Neurocomputing 493 (2022) 626–646

Contents lists available at ScienceDirect

Neurocomputing

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

Review the state-of-the-art technologies of semantic segmentation based on deep learning

Yujian Mo, Yan Wu *, Xinneng Yang, Feilin Liu, Yujun Liao

College of electronic and Information Engineering, Tongji University, Shanghai 201804, China

ARTICLE INFO

Article history:
Received 8 April 2021
Revised 23 December 2021
Accepted 2 January 2022
Available online 11 January 2022

Keywords:
Deep learning
Convolutional neural networks
Semantic segmentation
Real-time
Domain adaptation
Multi-modal fusion
Weakly-supervised

ABSTRACT

The goal of semantic segmentation is to segment the input image according to semantic information and predict the semantic category of each pixel from a given label set. With the gradual intellectualization of modern life, more and more applications need to infer relevant semantic information from images for subsequent processing, such as augmented reality, autonomous driving, video surveillance, etc. This paper reviews the state-of-the-art technologies of semantic segmentation based on deep learning. Because semantic segmentation requires a large number of pixel-level annotations, in order to reduce the fine-grained requirements of annotation and reduce the economic and time cost of manual annotation, this paper studies the works on weakly-supervised semantic segmentation. In order to enhance the generalization ability and robustness of the segmentation model, this paper investigates the works on domain adaptation in semantic segmentation. Many types of sensors are usually equipped in some practical applications, such as autonomous driving and medical image analysis. In order to mine the association between multi-modal data and improve the accuracy of the segmentation model, this paper investigates the works based on multi-modal data fusion semantic segmentation. The real-time performance of the model needs to be considered in practical application. This paper analyzes the key factors affecting the real-time performance of the segmentation model and investigates the works on real-time semantic segmentation. Finally, this paper summarizes the challenges and promising research directions of semantic segmentation tasks based on deep learning.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

The goal of semantic segmentation is to divide a given image into several visually meaningful or interesting areas for subsequent image analysis and visual understanding [169]. Semantic segmentation plays an important role in a broad range of applications, e.g., scene understanding, medical image analysis, robot perception and satellite image segmentation [170,169,168].

Before applying convolutional neural networks (CNN), researchers used random forest and conditional random field (CRF) to construct classifiers for semantic learning. In recent years, the deep learning method has yielded a new generation of segmentation models with remarkable performance improvements and has become the mainstream solution for semantic segmentation [65,64,196,136,63]. Long et al. [75,137] propose the fully convolutional network (FCN), which can be trained end-to-end. Compared with traditional methods, the FCN model has a 20% improvement in Pascal VOC 2012 dataset. Ronneberger et al. [130] propose the U-net for biomedical image segmentation. U-net contains a context path to learn context information and a spatial path to preserve spatial information. Table 1 lists the classical semantic segmentation models, e.g., U-net [130,201], FCN [75,137]. This paper comprehensively reviews the research of 2D semantic segmentation in weakly-supervised, domain adaptation, multi-modal data fusion and real-time.

Weakly-supervised semantic segmentation. The fully supervised semantic segmentation method based on CNNs needs a lot of time and economic cost to obtain pixel-level annotation, which restricts the further improvement of segmentation performance and the generalization ability of the model [21,115,97]. Therefore, researchers turn to weakly-supervised learning, which can take advantage of weak annotation forms and reduce the marking cost.

Domain adaptation semantic segmentation. Most machine learning techniques need to satisfy the assumption that the training and test set are independent and identically distributed. However, it is difficult to satisfy this assumption in practical application [96,25]. The purpose of the semantic segmentation method based on domain adaptation is to solve the problem of distribution mis-

---

* Corresponding author.
E-mail address: yanwu@tongji.edu.cn (Y. Wu).

https://doi.org/10.1016/j.neucom.2022.01.005
0925-2312/© 2022 Elsevier B.V. All rights reserved.

Table 1
Classical models of semantic segmentation based on deep learning.

Method	Publish	Year	Method	Publish	Year
FCN [75,137]	CVPR	2015	Dilated Convolutions [189]	ICLR	2016
U-net [130]	MICCAI	2015	RefineNet [91]	CVPR	2017
DeconvNet [113]	ICCV	2015	DUC [167]	WACV	2018
SegNet [160]	TPAMI	2015	ICNet [193]	ECCV	2018
ERFNet [129]	TITS	2018	BiSeNet [188]	ECCV	2018
PSPNet [194]	CVPR	2017	CCNet [70]	ICCV	2019
Deeplab v1 [20]	ICLR	2015	AdaptSegNet [155]	CVPR	2018
Deeplab v2 [21]	TPAMI	2018	EncNet [191]	CVPR	2018
Deeplab v3 [22]	arXiv	2016	Large Kernel Matters [120]	CVPR	2017
Deeplab v3+ [23]	ECCV	2018			

match between training data and test data so that the model can be well extended to practical applications [133,155,58,202].

Semantic segmentation based on multi-modal data fusion. The latest breakthrough of sensors, e.g., cameras, LiDAR, promotes the rapid development of semantic segmentation [86,30,110]. Semantic segmentation based on multi-modal data fusion has become a new research direction to use data of variety of sensors with complementary characteristics to enhance the performance of segmentation [157]. The sensing system based on a single camera cannot provide reliable 3D geometry and adapt to complex or harsh lighting conditions [161,128]. LiDAR can provide high-precision 3D geometry without changing the ambient light, but LiDAR is limited by low resolution, low refresh rate, severe weather conditions (rainstorm, fog and snow), and high cost [173]. By fusing different types of sensor data, the performance and robustness of semantic segmentation can be improved.

Real-time semantic segmentation. Real-time semantic segmentation is a challenging task as it considers both accuracy and inference speed simultaneously. However, under resource constraints, it is impossible to maintain accuracy and boost inference speed simultaneously. Therefore, real-time semantic segmentation aims to achieve a reasonable trade-off between accuracy and inference speed according to the application's requirements, e.g., autonomous driving [170,200,27,51].

In Section 2, this paper introduces the open dataset of semantic segmentation. From Section 3 to Section 6, the problems of weakly-supervised learning, domain adaptation, multi-modal data fusion and real-time in semantic segmentation are introduced respectively. Section 7 discusses the open challenges and promising directions. Finally, a conclusion is made in Section 87.

2. Datasets for semantic segmentation

This section surveys the datasets most commonly used for training and testing semantic segmentation models based on deep learning. According to whether the datasets take into account the changes of lighting conditions, weather and seasonal, this paper divides these datasets into two categories: no cross-domain datasets and cross-domain datasets, and provides the characteristics of each dataset. Most of the data in no cross-domain datasets are collected under normal daytime conditions. Cross-domain datasets collect data in various complex environments, e.g., nighttime, rainy, cloudy, etc. In addition, in the field of autonomous driving, synthetic data is usually used for training models. Therefore, this section also summarizes some synthetic datasets. Table 2 contains some common datasets for semantic segmentation tasks.

2.1. No cross-domain datasets

NYU Depth V2 [138] consists of about 1.5k densely labeled pairs of aligned RGB and depth images, comprising 464 different indoor scenes across 26 scene classes, gathered from a wide range of commercial and residential buildings in three cities. In the NYU Depth V2 dataset, each object is labeled with a class and an instance number (cup1, cup2, cup3, etc). Besides, NYU Depth V2 contains 407k unlabeled frames.

PASCAL-VOC 2012 [42] contains 20 classes. For the segmentation task, the train-val of PASCAL-VOC 2012 includes all corresponding pictures from 2007 to 2011 has 2913 pictures and 6929 objects in total, and the test only includes 2008–2011.

ADE20K [197,198] consists of 20k training images, 2k validation images, and all the images are exhaustively annotated with objects. Many objects are also annotated with their parts, and each object has additional information about whether it is occluded or cropped and other attributes.

ScanNet [36] is an RGB-D video dataset containing 2.5 million views in more than 1.5k scans. ScanNet annotated with 3D camera poses, surface reconstructions and instance-level semantic segmentations, collected by an easy-to-use and scalable RGB-D capture system that includes automated surface reconstruction and crowdsourced semantic annotation.

WoodScape [187] is the first autonomous driving fisheye dataset collected by 4 fisheye cameras on the vehicle. WoodScape provides semantic annotation of 40 classes at the instance level over 10k images and provides annotations for other tasks for images over 100k.

KITTI-2012 [77] contains real image data collected from urban, rural and expressway scenes. The diversity of its recording conditions is relatively low only during the daytime and on sunny days in a city [47,48]. There are up to 15 vehicles and 30 pedestrians in each image and various degrees of occlusion and truncation. Compared to the KITTI-2012, KITTI-2015 [104] comprises dynamic scenes for which a semi-automatic process establishes the ground truth. KITTI-360 [180] annotates both static and dynamic 3D scene elements with rough bounding primitives and transfers this information into the image domain, resulting in dense semantic and instance annotations for 3D point clouds and 2D images.

Inria Aerial Image Labeling Dataset [100] is a remote sensing image dataset for urban building detection. Its labels are divided into building and non-building, which are mainly used for semantic segmentation. The images cover dissimilar urban settlements, ranging from densely populated areas (e.g., San Francisco's financial district) to alpine towns (e.g., Lienz in Austrian Tyrol). Inria Aerial Image Labeling Dataset contains aerial orthophoto corrected color images with a spatial resolution of 0.3 m, and the coverage of Inria Aerial Image Labeling Dataset is 810 km^2.

Gaofen image dataset (GID) [153] is a large dataset for land use and land cover classification. GID contains 150 images from more than 60 different cities in China taken by gaofen-2 (GF-2) satellite, covering a geographical area of more than 50,000 km^2. The images in GID have high intra-class diversity and low inter-class separability. GID includes a panchromatic image with a spatial resolution of

Table 2
Some common datasets for semantic segmentation tasks. The top is no cross-domain datasets. The middle is cross-domain dataset. The bottom is synthetic datasets. "–" indicates that the corresponding result is not provided.

Dataset	Year	Classes	RGB	Depth
KITTI-2012 [77]	2012	8	400	–
NYU Depth v2 [138]	2012	894	≈1.5 k	≈1.5 k
PASCAL-VOC [42]	2012	20	≈3 k	–
Inria Aerial Image Labeling [100]	2017	2	–	–
GID [153]	2018	150	–	–
HippSeg [71]	2017	150	⩾25 k	–
ADE20K [197,198]	2017	150	⩾25 k	–
Pancreatic CT [131]	2015	–	–	–
ApolloScape [69]	2018	25	140 k	–
WoodScape [187]	2019	40	10 k	–
MVD [112]	2017	100	25 k	–
IDD [158]	2019	34	10 k	–
A2D2 [49]	2020	38	≈41 k	–
Cityscapes [34,35]	2016	19	5 k + 20 k	5 k + 20 k
IDDA [2]	2020	–	1000 k	–
Virtual KITTI [45]	2016	–	17 k	–

1 m and a multispectral image with a spatial resolution of 4 *m*, and the image size is 6908 × 7300 pixels. In addition, a multispectral provides images in blue, green, red and near-infrared bands.

HippSegDataset [71] contains T1-weighted MR images of 50 subjects, 40 of whom are patients with temporal lobe epilepsy and 10 are nonepileptic subjects. Hippocampus labels are provided for 25 subjects for training.

Pancreatic CT [131] includes 82 abdominal enhanced 3D CT scans (about 70 *s* after portal vein injection of contrast agent) in 53 male and 27 female subjects. Seventeen subjects are healthy kidney donors scanned before nephrectomy. The remaining 65 patients are chosen by radiologists from patients with neither major abdominal lesions nor pancreatic cancer. The age range of the subjects was 18–76 years, with an average age of 46.8 ± 16.7 years. The resolution of the CT scan is 512 × 512 pixels, the pixel sizes are different, and the slice thickness is between 1.5 and 2.5 *mm*.

In the task of semantic segmentation, research progress can be heavily linked to the existence of datasets [187], e.g., KITTI Vision Benchmark Suite [77]. These existing scene datasets are often smaller and cannot fully capture the variability and complexity of real-world inner-city traffic scenes, which inhibit further progress in the visual understanding of street scenes. Some datasets are recorded in more than one location in order to create the largest and most diverse dataset of street scenes with high-quality and coarse annotations [187,72,119,182,8,34,69].

ApolloScape [69] contains over 140*k* images and each with its per-pixel semantic mask. These images are captured in various traffic conditions; moving objects averages from tens to over one hundred. Besides, ApolloScape annotates each image with high-accuracy pose information at *cm* level accuracy. Moreover, in ApolloScape, 89*k* instance-level annotations for movable objects are further provided. Apollocar3D [140] contains 5*k* driving images and over 60*k* car instances, where each car is fitted with an industry-grade 3D CAD model with the absolute model size and semantically labeled key points. The annotation precision of Apolloscape exceeds that of KITTI and Cityscapes datasets of the same type.

2.2. Cross-domain datasets

Common semantic segmentation datasets, such as Cityscapes [34], focus on the scene in ideal weather, while images of rain, nighttime, snow, or fog are scarce. In order to increase the diversity of lighting conditions, some datasets collect data in both daytime and nighttime [107,10,72,13,5,55,32].

nuTonomy scenes (nuScenes) [10] is the first dataset to carry the full autonomous vehicle sensor suite: 6 cameras, 5 radars and 1 LiDAR. nuScenes comprises 1*k* scenes and each 20s long and fully annotated with 3D bounding-boxes for 23 classes and 8 attributes. nuScenes is a multi-modal dataset that contains data from nighttime and rainy conditions. Also, nuScenes contains object attributes, scene descriptions, object class and location.

Some datasets contain different weather conditions, e.g., sunny, rainy, cloudy, snowy [10,72,107,49,112,121,183,158].

