semantic segmentation through manual means. The methodology involves the extraction of image frames from the given videos, followed by the utilization of Adobe Photoshop to annotate the extracted images, thereby facilitating the generation of pixel-wise class labels. The dataset comprises a total

of 13 distinct classes of roadside objects, including but not limited to Road, Line, Pole, Tree, Grass and Light.

- **KITTI**

The KITTI dataset [200] is a popular benchmark dataset, specifically for tasks related to self-driving vehicles. These data have been collected by the Karlsruhe Institute of Technology and the Toyota Technological Institute in Chicago. Although the number of classes for 2D and 3D object detection tasks varies, for semantic segmentation tasks, KITTI has 19 classes such as Person, Truck, Terrain, Vegetation, Traffic light, Traffic sign, Train, etc.

- **GATECH**

The authors have presented a newly developed dataset that features pixel-level annotations for the purpose of conducting geometric scene analysis of video. The dataset [201] comprises a total of 20,000 frames from 160 outdoor videos. A portion of the videos have been sourced from YouTube, while the remainder were captured by the researchers during their urban excursions on foot or by vehicle. The duration of videos varies between 60 and 400 frames, while their resolution ranges from 320×480 to 600×800. The video content has partitioned into three main geometric classes: Sky, Support and Vertical. These classes also contain subclasses such as Buildings, Cars, Humans, Ground, Trains, Trees.

- **SIFT Flow**

The SIFT Flow dataset [202] is a subset of the LabelMe dataset [203]. This dataset comprises a total of 2688 images that have been fully annotated. Most of these images depict outdoor scenes, featuring various elements such as Bridge, Mountain, Road, Traffic light, Tower, Water. The 33 most prominent object categories have been identified based on the highest number of labeled pixels. Pixels that have not been assigned a label or have been labeled as a different object category are regarded as the 34th category, which is referred to as "unlabeled".

- **Stanford Background**

The Stanford Background dataset [204] comprises 715 outdoor scene images that have been selected from publicly available datasets, including Geometric Context [205], LabelMe [203], MSRC [206], and Pascal [207]. The criteria employed for image selection entailed a minimum resolution of 320 × 240 pixels, inclusion of at least one object in the foreground and proper positioning of the horizon within the image, regardless of its visibility. The annotations procured from Amazon Mechanical Turk exhibit a high level of quality. This dataset has been annotated with class segmentations, and it uses 8 classes: Building, Foreground Object, Grass, Mountain, Road, Sky, Tree and Water. The 'Foreground Object' classification encompasses all objects that are not classified under the initial seven categories.

Table 11. Summary of static scene parsing datasets.

Content	Dataset, Year	Number of Classes	Number of Images	Samples			Image Resolution
				Training	Validation	Testing	
Generic	ADE20K [189] (2017)	150	25k	20,210	2000	3352	2400 × 1800
	COCO Stuff [191] (2018)	171	164k	118k	5k	45k	variable
	Pascal VOC 2010 [207] (2010)	20	1928	771	289	868	500 × 400
	Pascal VOC 2012 [192] (2012)	21	4369	1464	1449	1456	variable
	Pascal Context [193] (2014)	59	10,103	4998	-	5105	variable

Table 11. *Cont.*

Content	Dataset, Year	Number of Classes	Number of Images	Samples			Image Resolution
				Training	Validation	Testing	
Indoor	NYUDv2 [194] (2012)	40	1449	795	-	654	480 × 640
	SUN RGBD [195] (2013)	37	10,335	5285	-	5050	variable
	BDD100K [196] (2020)	40	100k video frames	70k	10k	20k	1280 × 720
	CamVid [197] (2009)	32	701 video frames	367	100	233	960 × 720
	Cityscapes [198] (2016)	30	5000	2975	500	1525	2048 × 1024
	DTMR-DVR [199] (2020)	13	600 video frames	400	100	100	1280 × 960
	KITTI [200] (2012)	19	580	289	-	290	1226 × 370
Outdoor	GATECH [201] (2013)	84	20k video frames	13k	7k	7k	variable
	SIFT Flow [202] (2009)	33	2688	2488	-	200	256 × 256
	Stanford Background [204] (2009)	8	725	572	-	143	320 × 240

7. Evaluation

In this article, the studies are summarized by categorizing them according to their methods. These summaries also include summary tables containing the accuracy rates of the "*Mean Intersection over Union (MIoU)*" criteria type.

According to Table 12, the FCN-based methods have produced comparable outcomes for Pascal VOC test set. For all that, it can be observed that the best FCN-based approach is the "multi-scale and pyramid" approach because the use of multiple scales or a pyramid structure can make the model more robust to changes in the size or scale of objects in the image. This is particularly important in semantic segmentation tasks, where objects of interest can vary greatly in size. Moreover, thanks to the utilization of multi-scale or multi-level pyramidal image processing, the model can extract a more comprehensive array of features. These features can capture fine details, thereby augmenting the overall efficacy of the segmentation.

In addition, region-based methods have attained notably inferior rates in comparison to FCN-based methods. The main reasons for this are: Firstly, FCNs are trained in an end-to-end manner, enabling them to learn the capacity to map raw pixel values to semantic labels through a unified model. This approach has the potential to yield more precise segmentations in contrast to region-proposal techniques that split the problem into distinct stages (e.g., initial region identification followed by classification) due to the possibility of error propagation throughout the pipeline. Secondly, FCNs enables them to generate output masks of the same size as the input images, regardless of their dimensions. In contrast to region-based techniques, which frequently necessitate partitioning the input image into patches or resizing it to a predetermined size, this approach differs. Thirdly, FCNs are pixel-based methods, so they can predict the class of each pixel in the image. Region-based methods typically unite neighboring pixels into larger regions and may not accurately capture fine-grained details.

Table 12. Comparison of pioneering methods on Pascal VOC test set. All methods in the table use VGG-16 as the backbone network.

Method	Model Structure	Accuracy mIoU (%)
FCN-8s [27]	FCN	62.2
Fast R-CNN [18]	Region proposal	65.7
DeconvNet [35]	Encoder-decoder	70.5 *
DeepLab-v1 [31]	Dilated convolution	70.3 *
DilatedNet [63]	Dilated convolution	67.6 *
ParseNet [75]	Feature fusion	69.8 *
FeatMap-Net [82]	Multi-scale and pyramid	75.3 *
CRFasRNN [33]	RNN	72.0 *
BoxSup [119]	Weakly (box)	64.6 *
SEC [129]	Weakly (image)	51.7 *
Point-level [161]	Weakly (point)	42.9 *
ScribbleSup [159]	Weakly (scribble)	64.7 *

(Results marked with * have included the CRF method in their work).

By applying the CRF method to the related approaches as a post-processing step, the accuracy rate has generally increased by around 2.5-3.0%. According to some researchers, CRF improves object boundaries and contributes to increased performance, but it is too long to process and computationally expensive. According to [80], "Domain Transform (DT) filtering" is many times faster than CRF extraction.

Table 13 shows the comparison of VGG-16 and ResNet-101, which are the most utilized backbone convolution models in this field. According to this comparison, it has been observed that the use of ResNet-101 based on the network gives more successful results than the use of VGG-16.

