Neurocomputing 406 (2020) 302–321

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A Brief Survey on Semantic Segmentation with Deep Learning

Shijie Hao[a,b], Yuan Zhou[a,b], Yanrong Guo[a,b,*]

[a] Key Laboratory of Knowledge Engineering with Big Data Hefei University of Technology, Ministry of Education, Hefei 230009, China
[b] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China

ARTICLE INFO

Article history:
Received 3 July 2019
Revised 14 November 2019
Accepted 19 November 2019
Available online 13 April 2020

Keywords:
Semantic segmentation
Deep learning

ABSTRACT

Semantic segmentation is a challenging task in computer vision. In recent years, the performance of semantic segmentation has been greatly improved by using deep learning techniques. A large number of novel methods have been proposed. This paper aims to provide a brief review of research efforts on deep-learning-based semantic segmentation methods. We categorize the related research according to its supervision level, i.e., fully-supervised methods, weakly-supervised methods and semi-supervised methods. We also discuss the common challenges of the current research, and present several valuable growing research points in this field. This survey is expected to familiarize readers with the progress and challenges of semantic segmentation research in the deep learning era.

1. Introduction

The significant role of semantic segmentation. Semantic segmentation assigns a category label to each pixel of an image, which is a fundamental but challenging task in computer vision research. As semantic segmentation is able to provide the category information at the pixel level, many real-world applications benefit from this task, such as self-driving vehicles [1,2], pedestrian detection [3,4], defect detection [5], therapy planning [6,7], and computer-aided diagnosis [8,9]. The pixel-level semantic information helps intelligent systems to grasp spatial positions or make important judgments. In this context, semantic segmentation distinguishes itself from other common computer vision tasks. For example, object classification requires that a whole image is annotated with one or more semantic labels. Regarding object detection, the system needs to know where the target objects locate in the scene.

The pervasiveness of deep learning in semantic segmentation. A large number of segmentation methods have been proposed before the deep learning era, such as the partial differential equation based methods. With sufficient training data, the supervised learning strategy is able to greatly extend the capacity of a segmentation model, such as the random forest [10,11] and visual grammar [12,13] applied in natural scene understanding. Recently, the emergence of the deep learning technique has greatly promoted semantic segmentation research. For example, Long et al. proposed the pioneering Fully Convolutional Network (FCN) [14], which dramatically increased the segmentation accuracy. FCN has paved the way for deep-learning-based semantic segmentation. To date, numerous novel deep-learning-based methods have been proposed, which are based on different technical roadmaps and target different applications. Compared to the traditional methods, the deep-learning-based methods have shown remarkable improvement in effectiveness. Almost all of the state-of-the-art performances of public datasets have been achieved by deep-learning methods. Apart from the academic community, the industrial community is also making great efforts in developing advanced systems based on semantic segmentation techniques, such as various intelligent imaging devices. For example, a single phone camera has been able to produce a portrait with the depth-of-focus effect by incorporating the segmentation technique [15].

Difference between this survey and others. We compare our paper with six recent survey papers on semantic segmentation [16–21]. [16] mainly introduced the traditional learning-based semantic segmentation methods, such as the methods based on the support vector machine (SVM) and decision tree. In [18], the authors noted the emergence of the deep-learning-based semantic segmentation methods, such as region-proposal-based and FCN-based approaches. Zhao et al. [20] concentrated on the PASCAL VOC 2012 semantic segmentation challenge and analyzed the related methods as well as their results. Guo et al. [19] comprehensively introduced the methods based on region proposal and FCN, as well as the methods based on weak supervision. Further, in [19], their strengths, weaknesses and major challenges are further summarized. [17] and [21] provided a more comprehensive introduction to this field, including the network structures, the commonly used datasets and metrics, the state-of-the-art methods and some possible future directions. Our paper is different from the

* Corresponding author. Tel.: +(86)13965007001.

https://doi.org/10.1016/j.neucom.2019.11.118
0925-2312/

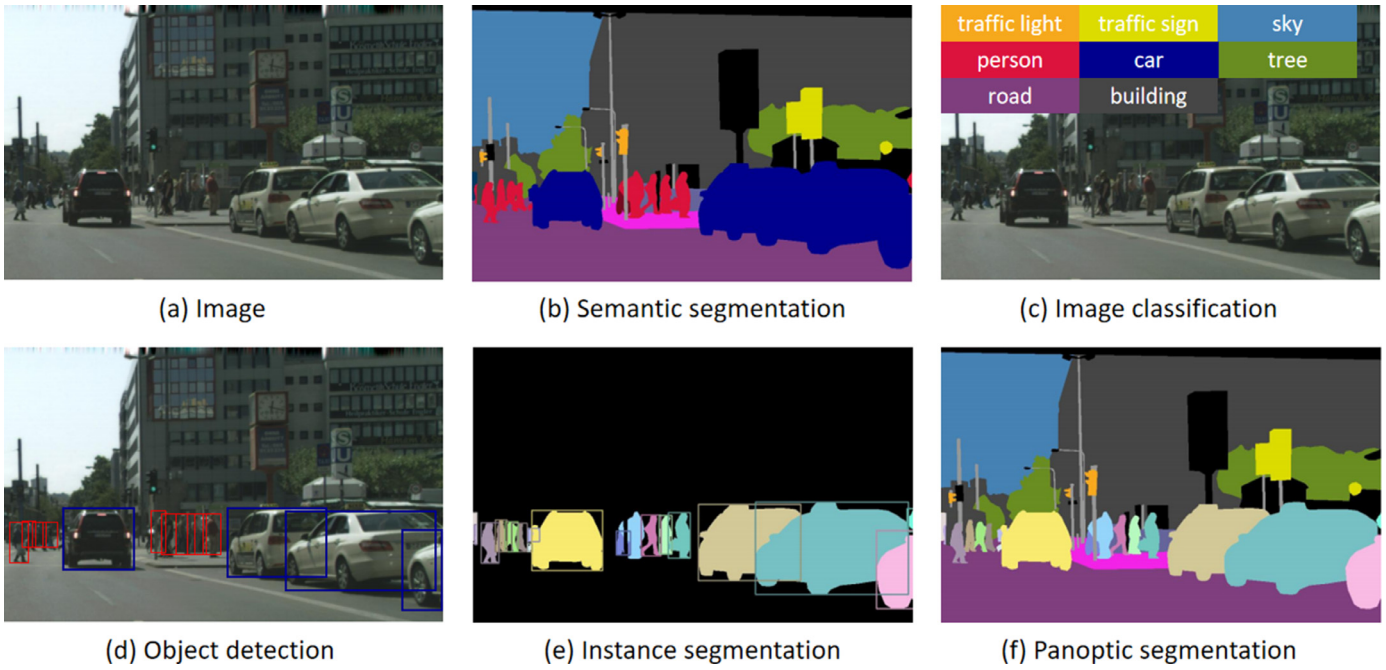


Fig. 1. An example of different vision tasks. The figure is partially borrowed from [22].

above surveys in the following aspects. First, our paper categorizes the methods based on a different perspective, that is, the supervision degree during the training process. Second, our paper particularly summarizes the methods focusing on real-time segmentation, which is less discussed in other surveys. Third, we discuss some potentially valuable research points raised very recently.

Organization and goals of this survey. The organization of this survey is summarized as follows. Section 2 briefly overviews the task of semantic segmentation, and the common deep network architectures. Section 3 reviews the deep-learning-based semantic segmentation methods based on the level of supervision. In Section 4, we introduce the commonly used datasets and evaluation metrics. In Section 5, we summarize the common challenges faced by the current methods, and enumerate several growing research points. Section 6 finally concludes our paper. We expect that this survey can help readers become familiar with deep-learning-based semantic segmentation from a new perspective, and provide some possible hints for a future in-depth investigation.

2. Overview

2.1. Semantic segmentation

Introduction of semantic segmentation. Given an image, i.e. Fig. 1 (a), semantic segmentation describes the task of assigning each pixel with a predefined category label, which is exemplified in Fig. 1 (b). In other words, semantic segmentation aims to partition an image into mutually exclusive subsets, in which each subset represents a meaningful region of the original image.

Comparisons with other computer vision tasks. We first compare the semantic segmentation task with two other fundamental tasks in computer vision, i.e., image classification and object detection. Then, we also introduce two newly developed tasks, i.e., instance segmentation and panoptic segmentation.

Image classification aims to assign one or more category labels for a whole image. In other words, an image classification algorithm tells us what objects exist in an image, e.g., semantic concepts such as a person, car, road and building are detected in Fig. 1(c). Object detection goes one step further. It needs to know not only what objects exist in an image but also their locations in the image scene. For example, in Fig. 1(d), a typical object detection algorithm locates the objects with annotated rectangles. Different from these two tasks, semantic segmentation has higher demands, as it aims at accurately partitioning each object region from the background region. Taking Fig. 1(b) as an example, the output of a segmentation algorithm not only locates all the target objects but also accurately delineates their boundaries. Obviously, the semantic segmentation is much more challenging than the above two tasks, as it is required to fully bridge the semantic gap between low-level features and high-level semantics.

Recently, two fine-grained new tasks, i.e., instance segmentation and panoptic segmentation, have emerged as the new research directions. Instance segmentation aims to detect each object as an individual in the image. For example, as shown in Fig. 1(e), the cars are labeled with different labels, and each label denotes an instance. Taking the task a step further, panoptic segmentation has the highest goal, which assigns a semantic label and an instance label to each pixel. Compared with tradition requirements of semantic segmentation, these two tasks are more challenging, as their concentration extends from “stuff” (traditional semantic segmentation) to “thing” (instance segmentation) and “stuff+thing” (panoptic segmentation).

2.2. Deep learning for semantic segmentation

In semantic segmentation, various methods have achieved promising results by using deep neural networks. In general, by feeding sufficient images and their pixelwise labeling maps as training data, a deep neural network is trained to learn a mapping between a semantic label and its diversified visual appearances. The learning process gradually bridges the inconsistency between high-level semantics and low-level features, making the network continually more aware of various semantic concepts [23]. In the following, we briefly review several deep architectures commonly used in semantic segmentation, as shown in Table 1.

VGG. The VGG network was proposed by Simonyan et al. [24] from the Visual Geometry Group at Oxford University. There

Table 1
The popular common deep architectures.

Architecture	Author	Publishing year	Source code	Representative methods
VGG[24]	Simonyan et al.	2014	https://github.com/tensorflow/models/tree/master/research/slim/nets/vgg.py	FCN[14], SegNet[25], DilatedNet[26]
ReNet[27]	Visin et al.	2015	N/A	ReSeg[28]
ResNet[29]	He et al.	2016	https://github.com/KaimingHe/deep-residual-networks	PSPNet[30], RefineNet[31], LKernel[32], EncNet[33] DeepLab[34–36], Mask-RCNN[37]
DenseNet[38]	Guo et al.	2017	https://github.com/liuzhuang13/DenseNet	DenseASPP[39], SDNet[40], FC-DenseNet[41]
ResNeXt[42]	Xie et al.	2017	https://github.com/facebookresearch/ResNeXt	DShortcut[43], ExFuseNet[44]
MobileNetV1[45]	Howard et al.	2017	https://github.com/tensorflow/models/blob/master/research/slim/nets/mobilenet_v1.py	FSTSL[46]
MobileNetV2[47]	Sandler et al.	2018	https://github.com/tensorflow/models/blob/master/research/slim/nets/mobilenet/mobilenet_v2.py	LWRF[48], Fast-SCNN[49]
MobileNetV3[50]	Howard et al.	2019	N/A	-

are different versions for VGG networks according to the layer number, such as VGG-13, VGG-16, and VGG-19. The key contribution of the VGG networks is that they paved the way to design deeper structures for better performance. VGG has been adopted as the backbone of several semantic segmentation models [14,25,26].

ReNet. ReNet [27] is highlighted by the manner in which it replaces convolutional layers with multi-direction recurrent neural networks (RNN), which provides an alternative way of constructing the network architecture. The representative semantic segmentation method based on ReNet is ReSeg [28].

ResNet. The deep residual network (ResNet) [29] successfully enables a much deeper network and achieves better performance in various vision tasks. Its key contribution lies in modeling the residual representation into the CNN network structure, which solves the difficulty of training a very deep network structure. A large number of semantic segmentation methods choose it as their backbone, e.g., the DeepLab family [34–36], and PSPNet [30], just to name but a few.

DenseNet. Different from the traditional strategy that makes a network deeper or wider, DenseNet [38] connects every layer to each other. Its advantage lies in the following aspects: 1) less parameters, 2) more reuse of features, and 3) a better training process that relieves the vanishing gradient and model degeneration issue. The representative semantic segmentation methods based on DenseNet include DenseASPP [39], FC-DenseNet [41], and SDNet [40].

ResNeXt. Aiming at enhancing the network performance while preserving the network complexity, ResNeXt [42] is highlighted in its homogeneous, multi-branch architecture that has only a few hyperparameters to set. The semantic segmentation methods DShortcut [43] and ExFuseNet [44] use ResNeXt as their backbone.

MobileNet. It is important to design networks for balancing accuracy and computational costs. In this context, various lightweight networks have been designed. MobileNetV1 [45] introduces depthwise convolution, which achieves a great improvement in efficiency. In the ImageNet classification challenge, it achieves 70.6% accuracy with 4.2M parameters. Addressing the limitation of MobileNetV1, MobileNetV2 [47] is based on the inverted residual structure. MobileNetV3 [50] achieves better performance with even less parameters via incorporating the attention mechanism. Semantic segmentation methods [46,48,49] based on MobileNetV1 and MobileNetV2 are potentially useful for real-time applications.

3. State-of-the-art methods

In this section, we extensively review the segmentation methods, which are grouped according to the supervision level, i.e., full supervision (Section 3.1), weak supervision (Section 3.2), and semi-supervision (Section 3.3).

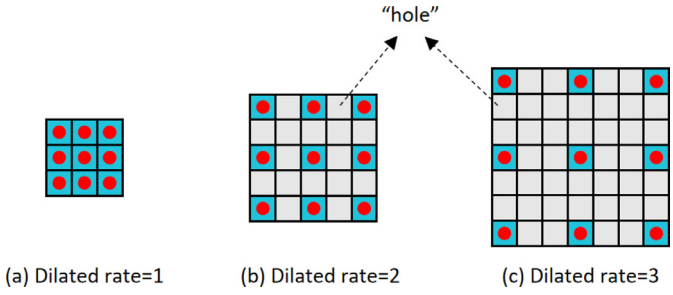


Fig. 2. Several dilated convolution kernels with different dilated rates.