Audi Autonomous Driving (A2D2) [49] consists of simultaneously recorded images and 3D point clouds, recorded the data of highways, country roads and cities in the south of Germany under complex weather conditions, e.g., cloudy, rainy and sunny. The sensor suite of A2D2 consists of 6 cameras, 5 LiDAR units, GPS, IMU, steering angle, brake, throttle, odometry, velocity, pitch and roll. In A2D2, 41*k* frames with semantic segmentation annotations and point cloud labels, of which 12*k* frames also have 3D bounding box annotations for objects within the front camera field of view. In addition, A2D2 contains 392*k* sequential frames of unannotated sensor data recorded in three cities in the south of Germany.

The Mapillary Vistas (MVD) [112] is a large-scale street-level image dataset for semantic segmentation of urban, countryside and off-road scenes and captured at various weather, season and daytime conditions. MVD contains 25*k* high-resolution images annotated into 100 object categories and 60 instance-specific class labels. And these images come from different devices, e.g., mobile phones, tablets, action cameras and professional capturing rigs.

Cityscapes [34,35] is a benchmark suite and large-scale dataset for pixel-level and instance-level semantic segmentation. Cityscapes is comprised of a large, diverse set of stereo video sequences recorded in the streets from 50 different cities, in which 5*k* images have high-quality pixel-level annotations and 20*k* additional images have coarse annotations. It is more complex than previous datasets in the aspects of dataset size, annotation richness, scene variability and complexity.

IDD [158] is a novel dataset for road scene understanding in unstructured environments and consists of 10*k* images, finely annotated with 34 classes collected from 182 drive sequences on Indian roads. The singular feature of IDD is that it corresponds to driving in less structured environments. The variety of traffic participants on Indian roads is larger, including novel classes, e.g., autorickshaws, animals. The within-class diversity is higher because vehicles span a larger range of manufacturing years and ply with larger variation in wear. Moreover, variations in weather and lighting and other ambient factors, e.g., air quality and dust, also span greater ranges.

2.3. Synthetic datasets

Modern computer vision algorithms typically require expensive data acquisition and accurate manual labeling [45]. Therefore, some scholars propose to leverage the recent progress to generate fully labeled, dynamic and photo-realistic proxy virtual worlds. Although synthetic datasets cannot completely replace real-world data, previous work has shown that they are a cost-effective supplement and have good portability [38].

Virtual KITTI [45] is an earlier published composite dataset for autonomous driving. Virtual KITTI is generated in a virtual world created by cloning a few seed real-world video sequences. Virtual KITTI dataset contains 35 photo-realistic synthetic videos (5 cloned from the KITTI, each video with 7 variations) and a total of approximately $17k$ high-resolution frames. Virtual KITTI2 [9] is a more photo-realistic and better-featured version and adds a stereo camera compared with Virtual KITTI.

IDDA (ItalDesign Dataset) [2] is a large synthetic dataset, consists of about 1million frames taken from the virtual world simulator CARLA. IDDA can be used for semantic segmentation with more than 100 different sources visual domains and has the challenges of domain shift between training and test data in various weather and viewpoint conditions in seven different city types.

Although many large-scale image datasets have been created for semantic segmentation, more challenging datasets are still needed in the future. For autonomous driving tasks, datasets containing fine annotations under various complex weather conditions are required.

3. Weakly-supervised semantic segmentation

A key bottleneck in building CNN-based segmentation models is that they typically require pixel-level annotated images during training [195,79]. Acquiring fully supervised data is an expensive, time-consuming effort. Hence, some researchers use weak annotations and propose weakly-supervised semantic segmentation methods to reduce the use of fully annotated data. Weak annotation (in the form of annotated bounding boxes, image-level labels, scribble annotation and point annotation), is far easier to collect than detailed pixel-level annotations. The newly published literatures also point out this promising research trend.This section collects and studies them systematically to get some insight in this field. Finally, this section classifies these papers according to the main kind of weakly-supervised labels, which are image-level labels, bounding-boxes, scribble annotation and point annotation. As showed in Fig. 1, in the leftmost figure, there are only bounding-boxes to indicate the position information of objects, which are standard annotating labels for object detection tasks. The middle one uses hand-drawing scribbles to indicate the position information of objects. The rightmost one's labels are even simpler, which only contain some points representing the rough middle of objects. Table 3 compares the effects of some weakly-supervised semantic segmentation methods. The ”Annotation type” in the first column from left side represents the type of weakly-supervised annotation.

3.1. Segmentation algorithm based on image-level labels

The main paradigm of using image-level labels to do semantic segmentation tasks is first generating the score map or heat map from the pretext task, such as the standard classification task, as the original rough mask. Then the researchers apply some clustering algorithm to refine and improve the generating mask iteratively until getting a satisfactory result. The last procedure is to feed the mask produced from the previous step as the fully annotated label to some pre-defined models that do standard supervised segmentation tasks. Inspired by the multi-instance learning (MIL) framework, Pinheiro et al. [122] infer object segmentation by leveraging only object class information and considering only minimal priors on the object segmentation task. Pathak et al. [118] utilize the MIL structure to predict multiple pixels for each image, in which each pixel represents a class. Pathak et al. [117] also propose a constrained convolutional neural network, which casts the image-level label as a constraint to the network. In more detail, during each training iteration, the network firstly predicts a target point in the latent representation space with the highest probability.

Zhang et al. [190] propose a causal inference framework that can be added to the weakly-supervised semantic segmentation methods. Their framework can generate better pixel-wise pseudo labels by addressing the issue that the score map is hard to distinguish between the boundaries. They propose a new method to mitigate the bias in image-level classification. Lin et al. [172] present a multi-path region mining module to generate pseudo-point-level labels from image-level labels. They explore the localization information for each class from different aspects of the object with various attention modules. Fan et al. [43] address the problem that the CAM-like (class activation map) method can only cover a part of objects. They argue that the critical problem affecting the performance is the mismatch of classification boundaries. Therefore, they propose an intra-class discriminator module to distinguish the difference between objects within the image to generate a pseudo-label with better boundaries.

Araslanov et al. [3] define three metrics to check the performance of a weakly-supervised method: local consistency, semantic fidelity and completeness, respectively. Using these metrics, they propose a single-stage method to train the model from image-level labels. Chen et al. [19] adopt a wise selection of training samples and a model evaluation criterion to do video object segmentation using only the action labels of the actors. Chang et al. [15] address the problem that the response maps only focus on the critical part of the object, which is not the best for the segmentation task. To enforce the network to focus on the whole image, they propose a self-supervised task with a sub-category loss that exploits the sub-category information and performs clustering on image features to generate the pseudo mask.

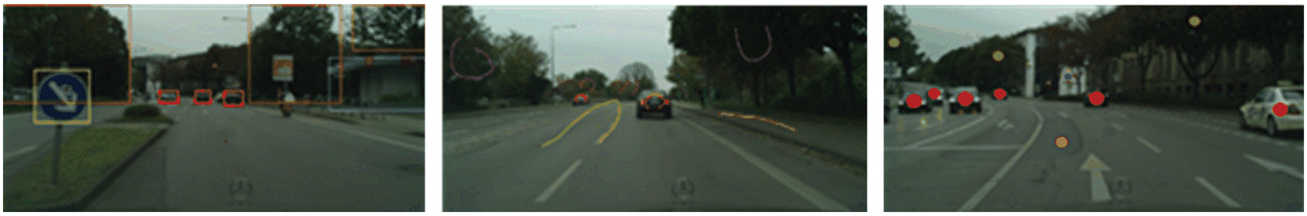

Fig. 1. In the leftmost figure, each bounding-boxes indicates one instance. The middle one uses hand-drawing scribbles to indicate the rough position information of instance, which is much easier to annotate. In the rightmost figure, each point represents one instance. Note that the instances belonging to the same category have the same color.

Table 3
The results of some weakly-supervised semantic segmentation methods. The "Annotation type" in the leftmost column represents the type of weakly-supervised annotation.

Annotation type	Mehtod	Publish	Dataset	mIoU%
image-level	Pinheiro et al. [122]	CVPR2015	Pascal VOC(test) 2012	40.6
	Pathak et al. [117]	ICCV2015	Pascal VOC(test) 2012	45.1
	Fan et al. [43]	CVPR2020	Pascal VOC 2012(test)	64.3
	Zhang et al. [190]	arXiv2020	Pascal VOC 2012(test)	66.7
	Zhang et al. [190]	arXiv2020	MS-COCO(val)	33.4
bounding-box	Xia et al. [178]	ICCV2013	Pascal VOC 2012(test)	48
	Dai et al. [37]	ICCV2015	Pascal VOC 2012(test)	64.6
	Papandreou et al. [115]	ICCV2015	Pascal VOC 2012(test)	62.2
scribble	Lin et al. [90]	CVPR2016	Pascal VOC 2012(val)	63.1
	Tang et al. [148]	CVPR2018	Pascal VOC 2012(val)	74.5
	Tang et al. [149]	ECCV2018	Pascal VOC 2012(val)	75
	Wang et al. [164]	IJCAI2019	Pascal VOC 2012(val)	76
point	Bearman et al. [6]	ECCV2016	Pascal VOC 2012(val)	46.1
	Qian et al. [126]	AAAI2019	PASCAL-Context(val)	30
	Qian et al. [126]	AAAI2019	ADE20K(val)	19.6
	McEver et al. [102]	arXiv2020	Pascal VOC 2012(val)	70.5

Based on multi-instance learning, pseudo pixels, iterative training, causal reasoning and other methods, the above methods improve the semantic segmentation performance based on image-level labels by supplementing the position information, shape information, and contour information of the target in the image. Weakly-supervised semantic segmentation using image-level labels is the mainstream of weakly-supervised segmentation methods, but image-level annotation provides little information. Therefore, introducing additional information, e.g., location information, the supervision information of constructing pseudo pixels, and modeling these information into the loss function is the key to further improving semantic segmentation performance.

3.2. Segmentation algorithm based on bounding-box

Aside from employing image-level information as a weak supervision annotation, some works are established based on bounding-box labels. Papandreou et al. [115] use a small part of the full supervised label and a major part of the weakly-supervised label to train the network model. Dai et al. [37] propose a model to harness the bounding-box label to facilitate the segmentation task, which is trained in an iterating fashion (Fig. 2). Khoreva et al. [76] focus on the work that manipulates the input label of images. They process the input label carefully instead of changing the training procedure. Xia et al. [178] adopt a voting scheme to infer each object's shape within the bounding-box. Firstly, they cut out the original image into different parts according to the bounding-box, and then they segment each part respectively. Finally, they merge the individual result into a complete one. These methods use bounding-box labels to generate rough segmentation maps of the target and then iteratively optimize the segmentation model. However, compared with the method based on the image-level label, it is reliant on image annotation quality.

3.3. Segmentation algorithm based on scribble

Compared with the bounding-box label, scribble annotations are easier to get. For instance, Lin et al. [90] create the scribble dataset for PASCAL VOC 2012. They combine the scribble annotations and proposed region generated from the previous model and train the model in an iterating manner. Finally, they get a well-performed segmentation network. Tang et al. [148] design a new loss for weakly-supervised segmentation tasks. The loss function is composed of two-part, one is the partial cross-entropy loss which only evaluates the seed with the known label, and the other one is normalized cut loss which focuses on the consistency of all the pixels. Wang et al. [164] harness a standard multi-task learning structure to use the scribble label. Their structure composes two sub-modules to predict each object's scribble and bounding, respectively. Their main contribution is mainly on the design of extracting high-level and low-level features simultaneously and combining them to get the final segmentation result. Vernaza et al. [159] propose a new label propagation mechanism to process the input label, named random-walk propagation. Random-walk

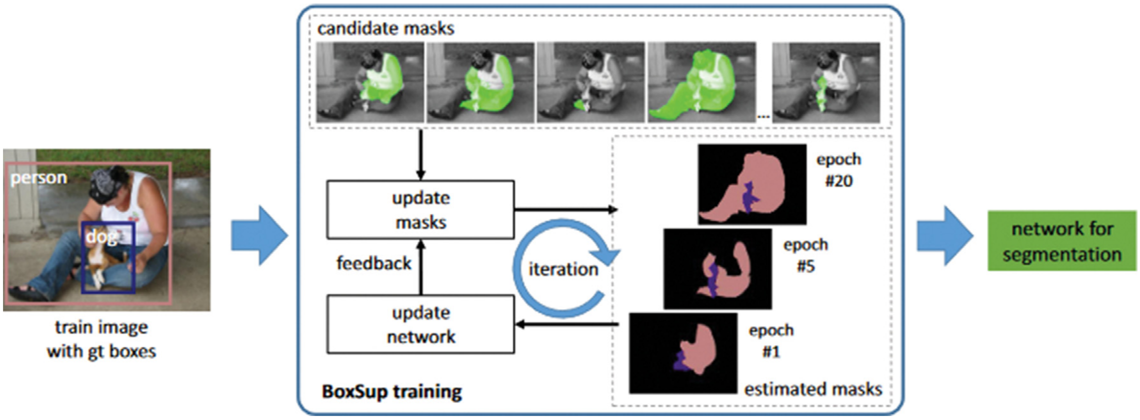

Fig. 2. The BoxSup model. From [37].

propagation can help to distill and optimize the segmentation result. In [159], Xu et al. introduce a method to use multiple forms of weak supervision labels, e.g., the scribble label, bounding-box.

3.4. Segmentation algorithm based on point

The most natural way for humans to refer to an object is by pointing, such as ”That cat over there” (point) or ”What is that over there?” (point) [126]. Some fields use the point as an effective means of communication, e.g., robotics and human–computer interaction for a long time. However, point annotation has not been well applied in semantic segmentation. Some scholars propose some new point-based semantic segmentation methods, where each class has only one label point, and they incorporate the point supervision into the training loss function. Qian et al. [126] propose a method based on a simple idea that the objects within the same class should have a similar representation, so they design a distance metric loss to optimize the model. The result shows that their method can exploit the label well and give satisfied segmentation results. Bearman et al. [6] present a time-efficient method that adopts both point-based label and objectness prior information to supervise the training process with image-level, point-level and objectness prior loss functions. McEver et al. [102] use the classical attention mechanism to generate an attention map, select the point that makes the most significant contribution as the predicted point representing the object, and then use a metric loss to supervise the model's training process. The final result validates the effectiveness of their method (Fig. 3).