Table 13. Comparison of VGG-16 and ResNet-101 backbone on PASCAL VOC test set for weakly supervised (image-level) methods.

Method	Backbone Accuracy mIoU (%)	
	VGG-16	ResNet-101
SeeNet [134]	60.7	62.8
DSRG [148]	60.4	63.2
Ficklenet [136]	61.9	65.3
MCOF [66]	57.6	61.2
OAA [149]	62.8	66.4
ICD [152]	63.9	68.0

Accordingly, the use of ResNet-101 as a backbone network provides an approximately 3.0% performance increase compared to the use of VGG-16.

Table 14 indicates that the FCN-based methods have produced comparable outcomes for Cityscapes dataset. Nevertheless, as can be seen in both Tables 12 and 14, weakly supervised based methods have attained notably inferior rates in comparison to FCN-based methods. The main two reasons for this are: Firstly, FCNs can accurately predict the classification of each individual pixel, thereby producing high-resolution segmentation masks. In contrast, weakly supervised techniques typically rely on labels that are less granular in nature, such as labels at the image level, which may constrain their level of

accuracy. Secondly, FCNs are typically trained using fully annotated data, resulting in improved segmentation accuracy. In contrast, weakly supervised techniques may not effectively leverage the complete set of available annotation data, which could result in lower performance.

Table 14. Comparison of state-of-the-art methods on Cityscapes dataset.

Method	Model Structure	Backbone	Accuracy mIoU (%)
CANet [45]	Encoder-decoder	ResNet-101	81.8
DeepLab-v3+ [48]	Encoder-decoder + Dilated convolutions	ResNet-101	82.1
SA-FFNet [77]	Feature fusion	ResNet-18	75.0
SpyGR [92]	Multi-scale and pyramid	ResNet-101	81.6
CGBNet [117]	Methods using RNN	ResNet-101	81.2
Weakly super. Two-stream Network [158]	Weakly supervised (Image-level labels)	VGG-16	47.2
DAFormer [180]	UDA	SegFormer	68.3

In addition, the "DAFormer" model, whose success rate is not very remarkable in Table 14, is a UDA method. UDA methods that have surfaced in recent times have reached a level of competitiveness with FCN-based methods. UDA techniques obviate the need for procuring and annotating copious quantities of labeled data in the target domain. In comparison to the FCN method, which necessitates a substantial quantity of annotated and labeled data, this approach presents noteworthy benefits.

Table 15 displays the mean accuracy values of all the studies analyzed in this research, across all the datasets considered. Upon the table, it can be observed that mean performance rate of the datasets just Pascal VOC and Cityscapes have surpassed 70%. The Pascal VOC and Cityscapes datasets are widely recognized as standard benchmark datasets, particularly for the task of semantic segmentation. This facilitates the comparative analysis of diverse methods and models. We think this is because the background complexity of Pascal VOC dataset is less than the others because it includes twenty foreground object classes and one background class. Moreover, it also includes high-quality images and labels. This facilitates the learning process of the model by utilizing data with lower levels of noise, leading to improved performance outcomes. The Cityscapes dataset is composed of images with high resolution and pixel-level annotations of high quality. This situation can enhance the robustness of the models trained on the dataset.

As per the data presented in the table, it can be observed that the COCO Stuff dataset exhibited the least performance. The COCO-Stuff dataset contains 80 'thing' and '91' 'stuff' classes. Thing classes include objects that are usually countable, discrete and have a well-defined structure. For example, 'person', 'car', 'bus', 'bird' 'train', etc. Stuff classes include regions in an image that do not have a well-defined structure, are more ambiguous, and are not countable. For example, 'dirt', 'clouds', 'grass', 'sky', 'water', etc. The current models have not attained the intended level of performance, as they must also possess the capability to recognize and segment regions that are less precisely defined. Consequently, it was contended that the models ought to incorporate the extra layer of complexity.

According to another poor-performing ADE20K dataset, for segmenting to be efficacious, not only objects such as doors and glasses, but also their parts, such as door handles and glass handles, must be recognized and localized. The dataset has 150 classes that include object, object part and stuff. As an illustration, a car is an 'object', a wheel that is a part of a car is an 'object part', and a rim that is a part of a wheel is a 'stuff'. Considering this scenario, it undoubtedly results in segmentation of higher quality. Nevertheless, this shows the high annotation complexity of ADE20K dataset.

Table 15. The performance comparison of the datasets.

Dataset	Average Accuracy mIoU (%)
ADE20K	44.8
COCO Stuff	36.2
Pascal VOC 2012	74.0
Pascal Context	48.2
NYUDv2	44.0
SUN RGBD	38.1
BDD100K	66.5
CamVid	63.7
Cityscapes	71.9
DTMR-DVR	60.4
KITTI	62.9
GATECH	53.5
SIFT Flow	45.9
Stanford Background	64.8

Another of the lowest-performing datasets is SUN RGBD. The task of performing semantic segmentation on 3D images presents greater challenges compared to its 2D counterpart. Therefore, we consider the performance of the SUN RGBD, which consists of 3D data, to be poor. Additionally, Sun RGBD is known to have noisy labels at the object level. This situation is likely to decrease the precision rate of the models.

Accuracy

The most popular performance evaluation metrics used for semantic segmentation are the ones mentioned below. In this article, the approaches discussed within the scope of semantic segmentation have been compared by taking mostly "*mIoU*" results from the main evaluation metrics. "*mIoU*" is the most widely used metric in semantic segmentation as it penalizes both over- and under-segmentation. In addition, "PA_{CC}" and "MA_{CC}" metrics are rarely used. These metrics can be obtained by the equations as follows:

Let $k+1$ is the number of semantic classes, p_{ii} is the number of correctly classified pixels, $\sum_{j=0}^{k} p_{ij}$ is the total number of pixels in class i, p_{ij} is the number of pixels which belong to class i but predicted to class j. p_{ji} is the number of pixels which belong to class j but predicted to class i.

Mean Intersection over Union (MIoU): The ratio of accurately classified pixels in a class over the union set of pixels predicted to this class and ground truth. Next, the average of all classes is calculated.

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

Pixel Accuracy (PA_{CC}): The ratio of the number of accurately classified pixels to their total number.

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

Mean Accuracy (MA_{CC}): The ratio of accurate pixels is calculated for each class. Next, the average of all classes is calculated.

$$MA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{3}$$

8. Discussion and Future Directions

According to the data observed from this study, as methods have developed and new methods have emerged, the success rate has increased. However, in general, an accuracy rate above 90% (MIoU) has not been reached yet. The reasons for this are the constraints and difficulties that exist in this field. It can be expected that the success rate will increase with the clear identification of these problems, the development of solutions, and the development of a suitable model. According to the literature reviewed in this study, the main challenges, limitations and future directions in the field of semantic segmentation are:

(1) Loss resolution: CNN reduces the resolution of the image it receives due to its structure. Lower-resolution images may not contain enough fine detail for accurate segmentation. For networks to maintain high resolution information through layers, the 'Encoder-decoder' structure has been introduced (Section 2.2.1). However, a significant amount of spatial information is lost during the down-sampling process in the encoder.