3.1. Supervised methods

In regard to the fully supervised semantic segmentation methods, there is a tacit assumption that sufficient labeled training data are available, including the original images and their corresponding pixelwise semantic-annotated images.

3.1.1. Context-based methods

The representative semantic segmentation methods that leverage context information are summarized in Table 2.

Generally, the context information, which is far beyond the pixel-level appearance, becomes aware of semantics, and provides a useful complementary source for building semantic segmentation models. In 2011, Lucchi et al. [62] stated that semantic segmentation accuracy can be boosted by appropriately introducing context information. In 2015, Yu et al. [26] proposed DilatedNet to aggregate multi-scale contexts based on dilated convolution. Different from the traditional convolution operator, dilated convolution introduces a flexible "dilated rate". As shown in Fig. 2, the convolution with a larger dilated rate has a larger receptive field while introducing no extra computations. As a special case, the dilated convolution with the dilated rate of 1 degenerates to the traditional convolution. In DilatedNet, the contexts are extracted from multiple scales, in which five different dilated rates, 1, 2, 4, 8, and 16, are used. For the PASCAL VOC 2012 test set, DilatedNet achieves 67.6% MIoU. By exploiting global pooling, [51] extended the strategy of enlarging the receptive field to the direct extraction of global features. The experiments in this work indicate that, apart from local features, the global context also helps to effectively produce more accurate and smooth segmentation results, i.e., 69.8% MIoU for the PASCAL VOC 2012 test set. [54] greatly improves the

Table 2
Collections for context-based methods.

Method	Author	Publishing year	Source code	Backbone	Contribution
DilatedNet[26]	Yu et al.	2015	https://github.com/fyu/dilation	VGG16	Dilated convolution
ParseNet[51]	Liu et al.	2015	https://github.com/weiliu89/caffe/tree/fcn	FCN32s or DeepLabV1[52]	Global context aggregation
Zoom-out[53]	Mostajabi et al.	2015	https://github.com /Harvey1973/Zoomout	VGG-16	Zoom-out feature construction
Piecewise[54]	Lin et al.	2016	N/A	VGG16	Cooperate CNN with CRF to capture the correlations between adjacent patches
PSPNet[30]	Zhao et al.	2017	https://github.com/hszhao/PSPNet	ResNet	Pyramid Pooling Module (PPM)
LKernel[32]	Peng et al.	2017	https://github.com/SConsul/Global_Convolutional_Network	ResNet152	Global Convolutional Network (GCN)
DeepLabV2[34]	Chen et al.	2017	https://github.com/DrSleep/tensorflow-deeplab-resnet	ResNet	Atrous Spatial Pyramid Pooling module (ASPP)
DeepLabV3[35]	Chen et al.	2017	https://github.com/rishizek/tensorflow-deeplab-v3-plus	ResNet	Extend ASPP by introducing global pooling
GCE[55]	Chen et al.	2017	https://github.com/hfslyc/GCPNet	FCN8s or DeepLabV2	Global context learned from scene similarity
EncNet[33]	Zhang et al.	2018	https://github.com/zhanghang1989/PyTorch-Encoding	ResNet	Context Encoding Module
CCL[56]	Ding et al.	2018	N/A	ResNet101	Context contrasted Local module (CCL) and Gate Sum block
OCNet[57]	Yuan et al.	2018	https://github.com/PkuRainBow/OCNet.pytorch	ResNet101	Object context pooling
DShortcut[43]	Bilinski et al.	2018	N/A	ResNeXt101	Dense decoder shortcut connections
DRNet[58]	Zhuang et al.	2018	https://github.com/zhuangyqin/DRN	ResNet38[59]	Dense Relation Network and Context-Restricted Loss
Dense ASPP[39]	Yang et al.	2018	https://github.com/zhulf0804/DenseASPP-Tensorflow	DenseNet	Dense Atrous Spatial Pyramid Pooling
CFNet[60]	Zhang et al.	2019	http://hangzh.com/	ResNet	Aggregated Co-occurrent Feature Module (ACF)
SVCNet[61]	Ding et al.	2019	N/A	ResNet101 and FCN-4s	Shape-Variant Convolution

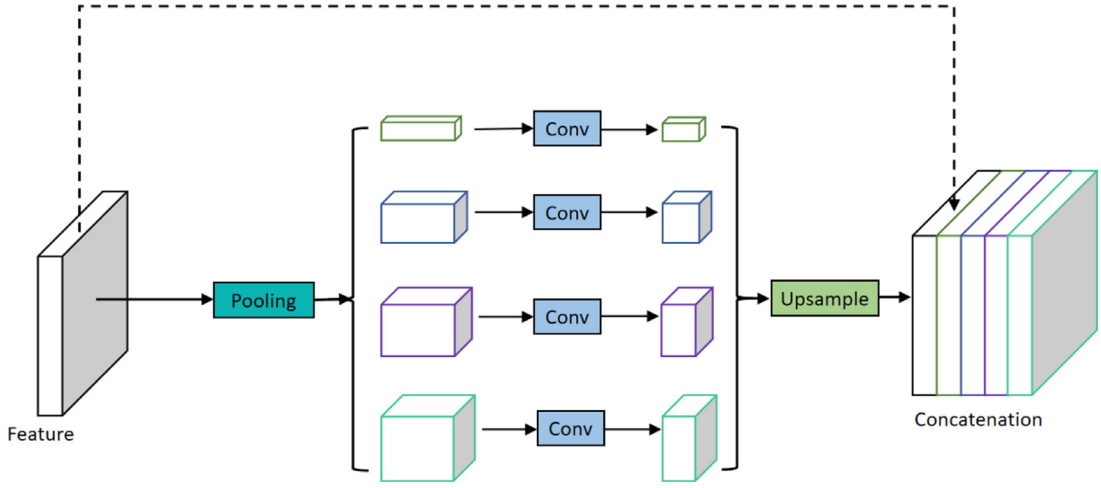


Fig. 3. The architecture of the Pyramid Pooling Module (PPM). The figure is reproduced from [30].

MIoU on the PASCAL VOC 2012 test set to 78.0% via combining CNN with CRF to extract the patch-wise context. [30] introduced a multi-scale context-aggregated module called the Pyramid Pooling Module (PPM), as shown in Fig. 3, which has inspired many of the following methods. For example, [34] proposed the Atrous Spatial Pyramid Pooling (ASPP) module via replacing the pooling and convolution in PPM with the atrous convolution. It achieves 79.7% MIoU on the PASCAL VOC test set. [35] further improves the ASPP module via adding an additional global pooling to the parallel branches, which substantially increases the MIoU on the PASCAL VOC 2012 test set to 85.7%. [39] proposed DenseASPP via densely connecting each parallel branch. In [56], a variant of ASPP was proposed, named the Context Contrasted Local (CCL) module, where subtraction is applied between the features of parallel branches to highlight local contexts. In [32], Peng et al. stated that the large kernel plays a critical role in semantic segmentation, and introduced the Global Convolution Network, as shown in Fig. 4. By exploiting the strategy of stacked convolution, it achieves a good trade-off between performance and computation costs. Concerning the context information at the global level, [55] proposed the Global Context Embedding Network (GCENet), which retrieves the context information from semantically similar images in the train-

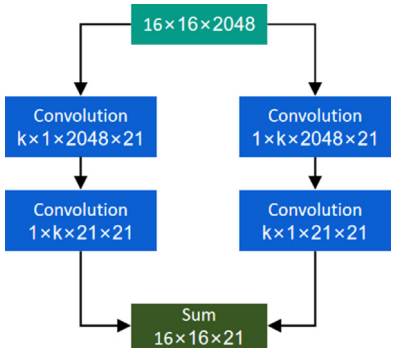


Fig. 4. The architecture of the Global Convolution Network (GCN). The figure is reproduced from [32].

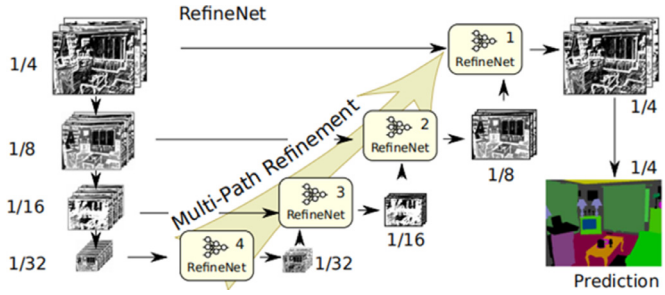


Fig. 5. The architecture of RefineNet. The figure is borrowed from [31].

ing set to supplement the context extracted from the local region. In [33], Zhang et al. proposed EncNet, which incorporates a CNN with the classic dictionary learning to extract global context. This method performs well on the PASCAL VOC 2012 test set by achieving 85.9% MIoU. Different from the previous works, [57] proposed object context pooling (OCP), which jointly exploits pixels with the same category, to represent the object-level context. In [43], the densely connected decoder block was proposed to aggregate multiple instances of context information. In [58], Zhuang et al. proposed the Dense Relation Network (DRN) and Context-Restricted Loss (CRL) to aggregate the global context with the local information. [60] proposed the co-occurrent feature module, which is able to provide fine-grained representation by leveraging the co-occurrent features. This method achieves 87.2% MIoU on the PASCAL VOC 2012 test set. In [61], Ding et al. pointed out that due to the diverse shapes and complex layout of objects in an image, context aggregation tends to be ineffective and inefficient. To solve this issue, this work utilizes a scale- and shape-variant semantic mask for each pixel to confine its contextual region based on the novel paired convolution and the shape-variant convolution.

3.1.2. Feature-enhancement-based methods

For a typical CNN pipeline, on one hand, features extracted in deep layers are more semantic-aware, but lose the spatial details, due to pooling and stride convolution. On the other hand, we can see that features from shallow-layers are more aware of details, such as strong edges. In this context, the appropriate cooperation of these two types of features has the potential to boost the performance of semantic segmentation. We name this strategy as feature-enhancement-based methods, and list the related papers in Table 3.

The Fully Convolutional Network (FCN) [14] pioneers feature enhancement via the skip-connection strategy. By applying the skip connection between the features for prediction and the features from the midlayer, the final resolution is increased from 1/32 to the 1/8, and the accuracy is improved approximately 3% in MIoU. It proves that leveraging the features from shallow and deep layers is beneficial for improving the semantic segmentation accuracy. [74] further demonstrates the importance of the skip connection in image segmentation. Ronneberger et al. [63] proposed the symmetrical encoder-decoder architecture called U-Net. Different from FCN, U-Net fully leverages the features from each layer by using dense skip connections. The features from each layer in the encoder part are connected to the symmetrical layers in the decoder part. U-Net has drawn extensive attention from the medical image analysis community [75,76]. For example, to address the limitation of ignoring the spatial information along the z-dimension in a 2D-based model, Li et al. [70] extend [63] by operating 2D-DenseUNet and 3D-DenseUNet in a cooperative manner. This method aims at learning intraslice and interslice features respectively, and fusing them via the fusion block. U-Net has also been extended and applied in other applications. For the task of natural image segmentation, [72] extends [63] into stacked U-Nets. [66] extends [63] by introducing the residual block. [64] proposed the hypercolumn representation, which uses the concatenated features from different layers in the CNN pipeline to make the final inference. Recently, RefineNet [31] was proposed to further leverage complementary features, in which a multi-path refinement process is built, and the spatial details of the feature maps are enhanced gradually, as shown in Fig. 5. However, the RefineNet is relatively computationally expensive, as noted in [48]. To extract high-level semantic features while maintaining spatial details, [68] separates the whole network into two substreams, i.e., the pooling stream and the residual stream. The pooling stream aims to extract high-level semantics (low resolution). The residual stream aims to maintain details (high resolution), and cooperate with the features learned by the pooling stream. [41] extends DenseNet [38] by introducing an upsampling path and the CRFs [52]. In [44], Zhang et al. proposed the bilateral enhancement strategy, i.e., a mutual enhancement process between low-level and high-level features. Based on DeepLabV3 [35], Chen et al. [36] proposed DeepLabV3-Plus via densely connecting the decoder component with the encoder. [71] proposed the Refinement Residual bBlock (RRB) and Channel Attention Block (CAB) to enhance features from the different layers. [73] proposed the HFCNet that utilizes the concatenated features from all the decoder layers. From the above methods, we can see that the methods based on feature enhancements mainly encompass designing a network to connect and fuse different kinds of features.

3.1.3. Deconvolution-based methods

The first deconvolution-based semantic segmentation method was proposed by Noh et al. [77] in 2015 and is called DeconvNet. As shown in Fig. 6, DeconvNet is based on the symmetrical encoder-decoder architecture. In the encoder part, the semantic features are gradually extracted while the resolution is lower due to max pooling. Moreover, the method stores the locations of maximum values, called pooling indices, in the sliding windows during the pooling process [23,77]. In the decoder component, the unpooling operator utilizes the saved indices to upsample the low-resolution feature map into a high-resolution feature map. Then, the deconvolution adopts the trainable filters to reproduce the dense feature maps. We further provide the comparison between pooling and unpooling or convolution and deconvolution in Fig. 7. The following work [25] scales the DeconvNet into a light one called SegNet. Fourure et al. [79] proposed GridNet, where the decoder component employs deconvolution to recover spatial resolution. [78] further extended SegNet into a Bayesian-based model, called Bayesian-SegNet. Fu et al. [40] proposed the Stacked

Table 3
Collections for feature-enhancement-based methods.