4. Domain adaptation in semantic segmentation

Fully convolutional models for semantic segmentation have been proven successful. Such models perform well in a supervised setting, but performance can be surprisingly poor under domain shifts that appear mild to a human observer [59]. For example, training on one city and testing on another in a different geographic region and/or weather condition may result in significantly degraded performance due to pixel-level distribution shift. Domain adaptation is a particular case of transfer learning. It uses the labeled data in one or more related source domains to perform new tasks in the target domain [166]. This section focuses on domain adaptation methods in semantic segmentation and divides the existing works in this field into three categories according to their adaptation level, i.e., input level, feature level, output level. Table 4 compares the results of some domain adaptation semantic segmentation methods, in which source domain is GTA5 and target domain is Cityscapes. Table 5 compares the results of some domain adaptation semantic segmentation methods, in which source domain is SYNTHIA and target domain is Cityscapes.

4.1. Input-level domain adaptation

Due to images between source domain and target domain carrying strong high-level semantic similarity in scene content and layout, a rich line of works use image translation or style transfer methods to map the data in one domain to the other while reserving the class-related feature of translated images. Then, the semantic segmentation models can be trained with the translated images with the original labels. The potential issue of these methods is the quality of the generated images, as semantic segmentation models are commonly demanding on the quality of input images, even pixel-level flaws could significantly influence the segmentation accuracy. To bypass these issues, lots of studies resort to enforce semantic consistency of image translations, thus raising the image quality in detail [150]. This section divides these methods into GAN-based (generative adversarial network) methods and style transfer methods based on the techniques of translate images.

A considerable amount of researches [88,58,29,152,199,127,82,185,50] use CycleGAN [203] architecture to address the input space's domain shifts. They firstly translate images from the source domain to the target domain with an image-to-image translation model and then add a discriminator on top of the features of the segmentation model to further decrease the domain gap. When the domain gap is reduced by the former step, the latter one can further decrease the domain shift. Unfortunately, the segmentation model relies on the quality of image-to-image translation. Once the image-to-image translation fails, nothing can be done to make it up in the following stages. In order to motivate the two steps promoting each other and reduce the domain gap, Li et al. [88] propose a new bidirectional closed-loop learning framework for domain adaptation of image semantic segmentations. The system involves two separated modules: image-to-image translation model and segmentation adaptation model, but the learning process involves two directions (i.e., ”translation-to-segmentation” and ”segmentation-to-translation”). Similar to bidirectional learning framework [88], in order to encourage the model to preserve the semantic information in the

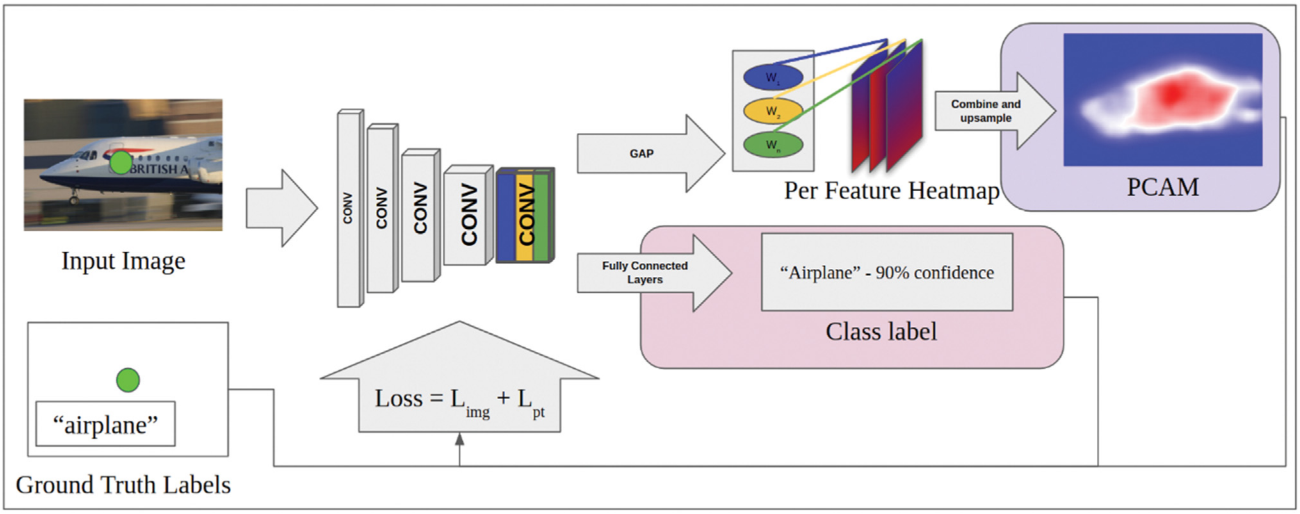

Fig. 3. The PCAMs model. From [102].

Table 4
Comparison results from GTA5 to Cityscapes. The top is the result of input-level domain adaptation methods. The middle is the result of feature-level domain adaptation methods. The bottom is the result of output-level domain adaptation methods.

Method	Publish	Backbone	mIoU(%)	Method	Publish	Backbone	mIoU(%)
Hoffman et al. [58]	ICML2018	MobileNet-v2	37.3	Hoffman et al. [58]	ICML2018	DRN-26	39.5
Li et al. [88]	CVPR2019	ResNet-101	48.5	Li et al. [88]	CVPR2019	VGG-16	41.3
Chen et al. [29]	CVPR2019	DRN-26	45.1	Chen et al. [29]	CVPR2019	FCN8s	38.1
Yang et al. [185]	CVPR2020	DRN-26	42.6	Gong et al. [50]	CVPR2019	ResNet-101	42.3
Yang et al. [186]	CVPR2020	ResNet-101	50.45	Yang et al. [186]	CVPR2020	VGG-16	42.2
Yang et al. [185]	CVPR2020	ResNet-101	50.5	Yang et al. [185]	CVPR2020	VGG-16	44.6
Zhou et al. [199]	arXiv2020	VGG-16	47.8				
Chen et al. [28]	CVPR2018	VGG-16	35.9	Zhu et al. [204]	ECCV2018	FCN8s	38.1
Murez et al. [111]	CVPR2018	DenseNet	35.7	Chang et al. [14]	CVPR2019	ResNet-101	45.4
Luo et al. [95]	ICCV2019	VGG-16	34.2	Luo et al. [95]	ICCV2019	ResNet-101	42.6
Saito et al. [133]	CVPR2018	VGG-16	28.0	Saito et al. [133]	CVPR2018	DRN-105	39.7
Tsai et al. [155]	CVPR2018	ResNet-101	42.4	Vu et al. [162]	CVPR2019	VGG-16	36.1
Chen et al. [24]	ICCV2019	ResNet-101	46.4	Lee et al. [78]	ICCV2019	ResNet-50	35.8
Spadotto et al. [141]	arXiv2020	ResNet-101	35.1	Vu et al. [162]	CVPR2019	ResNet-101	45.5

Table 5
Comparison results from SYNTHIA to CityScapes. The top is the result of input-level domain adaptation methods. The middle is the result of feature-level domain adaptation methods. The bottom is the result of output-level domain adaptation methods.

Method	Publish	Backbone	mIoU(%)	Method	Publish	Backbone	mIoU(%)
Yang et al. [186]	CVPR2020	ResNet-101	52.5	Yang et al. [186]	CVPR2020	VGG-16	40.5
Chen et al. [29]	CVPR2019	DRN-26	33.4	Chen et al. [29]	CVPR2019	FCN8s	38.2
Zhou et al. [199]	arXiv2020	VGG16	48.6	Yang et al. [185]	CVPR2020	VGG-16	41.1
Yang et al. [185]	CVPR2020	ResNet-101	46.2				
Chen et al. [28]	CVPR2018	VGG-16	36.2	Zhu et al. [204]	ECCV2018	FCN8s	34.2
Luo et al. [95]	ICCV2019	VGG-16	37.2	Luo et al. [95]	ICCV2019	ResNet-101	46.3
Chang et al. [14]	CVPR2019	ResNet-101	41.5				
Tsai et al. [155]	CVPR2018	ResNet-101	46.7	Saito et al. [133]	CVPR2018	DRN-105	37.3
Vu et al. [162]	CVPR2019	VGG-16	36.6	Vu et al. [162]	ICCV2019	VGG-16	36.8
Vu et al. [162]	CVPR2019	ResNet-101	48.0	Vu et al. [162]	ICCV2019	ResNet-101	42.6
Chen et al. [24]	ICCV2019	ResNet-101	41.4	Spadotto et al. [141]	arXiv2020	ResNet-101	34.6

process of distribution alignment, Hoffman et al. [58] reconstruct the original data from the translated version. They propose Cycle-Consistent Adversarial Domain Adaptation (CyCADA), which adapts representations at both the pixel-level and feature-level while enforcing local and global structural consistency through pixel cycle-consistency and semantic losses. They also use a reconstruction (cycle-consistency) loss to encourage the cross-domain transformation to preserve local structural information and a semantic loss to enforce semantic consistency (Fig. 4). Compared with CycleGAN, CyCADA can be applied in a variety of visual recognition and prediction settings. Besides, other cross-domain consistency loss functions are proposed to enforce domain adaptation models to produce consistent predictions [29,111].

However, these methods mentioned above focus on high-level features adaptation while neglecting low-level feature structure/information which is crucial for differentiating certain trivial patterns, so the image content can be mistakenly revised or blurred. In order to align low-level features, Qin et al. [127] design a novel generatively inferential co-training freamework which is based on cross-domain feature generation, and feature generation adapts the representation at low level by translating images across domains. Moreover, they adapt a channel attention layer to align

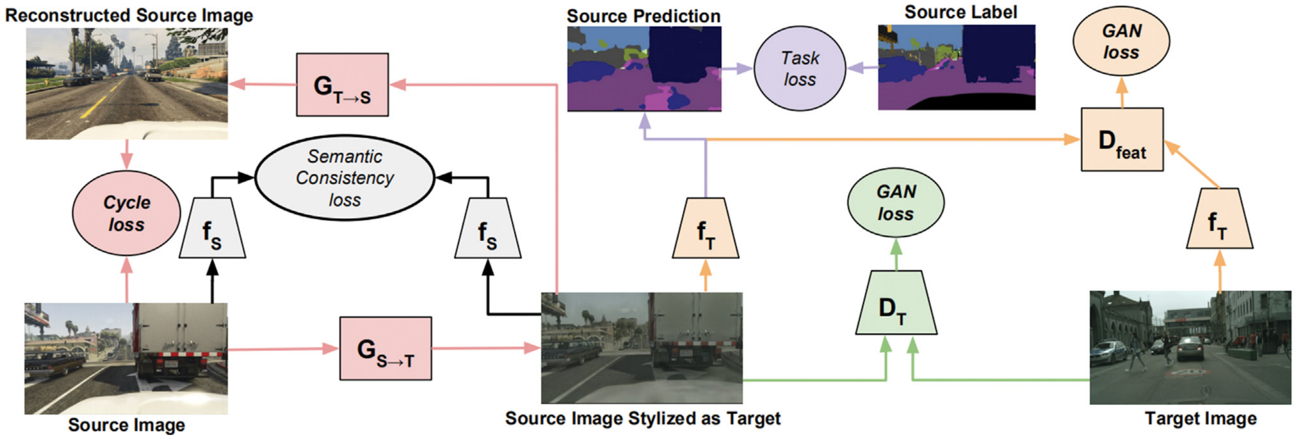

Fig. 4. The CyCADA model. From [58].

high-level real and generated image features. Li et al. [82] propose a soft gradient-sensitive loss to keep semantic boundaries and a semantic-aware discriminator to validate adaptation's fidelity.

On the other hand, style transfer methods avoid the costly computation of GAN-based methods by exploiting traditional neural style transfer techniques [192]. Yang et al. [186] propose a spectral transfer method based on Fourier Transform, which does not require any training. By swapping the low-frequency component of the source images' spectrum with the target images, the translated image is mapped to the target style without altering semantic content. Style transfer can be achieved by renormalizing feature maps of source images so they have first- and second-order feature statistics that match target images [68,156]. These renormalized feature maps are then fed into a generator network that produces stylized images. Wu et al. [174] use an image generator to align the distribution of mean and variances of feature maps of source domain with target domain at the pixel-level, since these statistics are simple to optimize and contain enough information to get a good stylization. Note that some authors match gram matrices [174,46] to make styles consistent. Compared with using several layers for alignment [87], they simply match one layer of feature maps from the visual geometry group (VGG) encoder, which is faster yet sufficient.

The input-level domain adaptation methods are intuitive and effective as they share the same idea of achieving domain invariance on visual appearance by mitigating the cross-domain discrepancy in the image layout and structure. But they also possess a sharing vulnerability of being over-reliant on image quality and computational costly.

4.2. Feature-level domain adaptation

An alternative approach to close the domain shifts lies in a distribution alignment of feature latent embeddings. In general, this is achieved by forcing the feature extractor to extract domain-invariant features, by adjusting the distribution of latent representations from source and target domains. In this way, the network classifier should be able to learn to segment both source and target representations from the common latent space, by relying solely on the supervision from source data [151].