To tackle this issue, the 'skip connections' technique which was proposed by [27] has been implemented. The fundamental idea behind this technique is to assist in the transfer of more complex details and spatial information from the encoder to the corresponding decoder layers, thereby enhancing the precision of localization and segmentation. The "U-Net" architecture [208] and the "DeepLab" model series [31,48,56,61,98] are prominent examples of this technique. Skip connections are utilized in these networks to connect the layers of the encoder and decoder, thereby integrating low-level feature maps with high-level ones. By utilizing features at multiple levels of abstraction, the network can generate more accurate segmentation predictions, thereby enhancing precision.

Despite the important contributions of the 'skip connections', it has challenges and limitations. First, this technique makes neural networks more complicated. The complexity of a system could be a problem if there are not enough resources or if there are applications that need to be processed in real time. Secondly, determining the suitable architecture can pose a challenge, given the complexity of deciding the optimal placement and quantity of skip connections. Making an inappropriate choice can result in suboptimal outcomes. Lastly, the utilization of 'skip connections' in models enables the acquisition of complex features while increasing the risk of overfitting, particularly in scenarios involving limited datasets.

Notwithstanding these limitations, skip connections persist as a pivotal component of this field, and current investigations are concentrated on enhancing their efficacy. There are a few potential future approaches that can be taken to address the restrictions that have been outlined above. To begin, the computing requirements might potentially be reduced by employing methods that generate sparser connections as an alternative to having each layer be dependent on all the other layers. Creating a model where the network "learns" which connections to employ during training can be quite effective as well. Second, it can be solved by automating the process of determining the optimal number and positioning of skip connections.

(2) Capturing boundary: The accurate identification of object boundaries is a crucial and formidable task in semantic segmentation, as it directly impacts the proper classification of individual pixels. The delimitation of boundaries in images can be facilitated using simple or specific objects. However, accurately defining boundaries can be difficult, particularly in complex images such as a street scene. Moreover, the depicted objects in the images may display a variety of sizes and differing levels of proximity to the camera. So, the task of delineating object boundaries can be challenging due to the significant variations in the apparent size and shape of similar objects. Some of the future directions in to overcome this problem:

At first, the post-processing and refinement modules can be used to capture more accurate boundaries and enhance fine-grained details. Algorithms such as CRF, MRF and Domain Transform have been utilized as post-processing techniques to enhance boundaries in certain architectures. Nevertheless, these techniques have certain challenges and limitations. The utilization of CRF and MRF results in a significant increase in the computa-

tional complexity of the model. This situation makes these algorithms less applicable in applications with limited computational capabilities or in real-time applications.

Next, the optimal performance of both CRF and MRF necessitates meticulous parameter tuning. The process can be laborious and necessitate a certain degree of proficiency in said models.

Afterwards, some research [31–33] has emphasized the integration of CRFs in the process of end-to-end training of deep neural networks. This has the potential to decrease the necessity for post-processing procedures and potentially result in enhanced performance.

According to [80], "Domain Transform" is many times faster than CRF extraction. Therefore, incorporating "Domain Transform", such as CRF, directly into the learning process may provide a direction for future research.

In short, in future studies, more research can be completed to come up with efficient approaches that reduce the cost of computation while still giving good performance.

(3) Generalization: Although a model may be trained to execute a specific task, its performance may not be consistent when applied to different datasets. For example, the "ENet" [46] study achieved an accuracy of 58.3% on the Cityscapes dataset while a significantly lower rate of 19.7% on SUN RGBD dataset. Many such examples can be encountered in the tables. The main reasons for this are:

Large dataset: In the field of deep learning, a significant amount of data is frequently associated with improved performance. Hence, a larger dataset has the potential to yield more successful results. In our opinion, contributing to the development of existing datasets for certain problems in this field should be as important as solving a problem. Some future directions for this problem:

First, data augmentation techniques can be used to enlarge the dataset. Data augmentation is a technique employed in deep learning with the aim of enhancing the scope and magnitude of the training data, without the need for additional data collection. It has many applications such as Cropping, Flipping and Rotation, Mixup, Cutmix [209]. Second, synthetic data, possessing predetermined ground truth labels, can be employed for the purpose of training. The advancement of domain adaptation techniques holds potential for effectively addressing the gap in question.

Nevertheless, it should be noted that these techniques are not suitable for all data types and cannot replace real-world data.

Scale Variation: The issue of scale variation poses a considerable obstacle in the context of semantic segmentation, given that the objects depicted in images may exhibit a range of sizes or be situated at varying distances from the camera. Thus, the datasets with objects that demonstrate significant variations can make it difficult for the model to correctly segment all objects. Future works can focus these directions:

First, the "multi-scale and pyramid networks" employ the input image at varying scales or resolutions and the outcomes are subsequently integrated. This method facilitates the neural network's ability to identify objects of varying sizes. Future research endeavors may concentrate on augmenting these structures to achieve superior multi-scale processing capabilities. Second, future work could include making networks whose receptive fields can change size based on the size of the object. This could help the model handle things of different sizes better.

Background Complexity: When there are a lot of mixed textures, colors, and overlapping things in the background of an image, the model can become confused. The difficulty is to distinguish the target object from these background components. Therefore, the success of two datasets can be different if one has simpler background images and the other has difficult and complex data. The future directions for dealing with these problems:

Firstly, the development of loss functions that penalize misclassification of difficult or ambiguous examples more than straightforward ones could result in models that are more adept at handling complex backgrounds. Secondly, the inclusion of supplementary data beyond RGB, such as depth data, could help to facilitate the discrimination of foreground entities from complex backgrounds.

Annotation Quality: The effectiveness of semantic segmentation is highly dependent on the accuracy of the annotated data. So, in-depth annotations with extensive labels and the level of detail required for pixel-level annotations have a significant impact. Creating more effective and intuitive tools for data annotation can aid in enhancing the speed and accuracy of the annotation process, and thus the quality of the annotations.

High resolution: Fine details are evident in high resolution images, while detailed structures of the object are lost in low resolution images. In this case, higher success can be achieved in datasets containing high-resolution images than in low-resolution datasets. For this reason, conducting tests on datasets that comprise high-resolution images would generate more robust outcomes for our forthcoming models.

Imbalanced classes: This common problem emerges when some classes within the dataset are represented significantly more frequently than others. Thence, the accuracy of the model may decrease in the dataset with underrepresented classes. Here are several solution strategies:

Initially, incorporating hierarchical relationships between classes into the learning process can enhance the model's capacity to generalize from majority to minority classes. Afterwards, changing the loss function to assign greater significance to minority classes can potentially alleviate the issue of imbalanced class distribution. More effective class balanced loss functions can be developed.

Consequently, future works should focus on developing models that exhibit strong generalization capabilities across diverse datasets, considering these reasons.

(4) Lighting and weather conditions: Real-world data is complex and variable because it includes good weather conditions as well as bad weather (e.g., rainy, snowy etc.) [210] and bad lighting conditions (e.g., foggy, dark, night etc.) [211]. These adverse conditions affect the appearance of objects in the scene, negatively affecting the accuracy of semantic segmentation. However, this is not the case in the real world, and it is necessary to verify the reliability and robustness of a trained model under different environmental conditions. There are a few possible ways to deal with this problem in the future.