Method	Author	Publishing year	Source code	Backbone	Contribution
FCN[14]	Long et al.	2015	https://fcn.berkeleyvision.org	VGG16	Fully convolutional layer
U-Net[63]	Ronneberger et al.	2015	http://lmb.informatik.unifreiburg.de/people/ronneber/u-net	FCN	U-shape framework
Hypercolumn[64]	Hariharan et al.	2015	https://github.com/AlexandreRobicquet/Cnn-Hypercolumn	SDS[65]	The hypercolumn presentation
FusionNet[66]	Quan et al.	2016	https://github.com/GunhoChoi/FusionNet-Pytorch	FRCNN	Fully residual convolutional neural network
UpNet[67]	Valada et al.	2016	http://deepscene.cs.uni-freiburg.de	VGG13	Early and late fusion framework
RefineNet[31]	Lin et al.	2017	https://github.com/guosheng/refinenet	ResNet	MultiPath Refinement Network
FC-DenseNet[41]	Jegou et al.	2017	https://github.com/SimJeg/FC-DenseNet	DenseNet	Extend [38] to semantic segmentation
FRRNet[68]	Pohlen et al.	2018	https://github.com/TobyPDE/FRRN	FRRN	Full-Resolution Residual Network
DeepLabV3-Plus[36]	Chen et al.	2018	https://github.com/tensorflow/models/tree/master/research/deeplab	DeepLabV2 and Xception[69]	Extend [52] to the encoder-decoder framework
H-DenseUNet[70]	Li et al.	2018	https://github.com/xmengli999/H-DenseUNet	U-Net	Hybrid Densely Connected U-Net
DFNet[71]	Yu et al.	2018	https://github.com/YuhuiMa/DFN-tensorflow	ResNet101	Discriminative Feature Network
SUNets[72]	Shah et al.	2018	https://github.com/shahsohil/sunets	U-Net	Stacked U-Nets
ExFuseNet[44]	Zhang et al.	2018	https://github.com/lxtGH/fuse_seg_pytorch	GCN[32] and ResNeXt101	Enhancing Feature Fusion Network
HFCNet[73]	Yang et al.	2019	N/A	VGG16	Highly Fused Convolutional Network and Soft Cost Functions

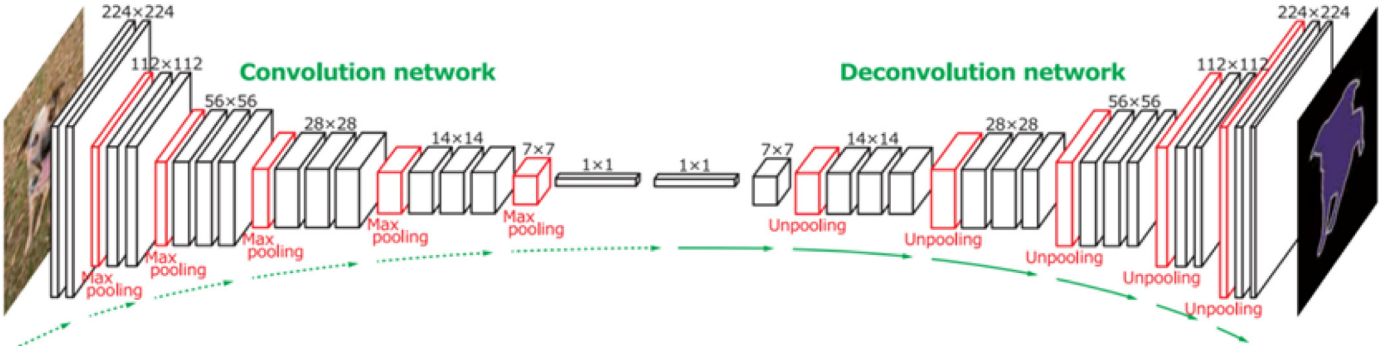


Fig. 6. The architecture of DeconvNet. The figure is borrowed from [77].

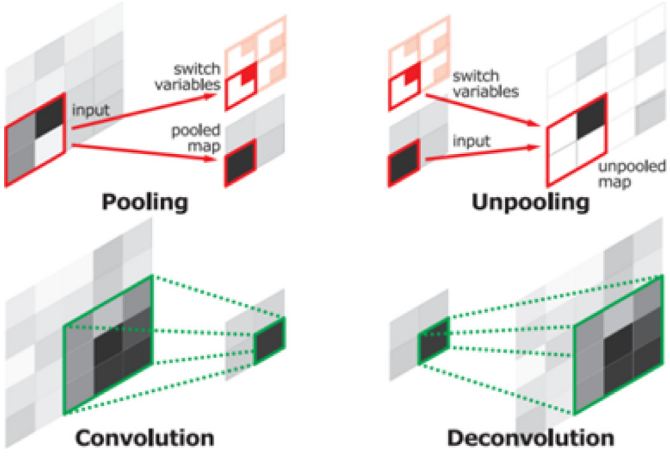


Fig. 7. The comparisons for pooling and unpooling or convolution and deconvolution. The figure is borrowed from [77].

Deconvolution Network (SDN) method. In the PASCAL VOC 2012 test set, SDN achieves 86.6% MIoU, which outperforms DeepLabV3. The main deconvolution-based methods are summarized in Table 4.

3.1.4. RNN-based methods

The Recurrent Neural Network (RNN) achieves promising results in processing sequential signals, such as text and speech [80,81], which also leads to the application in semantic segmentation. We list several main RNN-based segmentation methods in Table 5.

In 2014, Pinheiro et al. [82] proposed the Recurrent Convolutional Neural Network (RCNN) which is built on several plain CNN blocks. Each CNN block is fed with the previous prediction or the raw input image. The final prediction is obtained via a complex refining process. Poudel et al. [83] proposed Recurrent Fully Convolutional Networks (RFCNs) to segment the target object from MRI sequences, which validates the cooperation of FCN and RNN. [84] proposed the LSTM Recurrent Neural Network (LSTM-RNN), which is mainly composed of the 2D LSTM layer and the feedforward layer. This method addresses the limitation of failing to model the long-range dependencies, and extracting almost all of the features is based on the adjacent pixels or patches in the kernel-based methods. In this method, the 2D LSTM layer, containing four LSTM blocks, captures the long-range dependencies in four directions, i.e., left-top, left-bottom, right-top, and right-bottom. Then, the feedforward layer summarizes the outputs of all LSTM blocks. To decrease the computational cost, [84] separates the whole image into non-overlapping patches as the inputs. Visin et al. [28] proposed the ReSeg method, which is composed of three components, i.e., the local features extractor, the ReNet layers, and

Table 4
Collections for deconvolution-based methods.

Method	Author	Publishing year	Source code	Backbone	Contribution
DeconvNet[77]	Noh et al.	2015	https://github.com/HyeonwooNoh/DeconvNet	VGG16	Deconvolution Network
Bayesian-SegNet[78]	Kendall et al.	2015	http://mi.eng.cam.ac.uk/projects/segnet/	VGG16	Incorporate [25] with Bayesian model
SegNet[25]	Badrinarayanan et al.	2017	http://mi.eng.cam.ac.uk/projects/segnet/	VGG16	SegNet, an efficient framework
GridNet[79]	Fourure et al.	2017	https://github.com/Fourure/GridNet	GridNet	Residual Conv-Deconv Grid Network (GridNet)
SDNet[40]	Fu et al.	2019	https://github.com/tum271828/sdn-keras	DenseNet161	Stacked deconvolution network

Table 5
Collections for RNN-based methods.

Method	Author	Publishing year	Source code	Backbone	Contribution
RCNN[82]	Pinheiro et al.	2014	N/A	RCNN	Recurrent Convolutional Neural Network
LSTM-RNN[84]	Byeon et al.	2015	N/A	LSTM-RNN	LSTM Recurrent Neural Network
RFCNs[83]	Poudel et al.	2016	N/A	FCN	Recurrent Fully Convolutional Networks
ReSeg[28]	Visin et al.	2016	https://github.com/fvisin/reseg	ReNet[27] and VGG16	Introduce [27] to semantic segmentation
DAG-RNN[85]	Shuai et al.	2016	https://github.com/sallymmx/DAG-RNN	VGG16 and FCN	Directed Acyclic Graphs Recurrent Neural Network
DD-RNNs[86]	Fan et al.	2018	N/A	VGG and ADNet[89]	Dense Recurrent Neural Networks
ML-CRNNs[88]	Fan et al.	2018	N/A	VGG16	Multilevel Contextual Recurrent Neural Networks

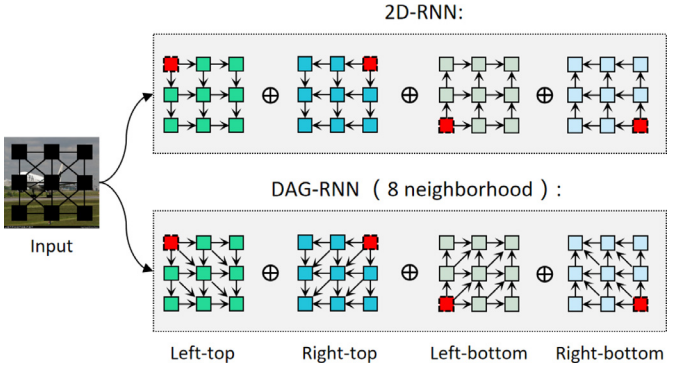


Fig. 8. The comparisons between 2D-RNN and DAD-RNN (8 neighborhood) in four directions, i.e., left-top, right-top, left-bottom, and right-bottom.

the deconvolution layers. Among them, the ReNet layers capture the long-range dependencies based on the local features. Extended from 2D-RNN, Shuai et al. [85] proposed the DAG-RNN method. By comparing DAG-RNN with 2D-RNN in Fig. 8, we observe that DAG-RNN contains more dense connections. To alleviate the dependencies fading problem due to the long-range paths, Fan et al. [86] proposed the Dense Recurrent Neural Network, which builds dense connections between each patch in the image. It adopts the attention model [87] to weigh more or less for the vital or irrelevant dependencies. [88] utilizes the recurrent neural network to capture multi-level dependencies, i.e., the local, global, and topic dependencies.

3.1.5. GAN-based methods

The Generative Adversarial Network (GAN) [90] has achieved success in many applications, such as text to image synthesis [91], style transfer [92], and image inpainting [93]. The pioneering research for applying GAN in semantic segmentation is ANet proposed by Luc et al. [94], which is made up of the segmentation network and the adversarial network. In the forward-propagation, the segmentation network (generator) partitions the input image into some non-overlapping regions, and the adversarial network (discriminator) distinguishes the generator output and the ground-truth label maps. In the back-propagation, the adversarial loss is alternatively applied to the generator and the discriminator, which acts in the min-max-game way. Due to the specific requirement of semantic segmentation, the pixel-level classification loss is used as a strong constraint.

Moeskops et al. [95] incorporate adversarial learning with dilated convolution [26] to realize brain MRI segmentation. [96] replaces the GAN model with the cGAN [97] in the brain tumor segmentation task. Of note, the discriminator output in the previous methods is just a binary value (e.g., 0/1 for fake/real). In contrast, [98] stated that this is not suitable for pixel-level segmentation. Instead, the SegAN network [98] was proposed to modify the discriminator, as shown in Fig. 9. The discriminator extracts the features of the different levels from both the ground truth and the prediction. Then, the L_1 loss is applied to measure the distance between these two parts, which is utilized to optimize the segmentation network in the back-propagation. To obtain better results, [99] additionally adds the refinement block to renew the outputs of the generator. Luo et al. [100] stated that the single classification loss at the pixel level produces inconsistency between local and global features for the segmentation model. The Macro-Micro Adversarial Network (MMAN) [100] shown in Fig. 10 was proposed to alleviate this problem. MMAN contains two discriminators, i.e., a macro discriminator and a micro discriminator. The former is applied to the high-level features (low resolution), aiming to force the model to better understand the global semantics. The latter is applied to the final outputs of the generator (high resolution), aiming at assisting the elimination of the local inconsistency. We summarize the reviewed methods in Table 6.

3.1.6. RGBD-based methods

RGBD images provide an extra information source (depth images) for semantic segmentation. However, the cooperation between the depth map and the photometric image is not straightfor-

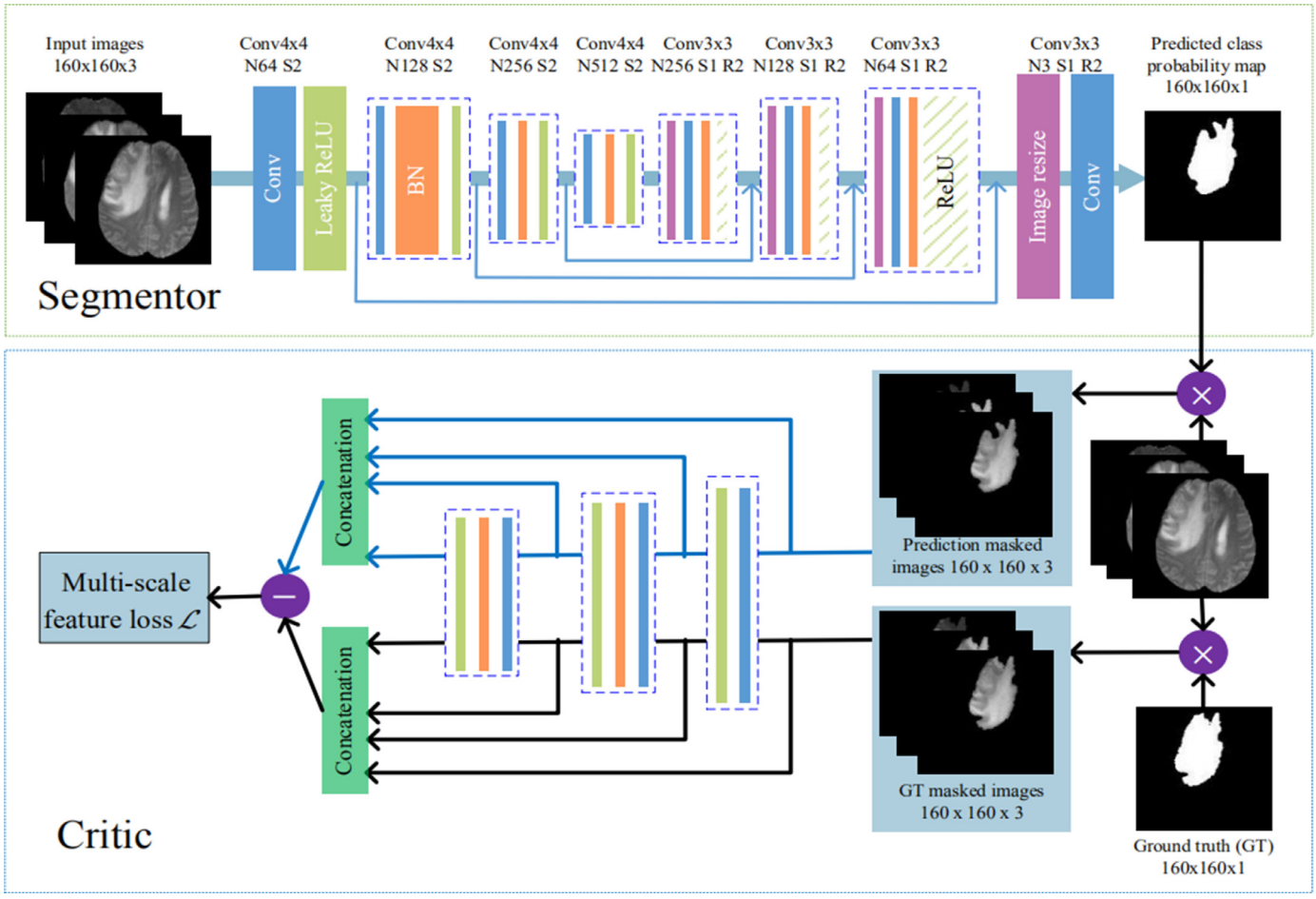


Fig. 9. The architecture of SegAN. The figure is borrowed from [98].