The intrinsic domain difference between source and target images usually causes a significant segmentation performance drop. From the perspective of representation, since the feature extraction model is trained on the source images, the convolutional filters tend to overfit to the style of source images, making them incompetent to extract informative features for target images. Besides, from the distribution perspective, source and target data suffers a considerable distribution mismatch, which makes the model biased to source domain. In order to learn the style of target distributions, Chen et al. [28] propose a target guided distillation approach, which is achieved by training the segmentation model to imitate a pretrained real style model using real images. Besides, they take advantage of the intrinsic spatial structure presented in urban scene images, and propose a spatial aware adaptation scheme to effectively align the distribution of two domains. Zhang et al. [192] facilitate domain shifts from the perspectives of both visual appearance-level and representation-level domain adaptation. They present the fully convolutional adaptation network, a novel deep architecture for semantic segmentation which combines appearance adaptation network (AAN) and representation adaptation network (RAN). AAN learns a transformation from one domain to the other in the pixel space and RAN is optimized in an adversarial learning manner to maximally fool the domain discriminator with the learnt source and target representations. Zhu et al. [204] introduce a conservative loss, which enables the network to learn features that are discriminative by gradient descent and are invariant to the change of domains via gradient ascend method. Murez et al. [111] add extra networks and losses to regularize the features extracted by the backbone encoder network. The extracted features are able to reconstruct the images in both domains and the distribution of features extracted from images in the two domains are indistinguishable. With the hypothesis that the structural content of images is the most informative and decisive factor to semantic segmentation and can be readily shared across domains, Chang et al. [14] propose a domain invariant structure extraction framework. They learn to disentangle the domain-invariant structure information of the image from its domain-specific texture information to discover the domain invariant structure features.

The strategy of aligning the two domains in latent feature space through adversarial learning has achieved much progress in image classification [179,109,44,105], but usually fails in semantic segmentation tasks in which the latent representations are over complex. Some of the task-independent nuisance factors might be easily involved in the encoded representation and mislead the domain alignment. Luo et al. [95] equip the adversarial network with a "significance-aware information bottle neck", to enable a significance-aware feature purification before the adversarial adaptation, which eases the feature alignment and stabilizes the adversarial training course. Huang et al. [67] align the distributions of activations of intermediate layers. Because this scheme exhibits two key advantages. First, matching across intermediate layers introduces more constraints for training the network in the target domain, making the optimization problem better conditioned. Second, the matched activations at each layer provide similar inputs to the next layer for both training and adaptation, and thus alleviate covariate shift.

The feature-level domain adaptation methods succeed in semantic segmentation as they are able to align network latent embeddings with either adversarial strategies or auxiliary losses. However, the computational burden of aligning high-dimensional feature spaces does exist when it comes to over complex feature representation such as segmentation tasks.

4.3. Output-level domain adaptation

Some other adaptation methods resort to the cross-domain distribution alignment over the segmentation output space to avoid dealing with an excessively convoluted latent space [123,186,192,174]. While retaining enough complexity and richness of semantic cues, prediction maps from the segmentation network output identify a low-dimensional space where adaptation can be performed effectively [151].

Considering semantic segmentation as structured outputs that contain spatial similarities between the source and target domains, Tsai et al. [155] propose to adopt adversarial learning in the output space of the segmentation network firstly. To further enhance the adapted model, they construct a multi-level adversarial network to effectively perform output space domain adaptation at different feature levels. Some works [176,40,59] use the domain discriminator's output to boost the segmentation network's performance in a self-training manner. Biasetton and Michieli et al. [7,106] propose an adversarial discriminative adaptation framework combined with a self-training strategy on the segmentation network output. They use the pseudo label generated by a segmentation network trained in the source domain to improve the generator through the segmentation confidence that estimated by the fully convolutional discriminator of the adversarial learning module. Spadotto et al. [141] introduce a novel unsupervised domain adaptation (UDA) framework where a standard supervised loss on labeled synthetic data is supported by an adversarial module and a self-training strategy aiming at aligning the two domain distributions.

The adversarial module contains a fully convolutional discriminators to discriminate segmentation maps coming from synthetic or real-world data. The self-training module exploits the confidence estimated by the discriminators on unlabeled data to select the regions used to reinforce the learning process. Based on the entropy minimization strategy introduced from the semi-supervised learning field, some works [162,163,24] explicitly promote the domain discriminator to output more confident predictions. Vu et al. [162,163] align weighted self-information distributions of target and source domains to minimize the entropy by having target's entropy distribution similar to the source. Chen et al. [24] propose a maximum squares loss to solve the unbalanced domain transfer when the entropy minimization strategy is applied to domain adaptive semantic segmentation.

Following a different idea, [132,133,171,78] resort to classifier discrepancy as a better domain critic that detects non-discriminative features between different classes. In particular, this line of methods replaces the original adversarial discriminator with a better critic based on the prediction discrepancy of multiple classifiers. Saito et al. [132] propose the adversarial dropout regularization to encourage the network to output more discriminative features away from the decision boundary. By using the dropout strategy on the two domain classifiers, a discrepancy loss is measured through the divergence between the two prediction maps at the output level. Formal adversarial training is then applied to achieve class-wise feature alignment. Saito et al. [133] further improve the framework by utilizing task-specific classifiers as discriminators that detect target samples far from the source's decision boundary, thus avoiding generating target features near the class boundaries. Similarly, Lee et al. [78] propose a modified version of the adversarial dropout module called drop to adapt with a single classifier that can output divergent predictions. In this way, the domain adapted model draws a robust decision boundary that avoids clusters.

To sum up, the output-level domain adaptation methods reach efficient cross-domain distribution alignment on the low-dimensional output space of segmentation network with self-training, entropy minimization or adversarial learning strategies. But with limited semantic information to encode in the output space, the adaptation process usually requires extra supervised signal to achieve the precise semantic alignment.

5. Semantic segmentation based on multi-modal data fusion

In some artificial intelligence applications, e.g., human action recognition [18,1,54,108,92,17,41,16], autonomous driving, several different sensors are usually used to collect data. It has become a new research direction to improve the performance of semantic segmentation algorithm by using sensor data with complementary characteristics. By fusing different types of sensor data, the performance and robustness of semantic segmentation algorithm can be improved [165]. Table 6 compares the results of some classic semantic segmentation methods based on multi-modal data fusion, e.g., RGB + T(Thermal), RGB + D(Depth).

5.1. Fuse RGB and thermal/depth images

When the illumination conditions are not satisfied, e.g., dim or dark light, semantic segmentation network performance based on single-modal data is easy to degrade. Some studies have found that thermal images generated by thermal imaging cameras are robust to challenging lighting conditions [89,18,1,54]. Utilizing thermal information can potentially improve the performance of semantic segmentation since the thermal channel has complementary information to RGB channels and encodes structural information of the scene. Some works integrate thermal and RGB information into semantic segmentation framework. The encoder part of these works is composed of two network branches, which extract features from RGB and thermal images simultaneously and fuse the thermal features into the RGB feature mapping with the deepening of the network.

Ha et al. [55] propose the multi-spectral fusion network (MFNet) for semantic segmentation of urban scenes using both RGB and thermal images. Similar as FuseNet [56], MFNet also adopts the Encoder-Decoder architecture and adopted dilated convolution to form a mini-inception block in the encoders. Two identical feature extractors are designed for RGB and thermal images, respectively. To fuse the feature maps from the RGB and thermal encoders, a short-cut block is designed to concatenate the two feature maps from the two encoders. The concatenated feature map is then added to the output of the corresponding last layer of the decoder. Sun et al. [143] propose the RGB-Thermal fusion network (RTFNet), which adopts the encoder-decoder design concept. RTFNet employs the residual network (ResNet) for feature extraction and develops a new decoder is to restore the feature map resolution. The data fusion is performed in the encoding stage by the element-wise summation with feature maps. Sun et al. [144] propose a multi-modal data fusion end-to-end network named FuseSeg (Fig. 5), which contains two modules, the first module extracts feature by two encoders in which the backbone is DenseNet (Densely Connected Convolutional Network) [66], and the second module restores the resolution by a decoder. The corresponding thermal and RGB feature maps are fused hierarchically through elementwise summation in the RGB encoder and fused with the related feature maps in the decoder again [130].

The auxiliary depth may reduce the uncertainty of the segmentation of objects having similar appearance information, because compared with RGB images, depth images contain more location and contour information that benefit the context-critical semantic segmentation. Recently encoder-decoder type fully convolutional CNN architectures have achieved a great success in the field of semantic segmentation, Hazirbas et al. [56] propose FuseNet by fusing RGB and depth data in an encoder-decoder structure (Fig. 6). In FuseNet, two encoders using VGG-16 [139] as backbones are designed to extract features from RGB and depth images, respectively. The feature maps from the depth encoder are gradually fused into the RGB encoder as the network goes deeper. Similarly, Sun et al. [142] design two independent branches to extract features for RGB and depth images separately RGB branch as the

Table 6
The results of some semantic segmentation methods based on multi-modal data fusion.

Method	Publish	Modals	Dataset	mIoU(%)
FuseNet [56]	ACCV2016	RGB + T	Ha et al. [55]	45.6
Ha et al. [55]	IROS2017	RGB + T	Ha et al. [55]	39.7
RTFNet [143]	IEEE Robot Autom Let 2019	RGB + T	Ha et al. [55]	53.2
Sun et al. [144]	IEEE Trans. Autom. Sci. Eng. 2020	RGB + T	Ha et al. [55]	54.5
Schneider et al. [135]	SCIA2017	RGB + D	Cityscapes(val)	69.1
Deng et al. [39]	arXiv2019	RGB + D	Cityscapes(val)	72
Sun et al. [142]	IEEE Robot Autom Let 2020	RGB + D	Cityscapes(val)	72.5

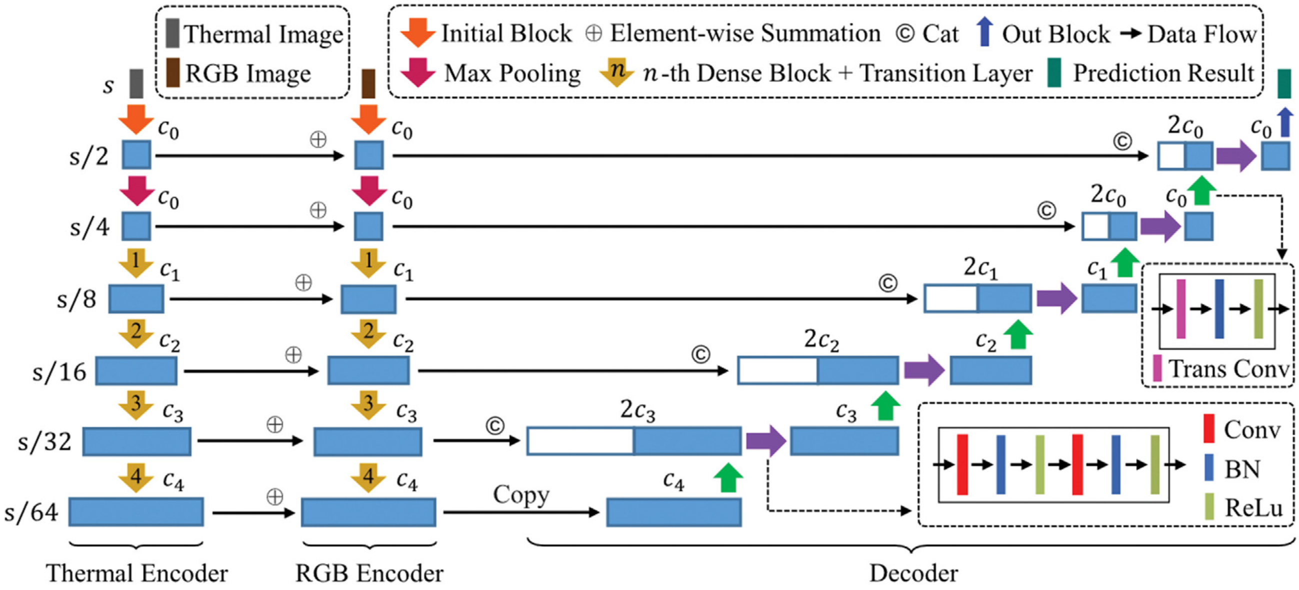
Thermal Image
RGB Image
Initial Block
⊕ Element-wise Summation
© Cat
Out Block
Data Flow
Max Pooling
n-th Dense Block + Transition Layer
Prediction Result
s
s/2
s/4
s/8
s/16
s/32
s/64
c_0
c_1
c_2
c_3
c_4
$2c_0$
$2c_1$
$2c_2$
$2c_3$
Copy
Trans Conv
Conv
BN
ReLu
Thermal Encoder
RGB Encoder
Decoder

Fig. 5. The FuseSeg model. From [144].