Firstly, "Unsupervised domain adaptation (UDA)" techniques (mentioned in Section 4.2) can be employed to mitigate the distributional discrepancy that arises from variations in weather and lighting conditions.

Secondly, "Meta-learning" strategies [212] involve training models not only to perform the task, but also to rapidly adapt to new circumstances. This could be especially beneficial for adapting to various environmental conditions.

Lastly, the various data types, such as infrared or lidar data, exhibit a relatively lower susceptibility to lighting and weather conditions as compared to regular images. Models that use this information in addition to standard images can be more resistant to environmental changes.

(5) Annotation difficulty: Semantic segmentation requires many detailed pixel-level annotations for training the model. As mentioned in this article, the main challenge in preparing a data set is the labeling part rather than collecting the data. This process is very time consuming and costly. Here are a few possible ways to tackle this problem in the future:

We think primarily that weaky supervised and unsupervised methods with little or no need for manual labeling should be focused on at least as much as fully supervised methods. Thus, significantly larger datasets can be efficiently generated with minimal reliance on human resources.

Next, the 'UDA' technique has focused on improving the ability of fully supervised learning models to adapt to the weakly supervised learning domain. It has not achieved as high a success rate as fully supervised learning methods, but despite being an unsupervised method, it presents a promising avenue for further investigation.

At last, the approaches using 'active learning' technique [213] can be increased. In this technique, the model is first trained on a limited dataset and then employed to estimate the classifications of unlabeled data. Subsequently, the model's instances of low confidence are

manually annotated and incorporated into the training set, thereby reducing the amount of manual labeling required. These developments can drastically reduce the dependence on high-quality, fully annotated datasets.

(6) Real-time processing: Depending on the semantic segmentation applications, such as autonomous driving, robotics and video surveillance, real-time processing may be required. However, it can be difficult to simultaneously achieve real-time performance and high accuracy due to the computational complexity of semantic segmentation models. Some popular methods "ENet" [46], "ICNet [214]", "LASNet" [215], "SFANet" [44], "ShelfNet" [216] and "BiSeNet" [217] have applied semantic segmentation methods in real time. For example, ShelfNet and BiSeNet have achieved comparable segmentation accuracy to state-of-the-art off-line models with a four to five times faster inference speed. The real-time performances of these frameworks on the Cityscapes dataset are 74.8% and 74.7% mIoU, respectively, while their non-real time performances are 79.0 and 78.9, respectively. According to the results, the real-time performance of these models is about 5% lower than the non-real-time performance. To improve performance, speed up processing time, and extend the applicability of semantic segmentation models to real-time systems, we may focus on the following directions in the future:

First, the 'approximate computing' technique [218,219], which aims to trade a balance between computational accuracy and speed of computation, can be used. While some applications necessitate high accuracy, many machine learning tasks can tolerate a certain amount of error or approximation without considerably affecting the overall quality of the result, according to the fundamental principle of this technique. Some ways in which approximate computing can be applied are common model compression techniques such as quantization, pruning and distillation [220]. Additionally, techniques such as 'skip connection' and 'early stop' may allow for some models to predict early before the entire model is calculated based on the outputs of the first layers. Thus, computational complexity can be reduced thanks to these paradigms in real-time applications where speed and efficiency are significantly more essential.

Second, the 'dynamic computation' technique [221] can be used to increase performance and reduce the amount of computation. Dynamic computation aims to apply computational resources selectively during model execution, as opposed to uniformly applying the same computations to the entire input. One of the most popular techniques involving dynamic computation is the 'attention mechanism' [222]. The 'attention mechanisms' can enable the neural network to selectively attend to distinct regions of the input image during various stages of computation. As an instance, a neural network could acquire the ability to prioritize the salient objects within an image while minimizing its attention towards the background.

Lastly, enhancing the runtime performance of models can be achieved by optimizing their performance for hardware platforms, such as GPUs and TPUs. Moreover, the implementation of techniques to distribute the processing workload among multiple CPUs, GPUs, or devices could facilitate the real-time processing of larger models.

The mentioned techniques represent promising directions in the field of real-time semantic partitioning. Nonetheless, it is crucial to consider the requirements of each application when determining whether to employ these methods. In addition, there are few real-time works in the literature. Therefore, future studies may focus on increasing the number of these works and improving real-time performance by considering the techniques mentioned.

(7) 3D semantic segmentation: The task of 3D semantic segmentation holds importance in various applications such as autonomous driving, robotics, and augmented reality. The depth dimension, which is added to the height and width dimensions, is crucial for comprehending the scene. It entails labeling every point in a 3D point cloud or voxel grid with a semantic label. Some of the challenges, limitations and future directions in this area are mentioned below:

The first of these, volumetric nature of 3D data, makes it larger and more complex than 2D image data. This situation requires more computational resources for processing. Therefore, real-time application of 3D semantic segmentation is a challenging task. Developing novel network architectures designed specifically for 3D data can aid in reducing computational complexity. The popular new models such as "PointNet" [223], "PointGrid" [224], "RandLA-Net" [225], "RangeNet" [226], "SEGCloud" [227], "MFFRand" [228], "LESS" [229] and "SQN" [230] developed in this field have tried to process more data and effectively construct 3D spatial relationships between points or voxels. However, we think that there is not enough work yet. In the future, 3D semantic segmentation models should be able to compete with 2D semantic segmentation models.

Another of these, the process of annotating 3D data, is costly and time-consuming, resulting in a lack of large, high-quality, labeled datasets for training models. The utilization of the 'active learning' technique [213] can optimize the utilization of restricted annotated data. The process of active learning involves training a model iteratively and then selecting the most informative examples for annotation using the model so that it can achieve more accuracy with fewer training labels.

Last, the recognition of objects in 3D data can pose a challenge for models due to variations in scale resulting from differences in distance. The development of scale-invariant models can be a potential solution to mitigate the problem of scale variation.

Finally, the progress in these domains may result in significant enhancements in semantic segmentation, facilitating the development of more precise and effective models that can process complex real-world images.

9. Conclusions

The paper offers an organized examination of roughly 150 CNN methods for semantic segmentation that have been developed over the past ten years. In addition, it has examined 15 popular datasets that comprise general, indoor, outdoor and street scenes. Moreover, this paper has referenced various contemporary methodologies, including SAM, UDA and traditional post-processing algorithms such as CRF, MRF and Random Walker. Furthermore, it has exhibited and discussed the outcomes of the frameworks and datasets in a tabulated format. The article ultimately addresses the main challenges and possible solutions and underlines some future research directions in semantic segmentation tasks.

In summary, it is necessary to enhance semantic segmentation models to effectively address real-world challenges, despite the existence of several successful models. In the coming years, it is possible that novel research projects will be proposed which may introduce innovative approaches and methodologies related to semantic segmentation. The review paper we are producing will serve as a fundamental reference for understanding forthcoming research work.