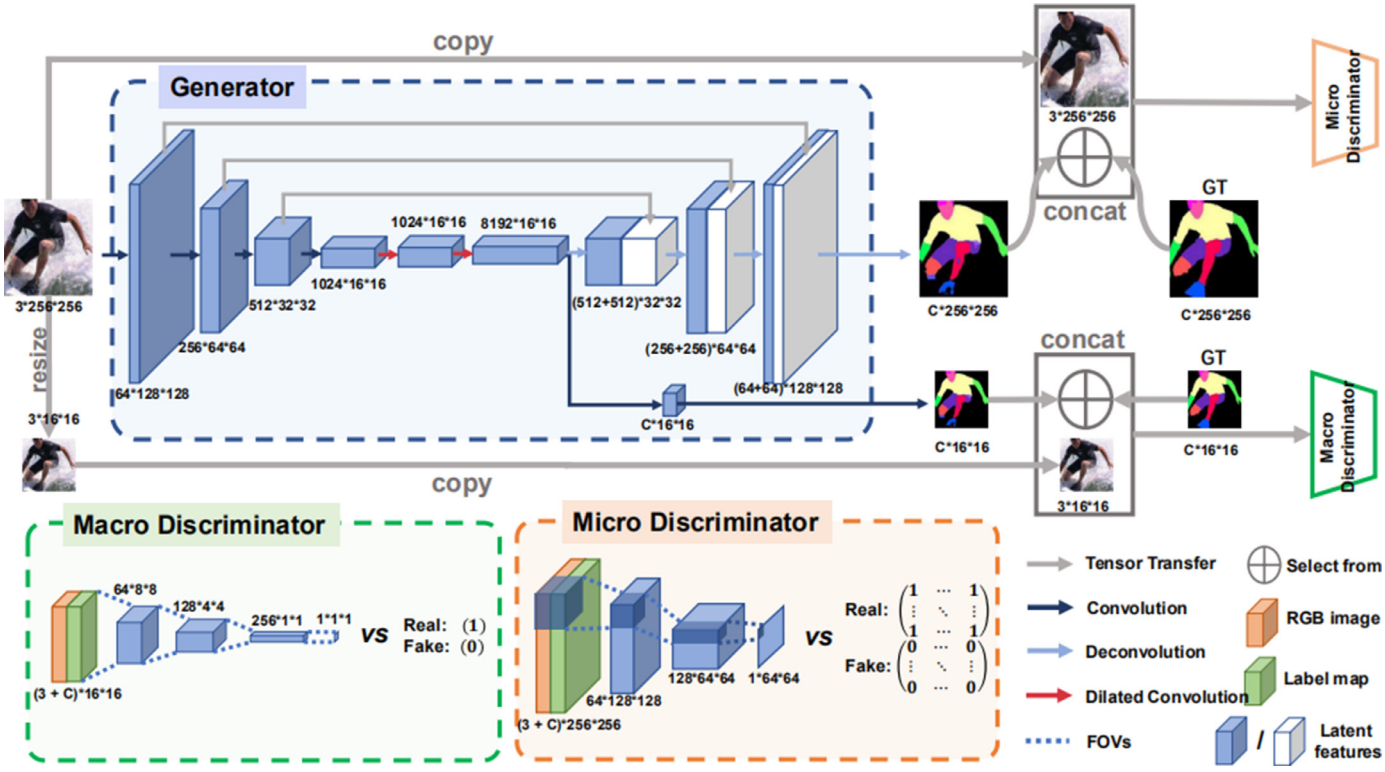


Fig. 10. The architecture of Macro-Micro Adversarial Network (MMAN). The figure is borrowed from [100].

Table 6
Collections for GAN-based methods.

Method	Author	Publishing year	Source code	Backbone	Contribution
ANet[94]	Luc et al.	2016	https://github.com/oyam/Semantic-Segmentation-using-Adversarial-Networks	DilatedNet[26] or MSNet[101]	Introduce the adversarial learning to realize semantic segmentation
ATDC[95]	Moeskops et al.	2017	N/A	FCN or DilatedNet	Incorporate the adversarial network with dilated convolution to realize the brain MRI segmentation
CANet[96]	Rezaei et al.	2017	N/A	U-Net and MkGAN[102]	Conditional adversarial network for semantic segmentation
SegAN[98]	Xue et al.	2018	https://github.com/YuanXue1993/SegAN	SegAN	SegAN and multiscale L_1 loss
CGRANet[99]	Rezaei et al.	2018	N/A	CGRANet	Conditional Generative Refinement Adversarial Network
MMANet[100]	Luo et al.	2018	https://github.com/RoyalVane/MMAN	DeepLab[34]	Macro-Micro Adversarial Network

Table 7
Collections for RGBD-based methods.

Method	Author	Publishing year	Source code	Backbone	Contribution
ISSD[103]	Couprie et al.	2013	N/A	MSCNN[101]	Use CNN to solve RGBD semantic segmentation problem.
FSS[104]	Höft et al.	2014	N/A	LOCS[120]	Utilize HOG[105] and HOD[106] to preprocess the photometric image and the depth map.
RF-RGBD[107]	Gupta et al.	2014	http://www.cs.berkeley.edu/~sgupta/eccv14/	VGG	HHA
FuseNet[109]	Hazirbas et al.	2016	https://github.com/tum-vision/fusenet	VGG16	Fusion-based CNN framework
LSTM-CF[112]	Li et al.	2016	https://github.com/Winstonxu123/LSTM-CF	DeepLabV1[52]	introduce LSTM to capture interchannel dependencies
LCSF[110]	Wang et al.	2016	N/A	VGG16	Feature Transfor-mation Network
CFN[117]	Lin et al.	2017	N/A	VGG16	Cascaded Feature Network
3DGNN[111]	Qi et al.	2017	https://github.com/yanx27/3DGNN_pytorch	DeepLab-LargeFov[52] or ResNet101	3D Graph Neural Networks
LSD-GF[113]	Cheng et al.	2017	N/A	DeepLabV2[34]	Locality-sensitive DeconvNet
RDFNet[115]	Park et al.	2017	https://github.com/SeongjinPark/RDFNet	ResNet	Extend [31] to RGBD semantic segmentation
STD2P[116]	He et al.	2017	https://github.com/SSAW14/STD2P	FCN	Spatiotemporal Data-Driven Pooling
DCNN[119]	Wang et al.	2018	https://github.com/laughtervv/DepthAwareCNN	DeepLabV1[52]	Depth-aware CNN
DRR[108]	Guo et al.	2018	N/A	GGR[121] and FCN	Cooperate depth estimation with RGBD semantic segmentation
Zig-Zag[118]	Lin et al.	2018	N/A	CFN[117]	Zig-Zag Network
ACNet[114]	Hu et al.	2019	https://github.com/anheidelonghu/ACNet	ResNet50	Attention Complementary Network

ward. We briefly review the semantic segmentation methods based on RGBD data, summarized in Table 7.

In the seminal research of [103], the geometric cues are simply utilized by directly concatenating the RGB channels and the depth channel. However, as the appearances of a photometric image and its depth map can be greatly different, a simple combination is not enough to fully leverage the rich information of RGBD data. To address this limitation, Höft et al. [104] proposed to preprocess the photometric image and the depth map with HOG [105] and HOD [106], receptively. Gupta et al. [107] proposed to encode the depth map into three channels (called HHA), i.e., the horizontal disparity, the height above the ground, and the angle with gravity. HHA is widely used in the following works, such as [14,108]. Conversely, Hazirbas et al. [109] proposed the FuseNet, which introduces an additional depth encoding block to extract the features from the depth map. In this method, the photometric features are densely enhanced by the depth features. Wang et al. [110] proposed the feature transformation network, which utilizes the common features extracted from the photometric and the depth image. Qi et al. [111] introduced the Graph Neural Network (GNN) for RGBD semantic segmentation. Li et al. [112] proposed a strategy of mixing HHA and FuseNet, and additionally introduced LSTM to capture the interchannel dependencies between photometric and depth features. Cheng et al. [113] and Hu et al. [114] introduced

Table 8
Collections for real-time methods.

Method	Author	Publishing year	Source code	Backbone	Contribution
ENet[122]	Paszke et al.	2016	https://github.com/TimoSaemann/ENet	ENet	Efficient Neural Network
SQ[124]	Treml et al.	2016	https://github.com/klickmal/speeding_up_semantic_segmentation	SqueezeNet[125]	Construct a lightweight network based on Squee-zeNet[125]
ERFNet[127]	Romera et al.	2017	https://github.com/Eromera/erfnet_pytorch	ERFNet	Efficient Residual Factor-ized Convolution Network (ERFNet)
LinkNet[139]	Chaurasia et al.	2017	https://github.com/e-lab/LinkNet	ResNet18	Propose a lightweight framework LinkNet
SS[130]	Wu et al.	2017	N/A	ResNet50	Propose spatial sparsity two-column network
ContextNet[131]	Poudel et al.	2018	https://github.com/klickmal/ContextNet	ContextNet	Separate feature extraction and resolution maintaining in subnetwork
BiSeNet[129]	Yu et al.	2018	https://github.com/ooooverflow/BiSeNet	Xception39[69] or ResNet18	Propose spatial path to generate high-resolution features and context path to extract semantics
DSNet[128]	Wang et al.	2018	https://github.com/s7ev3n/DSNet	DSNet	Driving Segmentation Network
ESPNet[133]	Mehta et al.	2018	https://sacmehta.github.io/ESPNet	ESPNet	Efficient spatial pyramid module
ICNet[140]	Zhao et al.	2018	https://github.com/hszhao/ICNet	ICNet	Image cascade network
ESSGG[138]	Vallurupalli et al.	2018	N/A	ERFNet	A novel training strategy for grouped-convolution-based methods
LWRF[48]	Nekrasov et al.	2018	https://github.com/DrSleep/light-weight-refinenet	ResNet or MobileNetV2	Extend [31] to achieve real-time semantic segmentation
C3Net[134]	Park et al.	2018	N/A	C3Net	Concentrated-Compreh-ensive Convolution
DABNet[132]	Li et al.	2019	https://github.com/Reagan1311/DABNet	DABNet	Depthwise Asymmetric Bottleneck Network
DFANet[141]	Li et at.	2019	N/A	Xception	Deep Feature Aggregat-ion Network
Fast-SCNN[49]	Poudel et al.	2019	https://github.com/Tramac/Fast-SCNN-pytorch	Fast-SCNN	Fast Semantic Segment-ation Network

the attention mechanism to weigh the features with different importance levels. Following the strategy of [109], [115] extends RefineNet [31] to the RGBD-segmentation version. The method introduces an additional branch to extract the depth features, and a multi-modal feature fusion block to combine the photometric and depth features. He et al. [116] proposed STD2P to boost the performance of semantic segmentation with RGBD data, which considers the features extracted from the image shot in a different view of the same scene. Lin et al. [117] proposed the Cascade Feature Network (CFN), which exploits the depth information to guide the process of the features learning. Lin et al. [118] then extended [117] by introducing the Zig-Zag block for better aggregating the features from different levels. In semantic segmentation, the objects that have similar appearances tend to be misrepresented as one object. Therefore, Wang et al. [119] proposed the depth-aware convolution and pooling, which identify these objects according to the depth channel.

3.1.7. Real-time methods

Although deep-learning-based semantic segmentation methods achieve high accuracy, their huge computational costs hinder the application in some real-time situations. The key challenge is how to greatly enhance the model efficiency while keeping the segmentation accuracy level. We review the real-time semantic segmentation methods from this perspective. The related methods are listed in Table 8.

ENet [122], based on the factorized convolution [123] and early downsampling strategy, is one of the earliest works that focuses on real-time semantic segmentation. Treml et al. [124] proposed using the lightweight encoder [125] and decoder [126] to speed up the process. Romera et al. [127] proposed ERFNet, which combines residual connection with factorized convolution to preserve accuracy under limited computational resources. Wang et al. [128] proposed Driving Importance-weighted Loss (DIL), aiming to alleviate the problem of class imbalance. [49,129–131] adopt the divide-and-conquer strategy, by using two subnetworks to address the tasks of capturing semantics and maintaining spatial details simultaneously. In these methods, the shallow network is used to maintain spatial details, and the deep network with the fast downsampling aims at capturing semantics. In this way, the balance between accuracy and efficiency can be achieved. Li et al. [132] proposed DABNet based on depthwise asymmetric convolution and dilated convolution. Mehta et al. [133] improved the depthwise separable convolution [69] by setting different dilation rates for aggregating the diverse contexts, named as the Efficient Spatial Pyramid (ESP) module. Park et al. [134] introduced Concentrated-Comprehensive Convolution to tackle the limitation of the simple combination of the dilated convolution and the depthwise separable convolution. Nekrasov et al. [48] extended the RefineNet [31] to real-time semantic segmentation by appropriately compressing the framework. Gamal et al. [135] proposed ShuffleSeg based on grouped convolution [136] and channel shuffling [137]. Vallurupalli et al. [138] proposed a novel training strategy. As shown in Fig. 11, it starts with dense convolutions, but gradually converts into grouped convolutions. [139–141] construct the lightweight architecture based on the feature reuse strategy. Notably, [141] has achieved a remarkable balance between accuracy and efficiency.