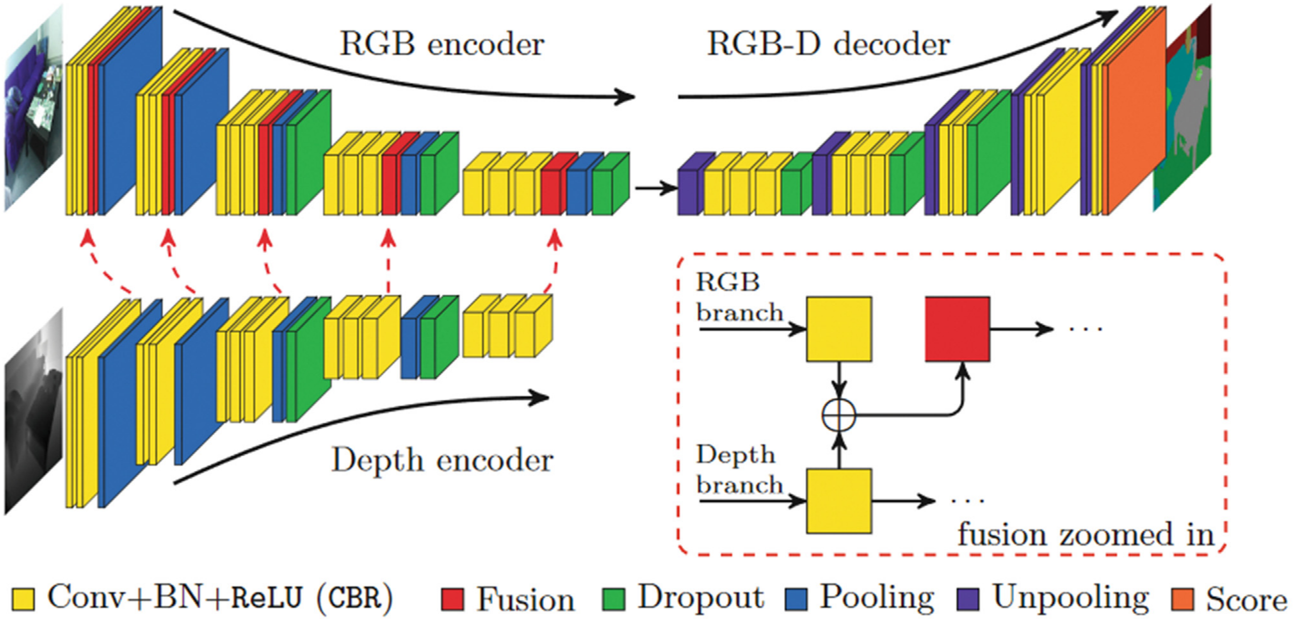
RGB encoder
RGB-D decoder
Depth encoder
RGB branch
Depth branch
. . .
fusion zoomed in
Conv+BN+ReLU (CBR)
Fusion
Dropout
Pooling
Unpooling
Score

Fig. 6. The FuseNet model. From [56].

main branch and Depth branch as the subordinate branch. They fuse the output features from Depth branch to RGB branch by the attention feature complementary (AFC) module. Then, they use the spatial pyramid pooling block gathers the fused RGB-D features from two branches and produce feature maps with multi-scale information. The AFC module contains the squeeze-and-excitation [74] block that can learn to use global information to emphasize informative channels and suppress less useful channels, which helps the AFC module exploit informative features from both branches effectively.

The performance of semantic segmentation can be potentially improved by fusing RGB and thermal/depth images. The above methods usually use two branches to extract the features of RGB and thermal/depth images respectively, and then fuse them at the feature level. However, the data of different modals have different properties, so it is necessary to fully mine the private information and shared information between different modals.

5.2. Fuse RGB images and LiDAR point clouds

RGB cameras can provide dense information over a long-range under good illumination and fair weather. However, RGB cameras are strongly affected by the level of illumination. LiDAR senses the environment by using their emitted pulses of laser light, therefore, LiDAR is only marginally affected by the external light conditions. Furthermore, LiDAR provides accurate range measurements. Based on this description of the advantages and also downsides of LiDAR and RGB cameras, it is easy to see that fusing the data of LiDAR and RGB cameras could give an enhanced total [83]. To enhance the image and introduce additional information, the LiDAR point cloud can be added to images.

Caltagirone et al. [11] project point cloud into the camera image plane and then up-sampled to obtain a set of dense 2D images encoding spatial information. Then they train several fully convolutional neural networks (FCNs) to carry out road detection by using three fusion strategies: early, late and cross fusion. Chen et al. [31] introduce a novel progressive LiDAR adaptation-aided road detection (PLARD) approach to adapt LiDAR information into visual image-based road detection. PLARD transforms the LiDAR data to the visual data space to align with the perspective view and adapts LiDAR features to visual features through a cascaded fusion structure. CRF models have been widely used in the semantic segmentation task based on the image or LiDAR point cloud data. Recently, some works use CRF to fuse multi-modal sensor data. Gu et al. [52] propose a data fusion system based on 3D LiDAR and a monocular camera for urban road detection. First, their framework projects the 3D point cloud of LiDAR into the camera's image frame to exploit both range and color information. Second, it uses a fusion method based on CRF to integrate the two road detection results. Gu et al. [53] propose a real-time road detection method based on the LiDAR-camera fusion strategy for the urban road detection task. First, they obtain road detection results based on the visual domain by geometric up-sample and use a transfer learning strategy to learn from a small annotated road dataset. Then, they operate LiDAR data at a small-sized LiDAR range image at high frequency. Last, they use a multi-modal CRF framework to fuse the dense and binary road detection results from LiDAR and the camera.

In the most methods mentioned above, LiDAR is typically projected onto the camera image plane. However some works think that they lead to a sparse representation of LiDAR which might lead to sub-optimal models. Wulff et al. [177] incorporate an early fusion of LiDAR and camera data into a multi-dimensional occupation grid representation as FCN input and project the camera image into the BEV representation and overlay them onto the LiDAR occupancy grid. Lv et al. [98] propose the two-stream fusion fully convolutional network (TSF-FCN) to fuse LIDAR point clouds and RGB images (Fig. 7). LIDAR stream aggregates multi-scale contextual information from LiDAR point clouds. RGB stream extracts features from RGB images. To combine the two streams, TSF-FCN converts the RGB stream feature maps into a bird-view representation for concatenating with the LIDAR stream. Hu et al. [62] propose a road detection method based on monocular images and LiDAR data and divide it into three modules. The first module extracts ground points, and the second module computes an image representation of illumination-invariant features simultaneously. Last, ground points are projected to the image plane and then used to compute a road probability map using a Gaussian model to finish the final classification and segmentation of the road area.

Because cameras and LiDARs obtain information in different ways, there are significant differences in the representation of the same object. The image is dense and regular, containing rich color and texture information, but the disadvantage is the scale problem because of the distance. Compared with images, point cloud data is sparse and irregular, making it impossible to use the traditional CNN model to process point cloud data. However, point cloud data contain three-dimensional geometry and depth information. Because the image and LiDAR data are complementary in theory, it is necessary to mine the inevitable relationship between image information and point cloud information to improve the segmentation algorithm's performance.

6. Real-time semantic segmentation

Due to the use of complex network architecture, the semantic segmentation method based on CNNs often has the problem of high computational complexity, which greatly limits the application in real-time processing of real scenes, e.g., autonomous driving, video surveillance, robot sensing. In U-shape structures, e.g., U-net, the encoder is usually a backbone, which is the most vital part of the whole framework and accounts for the dominant proportion of model size and computational budget, the pipeline of U-shape architecture is shown in Fig. 8(a) [4,175]. For real-time inference, some works adopt a lightweight backbone model and investigate how to improve the segmentation performance with limited computation. According to the adopted backbone, this section divides current real-time semantic segmentation methods

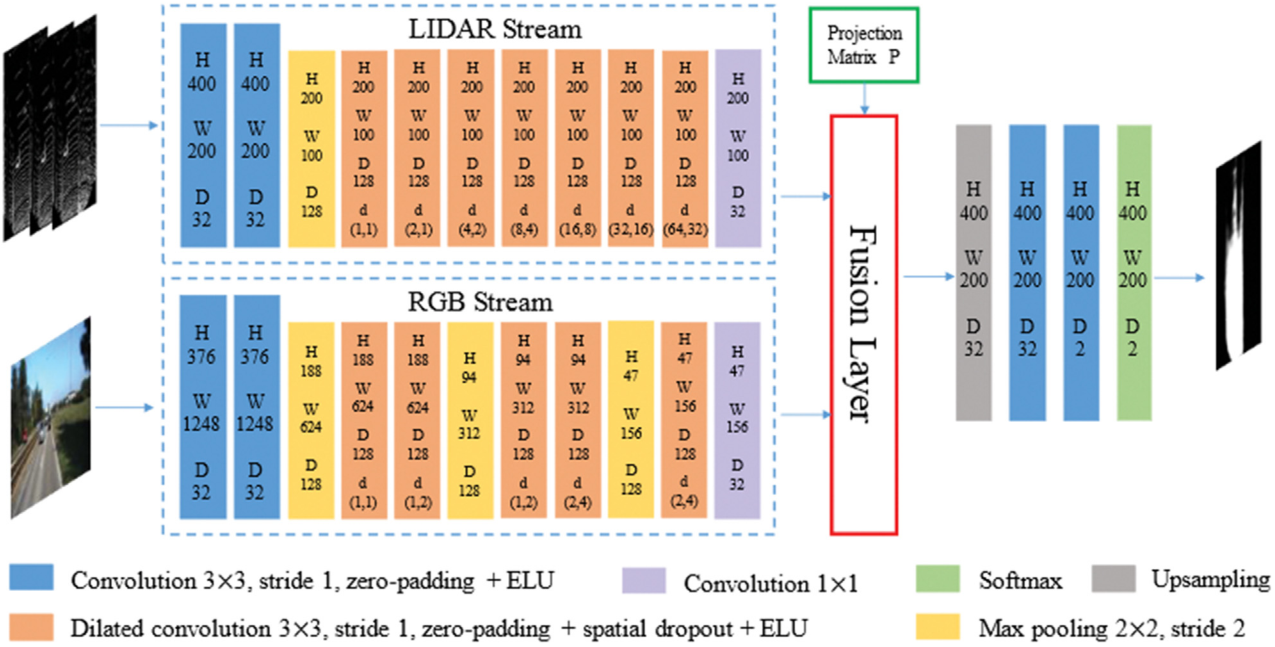
LIDAR Stream
H 400 W 200 D 32
H 400 W 200 D 32
H 200 W 100 D 128
H 200 W 100 D 128 d (1,1)
H 200 W 100 D 128 d (2,1)
H 200 W 100 D 128 d (4,2)
H 200 W 100 D 128 d (8,4)
H 200 W 100 D 128 d (16,8)
H 200 W 100 D 128 d (32,16)
H 200 W 100 D 128 d (64,32)
H 200 W 100 D 32
RGB Stream
H 376 W 1248 D 32
H 376 W 1248 D 32
H 188 W 624 D 128
H 188 W 624 D 128 d (1,1)
H 188 W 624 D 128 d (1,2)
H 94 W 312 D 128
H 94 W 312 D 128 d (1,2)
H 94 W 312 D 128 d (2,4)
H 47 W 156 D 128
H 47 W 156 D 128 d (2,4)
H 47 W 156 D 32
Projection Matrix P
Fusion Layer
H 400 W 200 D 32
H 400 W 200 D 32
H 400 W 200 D 2
H 400 W 200 D 2
Convolution 3×3, stride 1, zero-padding + ELU
Convolution 1×1
Softmax
Upsampling
Dilated convolution 3×3, stride 1, zero-padding + spatial dropout + ELU
Max pooling 2×2, stride 2

Fig. 7. The TSF-FCN model. From [98].

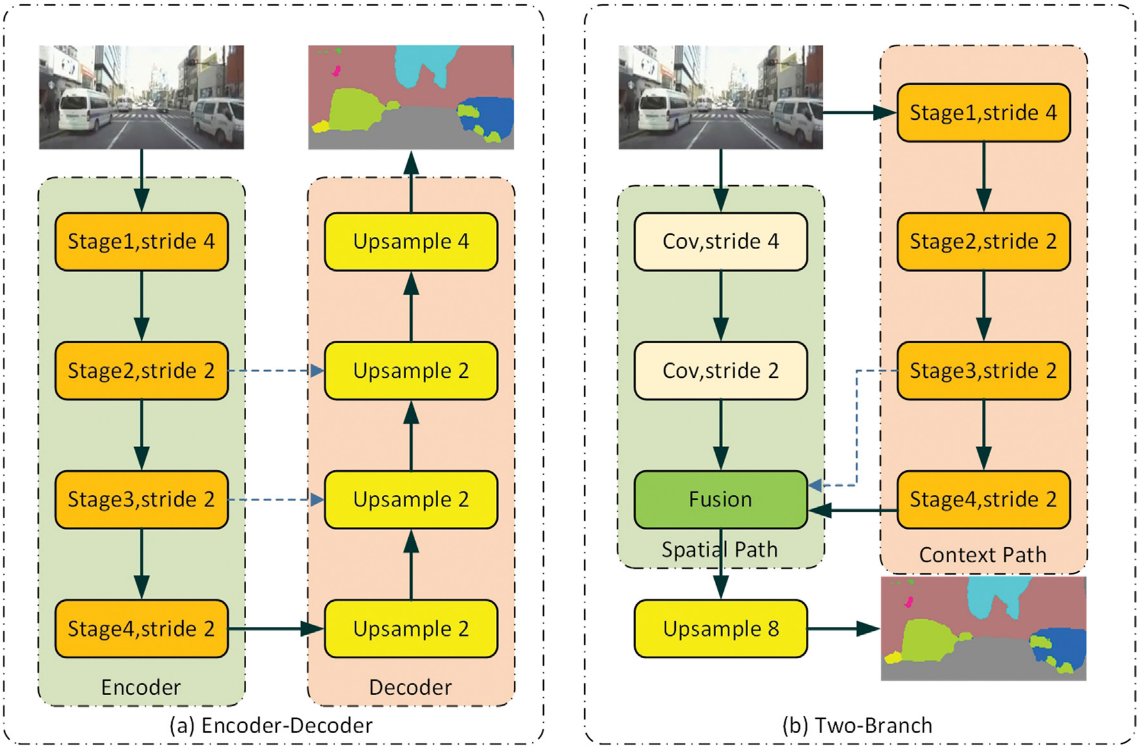

Fig. 8. The pipeline of encoder-decoder based method (a) and two-branch based method (b) for real-time semantic segmentation.

into lightweight classification model-based method, specialized backbone based method and two-branch architecture based method. The accuracy and efficiency results of real-time semantic segmentation methods are shown in Table 7. The top is the result of real-time semantic segmentation methods based on lightweight classification backbone. The middle is the result of real-time semantic segmentation methods based on specialized backbone. The bottom is the result of real-time semantic segmentation methods based on two-branch architecture.