Author Contributions: Conceptualization, B.E.S., M.S.G. and G.E.B.; Formal Analysis, B.E.S. and M.S.G.; Investigation, B.E.S., F.E., T.A. and K.A.; Writing—Original Draft Preparation, B.E.S., F.E., T.A. and K.A.; Writing—Review and Editing, T.A. and K.A.; Visualization, B.E.S.; Supervision, M.S.G. and G.E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zheng, W.; Liu, X.; Ni, X.; Yin, L.; Yang, B. Improving visual reasoning through semantic representation. *IEEE Access* **2021**, *9*, 91476–91486. [CrossRef]
2. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995.
3. Lateef, F.; Ruichek, Y. Survey on semantic segmentation using deep learning techniques. *Neurocomputing* **2019**, *338*, 321–348. [CrossRef]
4. Hao, S.; Zhou, Y.; Guo, Y. A brief survey on semantic segmentation with deep learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]
5. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A survey on deep learning techniques for image and video semantic segmentation. *Appl. Soft Comput.* **2018**, *70*, 41–65. [CrossRef]
6. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]
7. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 87–93. [CrossRef]
8. Zhang, M.; Zhou, Y.; Zhao, J.; Man, Y.; Liu, B.; Yao, R. A survey of semi-and weakly supervised semantic segmentation of images. *Artif. Intell. Rev.* **2020**, *53*, 4259–4288. [CrossRef]
9. Ulku, I.; Akagündüz, E. A survey on deep learning-based architectures for semantic segmentation on 2d images. *Appl. Artif. Intell.* **2022**, *36*, 2032924. [CrossRef]
10. Alokasi, H.; Ahmad, M.B. Deep learning-based frameworks for semantic segmentation of road scenes. *Electronics* **2022**, *11*, 1884. [CrossRef]
11. Yu, Y.; Wang, C.; Fu, Q.; Kou, R.; Huang, F.; Yang, B.; Yang, T.; Gao, M. Techniques and Challenges of Image Segmentation: A Review. *Electronics* **2023**, *12*, 1199. [CrossRef]
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
13. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
15. 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; pp. 1–9.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
18. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
20. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
21. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
22. Dai, J.; He, K.; Sun, J. Convolutional feature masking for joint object and stuff segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3992–4000.
23. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous detection and segmentation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 297–312.
24. Arbeláez, P.; Pont-Tuset, J.; Barron, J.T.; Marques, F.; Malik, J. Multiscale combinatorial grouping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 328–335.
25. Caesar, H.; Uijlings, J.; Ferrari, V. Region-based semantic segmentation with end-to-end training. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 381–397.
26. Shen, D.; Ji, Y.; Li, P.; Wang, Y.; Lin, D. Ranet: Region attention network for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 13927–13938.
27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
28. Elman, J.L. Finding structure in time. *Cognit. Sci.* **1990**, *14*, 179–211. [CrossRef]
29. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

30. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 2001. Available online: https://repository.upenn.edu/cis_papers/159/?ref=https://githubhelp.com (accessed on 5 April 2023).
31. 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.
32. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–14 December 2011; Volume 24.
33. Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P.H. Conditional random fields as recurrent neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1529–1537.
34. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
35. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
36. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
37. Kendall, A.; Badrinarayanan, V.; Cipolla, R. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv* **2015**, arXiv:1511.02680.
38. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
39. Gal, Y.; Ghahramani, Z. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv* **2015**, arXiv:1506.02158.
40. Treml, M.; Arjona-Medina, J.; Unterthiner, T.; Durgesh, R.; Friedmann, F.; Schuberth, P.; Mayr, A.; Heusel, M.; Hofmarcher, M.; Widrich, M. Speeding Up Semantic Segmentation for Autonomous Driving. 2016. Available online: https://openreview.net/forum?id=S1uHiFyyg (accessed on 5 April 2023).
41. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1925–1934.
42. Fourure, D.; Emonet, R.; Fromont, E.; Muselet, D.; Tremeau, A.; Wolf, C. Residual conv-deconv grid network for semantic segmentation. *arXiv* **2017**, arXiv:1707.07958.
43. Li, Q.; Wang, H.; Li, B.-Y.; Yanghua, T.; Li, J. IIE-SegNet: Deep semantic segmentation network with enhanced boundary based on image information entropy. *IEEE Access* **2021**, *9*, 40612–40622. [CrossRef]
44. Weng, X.; Yan, Y.; Chen, S.; Xue, J.-H.; Wang, H. Stage-aware feature alignment network for real-time semantic segmentation of street scenes. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4444–4459. [CrossRef]
45. Tang, Q.; Liu, F.; Zhang, T.; Jiang, J.; Zhang, Y.; Zhu, B.; Tang, X. Compensating for Local Ambiguity With Encoder-Decoder in Urban Scene Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 19224–19235. [CrossRef]
46. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
47. Islam, M.A.; Naha, S.; Rochan, M.; Bruce, N.; Wang, Y. Label refinement network for coarse-to-fine semantic segmentation. *arXiv* **2017**, arXiv:1703.00551.
48. 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 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
49. Amirul Islam, M.; Rochan, M.; Naha, S.; Bruce, N.D.; Wang, Y. Gated Feedback Refinement Network for Coarse-to-Fine Dense Semantic Image Labeling. *arXiv* **2018**, arXiv:1806.11266.
50. Bilinski, P.; Prisacariu, V. Dense decoder shortcut connections for single-pass semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6596–6605.
51. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1492–1500.
52. Fu, J.; Liu, J.; Wang, Y.; Zhou, J.; Wang, C.; Lu, H. Stacked deconvolutional network for semantic segmentation. *IEEE Trans. Image Process.* **2019**. [CrossRef]
53. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4700–4708.
54. Li, J.; Yu, J.; Yang, D.; Tian, W.; Zhao, L.; Hu, J. A Novel Semantic Segmentation Algorithm Using a Hierarchical Adjacency Dependent Network. *IEEE Access* **2019**, *7*, 150444–150452. [CrossRef]
55. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1251–1258.
56. 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]