3.2. Weakly-supervised methods

In fully-supervised methods, pixel-level annotations are indispensable for training the networks. Nevertheless, the annotation process is very demanding and labor-intensive. According to [142], annotating a single image in Cityscapes dataset costs more than 1.5 hours. In this context, it is valuable to conduct research on weakly-supervised semantic segmentation methods. For these methods,

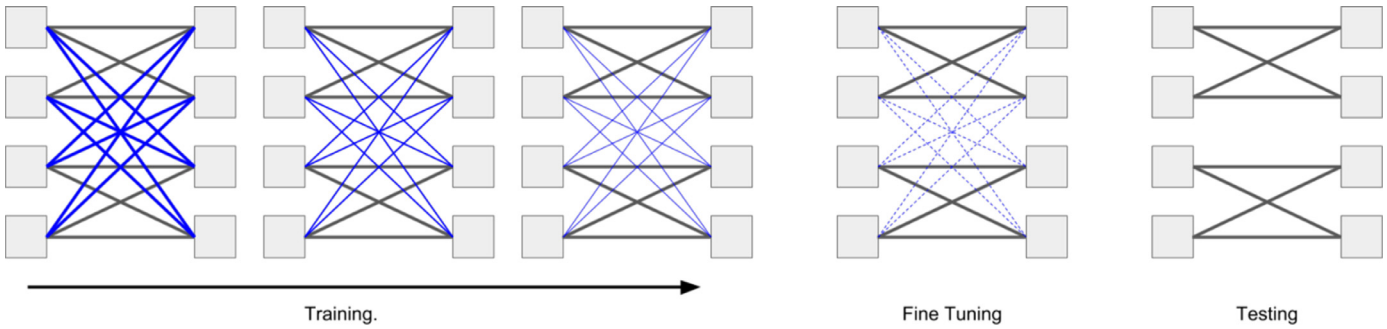


Fig. 11. The training strategy proposed by [138]. Each edge denotes a convolution filter. Particularly, the blue edge represents that the convolution filter is weighted by α which is decreased from 1 to 0 during the training time. The figure is borrowed from [138]

Table 9
Collections for the methods under the supervision of image tags.

Method	Author	Publishing year	Source code	Backbone	Contribution
MIL-FCN[145]	Pathak et al.	2014	N/A	VGG16	Use MIL[165] to realize semantic segmentation under image tags supervision
Aggregation[146]	Pinheiro et al.	2015	N/A	Overleaf[166]	Aggregation layer
CCNN[148]	Pathak et al.	2015	https://github.com/pathak22/ccnn	VGG16	Constrained Convolu-tional Neural Networks
SEC[149]	kolesnikov et al.	2016	https://github.com/kolesman/SEC	DeepLab-CRF-LargeFOV[52]	Seeding, expansion, and constrain-to-boundary loss
STC[150]	Wei et al.	2016	N/A	VGG16	A Simple-to-Complex Framework
AF[151]	Qi et al.	2016	N/A	VGG16 and DeepLabV1[52]	Incorporate semantic segmentation and object localization via introducing proposal aggregation and selection modules
LS[152]	Wei et al.	2016	N/A	DeepLabV1-CRF[52]	Extend [150] via introdu-cing [167] to improve the initial saliency map
AE-PSL[154]	Wei et al.	2017	N/A	DeepLab-CRF-LargeFOV[52]	Propose the Adversarial Erasing method to mine Object Region based on the iterative manner
WCV[164]	Hong et al.	2017	N/A	VGG16	Incorporate image-level with web-crawled videos to train network
WebS[163]	Jin et al.	2017	N/A	VGG16	Propose the three-stage training pipeline
AffinityNet[156]	Ahn et al.	2018	https://github.com/BeautyOfWeb/Affini-tyNet	DeepLabV1[52] or ResNet-38	Propose AffinityNet using semantic affinity to extend the local responses
MCOF[155]	Wang et al.	2018	N/A	VGG16 and ResNet101	MCOF a bottom-up and top-down framework

the key challenge is how to achieve satisfying segmentation accuracy only with weakly supervised annotations. Compared to fully supervised annotations, weakly supervised annotations can be tags, scribbles and bounding boxes, which involve much less manual labeling efforts. In this part, we review these methods according to their supervision types (i.e., image tags, scribbles, and bounding boxes). The key of these methods is fully leveraging the limited supervised supervision for generating acceptable pixelwise masks for training. The mentioned methods are summarized in Table 9, Table 10 and Table 11 according to different supervisions.

3.2.1. Methods based on tag-level supervision

Image tags only provide object categories existing in an image, which can be collected manually or with proper automated annotation techniques [143,144]. However, they only provide supervision to a learning system to a fairly low degree. For example, the output of a segmentation network can be seen as an H $\times$ W $\times$ C probability map (H, W, and C denote height, width, and the number of categories, respectively), while the one-hot encoded ground truth is the size of 1 $\times$ 1 $\times$ C at the tag-level. Therefore, the low dimension of the ground truth poses a challenge for modeling the semantic segmentation task. Methods based on Multi-Instance Learning (MIL), pseudo-supervised learning, and web retrieval are proposed to contend with this problem.

MIL-based method. Pathak et al. [145] proposed to implement MIL with FCN cooperatively via building the multi-class MIL loss. First, the 1 $\times$ 1 $\times$ C global class-aware vector is obtained by pixel-wisely extracting the maximum value along the *C* direction. Then, the MIL loss is built with the obtained vectors by using the cross-entropy function. Different from [145] that uses the max function to extract the class-aware vectors, [146] adopted a more effective smooth version called log-sum-exp (LSE) [147]. In the test time, [146] considered two more priors, i.e., the tag-level prior and smoothing priors, to produce more fine-grained results. In the

Table 10
Collections for the methods based on scribble-level supervision.

Method	Author	Publishing year	Source code	Backbone	Contribution
SSWPS[168]	Bearman et al.	2015	https://github.com/abearman/whats-the-point1	VGG16	Propose the method solv-ing point-level supervised semantic segmentation
ScribbleSup[170]	Lin et al.	2016	N/A	DeepLabV1-MSc-CRF-LargeFOV[52]	Introduce the graphical model to extend the information from the scribbles to unmarked pixels
RAWKS[172]	Vernaza et al.	2017	N/A	ResNet101	Introduce the random walk to propagate the scribbles
GraphNet[171]	Pu et al.	2018	https://github.com/MengyangPu/Grap-hNet	VGG16	Graph Convolutio-nal Network

Table 11
Collections for the methods based on bounding-box-level supervision.

Method	Author	Publishing year	Source code	Backbone	Contribution
WSSL[176]	Papandreou et al.	2015	https://bitbucket.org/deeplab/deeplab-public	DeepLabV1[52]	Expectation-Maximi-zation algorithm
BoxSup[177]	Dai et al.	2015	N/A	DeepLabV1-CRF[52]	Propose an iterative training strategy coope-ratively with region proposal methods
SimpleDoseIt[181]	Khoreva et al.	2017	N/A	DeepLabV2[34]	Introduce GrabCut[174] to produce pseudo mask
BCRMFRGL[182]	Song et al.	2019	N/A	DeepLabV1[52]	Box-driven Classwise Region Masking and Filling Rate Guided Loss

method of [148], Pathak et al. introduce several extra constraints (e.g., suppression constraint, foreground constraint, and size constraint) to train the neural network, apart from the tag information. Although these methods successfully utilize image tags to realize semantic segmentation, their accuracy is still far behind the performance of fully supervised methods.

Pseudo-supervised-based method. With the tags at hand, pseudo-supervised methods aim to generate coarse pixel-level supervision for the segmentation model. The typical method is generally composed of two parts, i.e., the segmentation component and the mask generating component. The former is based on the fully convolutional network. The latter aims to provide a relatively clean mask under the guidance of tag-level supervision. The mask is then used as the pixel-level supervision needed by the training process. Therefore, determining how to obtain an acceptable mask is the keypoint of the research.

In [149], three different kinds of loss functions were proposed, i.e., seeding, expansion and constrain-to-boundary loss, which regularize the network to generate higher-quality masks, and thus improve the segmentation network training. Wei et al. [150] proposed a novel simple-to-complex method. In this method, the initial DCNN and the enhanced DCNN progressively drive a simple saliency map toward a better pixel-level supervision, which is then used in the segmentation network. By introducing a proposal aggregation block, Qi et al. [151] converted the mask generating into the task of regional proposal classification, in which the key idea is aggregating the classified proposals. Wei et al. [152] extended [150] by replacing DRFI [153] with the approach used in [151], which provides a higher-quality initial supervision. In the research of visual classification, we can see that a well-trained network for classification is able to capture salient image regions. To leverage this property, Wei et al. [154] proposed a progressively adversarial erasing learning method, which can be summarized into two procedures. First, the image is fed into the classification network to detect the most discriminating region. Second, the detected region is erased from the raw image. After that, the erased image is fed into the classification network. As the repetitive progress continues, a small salient region gradually grows into a larger meaningful region. Similar to [154], Wang et al. [155] proposed an iterative bottom-up and top-down framework named MCOF, which repeatedly exploits the common object features from the initial salient region. Ahn et al. [156] proposed AffinityNet, which depends on semantic affinity to propagate the local response produced by the classification network.

Web-retrieval-based method. The key to the pseudo-supervised methods is the high-quality mask generation, which is difficult due to the property of some popular datasets. The images in datasets such as PASCAL VOC [157] and ADE20K [158] usually contain objects of multiple classes. In addition, their scenes are often complicated. These factors hinder an effective mask generation for the pseudo-supervised methods. To address this issue, it is helpful to apply the off-the-shelf image retrieval techniques [159–162] for data augmentation. Following this strategy, Jin et al. [163] and Hong et al. [164] additionally collected images or video frames by means of web image retrieval. These images are simple and clean, which is a suitable approach for learning the annotation mask.

3.2.2. Methods based on scribble-level supervision

Despite the feasibility of the tag-based methods, their accuracy is still not satisfying, due to the very limited tag-level information. Scribbles compromise between image tags and pixelwise annotations. Compared to image tags, scribbles use limited pixels to provide location information. Compared to pixelwise annotations, scribbles cost much less manual labeling efforts. To some extent, scribbles can be seen as the combination of image tags and a set of fully annotated pixels. In this part, we mainly review the methods based on two kinds of scribbles, i.e., point scribbles and line scribbles.

Point scribble. As a representative method based on point scribbles, [168] introduced a novel loss function to guide the network training, which is composed of two parts, i.e., the loss for the tag-level inference and the loss calculated from the point scribbles. To produce more accurate results, [168] further incorporates

Table 12
Collections for the methods based on domain adaptation.

Method	Author	Publishing year	Source code	Backbone	Contribution
UDABLS[184]	kamnitsas et al.	2016	https://biomedia.doc.ic.ac.uk/software/deepm-edic/	3D-MSCNN[192]	Introduce an adversari-al training strategy
CDA[190]	Zhang et al.	2017	https://github.com/YangZhang4065/AdaptationSeg	FCN-8s	Propose a curriculum-style learning approach
UDASS[191]	Zou et al.	2018	https://github.com/hfslyc/AdvSemiSeg	FCN-8s	Propose the UDA framework based on the self-training strategy
DT[185]	Huang et al.	2018	https://rsents.github.io/dam.html	ERFNet[127]	Propose an adversarial learning based on multiple discriminators
CGANet[186]	Hong et al.	2018	N/A	FCN-8s	Construct the segmentor based on residual connect-ion.
LFSD[187]	Sankaranar-ayanan et al.	2018	https://goo.gl/3Jsu2s	VGG16	Incorporate the segment-ation network with GAN model
LASOS[188]	Tsai et al.	2018	https://github.com/wasidennis/Adapt-SegNet	DeepLabV2[34]	Propose output space domain adaptation
TCLDS[189]	Luo et al.	2019	https://github.com/RoyalVane/CLAN	DeepLabV2[34]	The method of category-level domain adaptation

a generic objectness prior [169] (i.e., the probability that a pixel belongs to an object) in the loss function.

Line scribble. Most methods based on line scribbles include two components, i.e., the scribble propagation block and the segmentation network. The scribble propagation block aims to propagate the scribbles to other unlabeled pixels, and thus automatically produces full pixel-level annotations, which are used for the segmentation network training. Similar to the pseudo-supervised tag-level methods, the scribble propagation block is the key issue. Lin et al. [170] exploited the graph model to propagate the scribble to the unmarked pixels based on the constraints of spatial, appearance, and semantic context characteristics. Pu et al. [171] introduced the graph convolutional network (GraphNet) model. In the model, the scribbles are first embedded into the graph, which is then fed into the network to produce the pseudo mask. Vernaza et al. [172] introduced the random walk [173] to realize the label propagation.

3.2.3. Methods based on bounding-box-level supervision

Bounding box annotation provides the completed location of a whole object, as well as its semantic tag. The key idea is to extract pseudo masks from the bounding box annotation by adopting an unsupervised segmentation algorithm, such as GrabCut [174] and CRF [175].

Papandreou et al. [176] extracted pseudo masks by applying CRF [175] within bounding boxes. Then, the Expectation-Maximization (EM) algorithm is applied to refine the produced pseudo mask. Dai et al. [177] turned to region proposal methods [178–180] to generate the candidate pseudo mask. Based on this, the segmentation network is trained in an iterative way. Specifically, the segmentation network is first trained under the supervision of candidate masks, and then picks the better masks for the next training iteration. In [181], the pseudo mask is obtained by applying GrabCut [174] to bounding boxes. In addition to using CRF, [182] further proposed the box-driven classwise masking model (BCM), as well as the filling rate guided adaptive loss (FR-Loss), aiming to eliminate the wrongly labeled regions in the pseudo mask.

3.3. Semi-supervised methods

Different from the weakly-supervised situation, the semi-supervised semantic segmentation assumes that there is only a small number of fully annotated training images, instead of a large number of training images with weak supervision. The mentioned methods are summarized in Table 12 and Table 13.

3.3.1. Methods based on domain adaptation

In this part, we focus on the semantic segmentation methods based on the idea of adapting the model trained on the source domain to the target domain. This is a typical semi-supervised scenario, in which the source domain contains a large number of fully annotated data, while the labeled data in the target domain are few or totally unavailable. This research is valuable when we can obtain sufficient labeled photo-realistic synthetic data (e.g., SYNTHIA [183]) as the source domain. In real-world applications, such dataset can be collected at a relatively low cost. The key challenge in this research point is to solve the domain discrepancy between the synthetic and the real data.

Kamnitsas et al. [184] proposed an adversarial training strategy to solve this issue. Its contribution lies in building the Multi-Connected Adversarial Network (MCANet), which consists of the segmentor and the domain discriminator. During the training process, the segmentor aims to provide accurate segmentation and fool the domain discriminator by producing target-like features. The domain discriminator constantly distinguishes whether its input is from the target domain or from the source domain. Some works followed this roadmap and made some further improvements. For example, Huang et al. [185] introduced multiple domain discriminators that are applied on different intermediate layers. The discriminators provide more constraints for the training process while better guiding the optimization process. Hong et al. [186] modify the segmentor by formulating a residual network for enhancing feature representation. Differently, the model proposed in [187] is composed of three parts, i.e., the segmentor, the generator, and discriminator. Based on the observation that output spaces share more similarities and can be easily adapted, Tsai et al. [188] use the adversarial learning for directly adapting the structures of the output space. In [189], Luo et al. pointed out a common problem that in traditional adversarial learning methods, the adaptation is conducted from the global perspective. Therefore, misalignments are unavoidable; e.g., some categories that originally have correct alignment can possibly be adapted in a wrong way, as shown in Fig. 12. To address this issue, they proposed a category-level adversarial network to extend the global adaptation into a class-level adjustment. Zhang et al. [190] and Zou et al. [191] proposed the curriculum domain adaptation based on the iterative training strategy.