6.1. Lightweight classification model-based method

With the massive development of CNN in image recognition, it is a natural way to apply these modern CNN models as feature extractors in many computer vision tasks, such as object detection, semantic segmentation, human pose estimation and so on [129]. If the network model is initialized from scratch, it is possible to miss a huge regularization opportunity offered by knowledge transfer from larger and more diverse recognition datasets, incuring a com-

Table 7
Accuracy and efficiency results for state-of-the-art real-time semantic segmentation methods on Cityscapes test dataset. "–" indicates that the corresponding result is not provided. The top is lightweight classification backbone based encoder-decoder methods. The middle is specialized backbone based encoder-decoder methods. The bottom is two-branch architecture based methods.

Method	Publish	Backbone	Pretrain	Input Size	Params	FLOPs	GPU	FPS	mIoU(%)
DFANet[81]	CVPR2019	Xception	ImageNet	1024 × 1024	7.8 M	3.4G	TitanX	100	71.3
SwiftNet [114]	CVPR2019	ResNet-18	ImageNet	1024 × 2048	11.8 M	104G	1080Ti	39.9	75.5
ShelfNet[205]	ICCVW2019	ResNet-18	ImageNet	1024 × 2048	–	–	1080Ti	36.9	74.8
FANet[61]	ECCVW2020	ResNet-18	ImageNet	1024 × 2048	–	49.0G	TitanX	72.0	74.4
SFNet[85]	ECCV2020	DF1	ImageNet	1024 × 2048	9.0 M	–	1080Ti	74.0	74.5
SFNet[85]	ECCV2020	DF2	ImageNet	1024 × 2048	10.5 M	–	1080Ti	53.0	77.8
ENet[116]	arXiv2016	–	No	360 × 640	0.4 M	3.8G	TitanX	135	58.3
ERFNet[129]	TITS2018	–	No	512 × 1024	2.1 M	–	TitanX M	41.7	68.0
ESPNet[103]	ECCV2018	–	No	512 × 1024	0.4 M	–	TitanX	112	60.3
EDANet[94]	MMAsia2019	–	No	512 × 1024	0.7 M	–	1080Ti	109	67.3
LEDNet[170]	ICIP2019	–	No	512 × 1024	0.9 M	11.5G	1080Ti	71.0	69.2
DABNet[80]	BMVC2019	–	No	1024 × 2048	0.8 M	–	1080Ti	27.7	70.1
FPENet[93]	BMVC2019	–	No	768 × 1536	0.4 M	12.8G	TitanV	55.0	70.1
LRNNet[73]	ICMEW2020	–	No	512 × 1024	0.7 M	8.6G	1080Ti	71.0	72.2
DDPNet[184]	ACCV2020	–	No	768 × 1536	2.5 M	13.2G	1080Ti	85.4	74.0
DDPNet[184]	ACCV2020	–	No	1024 × 2048	2.5 M	23.5G	1080Ti	52.6	75.3
ContextNet[101]	BMVC2018	–	No	1024 × 2048	0.9 M	–	TitanX M	41.9	66.1
GUN[101]	BMVC2018	DRN-D-22	ImageNet	512 × 1024	–	–	TitanX	33.3	70.4
FastSCNN[125]	BMVC2019	–	No	1024 × 2048	1.1 M	–	TitanX	123	68.0
ICNet[193]	ECCV2018	PSPNet-50	ImageNet	1024 × 2048	26.5 M	28.3G	TitanX M	30.3	69.5
BiSeNet[188]	ECCV2018	Xception-39	ImageNet	768 × 1536	5.8 M	14.8G	TitanX	106	68.4
BiSeNet[188]	ECCV2018	ResNet-18	ImageNet	768 × 1536	12.9 M	55.3G	TitanX	65.5	74.7

paratively large overfitting risk. In order to boost the segmentation performance, the backbone of the semantic segmentation network is often pre-trained on a large-scale classification dataset, e.g., ImageNet [27]. SwiftNet [114] shows that ImageNet pre-training represents an important ingredient for reaching highly accurate predictions. However, this approach may decrease the model capacity and limit the size of the receptive field for features, therefore decreasing the model's discriminative ability. Some works are proposed to capture rich spatial contextual information.

Li et al. [81] propose the deep feature aggregation network (DFANet), which adopts a lightweight Xception [33] model as the backbone with little modification (Fig. 9). In order to enhance the model learning capacity and increase the receptive field simultaneously, DFANet reuses high-level features extracted from the backbone to bridge the gap between semantic information and structure details and combines features of different stages to enhance feature representation ability. SwiftNet [114] adopts ResNet-18 [57] and MobileNet V2 [134] as backbones and a lightweight up-sampling module with lateral connections as the decoder. SwiftNet also uses the spatial pyramid pooling and pyramid fusion methods to expand the receptive field and ensure real-time. Zhuang et al. [205] propose the ShelfNet for accurate and fast semantic segmentation, which adopts ResNet as a backbone. They reduce the computation burden by reducing channel number and propose a shared-weight strategy in the residual block, which reduces parameter number without sacrificing performance. Unlike a single encoder-decoder structure for semantic segmentation, ShelfNet has multiple encoder-decoder branch pairs with skip connection at each spatial level, which looks like a shelf with multiple columns that can be viewed as an ensemble of multiple deep and shallow paths.

The original self-attention mechanism has been shown to be beneficial for various vision tasks because of its ability to capture non-local context from the input feature maps. Therefore Hu et al. [61] propose the fast attention network (FANet) to capture rich spatial contextual information, which adopts ResNet as a backbone and further applies an additional down-sampling process in intermediate features to use rich context information and full-resolution spatial information under small computational costs. In the fast attention module, FANet converts the attention process to a series of matrix multiplication by replacing the softmax function with the cosine similarity function.

In order to effectively transfer semantic information from deep layers to shallow layers, Li et al. [85] propose the SFNet to learn the semantic flow between two network layers of different resolutions. They propose the concept of semantic flow to represent the relationship between two feature maps of arbitrary resolutions from the same image. They design a flow alignment module (FAM) that takes feature maps from adjacent levels as input and aligns these two feature maps according to semantic flow. SFNet adopts ResNet [57], ShuffleNet V2 [99] and DF [84] as the backbone and builds a feature pyramid aligned network with multiple FAMs in the decoder.

Due to the differences between semantic segmentation and image classification tasks, the backbone designed for image classification tasks may not be the best choice for semantic segmentation tasks. Compared with image classification, the image size to be processed by semantic segmentation is much larger. For example, CIFAR [154] and Cityscapes [34,35] are datasets commonly used for image recognition and semantic segmentation, respectively. The image size of Cityscapes (1024 × 2048) is much larger than cifar (32 × 32). On the other hand, semantic segmentation needs to aggregate multi-scale context and global information to classify each pixel of the image. Therefore, the receptive field of backbone should be large enough to cover the whole image, which can not be guaranteed in the classification network without global average pool.

6.2. Specialized backbone based method

Unlike image recognition, semantic segmentation needs multi-scale context information to make a correct identification, which is not guaranteed in a typical classification network. Therefore, many researchers focus on designing a specialized backbone for real-time semantic segmentation. Paszke et al. [116] adopt an initial block and a bottleneck module modified from ResNet to construct a novel architecture as a backbone. They propose a relatively small decoder to up-sample the output of the encoder. To maintain a real-time speed, Paszke [116] proposes the efficient neural network (ENet) to resize the input image and applies early down-sampling to reduce computational cost. LEDNet [170] adopts an asymmetric encoder-decoder architecture to accelerate the inference process. In the encoder, a novel basic block consists of skip connection and convolution with the channel split and shuffle. Channel split and 1D factorized convolution are used to reduce computational costs. Channel shuffle is used to enhance information exchange between feature maps. Finally, they adopt an attention pyramid network (APN) to produce a semantic prediction in the decoder. DABNet [80] proposes the depth-wise asymmetric bottleneck (DAB) module, which combines depth-wise convolution and asymmetric convolution with dilated convolution. Due to the DAB module, DABNet can efficiently utilize context information and maintain fast inference speed at the same time. Liu et al. [93] (Fig. 10) introduce a feature pyramid encoding block to encode

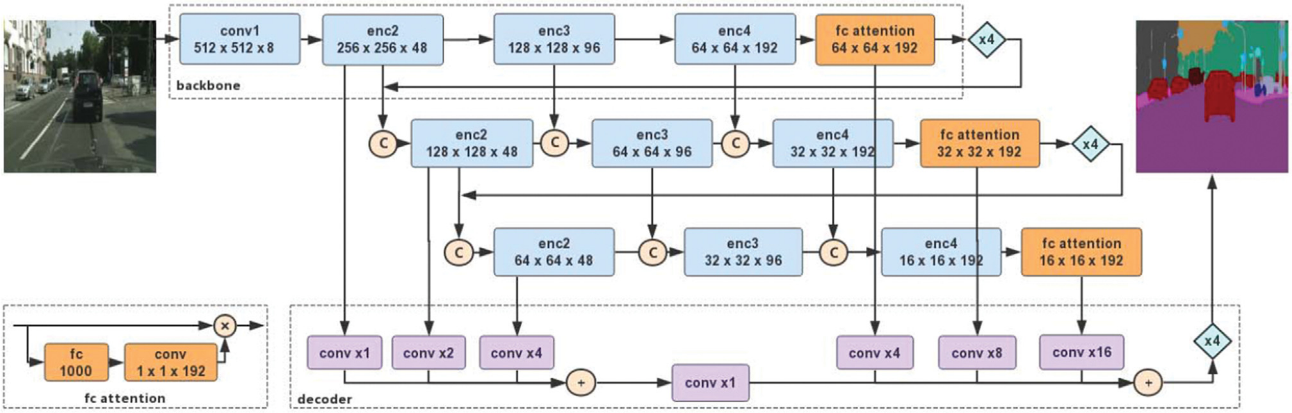

Fig. 9. The DFANet model. From [81]..

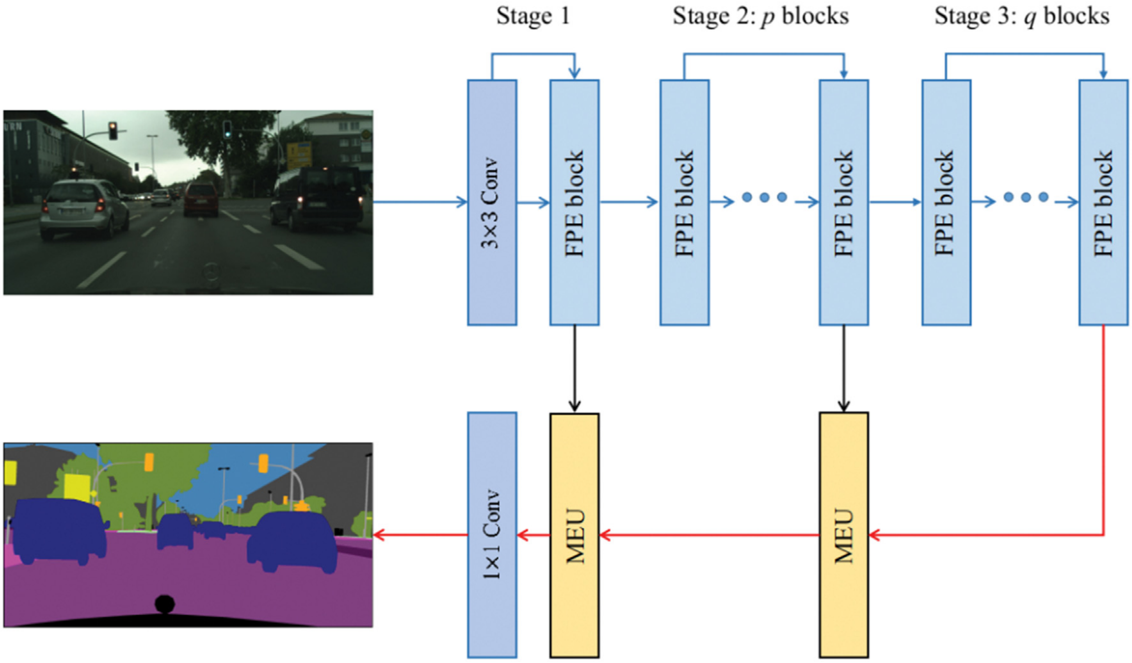

Fig. 10. The FPENet model. From [93].

multi-scale features in the encoder, which combines a pyramid of depth-wise convolution with different dilation rates to reduce computational complexity.

Convolutional factorization decomposes the convolutional operation into multiple steps to reduce the computational complexity while simultaneously allowing the network to learn the representations from a large effective receptive field, and has successfully shown its potential in reducing the computational complexity of deep CNN networks [146,147,145,181,33,60,103]. Romera et al. [129] design a novel factorized residual layer with residual connection and convolutional factorization, which is named ERFNet (efficient residual factorized network). They use the residual connection to facilitate the training process of network and convolutional factorization to reduce the model size and computational cost. A factorized residual layer is adopted as the basic block for the encoder, and the initial block proposed by ENet [116] is used as a down-sampling block. Similar to ENet [116], ERFNet has a small decoder only used for recovering image size. Mehta et al. [103] propose the efficient spatial pyramid module (ESP) as a basic block to build a novel backbone. ESP module is a form of factorized convolution which consists of a point-wise convolution and a spatial pyramid of dilated convolutions. ESP module has a large receptive field and a low computational cost. In addition, hierarchical feature fusion is proposed to address gridding artifacts in the ESP module. In the decoder, ESPNet uses a strategy of reduce-up-sample-merge to recover spatial information gradually. EDANet [94] applies asymmetric convolution to build a novel efficient dense module, which decomposes a standard 2D convolution into two 1D convolutions. As a variant of DenseNet [66], EDANet can gather features extracted from different layers and aggregate multi-scale context information. DDPNet [184] designs a novel lightweight backbone with dense connectivity and a dual-path module to aggregate multi-scale context information. A skip architecture with the proposed up-sampling module is adopted as the decoder, which leverages context information to refine semantic outputs. Jiang et al. [73] introduce a lightweight factorized convolution block (FCB) and an efficient reduced non-local module, which is named SVN. FCB, 1D factorized convolution, and depth-wise separable convolution with a large dilation rate are used to deal with short-range features and long-range features. In the SVN module, a regional singular vector is used to model long-range dependencies and maintain low computational computation and memory cost. Chen et al. [26] design a low-rank-to–high-rank context reconstruction framework and introduce the tensor generation module, which generates a number of rank-1 tensors to capture fragments of context feature. Then they use these rank-1 tensors to recover the high-rank context features through the tensor reconstruction module.