57. Lu, Y.; Liu, H. Semantic segmentation with step-by-step upsampling of the fusion context. In Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 28–30 June 2021; pp. 156–161.
58. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7262–7272.
59. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
60. Jiang, D.; Qu, H.; Zhao, J.; Zhao, J.; Liang, W. Multi-level graph convolutional recurrent neural network for semantic image segmentation. *Telecommun. Syst.* **2021**, *77*, 563–576. [CrossRef]
61. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
62. Wu, L.; Xiao, J.; Zhang, Z. Improved Lightweight DeepLabv3+ Algorithm Based on Attention Mechanism. In Proceedings of the 2022 14th International Conference on Advanced Computational Intelligence (ICACI), Wuhan, China, 15–17 July 2022; pp. 314–319.
63. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
64. Yu, A.; Palefsky-Smith, R.; Bedi, R. Deep reinforcement learning for simulated autonomous vehicle control. *Course Proj. Rep. Winter* **2016**, *2016*, 1–7.
65. Lv, L.; Li, X.; Jin, J.; Li, X. Image semantic segmentation method based on atrous algorithm and convolution CRF. In Proceedings of the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 19–20 October 2019; pp. 160–165.
66. Wang, X.; You, S.; Li, X.; Ma, H. Weakly-supervised semantic segmentation by iteratively mining common object features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1354–1362.
67. Jin, R.; Yu, T.; Han, X.; Liu, Y. The Segmentation of Road Scenes Based on Improved ESPNet Model. *Secur. Commun. Netw.* **2021**, *2021*, 1681952. [CrossRef]
68. Jiang, J.; Zhang, Z.; Huang, Y.; Zheng, L. Incorporating depth into both cnn and crf for indoor semantic segmentation. In Proceedings of the 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 24–26 November 2017; pp. 525–530.
69. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
70. Zhong, M.; Verma, B.; Affum, J. Multi-Receptive Atrous Convolutional Network for Semantic Segmentation. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
71. Zhao, L.; Wang, Y.; Duan, Z.; Chen, D.; Liu, S. Multi-Source Fusion Image Semantic Segmentation Model of Generative Adversarial Networks Based on FCN. *IEEE Access* **2021**, *9*, 101985–101993. [CrossRef]
72. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
73. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
74. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–284.
75. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
76. Park, S.-J.; Hong, K.-S.; Lee, S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4980–4989.
77. Sun, J.; Li, Y. Multi-feature fusion network for road scene semantic segmentation. *Comput. Electr. Eng.* **2021**, *92*, 107155. [CrossRef]
78. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1915–1929. [CrossRef]
79. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
80. Chen, L.-C.; Barron, J.T.; Papandreou, G.; Murphy, K.; Yuille, A.L. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4545–4554.
81. Mostajabi, M.; Yadollahpour, P.; Shakhnarovich, G. Feedforward semantic segmentation with zoom-out features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3376–3385.
82. Lin, G.; Shen, C.; Van Den Hengel, A.; Reid, I. Efficient piecewise training of deep structured models for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3194–3203.

83. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; pp. 2169–2178.
84. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2393–2402.
85. Zheng, W.; Liu, X.; Yin, L. Research on image classification method based on improved multi-scale relational network. *PeerJ Comput. Sci.* **2021**, *7*, e613. [CrossRef]
86. Zheng, W.; Tian, X.; Yang, B.; Liu, S.; Ding, Y.; Tian, J.; Yin, L. A few shot classification methods based on multiscale relational networks. *Appl. Sci.* **2022**, *12*, 4059. [CrossRef]
87. Lu, S.; Ding, Y.; Liu, M.; Yin, Z.; Yin, L.; Zheng, W. Multiscale feature extraction and fusion of image and text in VQA. *Int. J. Comput. Intell. Syst.* **2023**, *16*, 54. [CrossRef]
88. Adelson, E.H.; Anderson, C.H.; Bergen, J.R.; Burt, P.J.; Ogden, J.M. Pyramid methods in image processing. *RCA Eng.* **1984**, *29*, 33–41.
89. 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]
90. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 2881–2890.
91. Zhou, Y.; Sun, X.; Zha, Z.-J.; Zeng, W. Context-reinforced semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4046–4055.
92. Li, X.; Yang, Y.; Zhao, Q.; Shen, T.; Lin, Z.; Liu, H. Spatial pyramid based graph reasoning for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8950–8959.
93. Zheng, W.; Yin, L.; Chen, X.; Ma, Z.; Liu, S.; Yang, B. Knowledge base graph embedding module design for Visual question answering model. *Pattern Recognit.* **2021**, *120*, 108153. [CrossRef]
94. Grangier, D.; Bottou, L.; Collobert, R. Deep convolutional networks for scene parsing. In Proceedings of the ICML 2009 Deep Learning Workshop, Montreal, QC, Canada, 14–18 June 2009.
95. Ghiasi, G.; Fowlkes, C.C. Laplacian pyramid reconstruction and refinement for semantic segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 519–534.
96. Sharma, A.; Tuzel, O.; Liu, M.-Y. Recursive context propagation network for semantic scene labeling. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27.
97. Sharma, A.; Tuzel, O.; Jacobs, D.W. Deep hierarchical parsing for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 530–538.
98. Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.
99. Raj, A.; Maturana, D.; Scherer, S. *Multi-Scale Convolutional Architecture for Semantic Segmentation*; Tech. Rep. CMU-RITR-15-21; Robotics Institute, Carnegie Mellon University: Pittsburgh, PA, USA, 2015.
100. Roy, A.; Todorovic, S. A multi-scale cnn for affordance segmentation in rgb images. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 186–201.
101. Chandra, S.; Kokkinos, I. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 402–418.
102. Lin, G.; Shen, C.; Van Den Hengel, A.; Reid, I. Exploring context with deep structured models for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1352–1366. [CrossRef] [PubMed]
103. Li, X.; Liu, Z.; Luo, P.; Change Loy, C.; Tang, X. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3193–3202.
104. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA USA, 4–9 February 2017.
105. Ji, J.; Lu, X.; Luo, M.; Yin, M.; Miao, Q.; Liu, X. Parallel fully convolutional network for semantic segmentation. *IEEE Access* **2020**, *9*, 673–682. [CrossRef]
106. Shen, F.; Gan, R.; Yan, S.; Zeng, G. Semantic segmentation via structured patch prediction, context crf and guidance crf. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1953–1961.
107. Pinheiro, P.; Collobert, R. Recurrent convolutional neural networks for scene labeling. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
108. Byeon, W.; Breuel, T.M.; Raue, F.; Liwicki, M. Scene labeling with lstm recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3547–3555.

109. Visin, F.; Ciccone, M.; Romero, A.; Kastner, K.; Cho, K.; Bengio, Y.; Matteucci, M.; Courville, A. Reseg: A recurrent neural network-based model for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 27–30 June 2016; pp. 41–48.
110. Visin, F.; Kastner, K.; Cho, K.; Matteucci, M.; Courville, A.; Bengio, Y. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv* **2015**, arXiv:1505.00393.
111. Shuai, B.; Zuo, Z.; Wang, B.; Wang, G. Scene segmentation with dag-recurrent neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1480–1493. [CrossRef]
112. Arnab, A.; Jayasumana, S.; Zheng, S.; Torr, P.H. Higher order conditional random fields in deep neural networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 524–540.
113. Shuai, B.; Zuo, Z.; Wang, B.; Wang, G. Dag-recurrent neural networks for scene labeling. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016; pp. 3620–3629.
114. Fan, H.; Ling, H. Dense recurrent neural networks for scene labeling. *arXiv* **2018**, arXiv:1801.06831.
115. Fan, H.; Mei, X.; Prokhorov, D.; Ling, H. Multi-level contextual rnns with attention model for scene labeling. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 3475–3485. [CrossRef]
116. Zhang, Y.; Li, X.; Lin, M.; Chiu, B.; Zhao, M. Deep-recursive residual network for image semantic segmentation. *Neural Comput. Appl.* **2020**, *32*, 12935–12947. [CrossRef]
117. Ding, H.; Jiang, X.; Shuai, B.; Liu, A.Q.; Wang, G. Semantic segmentation with context encoding and multi-path decoding. *IEEE Trans. Image Process.* **2020**, *29*, 3520–3533. [CrossRef]
118. Xia, W.; Domokos, C.; Dong, J.; Cheong, L.-F.; Yan, S. Semantic segmentation without annotating segments. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2176–2183.
119. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1635–1643.
120. Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; Schiele, B. Simple does it: Weakly supervised instance and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 876–885.
121. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3136–3145.
122. Carreira, J.; Sminchisescu, C. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1312–1328. [CrossRef]
123. Xu, X.; Meng, F.; Li, H.; Wu, Q.; Ngan, K.N.; Chen, S. A new bounding box based pseudo annotation generation method for semantic segmentation. In Proceedings of the 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), Macau, China, 1–4 December 2020; pp. 100–103.
124. Oh, Y.; Kim, B.; Ham, B. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6913–6922.
125. Ma, T.; Wang, Q.; Zhang, H.; Zuo, W. Delving deeper into pixel prior for box-supervised semantic segmentation. *IEEE Trans. Image Process.* **2022**, *31*, 1406–1417. [CrossRef]
126. Maron, O.; Lozano-Pérez, T. A framework for multiple-instance learning. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 2–5 December 1997; Volume 10.
127. Pathak, D.; Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional multi-class multiple instance learning. *arXiv* **2014**, arXiv:1412.7144.
128. Pinheiro, P.O.; Collobert, R. From image-level to pixel-level labeling with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1713–1721.
129. Kolesnikov, A.; Lampert, C.H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 695–711.
130. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
131. Sun, K.; Shi, H.; Zhang, Z.; Huang, Y. Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 7283–7292.
132. Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; Yan, S. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1568–1576.
133. Li, K.; Wu, Z.; Peng, K.-C.; Ernst, J.; Fu, Y. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9215–9223.
134. Hou, Q.; Jiang, P.; Wei, Y.; Cheng, M.-M. Self-erasing network for integral object attention. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Volume 31.

135. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7268–7277.
136. Lee, J.; Kim, E.; Lee, S.; Lee, J.; Yoon, S. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5267–5276.
137. Zeng, Y.; Zhuge, Y.; Lu, H.; Zhang, L. Joint learning of saliency detection and weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 7223–7233.
138. Pathak, D.; Krahenbuhl, P.; Darrell, T. Constrained convolutional neural networks for weakly supervised segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 1796–1804.
139. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
140. Saleh, F.; Aliakbarian, M.S.; Salzmann, M.; Petersson, L.; Gould, S.; Alvarez, J.M. Built-in foreground/background prior for weakly-supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 413–432.
141. Saleh, F.S.; Aliakbarian, M.S.; Salzmann, M.; Petersson, L.; Alvarez, J.M.; Gould, S. Incorporating network built-in priors in weakly-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1382–1396. [CrossRef] [PubMed]
142. Papandreou, G.; Chen, L.-C.; Murphy, K.P.; Yuille, A.L. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1742–1750.
143. Qi, X.; Liu, Z.; Shi, J.; Zhao, H.; Jia, J. Augmented feedback in semantic segmentation under image level supervision. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 90–105.
144. Wei, Y.; Liang, X.; Chen, Y.; Jie, Z.; Xiao, Y.; Zhao, Y.; Yan, S. Learning to segment with image-level annotations. *Pattern Recognit.* **2016**, *59*, 234–244. [CrossRef]
145. Wei, Y.; Xia, W.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. CNN: Single-label to multi-label. *arXiv* **2014**, arXiv:1406.5726.
146. Ahn, J.; Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4981–4990.
147. Wu, Z.; Shen, C.; Van Den Hengel, A. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]
148. Huang, Z.; Wang, X.; Wang, J.; Liu, W.; Wang, J. Weakly-supervised semantic segmentation network with deep seeded region growing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7014–7023.
149. Jiang, P.-T.; Hou, Q.; Cao, Y.; Cheng, M.-M.; Wei, Y.; Xiong, H.-K. Integral object mining via online attention accumulation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2070–2079.
150. Jo, S.; Yu, I.-J. Puzzle-cam: Improved localization via matching partial and full features. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 639–643.
151. Chang, R.-H.; Guo, J.-M.; Seshathiri, S. Saliency Guidance and Expansion Suppression on PuzzleCAM for Weakly Supervised Semantic Segmentation. *Electronics* **2022**, *11*, 4068. [CrossRef]
152. Fan, J.; Zhang, Z.; Song, C.; Tan, T. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June 2020; pp. 4283–4292.
153. Shimoda, W.; Yanai, K. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 218–234.
154. Hong, S.; Yeo, D.; Kwak, S.; Lee, H.; Han, B. Weakly supervised semantic segmentation using web-crawled videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7322–7330.
155. Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.-M.; Feng, J.; Zhao, Y.; Yan, S. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 2314–2320. [CrossRef]
156. Jin, B.; Ortiz Segovia, M.V.; Susstrunk, S. Webly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3626–3635.
157. Luo, P.; Wang, G.; Lin, L.; Wang, X. Deep dual learning for semantic image segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2718–2726.
158. Saleh, F.; Aliakbarian, M.S.; Salzmann, M.; Petersson, L.; Alvarez, J.M. Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2125–2135.

159. Lin, D.; Dai, J.; Jia, J.; He, K.; Sun, J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3159–3167.
160. Vernaza, P.; Chandraker, M. Learning random-walk label propagation for weakly-supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7158–7166.
161. Bearman, A.; Russakovsky, O.; Ferrari, V.; Fei-Fei, L. What's the point: Semantic segmentation with point supervision. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 549–565.
162. Pu, M.; Huang, Y.; Guan, Q.; Zou, Q. GraphNet: Learning image pseudo annotations for weakly-supervised semantic segmentation. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 26 October 2018; pp. 483–491.
163. Tang, M.; Djelouah, A.; Perazzi, F.; Boykov, Y.; Schroers, C. Normalized cut loss for weakly-supervised cnn segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1818–1827.
164. Tang, M.; Perazzi, F.; Djelouah, A.; Ben Ayed, I.; Schroers, C.; Boykov, Y. On regularized losses for weakly-supervised cnn segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 507–522.
165. Wang, B.; Qi, G.; Tang, S.; Zhang, T.; Wei, Y.; Li, L.; Zhang, Y. Boundary perception guidance: A scribble-supervised semantic segmentation approach. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Macau, China, 10–16 August 2019.
166. Xu, J.; Zhou, C.; Cui, Z.; Xu, C.; Huang, Y.; Shen, P.; Li, S.; Yang, J. Scribble-supervised semantic segmentation inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15354–15363.
167. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
168. Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C. Segment anything model for medical images? *arXiv* **2023**, arXiv:2304.14660.
169. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment anything model for medical image analysis: An experimental study. *arXiv* **2023**, arXiv:2304.10517.
170. Piva, F.J.; de Geus, D.; Dubbelman, G. Empirical Generalization Study: Unsupervised Domain Adaptation vs. Domain Generalization Methods for Semantic Segmentation in the Wild. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 499–508.
171. Shafahi, A.; Najibi, M.; Ghiasi, M.A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L.S.; Taylor, G.; Goldstein, T. Adversarial training for free! In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
172. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. *Proc. Int. Conf. Mach. Learn.* **2018**, *80*, 1989–1998.
173. Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; Chandraker, M. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7472–7481.
174. Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2517–2526.
175. Araslanov, N.; Roth, S. Self-supervised augmentation consistency for adapting semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15384–15394.
176. Zou, Y.; Yu, Z.; Kumar, B.; Wang, J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 289–305.
177. Jiang, Z.; Li, Y.; Yang, C.; Gao, P.; Wang, Y.; Tai, Y.; Wang, C. Prototypical contrast adaptation for domain adaptive semantic segmentation. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 36–54.
178. Lai, X.; Tian, Z.; Xu, X.; Chen, Y.; Liu, S.; Zhao, H.; Wang, L.; Jia, J. DecoupleNet: Decoupled network for domain adaptive semantic segmentation. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 369–387.
179. Hoyer, L.; Dai, D.; Van Gool, L. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 372–391.
180. Hoyer, L.; Dai, D.; Van Gool, L. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9924–9935.
181. Tranheden, W.; Olsson, V.; Pinto, J.; Svensson, L. Dacs: Domain adaptation via cross-domain mixed sampling. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 1379–1389.