Table 13
Collections for the methods based on few-shot learning.

Method	Author	Publishing year	Source code	Backbone	Contribution
OSVOS[193]	Caelles et al.	2017	http://www.vision.ee.ethz.ch/~cvlsegment-ation/osvos/	VGG	The method of one-shot video object segmentation
PL[194]	Dong et al.	2018	N/A	VGG16	N-way k-shot semantic segmentation method
co-FCN[195]	Rakelly et al.	2018	N/A	FCN	Propose co-FCN a conditi-onal network for few-shot semantic segmentation
GuidedNet[196]	Rakelly et al.	2018	http://github.com/shelhamer/revolver	VGG16	Guided Network
OSLSS[197]	Shaban et al.	2017	https://github.com/lzzcd001/OSLSM	VGG16 and FCN-32	Design a two-branch network
SG-one[198]	Zhang et al.	2018	https://github.com/xiaomengyc/SG-One	VGG16	Similarity-guided one-shot semantic segmenta-tion network

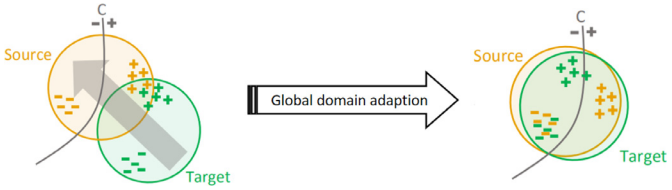


Fig. 12. The misalignments caused by traditional global domain adaptation. The figure is borrowed from [189].

3.3.2. Methods based on few-shot learning

Few-shot semantic segmentation aims to segment objects with only a few fully annotated instances. The most widely adopted technical roadmap is constructing novel structures to subtly leverage extra useful information as much as possible. Caelles et al. [193] combined a pretrained network and a parent network during the training process, which target learning sufficient prior information from the general datasets ImageNet and DAVIS. Dong et al. [194] designed an N-way k-shot semantic segmentation method, which has two branches to cooperate with each other, i.e., a prototype learner and a segmentor. These two branches are fused based on the learned weights. Rakelly et al. [195] proposed co-FCN to solve the FCN fine-tuning issue in the few-shot situation. The co-FCN method is able to align the few-shot training and testing paradigms, which requires no optimization at the testing stage. Rakelly et al. [196] proposed the guided network, which extracts a latent representation and uses it to guide the segmentation process. Shaban et al. [197] also designed a two-branch framework. One branch is used to generate parameter vectors that aims to benefit the other segmentation branch. In addition, meta-learning was also introduced into the framework. Zhang et al. [198] proposed a similarity-guided one-shot semantic segmentation network (SG-One). Its key contribution is to build a novel framework that efficiently learns pixel-level similarity and guides the segmentation network. From the above method, we can see that the various auxiliary structures are designed to provide extra information for the backbone segmentation network, e.g., the latent prior distribution, and enhanced similarity.

4. Datasets and metrics

In this section, we first list several popular semantic segmentation datasets. Then, we introduce the commonly used metrics for evaluating semantic segmentation models.

4.1. Datasets

ADE20K[1] [199]. ADE20K is a standard scene parsing dataset, which contains 20,210 images for training and 2000 images for validation. Currently, the test set has not yet been published. This dataset is challenging, as it involves 150 object categories, including various kinds of objects (e.g., dog, cow, and person) and stuff (e.g., road and sky).

Cityscapes[2] [142]. The Cityscapes dataset, collected from 50 different European cities, is widely used in semantic segmentation. It contains 20,000 coarse-annotated images and 5000 fine-annotated images. The coarse-annotated data are usually used in the pre-training stage to promote the model's generalization. The fine-annotated set is split into 2975, 500, and 1525 for training, validation, and testing, respectively. The ground truth of the test set is withheld by the organizer, and the online evaluation server[3] is provided. Of note, this dataset is also highlighted by its high image resolution (e.g., 2048 $\times$ 1024 resolution).

CamVid[4] [200]. The Camvid dataset contains five different video sequences of driving scenes, in which 701 frames are manually annotated with 32 semantic classes. In [201], Sturgess et al. further split it into the training, validation, and testing set with 367, 100, and 233 images, respectively. In addition, [201] only adopts 11 semantic classes, while the remaining 21 classes are merged as a void.

NYUDv2[5] [202]. The images in the NYUDv2 dataset are captured by a Microsoft Kinect device. This dataset has 40 classes, and contains 1449 RGBD indoor images, in which 795 images are for training and 654 images are for testing. This dataset is a twofold challenge, i.e., various indoor object sizes and the small dataset scale.

PASCAL VOC 2012[6] [157] PASCAL VOC 2012 consists of 1464, 1449, and 1456 images for training, validation, and testing, respectively. It includes 20 object classes and one background class. Similar to Cityscapes, the ground truth of test set is publicly available. The performance evaluation is performed in an online way[7]. An augmented version of PASCAL VOC 2012 [203] is more frequently used, which has 10,582 images in the training set.

[1] https://groups.csail.mit.edu/vision/datasets/ADE20K/
[2] https://www.cityscapes-dataset.com/
[3] https://www.cityscapes-dataset.com/submit/
[4] http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/
[5] https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html
[6] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/
[7] http://host.robots.ox.ac.uk:8080/

PASCAL Context[8] [204]. PASCAL Context is an extension of the PASCAL VOC 2010 segmentation task, which provides additional annotations for the whole scene. In this dataset, there are 10,103 images for training and validation, and 9637 images for testing. This dataset involves more than 400 semantic classes. However, with only 59 semantic classes frequently used, the distribution of the label set is very sparse.

PASCAL Part[9] [205]. PASCAL Part is an extension of the PASCAL VOC 2010 object detection task, which additionally provides the segmentation mask for each body part of the target object. Analogous to PASCAL Context, it also contains 10,103 images for training and validation and 9637 images for testing.

SUNRGBD[10] [206]. SUNRGBD is a large-scale RGBD dataset. It contains 10,335 images in total of 37 categories, of which 5285 images are for training and 5050 images are for testing. Of note, in this dataset, 4943 images are captured by the authors themselves, and the remaining 5392 images are borrowed from NYUDv2[202], B3DO[207], and SUN3D[208]. Due to the possession of this relatively sufficient training data, this dataset is widely used in the task of RGBD semantic segmentation.

SYNTHIA[11] [183]. SYNTHIA is a large-scale dataset for the application of training the autonomous car's vision system. The dataset has over 200,000 virtual-world images, and involves 11 categories, i.e., sky, building, road, sidewalk, fence, vegetation, lane-marking, pole, car, traffic signs, pedestrians, cyclists, and miscellaneous.

4.2. Evaluation metrics

Next, we focus on the evaluation metrics from two perspectives, i.e., accuracy and efficiency. For each perspective, the commonly used metrics are introduced in the following.

4.2.1. Metrics for accuracy

We assume that the category number is $k+1$ in total, including k classes and one background. Y is denoted as the pixel number. For example, Y_{ij} means there are Y_{ij} pixels, which are predicted as the jth class, but actually belong to the ith class. In other words, Y_{ii}, Y_{ij}, and Y_{ji} denote the number of True Positive (TP), False Positive (FP), and False Negative (FN), respectively. In the following, we review five metrics for evaluating semantic segmentation accuracy, including Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), Intersection over Union (IoU), Mean Intersection over Union (MIoU), and Frequency-Weighted Intersection over Union (FWIoU).

Pixel Accuracy (PA). PA denotes the ratio between the number of correctly classified pixels and the total number. In other words, it can be seen as TP/(TP+FN). PA can be obtained based on Eq. 1:

$$PA = \frac{\sum_{i=0}^{k} Y_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} Y_{ij}} \tag{1}$$

Mean Pixel Accuracy (MPA). MPA is an extension of PA, which averages the per-class Pixel Accuracy. The MAP can be obtained by Eq. 2:

$$MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{Y_{ii}}{\sum_{j=0}^{k} Y_{ij}} \tag{2}$$

Intersection over Union (IoU). IoU means the rate between the intersection and union. In other words, it is equal to TP/(TP+FP+FN), which is computed based on Eq. 3:

$$IoU = \frac{\sum_{i=0}^{k} Y_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} Y_{ij} + \sum_{i=0}^{k}\sum_{j=0}^{k} Y_{ji} - \sum_{i=0}^{k} Y_{ii}} \tag{3}$$

Mean Intersection over Union (MIoU). MIoU denotes the average of the pre-class IoU, as shown in Eq. 4:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{Y_{ii}}{\sum_{j=0}^{k} Y_{ij} + \sum_{j=0}^{k} Y_{ji} - Y_{ii}} \tag{4}$$

Frequency-Weighted Intersection over Union (FWIoU). FWIoU is an extension of MIoU, which uses the occurrence frequency to adjust the importance of each class, as shown in Eq. 5:

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

4.2.2. Metrics for efficiency

In semantic segmentation, it is also important to evaluate the model's efficiency for some real-time applications.

Model complexity. Model parameters and Floating Point Operations (FLOPs) are two metrics that are widely used for measuring model complexity. Models with large model parameters and FLOPs [14,77,122] tend to have low implementation efficiency. Of note, these two metrics are independent of the implementation environment.

Implementation speed. Implementation speed is another important aspect to measure the model's efficiency. Runtime and Frame Per Second (FPS) are widely used. However, these two metrics depend on the hardware and software environment.

5. Common challenging issues and growing research directions

5.1. Common challenging issues

Balance between accuracy and efficiency. Accuracy and efficiency are both critical for evaluating a semantic segmentation method. However, the gains in these two aspects are still contradictory to each other for all the current semantic segmentation methods. In this situation, the models with high accuracy (e.g., PSPNet [30] and DeepLab [34]) tend to have a low efficiency. Correspondingly, the models with good efficiency (e.g., SegNet [25] and ENet [122]) fail to provide sufficiently accurate segmentation results. To our best knowledge, the DFANet method [141] currently performs the best in achieving the balance between accuracy and efficiency. Nevertheless, there is still a gap in accuracy between DFANet and the other top-performance methods.

Dependency on high-quality training data. For accurate semantic segmentation, high-quality training data are considered to be prerequisite. However, obtaining high-quality training data, i.e., sufficient labeled images with pixel-level annotation, is an inevitably laborious and time-consuming task. This heavy dependency has become another common challenge of semantic segmentation. Aiming at alleviating this issue, various subtle weakly- and semi-supervised semantic segmentation methods have been designed under the assumption that only low-quality training data are available. However, compared to fully supervised methods, the performances of the weakly- and semi-supervised methods are still far from satisfactory, which indicates the need for further study.

Domain gap across different datasets. A domain gap widely exists in semantic segmentation and other vision tasks [209]. For example, PSPNet [30] achieves 85.4% MIoU on the PASCAL VOC 2012 dataset. However, it only achieves 44.94% MIoU on the ADE20K dataset, which is more a challenging dataset with 150 semantic classes and complex scenes. Generally, since various datasets are designed for different applications, they can be different from each other in terms of category number, scene appearance, dataset size, object size, and so on. In this situation, the

[8] https://cs.stanford.edu/~roozbeh/pascal-context/
[9] https://www.cs.stanford.edu/~roozbeh/pascal-parts/pascal-parts.html
[10] http://rgbd.cs.princeton.edu/
[11] http://synthia-dataset.net/

differences contribute considerably to the gap between heterogeneous domains. Therefore, researchers and developers are encouraged to consider the domain gap issue when applying semantic segmentation techniques in vision applications.

5.2. Growing directions

Real-time semantic segmentation. A large number of methods aiming at real-time semantic segmentation have been proposed. Although impressive performance has been achieved in terms of both accuracy and efficiency, there still exists a large space for improvement. Taking the state-of-the-art DFANet [141] as an example, DFANet is still 10% lower in terms of MIoU than PSPNet in Cityscapes. Therefore, the Pareto improvement on accuracy and speed is expected in this research direction.

Unsupervised segmentation Of note, there have been some initial attempts [210,211] on the task of unsupervised deep image segmentation. These methods pave the way for incorporating different deep neural network structures with well-developed unsupervised learning frameworks, such as spectral clustering [212,213] and subspace learning [214]. Another crucial factor in this research direction is the refinement of initially segmented results. It is a non-trivial task, as the semantic gap still exists between the segmented regions.

Occluded objects segmentation. Regarding human vision, a person can quickly identify and recover the occluded parts. However, this process cannot be well imitated by current computer vision systems. One important reason is that most current segmentation algorithms aim at a hard partition without the ability to transfer knowledge. Solving this issue is non-trivial. Ehsani et al. [215] constructed a dataset called DYCE[12], and proposed the SeGAN method, in which the occluded parts are synthesized from GAN. Purkait et al. [216] proposed to realize this task based on depth information, and they constructed data for this task based the SUNCG dataset [217]. Of note, PASCAL3D+[13] [218] is another available dataset that can be used in the study of this field.

Instance or panoptic semantic segmentation. Instance or panoptic semantic segmentation can be seen as a fine-grained extension of traditional semantic segmentation. These two tasks are more challenging, but they are also potentially very useful in various real-world applications. The most representative work about instance segmentation is Mask-RCNN [37] proposed by He et al. in 2017. Panoptic segmentation is a new task introduced by Kirillov et al. [22] in 2019. Due to the short emergence time and the great research value, this task deserves an in-depth investigation in the future.

Video semantic segmentation. Currently, the main direction of semantic segmentation focuses on the single-image level. However, the real-world applications for the visual recognition and understanding (e.g., self-driving cars) are usually based on video sequences, where image frames are highly correlated to each other. Therefore, in video semantic segmentation, the mainstream methods utilize the interframe correlation to improve accuracy (e.g., [219,220]). Compared to image semantic segmentation, video semantic segmentation potentially has more practical significance.