Due to the great difference between semantic segmentation and image recognition, image recognition belongs to image-level classification, while semantic segmentation belongs to pixel-level classification. Directly using an image recognition network as the backbone network of the encoder may not be the best choice. Therefore, some work designs a more suitable backbone for real-time semantic segmentation, but the disadvantage of this kind of method is that it needs to be pre-trained on large-scale data sets, otherwise the effect will be poor.

6.3. Two-branch architecture based method

The above methods adopt an encoder to generate context information and a decoder to recover spatial information. Some other methods design a two-branch architecture to aggregate context information and spatial information, the pipeline of two-branch architecture is shown in Fig. 8(b). Compared with single-branch networks, two-branch employs a deeper branch to capture global context and a shallow branch to learn spatial details at full input resolution, and the final semantic segmentation result is then provided by merging the two. Because the computational cost of deeper networks is overcome with small input size, and execution on the full resolution is only employed for few layers, real-time performance is possible. ICNet [193] develops a novel architecture for real-time semantic segmentation, which obtains context information from a low-resolution input and spatial information from a high-resolution input (Fig. 11). In addition, a cascade feature fusion unit is proposed to combine cascade features from different resolu-

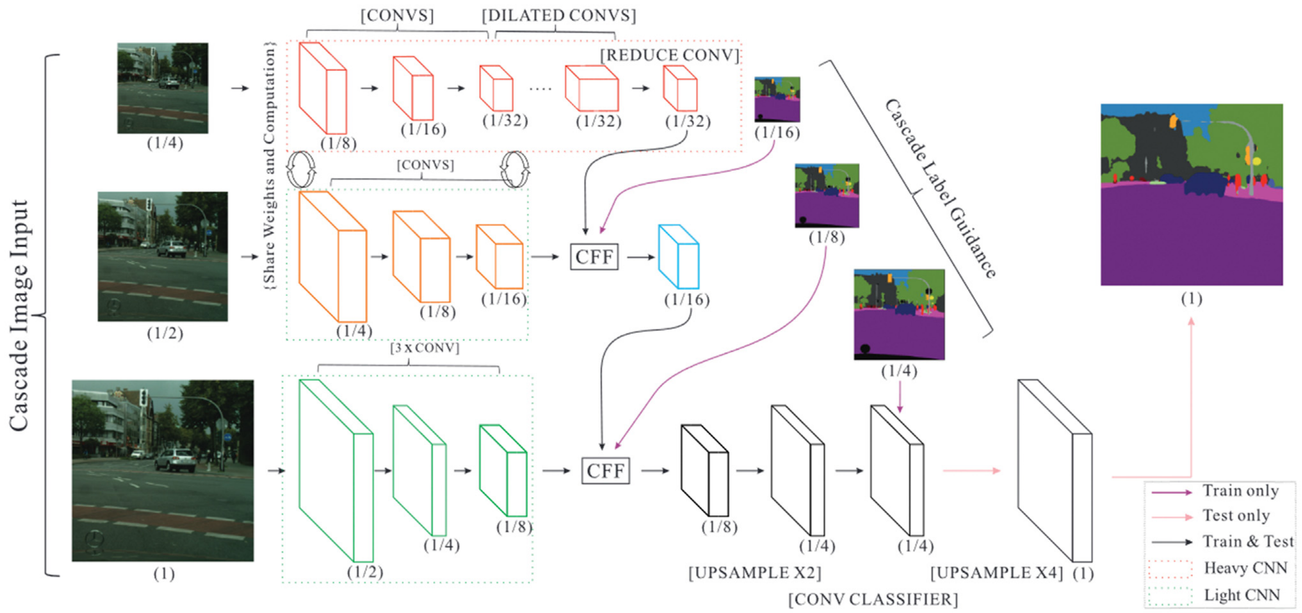

Fig. 11. The ICNet model. From [193].

tion inputs, and cascade label guidance is offered to enhance the learning procedure in each branch. BiSeNet [188] is another classic work that attempts to utilize high-resolution images and low-resolution images to achieve comparable performance. In detail, a context path with a lightweight model is designed to obtain context information, and a spatial path with only three convolution layers is adopted to preserve spatial information at the same time. FFM (feature fusion module) is introduced to combine feature representation from two paths to improve accuracy further. ARM (attention refinement module) is proposed to refine the features in the context path. With a similar design principle, GUN (guided up-sampling network) [101] and ContextNet [124] also design a two-branch architecture for real-time semantic segmentation. FSCNN [125] merges two-branch architecture with encoder-decoder architecture, which shares low-level features for two branches. By adopting skip connection and learning to down-sample module, FSCNN constructs a network similar to two-branch architecture but without an actual spatial branch. Finally, a feature fusion module is employed to combine features from different levels.

These methods design a two branch network structure. The core idea is to use a smaller resolution image to pass through a complete semantic segmentation network to capture global context, and then another full input resolution image to pass through a shallow network to learn spatial details. The final semantic segmentation map is then provided by merging the result of two branch. These real-time semantic segmentation methods based on two-branch network need to design a reasonable backbone network for each branch, and also need to consider how to efficiently integrate the prediction results of the two branches.

7. Open challenges and promising directions

Without a doubt, image segmentation has benefited greatly from deep learning, but there is still a long way to go and many problems need further study. This paper discusses some promising directions that will advance the performance of semantic segmentation algorithms.

1. How to get more challenging datasets? In real applications, the lighting and weather conditions are complex and changeable, which limits the generalization ability of the segmentation model. Therefore, there will be more datasets containing complex scene changes and weather changes in the future.
2. How to achieve the optimal balance between accuracy and inference speed? For lightweight classification backbone-based encoder-decoder methods, making lightweight classification models more suitable for semantic segmentation may be an interesting problem. For specialized backbone-based encoder-decoder methods, how to design a better-specialized backbone for fast and accurate semantic segmentation and improve accuracy with pre-training or data enhancement are very important.
3. How to improve the performance of the domain of adaptive semantic segmentation method? Compared to supervised semantic segmentation models, the segmentation accuracy of the UDA method in the target domain is still far behind that of the oracle. With more novel formulations being proposed, future work in this field can be further expanded in different adaptation settings like partial DA or openset DA. Moreover, strategies that have been proved successful in related tasks can also be introduced to this field, as self-training and curriculum learning did.
4. How to improve the accuracy of the model through context knowledge? Referring to the CRFs method in RNNs, it is a feasible research route to create an end-to-end method to improve the accuracy of real-world scenes. Multi-scale and feature blending also shows significant progress. These studies are essential steps towards the ultimate goal, but there are still many problems and a lot of research to be done.
5. How to explore the time domain correlation between video and image sequences? There are already a few methods for semantic segmentation of video and sequence, which use time information to improve efficiency and accuracy. However, there is no way to address relevance issues. For a system that processes video image segmentation, it is essential to achieve better processing results per frame and to predict the label of each pixel by using interframe information.

6. How to improve the model performance of the weakly-supervised semantic segmentation method? The performance trained from the weakly-supervised learning method is still cannot surpass the model trained from the fully supervised learning methods. Although great effort has been put into this research field, there is still some defective parts, for example, few works are focusing on increasing the model robustness trained from the weakly-supervised method, and domain shift is another issue, many models are dedicated to some specific task or scene, when facing new scene or data domain, the models will suffer a big drop in performance. Future research can focus on the improvement of the model adaptation ability and try to address the problem of model lacking ability to face adversarial examples.
7. How to solve the catastrophic forgetting problem of deep structure? Although deep structure performs well in many tasks, the existing semantic segmentation methods are still unable to gradually update its internal classification model (catastrophic forgetting) when new categories are found [12]. In order to solve these problems, some scholars have begun to combine incremental learning method with semantic segmentation method. However, there is little research in this field, so it needs more attention.

8. Conclusion

In recent years, semantic segmentation methods based on deep learning have made great progress, especially in weakly-supervised semantic segmentation, domain adaptation in semantic segmentation, semantic segmentation based on multi-modal data fusion, real-time semantic segmentation and so on. In order to let researchers quickly understand the research status of semantic segmentation and find the future research direction, this paper reviews the state-of-the-art technologies of semantic segmentation based on deep learning, which have achieved impressive performance in various segmentation tasks and benchmarks. This paper summarizes the latest research progress of semantic segmentation tasks in weakly-supervised, domain adaptation, multi-modal data fusion and real-time. This paper also carefully describes the datasets commonly used in semantic segmentation and explains their uses and characteristics, so that researchers can easily select the dataset that is most suitable for their needs. In addition, this paper summarizes the challenges and promising research directions of semantic segmentation tasks based on deep learning in the next few years. Recently, transformer-based methods are prevalent and have made great success in semantic segmentation. However, the transformer-based semantic segmentation methods have not been reviewed. In real-world applications, e.g., autonomous driving, the form of data is generally video rather than static pictures. However, the exploration of video semantic segmentation is still limited. Therefore, we will systematically study the transformer-based semantic segmentation and video semantic segmentation, and form a new and meaningful overview document.

Declaration of Competing Interest

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

Acknowledgement

This work was supported by the National Nature Science Foundation of China (No. U19A2069).

References

[1] Z. Ahmad, N.M. Khan, Multidomain multimodal fusion for human action recognition using inertial sensors, in: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), IEEE, 2019, pp. 429–434, https://doi.org/10.1109/BigMM.2019.00026.

[2] E. Alberti, A. Tavera, C. Masone, B. Caputo, Idda: A large-scale multi-domain dataset for autonomous driving, IEEE Robotics and Automation Letters 5 (2020) 5526–5533, https://doi.org/10.1109/LRA.2020.3009075.

[3] N. Araslanov, S. Roth, Single-stage semantic segmentation from image labels, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 4252–4261, https://doi.org/10.1109/cvpr42600.2020.00431.

[4] 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 (2017) 2481–2495, https://doi.org/10.1109/TPAMI.2016.2644615.

[5] D. Barnes, M. Gadd, P. Murcutt, P. Newman, I. Posner, The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 6433–6438, https://doi.org/10.1109/ICRA40945.2020.9196884.

[6] Bearman, A., Russakovsky, O., Ferrari, V., Li, F., 2016. What's the point: Semantic segmentation with point supervision, in: 2016 European Conference on Computer Vision (ECCV), Springer. pp. 549–565. DOI: 10.1007/978-3-319-46478-7_34..

[7] M. Biasetton, U. Michieli, G. Agresti, P. Zanuttigh, Unsupervised domain adaptation for semantic segmentation of urban scenes, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2019, pp. 1211–1220, https://doi.org/10.1109/CVPRW.2019.00160.

[8] M. Braun, S. Krebs, F. Flohr, D.M. Gavrila, Eurocity persons: A novel benchmark for person detection in traffic scenes, IEEE transactions on pattern analysis and machine intelligence 41 (2019) 1844–1861, https://doi.org/10.1109/TPAMI.2019.2897684.

[9] Cabon, Y., Murray, N., Humenberger, M., 2020. Virtual kitti2. arXiv preprint arXiv:2001.10773v1..

[10] H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, in: nuscenes: A multimodal dataset for autonomous driving, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 11618–11628, https://doi.org/10.1109/cvpr42600.2020.01164.

[11] L. Caltagirone, M. Bellone, L. Svensson, M. Wahde, Lidar-camera fusion for road detection using fully convolutional neural networks, Robotics and Autonomous Systems 111 (2019) 125–131, https://doi.org/10.1016/j.robot.2018.11.002.

[12] F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, B. Caputo, Modeling the background for incremental learning in semantic segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 9230–9239, https://doi.org/10.1109/CVPR42600.2020.00925.

[13] Chang, M., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J., 2019a. Argoverse: 3d tracking and forecasting with rich maps, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 8740–8749. DOI: 10.1109/CVPR.2019.00895..

[14] W. Chang, H. Wang, W. Peng, W. Chiu, All about structure: Adapting structural information across domains for boosting semantic segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 1900–1909, https://doi.org/10.1109/CVPR.2019.00200.

[15] Y. Chang, Q. Wang, W.C. Hung, R. Piramuthu, Y.H. Tsai, M. Yang, Weakly-supervised semantic segmentation via sub-category exploration, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 8988–8997, https://doi.org/10.1109/cvpr42600.2020.00901.

[16] C. Chen, R. Jafari, N. Kehtarnavaz, Action recognition from depth sequences using depth motion maps-based local binary patterns, in: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2015, pp. 1092–1099, https://doi.org/10.1109/WACV.2015.150.

[17] C. Chen, R. Jafari, N. Kehtarnavaz, Improving human action recognition using fusion of depth camera and inertial sensors, IEEE Transactions on Human-Machine Systems 45 (2015) 51–61, https://doi.org/10.1109/THMS.2014.2362520.

[18] C. Chen, N. Kehtarnavaz, R. Jafari, A medication adherence monitoring system for pill bottles based on a wearable inertial sensor, in: Conference of the IEEE Engineering in Medicine and Biology Society IEEE, 2014, pp. 4983–4986, https://doi.org/10.1109/EMBC.2014.6944743.

[19] J. Chen, Z. Li, J. Luo, C. Xu, Learning a weakly-supervised video actor-action segmentation model with a wise selection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 9898–9908, https://doi.org/10.1109/cvpr42600.2020.00992.