182. Wang, Q.; Dai, D.; Hoyer, L.; Van Gool, L.; Fink, O. Domain adaptive semantic segmentation with self-supervised depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 8515–8525.
183. Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; Wen, F. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12414–12424.
184. Wang, Q.; Fink, O.; Van Gool, L.; Dai, D. Continual test-time domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Waikola, HI, USA, 4–8 January 2022; pp. 7201–7211.
185. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
186. Teichmann, M.T.; Cipolla, R. Convolutional CRFs for semantic segmentation. *arXiv* **2018**, arXiv:1805.04777.
187. Vemulapalli, R.; Tuzel, O.; Liu, M.-Y.; Chellapa, R. Gaussian conditional random field network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3224–3233.
188. Liu, Z.; Li, X.; Luo, P.; Loy, C.C.; Tang, X. Deep learning markov random field for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1814–1828. [CrossRef]
189. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ade20k dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 633–641.
190. 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 European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
191. Caesar, H.; Uijlings, J.; Ferrari, V. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Las Vegas, NV, USA, 27–30 June 2016, pp. 1209–1218.
192. Everingham, M.; Winn, J. The PASCAL visual object classes challenge 2012 (VOC2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn. Tech. Rep.* **2012**, *2007*, 1–45.
193. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; Yuille, A. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 891–898.
194. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 746–760.
195. Xiao, J.; Owens, A.; Torralba, A. Sun3d: A database of big spaces reconstructed using sfm and object labels. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1625–1632.
196. Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2636–2645.
197. 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]
198. 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 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
199. Pfeiffer, D.; Gehrig, S.; Schneider, N. Exploiting the power of stereo confidences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 297–304.
200. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
201. Hussain Raza, S.; Grundmann, M.; Essa, I. Geometric context from videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3081–3088.
202. Liu, C.; Yuen, J.; Torralba, A. Nonparametric scene parsing: Label transfer via dense scene alignment. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1972–1979.
203. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image. *Int. J. Comput. Vis.* **2005**, *77*, 157–173. [CrossRef]
204. Gould, S.; Fulton, R.; Koller, D. Decomposing a scene into geometric and semantically consistent regions. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 1–8.
205. Hoiem, D.; Efros, A.A.; Hebert, M. Recovering surface layout from an image. *Int. J. Comput. Vis.* **2007**, *75*, 151–172. [CrossRef]
206. Shotton, J.; Winn, J.; Rother, C.; Criminisi, A. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. J. Comput. Vis.* **2009**, *81*, 2–23. [CrossRef]
207. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
208. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.

209. Xu, M.; Yoon, S.; Fuentes, A.; Park, D.S. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognit.* **2023**, *137*, 109347. [CrossRef]
210. Py, E.; Gherbi, E.; Pinto, N.F.; Gonzalez, M.; Hajri, H. Real-time Weather Monitoring and Desnowification through Image Purification. In Proceedings of the AAAI 2023 Spring Symposium Series, San Francisco, CA, USA, 27–29 March 2023.
211. Wang, H.; Chen, Y.; Cai, Y.; Chen, L.; Li, Y.; Sotelo, M.A.; Li, Z. SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21405–21417. [CrossRef]
212. Vanschoren, J. Meta-learning. In *Automated Machine Learning: Methods, Systems, Challenges*; Springer: New York, NY, USA, 2019; pp. 35–61.
213. Cohn, D.A.; Ghahramani, Z.; Jordan, M.I. Active learning with statistical models. *J. Artif. Intell. Res.* **1996**, *4*, 129–145. [CrossRef]
214. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 405–420.
215. Chen, Y.; Zhan, W.; Jiang, Y.; Zhu, D.; Guo, R.; Xu, X. LASNet: A Light-Weight Asymmetric Spatial Feature Network for Real-Time Semantic Segmentation. *Electronics* **2022**, *11*, 3238. [CrossRef]
216. Zhuang, J.; Yang, J.; Gu, L.; Dvornek, N. Shelfnet for fast semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
217. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
218. Agrawal, A.; Choi, J.; Gopalakrishnan, K.; Gupta, S.; Nair, R.; Oh, J.; Prener, D.A.; Shukla, S.; Srinivasan, V.; Sura, Z. Approximate computing: Challenges and opportunities. In Proceedings of the 2016 IEEE International Conference on Rebooting Computing (ICRC), San Diego, CA, USA, 17–19 October 2016; pp. 1–8.
219. Zhang, Q.; Wang, T.; Tian, Y.; Yuan, F.; Xu, Q. ApproxANN: An approximate computing framework for artificial neural network. In Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 9–13 March 2015; pp. 701–706.
220. Kim, J. Quantization Robust Pruning With Knowledge Distillation. *IEEE Access* **2023**, *11*, 26419–26426. [CrossRef]
221. Looks, M.; Herreshoff, M.; Hutchins, D.; Norvig, P. Deep learning with dynamic computation graphs. *arXiv* **2017**, arXiv:1702.02181.
222. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6688–6697.
223. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
224. Le, T.; Duan, Y. Pointgrid: A deep network for 3d shape understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9204–9214.
225. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
226. Milioto, A.; Vizzo, I.; Behley, J.; Stachniss, C. Rangenet++: Fast and accurate lidar semantic segmentation. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4213–4220.
227. Tchapmi, L.; Choy, C.; Armeni, I.; Gwak, J.; Savarese, S. Segcloud: Semantic segmentation of 3d point clouds. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 537–547.
228. Miao, Z.; Song, S.; Tang, P.; Chen, J.; Hu, J.; Gong, Y. MFFRand: Semantic Segmentation of Point Clouds Based on Multi-Scale Feature Fusion and Multi-Loss Supervision. *Electronics* **2022**, *11*, 3626. [CrossRef]
229. Liu, M.; Zhou, Y.; Qi, C.R.; Gong, B.; Su, H.; Anguelov, D. Less: Label-efficient semantic segmentation for lidar point clouds. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 70–89.
230. Hu, Q.; Yang, B.; Fang, G.; Guo, Y.; Leonardis, A.; Trigoni, N.; Markham, A. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 600–619.