6. Conclusion

In this paper, we briefly review the deep-learning-based semantic segmentation methods from a different perspective, which are divided according to the supervision level. For each reviewed method, we provide a detailed record its publishing year, the code's URL, and backbone, etc. We also discuss the common challenges and several possible directions in this field. We hope this survey can provide readers with a sketch of the research on semantic segmentation in the deep learning era. Of note, due to the limited space and authors' knowledge, we have not included all of the novel semantic segmentation methods in this paper, such as the 3D mesh and point cloud segmentation methods.

[12] https://homes.cs.washington.edu/~kianae/weights_segan_cvpr18.tar.gz

[13] http://cs.stanford.edu/cs/cvgl/PASCAL3D+_release1.1.zip

Declaration of Competing Interest

The authors declare that they have no known competing financialinterestsor personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The research was supported by the National Key R&D Program of China under Grant No. 2017YFC0820604, Anhui Provincial Natural Science Foundation under Grant No. 1808085QF188, and National Nature Science Foundation of China under Grant Nos. 61702156, 61772171 and 61876056.

References

[1] B. Li, S. Liu, W. Xu, W. Qiu, Real-time object detection and semantic segmentation for autonomous driving, in: Proceedings of the Multispectral Image Processing and Pattern Recognition: Automatic Target Recognition and Navigation, 10608, International Society for Optics and Photonics, 2018, p. 106080P.
[2] Y.-H. Tseng, S.-S. Jan, Combination of computer vision detection and segmentation for autonomous driving, in: 2018 IEEE/ION Position, Location and Navigation Symposium, IEEE, 2018, pp. 1047–1052.
[3] F. Flohr, D. Gavrila, et al., Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues., BMVC, 2013.
[4] G. Brazil, X. Yin, X. Liu, Illuminating pedestrians via simultaneous detection & segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4950–4959.
[5] X. Tao, D. Zhang, W. Ma, X. Liu, D. Xu, Automatic metallic surface defect detection and recognition with convolutional neural networks, Applied Sciences 8 (9) (2018) 1575.
[6] Z. Wang, L. Wei, L. Wang, Y. Gao, W. Chen, D. Shen, Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning, IEEE Transactions on Image Processing 27 (2) (2018) 923–937.
[7] Y. Guo, Y. Gao, D. Shen, Deformable MR prostate segmentation via deep feature learning and sparse patch matching, IEEE Trans. Med. Imaging 35 (4) (2016) 1077–1089.
[8] X. Zhu, H. Suk, S. Lee, D. Shen, Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification, IEEE Trans. Biomed. Engineering 63 (3) (2016) 607–618.
[9] X. Zhu, H. Suk, D. Shen, A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis, NeuroImage 100 (2014) 91–105.
[10] F. Schroff, A. Criminisi, A. Zisserman, Object class segmentation using random forests, in: Proceedings of the British Machine Vision Conference, 2008, pp. 54.1–54.10.
[11] Y. Gao, Y. Shao, J. Lian, A.Z. Wang, R.C. Chen, D. Shen, Accurate segmentation of CT male pelvic organs via regression-based deformable models and multi-task random forests, IEEE Trans. Med. Imaging 35 (6) (2016) 1532–1543.
[12] F. Han, S.-C. Zhu, Bottom-up/top-down image parsing with attribute grammar, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (1) (2009) 59–73.
[13] S.-C. Zhu, D. Mumford, et al., A stochastic grammar of images, Foundations and Trends® in Computer Graphics and Vision 2 (4) (2007) 259–362.
[14] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[15] N. Wadhwa, R. Garg, D.E. Jacobs, B.E. Feldman, N. Kanazawa, R. Carroll, Y. Movshovitz-Attias, J.T. Barron, Y. Pritch, M. Levoy, Synthetic depth-of-field with a single-camera mobile phone, ACM Transactions on Graphics 37 (4) (2018) 64.
[16] M. Thoma, A survey of semantic segmentation, arXiv preprint arXiv:1602.06541 (2016).
[17] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, J. Garcia-Rodriguez, A review on deep learning techniques applied to semantic segmentation, arXiv preprint arXiv:1704.06857 (2017).
[18] B. Zhao, J. Feng, X. Wu, S. Yan, A survey on deep learning-based fine-grained object classification and semantic segmentation, International Journal of Automation and Computing 14 (2) (2017) 119–135.
[19] Y. Guo, Y. Liu, T. Georgiou, M.S. Lew, A review of semantic segmentation using deep neural networks, International Journal of Multimedia Information Retrieval 7 (2) (2018) 87–93.

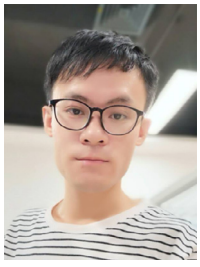
[20] Q. Geng, Z. Zhou, X. Cao, Survey of recent progress in semantic image segmentation with cnns, Science China Information Sciences 61 (5) (2018) 051101.
[21] F. Lateef, Y. Ruichek, Survey on semantic segmentation using deep learning techniques, Neurocomputing 338 (2019) 321–348.
[22] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, Panoptic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9404–9413.
[23] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[24] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[25] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (12) (2017) 2481–2495.
[26] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122 (2015).
[27] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, Y. Bengio, Renet: A recurrent neural network based alternative to convolutional networks, arXiv preprint arXiv:1505.00393 (2015).
[28] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, A. Courville, Reseg: A recurrent neural network-based model for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 41–48.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[30] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[31] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1925–1934.
[32] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters–improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4353–4361.
[33] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7151–7160.
[34] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (4) (2017) 834–848.
[35] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587 (2017).
[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 801–818.
[37] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[38] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[39] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, Denseaspp for semantic segmentation in street scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3684–3692.
[40] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, H. Lu, Stacked deconvolutional network for semantic segmentation, IEEE Transactions on Image Processing (2019). 1–1
[41] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, Y. Bengio, The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 11–19.
[42] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.
[43] P. Bilinski, V. Prisacariu, Dense decoder shortcut connections for single-pass semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6596–6605.
[44] Z. Zhang, X. Zhang, C. Peng, X. Xue, J. Sun, Exfuse: Enhancing feature fusion for semantic segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 269–284.
[45] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).
[46] J. Xie, B. Shuai, J.-F. Hu, J. Lin, W.-S. Zheng, Improving fast segmentation with teacher-student learning, arXiv preprint arXiv:1810.08476 (2018).
[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[48] V. Nekrasov, C. Shen, I. Reid, Light-weight refinenet for real-time semantic segmentation, arXiv preprint arXiv:1810.03272 (2018).
[49] R.P. Poudel, S. Liwicki, R. Cipolla, Fast-scnn: fast semantic segmentation network, arXiv preprint arXiv:1902.04502 (2019).
[50] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, arXiv preprint arXiv:1905.02244 (2019).
[51] W. Liu, A. Rabinovich, A.C. Berg, Parsenet: Looking wider to see better, arXiv preprint arXiv:1506.04579 (2015).
[52] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062 (2014).
[53] M. Mostajabi, P. Yadollahpour, G. Shakhnarovich, Feedforward semantic segmentation with zoom-out features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3376–3385.
[54] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3194–3203.
[55] W.-C. Hung, Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, M.-H. Yang, Scene parsing with global context embedding, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2631–2639.
[56] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, G. Wang, Context contrasted feature and gated multi-scale aggregation for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2393–2402.
[57] Y. Yuan, J. Wang, Ocnet: Object context network for scene parsing, arXiv preprint arXiv:1809.00916 (2018).
[58] Y. Zhuang, F. Yang, L. Tao, C. Ma, Z. Zhang, Y. Li, H. Jia, X. Xie, W. Gao, Dense relation network: Learning consistent and context-aware representation for semantic image segmentation, in: 2018 25th IEEE International Conference on Image Processing, IEEE, 2018, pp. 3698–3702.
[59] Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, Pattern Recognition 90 (2019) 119–133.
[60] H. Zhang, H. Zhang, C. Wang, J. Xie, Co-occurrent features in semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 548–557.
[61] H. Ding, X. Jiang, B. Shuai, A.Q. Liu, G. Wang, Semantic correlation promoted shape-variant context for segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8885–8894.
[62] A. Lucchi, Y. Li, X. Boix, K. Smith, P. Fua, Are spatial and global constraints really necessary for segmentation? in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 9–16.
[63] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.
[64] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 447–456.
[65] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in: European Conference on Computer Vision, Springer, 2014, pp. 297–312.
[66] T.M. Quan, D.G. Hildebrand, W.-K. Jeong, Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics, arXiv preprint arXiv:1612.05360 (2016).
[67] A. Valada, G. Oliveira, T. Brox, W. Burgard, Towards robust semantic segmentation using deep fusion, Robotics: Science and Systems Workshop, Are the Sceptics Right? Limits and Potentials of Deep Learning in Robotics, 2016.
[68] T. Pohlen, A. Hermans, M. Mathias, B. Leibe, Full-resolution residual networks for semantic segmentation in street scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4151–4160.
[69] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[70] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes, IEEE transactions on medical imaging 37 (12) (2018) 2663–2674.
[71] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Learning a discriminative feature network for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1857–1866.
[72] S. Shah, P. Ghosh, L.S. Davis, T. Goldstein, Stacked u-nets: a no-frills approach to natural image segmentation, arXiv preprint arXiv:1804.10343 (2018).
[73] T. Yang, Y. Wu, J. Zhao, L. Guan, Semantic segmentation via highly fused convolutional network with multiple soft cost functions, Cognitive Systems Research 53 (2019) 20–30.
[74] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, C. Pal, The importance of skip connections in biomedical image segmentation, in: Deep Learning and Data Labeling for Medical Applications, Springer, 2016, pp. 179–187.
[75] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer, 2016, pp. 424–432.
[76] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.

[77] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[78] A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, arXiv preprint arXiv:1511.02680 (2015).
[79] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, C. Wolf, Residual conv-deconv grid network for semantic segmentation, arXiv preprint arXiv:1707.07958 (2017).
[80] H. Sak, A. Senior, F. Beaufays, Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, arXiv preprint arXiv:1402.1128 (2014).
[81] R. Messina, J. Louradour, Segmentation-free handwritten chinese text recognition with lstm-rnn, in: 2015 13th International Conference on Document Analysis and Recognition, IEEE, 2015, pp. 171–175.
[82] P.H. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene labeling, in: 31st International Conference on Machine Learning, 2014. CONF
[83] R.P. Poudel, P. Lamata, G. Montana, Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation, in: Reconstruction, segmentation, and analysis of medical images, Springer, 2016, pp. 83–94.
[84] W. Byeon, T.M. Breuel, F. Raue, M. Liwicki, Scene labeling with lstm recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3547–3555.
[85] B. Shuai, Z. Zuo, B. Wang, G. Wang, Dag-recurrent neural networks for scene labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3620–3629.
[86] H. Fan, H. Ling, Dense recurrent neural networks for scene labeling, arXiv preprint arXiv:1801.06831 (2018).
[87] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).
[88] H. Fan, X. Mei, D. Prokhorov, H. Ling, Multi-level contextual rnns with attention model for scene labeling, IEEE Transactions on Intelligent Transportation Systems 19 (11) (2018) 3475–3485.
[89] M.D. Zeiler, G.W. Taylor, R. Fergus, et al., Adaptive deconvolutional networks for mid and high level feature learning., in: Proceedings of the International Conference on Computer Vision, 1, 2011, p. 6.
[90] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[91] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, arXiv preprint arXiv:1605.05396 (2016).
[92] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[93] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5505–5514.
[94] P. Luc, C. Couprie, S. Chintala, J. Verbeek, Semantic segmentation using adversarial networks, arXiv preprint arXiv:1611.08408 (2016).
[95] P. Moeskops, M. Veta, M.W. Lafarge, K.A. Eppenhof, J.P. Pluim, Adversarial training and dilated convolutions for brain mri segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2017, pp. 56–64.
[96] M. Rezaei, K. Harmuth, W. Gierke, T. Kellermeier, M. Fischer, H. Yang, C. Meinel, A conditional adversarial network for semantic segmentation of brain tumor, in: International MICCAI Brainlesion Workshop, Springer, 2017, pp. 241–252.
[97] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).
[98] Y. Xue, T. Xu, H. Zhang, L.R. Long, X. Huang, Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation, Neuroinformatics 16 (3-4) (2018) 383–392.
[99] M. Rezaei, H. Yang, C. Meinel, Conditional generative refinement adversarial networks for unbalanced medical image semantic segmentation, arXiv preprint arXiv:1810.03871 (2018).
[100] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, Y. Yang, Macro-micro adversarial network for human parsing, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 418–434.
[101] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2012) 1915–1929.
[102] C. Li, M. Wand, Precomputed real-time texture synthesis with markovian generative adversarial networks, in: European Conference on Computer Vision, Springer, 2016, pp. 702–716.
[103] C. Couprie, C. Farabet, L. Najman, Y. LeCun, Indoor semantic segmentation using depth information, arXiv preprint arXiv:1301.3572 (2013).
[104] N. Höft, H. Schulz, S. Behnke, Fast semantic segmentation of rgb-d scenes with gpu-accelerated deep neural networks, in: Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), Springer, 2014, pp. 80–85.
[105] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 2005, pp. 886–893vol. 1.
[106] L. Spinello, K.O. Arras, People detection in rgb-d data, in: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2011, pp. 3838–3843.
[107] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from rgb-d images for object detection and segmentation, in: European Conference on Computer Vision, Springer, 2014, pp. 345–360.
[108] Y. Guo, T. Chen, Semantic segmentation of rgbd images based on deep depth regression, Pattern Recognition Letters 109 (2018) 55–64.
[109] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, in: Asian conference on computer vision, Springer, 2016, pp. 213–228.
[110] J. Wang, Z. Wang, D. Tao, S. See, G. Wang, Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks, in: European Conference on Computer Vision, Springer, 2016, pp. 664–679.
[111] X. Qi, R. Liao, J. Jia, S. Fidler, R. Urtasun, 3d graph neural networks for rgbd semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5199–5208.
[112] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, L. Lin, Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling, in: European Conference on Computer Vision, Springer, 2016, pp. 541–557.
[113] Y. Cheng, R. Cai, Z. Li, X. Zhao, K. Huang, Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3029–3037.
[114] X. Hu, K. Yang, L. Fei, K. Wang, Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation, arXiv preprint arXiv:1905.10089 (2019).
[115] S.-J. Park, K.-S. Hong, S. Lee, Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4980–4989.
[116] Y. He, W.-C. Chiu, M. Keuper, M. Fritz, Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4837–4846.
[117] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, H. Huang, Cascaded feature network for semantic segmentation of rgb-d images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1311–1319.
[118] D. Lin, H. Huang, Zig-zag network for semantic segmentation of rgb-d images, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019). 1–1
[119] W. Wang, U. Neumann, Depth-aware cnn for rgb-d segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 135–150.
[120] H. Schulz, S. Behnke, Learning object-class segmentation with convolutional neural networks, in: Proceedings of the European Symposium on Artificial Neural Networks, 2012.
[121] Y. Luo, H. Jiao, L. Qi, J. Dong, S. Zhang, H. Yu, Augmenting depth estimation from deep convolutional neural network using multi-spectral photometric stereo, in: 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017, pp. 1–6.
[122] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, arXiv preprint arXiv:1606.02147 (2016).
[123] J. Jin, A. Dundar, E. Culurciello, Flattened convolutional neural networks for feedforward acceleration, arXiv preprint arXiv:1412.5474 (2014).
[124] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich, et al., Speeding up semantic segmentation for autonomous driving, in: MLITS, NIPS Workshop, 2, 2016, p. 7.
[125] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size, arXiv preprint arXiv:1602.07360 (2016).
[126] P.O. Pinheiro, T.-Y. Lin, R. Collobert, P. Dollár, Learning to refine object segments, in: European Conference on Computer Vision, Springer, 2016, pp. 75–91.
[127] E. Romera, J.M. Alvarez, L.M. Bergasa, R. Arroyo, Erfnet: Efficient residual factorized convnet for real-time semantic segmentation, IEEE Transactions on Intelligent Transportation Systems 19 (1) (2017) 263–272.
[128] W. Wang, Z. Pan, Dsnet for real-time driving scene semantic segmentation, arXiv preprint arXiv:1812.07049 (2018).
[129] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 325–341.
[130] Z. Wu, C. Shen, A.v.d. Hengel, Real-time semantic image segmentation via spatial sparsity, arXiv preprint arXiv:1712.00213 (2017).
[131] R.P. Poudel, U. Bonde, S. Liwicki, C. Zach, Contextnet: Exploring context and detail for semantic segmentation in real-time, arXiv preprint arXiv:1805.04554 (2018).
[132] G. Li, I. Yun, J. Kim, J. Kim, Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation, arXiv preprint arXiv:1907.11357 (2019).
[133] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, H. Hajishirzi, Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 552–568.
[134] H. Park, L.L. Sjösund, Y. Yoo, N. Kwak, Extremec3net: Extreme lightweight portrait segmentation networks using advanced c3-modules, arXiv preprint arXiv:1908.03093 (2019).
[135] M. Gamal, M. Siam, M. Abdel-Razek, Shuffleseg: Real-time semantic segmentation network, arXiv preprint arXiv:1803.03816 (2018).

[136] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[137] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.
[138] N. Vallurupalli, S. Annamaneni, G. Varma, C. Jawahar, M. Mathew, S. Nagori, Efficient semantic segmentation using gradual grouping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 598–606.
[139] A. Chaurasia, E. Culurciello, Linknet: Exploiting encoder representations for efficient semantic segmentation, in: 2017 IEEE Visual Communications and Image Processing, IEEE, 2017, pp. 1–4.
[140] H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, Icnet for real-time semantic segmentation on high-resolution images, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 405–420.
[141] H. Li, P. Xiong, H. Fan, J. Sun, Dfanet: Deep feature aggregation for real-time semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9522–9531.
[142] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[143] M. Wang, B. Ni, X. Hua, T. Chua, Assistive tagging: A survey of multimedia tagging with human-computer joint exploration, ACM Comput. Surv. 44 (4) (2012) 25:1–25:24.
[144] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, X. Wu, Image annotation by multiple-instance learning with discriminative feature mapping and selection, IEEE Trans. Cybernetics 44 (5) (2014) 669–680.
[145] D. Pathak, E. Shelhamer, J. Long, T. Darrell, Fully convolutional multi-class multiple instance learning, arXiv preprint arXiv:1412.7144 (2014).
[146] P.O. Pinheiro, R. Collobert, From image-level to pixel-level labeling with convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1713–1721.
[147] S. Boyd, L. Vandenberghe, Convex optimization, Cambridge university press, 2004.
[148] D. Pathak, P. Krahenbuhl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1796–1804.
[149] A. Kolesnikov, C.H. Lampert, Seed, expand and constrain: Three principles for weakly-supervised image segmentation, in: European Conference on Computer Vision, Springer, 2016, pp. 695–711.
[150] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, S. Yan, Stc: A simple to complex framework for weakly-supervised semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (11) (2016) 2314–2320.
[151] X. Qi, Z. Liu, J. Shi, H. Zhao, J. Jia, Augmented feedback in semantic segmentation under image level supervision, in: European Conference on Computer Vision, Springer, 2016, pp. 90–105.
[152] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, S. Yan, Learning to segment with image-level annotations, Pattern Recognition 59 (2016) 234–244.
[153] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: A discriminative regional feature integration approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2083–2090.
[154] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, S. Yan, 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, 2017, pp. 1568–1576.
[155] X. Wang, S. You, X. Li, H. Ma, Weakly-supervised semantic segmentation by iteratively mining common object features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1354–1362.
[156] J. Ahn, S. Kwak, 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, 2018, pp. 4981–4990.
[157] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2) (2010) 303–338.
[158] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 633–641.
[159] M. Wang, H. Li, D. Tao, K. Lu, X. Wu, Multimodal graph-based reranking for web image search, IEEE Transactions on Image Processing 21 (11) (2012) 4649–4661.
[160] X. Zhu, X. Li, S. Zhang, Z. Xu, L. Yu, C. Wang, Graph pca hashing for similarity search, IEEE Transactions on Multimedia 19 (9) (2017) 2033–2044.
[161] Y. Hao, T. Mu, R. Hong, M. Wang, N. An, J.Y. Goulermas, Stochastic multiview hashing for large-scale near-duplicate video retrieval, IEEE Transactions on Multimedia 19 (1) (2016) 1–14.
[162] Y. Hao, T. Mu, J.Y. Goulermas, J. Jiang, R. Hong, M. Wang, Unsupervised t-distributed video hashing and its deep hashing extension, IEEE Transactions on Image Processing 26 (11) (2017) 5531–5544.
[163] B. Jin, M.V. Ortiz Segovia, S. Susstrunk, Webly supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3626–3635.
[164] S. Hong, D. Yeo, S. Kwak, H. Lee, B. Han, Weakly supervised semantic segmentation using web-crawled videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7322–7330.
[165] T.G. Dietterich, R.H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial intelligence 89 (1-2) (1997) 31–71.
[166] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv preprint arXiv:1312.6229 (2013).
[167] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, S. Yan, Cnn: Single-label to multi-label, arXiv preprint arXiv:1406.5726 (2014).
[168] A. Bearman, O. Russakovsky, V. Ferrari, L. Fei-Fei, Whats the point: Semantic segmentation with point supervision, in: European Conference on Computer Vision, Springer, 2016, pp. 549–565.
[169] B. Alexe, T. Deselaers, V. Ferrari, Measuring the objectness of image windows, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11) (2012) 2189–2202.
[170] D. Lin, J. Dai, J. Jia, K. He, J. Sun, Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3159–3167.
[171] M. Pu, Y. Huang, Q. Guan, Q. Zou, Graphnet: Learning image pseudo annotations for weakly-supervised semantic segmentation, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM, 2018, pp. 483–491.
[172] P. Vernaza, M. Chandraker, Learning random-walk label propagation for weakly-supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7158–7166.
[173] L. Grady, Random walks for image segmentation, IEEE Transactions on Pattern Analysis & Machine Intelligence (11) (2006) 1768–1783.
[174] C. Rother, V. Kolmogorov, A. Blake, Grabcut: Interactive foreground extraction using iterated graph cuts, in: ACM transactions on graphics (TOG), 23, ACM, 2004, pp. 309–314.
[175] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with gaussian edge potentials, in: Advances in neural information processing systems, 2011, pp. 109–117.
[176] G. Papandreou, L.-C. Chen, K.P. Murphy, A.L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1742–1750.
[177] J. Dai, K. He, J. Sun, Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1635–1643.
[178] J. Carreira, C. Sminchisescu, Cpmc: Automatic object segmentation using constrained parametric min-cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (7) (2011) 1312–1328.
[179] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, International journal of computer vision 104 (2) (2013) 154–171.
[180] P. Arbeláez, J. Pont-Tuset, J.T. Barron, F. Marques, J. Malik, Multiscale combinatorial grouping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 328–335.
[181] A. Khoreva, R. Benenson, J. Hosang, M. Hein, B. Schiele, Simple does it: Weakly supervised instance and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 876–885.
[182] C. Song, Y. Huang, W. Ouyang, L. Wang, Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3136–3145.
[183] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A.M. Lopez, The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3234–3243.
[184] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al., Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: International conference on information processing in medical imaging, Springer, 2017, pp. 597–609.
[185] H. Huang, Q. Huang, P. Krahenbuhl, Domain transfer through deep activation matching, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 590–605.
[186] W. Hong, Z. Wang, M. Yang, J. Yuan, Conditional generative adversarial network for structured domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1335–1344.
[187] S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, R. Chellappa, Learning from synthetic data: Addressing domain shift for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3752–3761.
[188] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, M. Chandraker, Learning to adapt structured output space for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7472–7481.
[189] Y. Luo, L. Zheng, T. Guan, J. Yu, Y. Yang, Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2507–2516.

[190] Y. Zhang, P. David, B. Gong, Curriculum domain adaptation for semantic segmentation of urban scenes, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2020–2030.
[191] Y. Zou, Z. Yu, B. Vijaya Kumar, J. Wang, Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 289–305.
[192] K. Kamnitsas, C. Ledig, V.F. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation, Medical image analysis 36 (2017) 61–78.
[193] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. Van Gool, One-shot video object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 221–230.
[194] N. Dong, E. Xing, Few-shot semantic segmentation with prototype learning., in: Proceedings of the British Machine Vision Conference, 1, 2018, p. 6.
[195] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, S. Levine, Conditional networks for few-shot semantic segmentation, in: Proceedings of the British Machine Vision Conference, 2018.
[196] K. Rakelly, E. Shelhamer, T. Darrell, A.A. Efros, S. Levine, Few-shot segmentation propagation with guided networks, arXiv preprint arXiv:1806.07373 (2018).
[197] A. Shaban, S. Bansal, Z. Liu, I. Essa, B. Boots, One-shot learning for semantic segmentation, arXiv preprint arXiv:1709.03410 (2017).
[198] X. Zhang, Y. Wei, Y. Yang, T. Huang, Sg-one: Similarity guidance network for one-shot semantic segmentation, arXiv preprint arXiv:1810.09091 (2018).
[199] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ade20k dataset, International Journal of Computer Vision 127 (3) (2019) 302–321.
[200] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: A high--definition ground truth database, Pattern Recognition Letters 30 (2) (2009) 88–97.
[201] P. Sturgess, K. Alahari, L. Ladicky, P.H. Torr, Combining appearance and structure from motion features for road scene understanding, in: Proceedings of the British Machine Vision Conference, 2009.
[202] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[203] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 991–998.
[204] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 891–898.
[205] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, A. Yuille, Detect what you can: Detecting and representing objects using holistic models and body parts, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1971–1978.
[206] S. Song, S.P. Lichtenberg, J. Xiao, Sun rgb-d: A rgb-d scene understanding benchmark suite, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 567–576.
[207] A. Janoch, S. Karayev, Y. Jia, J.T. Barron, M. Fritz, K. Saenko, T. Darrell, A category-level 3d object dataset: Putting the kinect to work, in: Consumer depth cameras for computer vision, Springer, 2013, pp. 141–165.
[208] J. Xiao, A. Owens, A. Torralba, Sun3d: A database of big spaces reconstructed using sfm and object labels, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1625–1632.
[209] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153.
[210] X. Xia, B. Kulis, W-net: A deep model for fully unsupervised image segmentation, arXiv preprint arXiv:1711.08506 (2017).
[211] X. Ji, J.F. Henriques, A. Vedaldi, Invariant information distillation for unsupervised image segmentation and clustering, arXiv preprint arXiv:1807.06653 (2018).
[212] X. Zhu, S. Zhang, R. Hu, W. He, C. Lei, P. Zhu, One-step multi-view spectral clustering, IEEE Transactions on Knowledge and Data Engineering 31 (10) (2019) 2022–2034.
[213] X. Zhu, S. Zhang, Y. Li, J. Zhang, L. Yang, Y. Fang, Low-rank sparse subspace for spectral clustering, IEEE Transactions on Knowledge and Data Engineering 31 (8) (2019) 1532–1543.
[214] T. Mu, J.Y. Goulermas, S. Ananiadou, Data visualization with structural control of global cohort and local data neighborhoods, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (6) (2017) 1323–1337.
[215] K. Ehsani, R. Mottaghi, A. Farhadi, Segan: Segmenting and generating the invisible, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6144–6153.
[216] P. Purkait, C. Zach, I. Reid, Seeing behind things: Extending semantic segmentation to occluded regions, arXiv preprint arXiv:1906.02885 (2019).
[217] S. Song, F. Yu, A. Zeng, A.X. Chang, M. Savva, T. Funkhouser, Semantic scene completion from a single depth image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1746–1754.
[218] Y. Xiang, R. Mottaghi, S. Savarese, Beyond pascal: A benchmark for 3d object detection in the wild, in: IEEE Winter Conference on Applications of Computer Vision, IEEE, 2014, pp. 75–82.
[219] A. Kundu, V. Vineet, V. Koltun, Feature space optimization for semantic video segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3168–3175.
[220] M. Fayyaz, M.H. Saffar, M. Sabokrou, M. Fathy, F. Huang, R. Klette, Stfcn: spatio-temporal fully convolutional neural network for semantic segmentation of street scenes, in: Asian Conference on Computer Vision, Springer, 2016, pp. 493–509.

Dr. Shijie Hao is an associate professor at Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, and he is also with School of Computer and Information, Hefei University of Technology (HFUT). He received his Ph.D. degree at HFUT in 2012. His research interests include image processing and multimedia content analysis.



Yuan Zhou received his B.E. from Anhui Polytechnic University, Wuhu, in 2017. He is currently a master student of Department of Computer Science and Engineering at Hefei University of Technology, Hefei,China. His current research interest is computer version.

Dr. Yanrong Guo is an associate professor at Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, and she is also with School of Computer and Information, Hefei University of Technology (HFUT). She received her Ph.D. degree at HFUT in 2013. Her research interests include computer vision and pattern recognition.