[20] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A., 2014b. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062v4..

[21] L. 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 (2018) 834–848, https://doi.org/10.1109/TPAMI.2017.2699184.

[22] Chen, L., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587v3..

[23] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., Encoder-decoder with atrous separable convolution for semantic image segmentation, in: 2018 European Conference on Computer Vision (ECCV), Springer. p. 833–851. DOI: 10.1007/978-3-030-01234-2_49..

[24] M. Chen, H. Xue, D. Cai, Domain adaptation for semantic segmentation with maximum squares loss, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 2090–2099, https://doi.org/10.1109/ICCV.2019.00218.

[25] S. Chen, X. Jia, J. He, Y. Shi, J. Liu, Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation, ArXiv (2021), abs/2103.04705.

[26] W. Chen, X. Zhu, R. Sun, J. He, R. Li, X. Shen, B. Yu, Tensor low-rank reconstruction for semantic segmentation, ArXiv (2020), abs/2008.00490.

[27] X. Chen, Y. Wang, Y. Zhang, P. Du, C. Xu, C. Xu, Multi-task pruning for semantic segmentation networks, ArXiv (2020), abs/2007.08386.

[28] Y. Chen, W. Li, L.V. Gool, Road: Reality oriented adaptation for semantic segmentation of urban scenes, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 7892–7901, https://doi.org/10.1109/CVPR.2018.00823.

[29] Y. Chen, Y. Lin, M. Yang, J.B. Huang, Crdoco: Pixel-level domain transfer with cross-domain consistency, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 1791–1800, https://doi.org/10.1109/CVPR.2019.00189.

[30] Y. Chen, J. Wang, J. Li, C. Lu, Z. Luo, H. Xue, C. Wang, Lidar-video driving dataset: Learning driving policies effectively, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 5870–5878, https://doi.org/10.1109/CVPR.2018.00615.

[31] Z. Chen, J. Zhang, D. Tao, Progressive lidar adaptation for road detection, IEEE/CAA Journal of Automatica Sinica 6 (2019) 693–702, https://doi.org/10.1109/JAS.2019.1911459.

[32] Y. Choi, N. Kim, S. Hwang, K. Park, J.S. Yoon, K. An, I.S. Kweon, Kaist multi-spectral day/night data set for autonomous and assisted driving, IEEE Transactions on Intelligent Transportation Systems 19 (2018) 934–948, https://doi.org/10.1109/TITS.2018.2791533.

[33] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2017, pp. 1800–1807, https://doi.org/10.1109/CVPR.2017.195.

[34] 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: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 3213–3223, https://doi.org/10.1109/CVPR.2016.350.

[35] Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2015. The cityscapes dataset, in: CVPR Workshop on The Future of Datasets in Vision..

[36] A. Dai, A.X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2432–2443, https://doi.org/10.1109/CVPR.2017.261.

[37] J. Dai, K. He, J. Sun, Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 1635–1643, https://doi.org/10.1109/ICCV.2015.191.

[38] C.R. De Souza, A. Gaidon, Y. Cabon, A.M. López, Procedural generation of videos to train deep action recognition networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2594–2604, https://doi.org/10.1109/CVPR.2017.278.

[39] L. Deng, M. Yang, T. Li, Y. He, C. Wang, Rfbnet: Deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation, ArXiv (2019), abs/1907.00135.

[40] Dundar, A., Liu, M., Wang, T., Zedlewski, J., Kautz, J., 2018. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. arXiv preprint arXiv:1807.09384v1..

[41] N.E.D. Elmadany, Y. He, L. Guan, Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis, IEEE Transactions on Multimedia 21 (2019) 1317–1331, https://doi.org/10.1109/TMM.2018.2875510.

[42] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results VOC2012 (2012), URL: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[43] J. Fan, Z. Zhang, C. Song, T. Tan, Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 4282–4291, https://doi.org/10.1109/cvpr42600.2020.00434.

[44] V. Fischer, M. Kumar, J. Metzen, T. Brox, Adversarial examples for semantic image segmentation, ArXiv (2017), abs/1703.01101.

[45] A. Gaidon, Q. Wang, Y. Cabon, E. Vig, Virtualworlds as proxy for multi-object tracking analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 4340–4349, https://doi.org/10.1109/CVPR.2016.470.

[46] Gatys, L., Ecker, A., Bethge, M., Image style transfer using convolutional neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. p. 2414–2423. DOI: 10.1109/CVPR.2016.265..

[47] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, The International Journal of Robotics Research 32 (2013) 1231–1237, https://doi.org/10.1177/0278364913491297.

[48] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 3354–3361, https://doi.org/10.1109/CVPR.2012.6248074.

[49] Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham V.H. and Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., Schuberth, P., 2020. A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320v1..

[50] R. Gong, W. Li, Y. Chen, L. Van Gool, Dlow: Domain flow for adaptation and generalization, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 2472–2481, https://doi.org/10.1109/CVPR.2019.00258.

[51] Grönroos, S., Virpioja, S., Kurimo, M., 2020. Morfessor em+prune: Improved subword segmentation with expectation maximization and pruning, in: LREC..

[52] S. Gu, T. Lu, Y. Zhang, J.M. Alvarez, J. Yang, H. Kong, 3-d lidar + monocular camera: An inverse-depth-induced fusion framework for urban road detection, IEEE Transactions on Intelligent Vehicles 3 (2018) 351–360, https://doi.org/10.1109/TIV.2018.2843170.

[53] S. Gu, Y. Zhang, J. Tang, J. Yang, J.M. Alvarez, H. Kong, Integrating dense lidar-camera road detection maps by a multi-modal crf model, IEEE Transactions on Vehicular Technology 68 (2019) 11635–11645, https://doi.org/10.1109/TVT.2019.2946100.

[54] M. Guo, Z. Wang, N. Yang, Z. Li, T. An, A multisensor multiclassifier hierarchical fusion model based on entropy weight for human activity recognition using wearable inertial sensors. IEEE T, Hum.-Mach. Syst. 49 (2019) 105–111, https://doi.org/10.1109/THMS.2018.2884717.

[55] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, T. Harada, Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 5108–5115, https://doi.org/10.1109/IROS.2017.8206396.

[56] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, in: 2016 Asian Conference on Computer Vision (ACCV), Springer, 2016, pp. 213–228, https://doi.org/10.1007/978-3-319-54181-5_14.

[57] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[58] J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A.A. Efros, T. Darrell, Cycada: Cycle-consistent adversarial domain adaptation, 2017, arXiv preprint arXiv:1711.03213v3.

[59] Hoffman, J., Wang, D., Yu, F., Darrell, T., 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649v1..

[60] A. 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 (2017), abs/1704.04861.

[61] P. Hu, F. Perazzi, F.C. Heilbron, O. Wang, Z. Lin, K. Saenko, S. Sclaroff, Real-time semantic segmentation with fast attention, IEEE Robotics and Automation Letters 6 (2021) 263–270, https://doi.org/10.1109/LRA.2020.3039744.

[62] Hu, X., Rodríguez, F.S.A., Gepperth, A., 2014. A multi-modal system for road detection and segmentation, in: 2014 IEEE Intelligent Vehicles Symposium Proceedings (IV), IEEE. pp. 1365–1370. DOI: 10.1109/IVS.2014.6856466..

[63] Huang, D., 1996. Systematic Theory of Neural Networks for Pattern Recognition (in Chinese). Publishing House of Electronic Industry of China..

[64] D. Huang, Radial basis probabilistic neural networks: Model and application, International Journal of Pattern Recognition and Artificial Intelligence 13 (1999) 1083–1102.

[65] D. Huang, J. Du, A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks, IEEE Transactions on Neural Networks 19 (2008) 2099–2115.

[66] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 4700–4708.

[67] H. Huang, Q. Huang, P. Krähenbühl, Domain transfer through deep activation matching, in: 2018 European Conference on Computer Vision (ECCV), Springer, 2018, pp. 611–626, https://doi.org/10.1007/978-3-030-01270-0_36.

[68] Huang, X., Belongie, S., Arbitrary style transfer in real-time with adaptive instance normalization, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE. p. 1510–1519. DOI: 10.1109/ICCV.2017.167..

[69] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, R. Yang, in: The apolloscape dataset for autonomous driving, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2018, pp. 1067–10676, https://doi.org/10.1109/CVPRW.2018.00141.

[70] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Shi, H., Liu, W., Ccnet: Criss-cross attention for semantic segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE. p. 603–612. DOI: 10.1109/ICCV.2019.00069..

[71] K. Jafari-Khouzani, K. Elisevich, S. Patel, H. Soltanian-Zadeh, Dataset of magnetic resonance images of nonepileptic subjects and temporal lobe epilepsy patients for validation of hippocampal segmentation techniques, Neuroinformatics 9 (2011) 335–346, https://doi.org/10.1007/s12021-010-9096-4.

[72] Jasch, M., Fröhlich, B., Weber, T., Franke, U., Pollefeys, M., Rätsc, M., 2017. Multimodal neural networks: Rgb-d for semantic segmentation and object detection, in: Image Analysis, Springer. pp. 98–109. DOI: 10.1007/978-3-319-59126-1_9..

[73] W. Jiang, Z. Xie, Y. Li, C. Liu, H. Lu, Lrnnet: A light-weighted network with efficient reduced non-local operation for real-time semantic segmentation, in: 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), IEEE, 2020, pp. 1–6, https://doi.org/10.1109/ICMEW46912.2020.9106038.

[74] H. Jie, S. Li, S. Gang, S. Albanie, Squeeze-and-excitation networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2020) 2011–2023, https://doi.org/10.1109/TPAMI.2019.2913372.

[75] Jonathan Long, Evan Shelhamer, T.D., Fully convolutional networks for semantic segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. p. 3431–3440. DOI: 10.1109/CVPR.2015.7298965..

[76] A. Khoreva, R. Benenson, J. Hosang, M. Hein, B. Schiele, Simple does it: Weakly supervised instance and semantic segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 1665–1674, https://doi.org/10.1109/CVPR.2017.181.

[77] S. Kuutti, R. Bowden, Y. Jin, P. Barber, S. Fallah, A survey of deep learning applications to autonomous vehicle control, IEEE Transactions on Intelligent Transportation Systems 22 (2021) 712–733, https://doi.org/10.1109/TITS.2019.2962338.

[78] S. Lee, D. Kim, N. Kim, S. Jeong, Drop to adapt: Learning discriminative features for unsupervised domain adaptation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 91–100, https://doi.org/10.1109/ICCV.2019.00018.

[79] B. Li, C. Zheng, D. Huang, Locally linear discriminant embedding: An efficient method for face recognition, Pattern Recognition 41 (2008) 3813–3821, https://doi.org/10.1016/j.patcog.2008.05.027.

[80] Li, G., Yun, I., Kim, J.H., Kim, J., 2019a. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. arXiv preprint arXiv:1907.11357..

[81] H. Li, P. Xiong, H. Fan, J. Sun, Dfanet: Deep feature aggregation for real-time semantic segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 9522–9531, https://doi.org/10.1109/CVPR.2019.00975.

[82] Li, P., Liang, X., Jia, D., Xing, E., 2018. Semantic-aware grad-gan for virtual-to-real urban scene adaption. arXiv preprint arXiv:1801.01726v2..

[83] Q. Li, L. Chen, M. Li, S. Shaw, A. Nuchter, A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios, IEEE Transactions on Vehicular Technology 63 (2014) 540–555, https://doi.org/10.1109/TVT.2013.2281199.

[84] X. Li, Y. Zhou, Z. Pan, J. Feng, Partial order pruning: for best speed/accuracy trade-off in neural architecture search, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 9145–9153, https://doi.org/10.1109/CVPR.2019.00936.

[85] Li, X.T., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tan, S., Tong, Y., 2020. Semantic flow for fast and accurate scene parsing, in: 2020 European Conference on Computer Vision (ECCV), Springer. pp. 775–793..

[86] Y. Li, L. Ma, Z. Zhong, F. Liu, M.A. Chapman, D. Cao, J. Li, Deep learning for lidar point clouds in autonomous driving: A review, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 3412–3432, https://doi.org/10.1109/TNNLS.2020.3015992.

[87] Li, Y., Wang, N., Liu, J., Hou, X., Demystifying neural style transfer, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization. p. 2230–2236. 10.24963/ijcai.2017/310..

[88] Y. Li, L. Yuan, N. Vasconcelos, Bidirectional learning for domain adaptation of semantic segmentation, in: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 6929–6938, https://doi.org/10.1109/CVPR.2019.00710.

[89] Q. Lian, L. Duan, F. Lv, B. Gong, Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 6757–6766, https://doi.org/10.1109/ICCV.2019.00686.

[90] D. Lin, J. Dai, J. Jia, K. He, J. Sun, Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 3159–3167, https://doi.org/10.1109/CVPR.2016.344.

[91] Lin, G., Milan, A., Shen, C., Reid, I., Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. p. 5168–5177. DOI: 10.1109/CVPR.2017.549..

[92] K. Liu, C. Chen, R. Jafari, N. Kehtarnavaz, Fusion of inertial and depth sensor data for robust hand gesture recognition, IEEE Sensors Journal 14 (2014) 1898–1903, https://doi.org/10.1109/JSEN.2014.2306094.

[93] Liu, M., Yin, H., 2019. Feature pyramid encoding network for real-time semantic segmentation. arXiv preprint arXiv:1909.08599..

[94] S. Lo, H. Hang, S. Chan, J. Lin, Efficient dense modules of asymmetric convolution for real-time semantic segmentation, in: 2019 Proceedings of the ACM Multimedia Asia ACM, 2019, pp. 1–6.

[95] Y. Luo, P. Liu, T. Guan, J. Yu, Y. Yang, Significance-aware information bottleneck for domain adaptive semantic segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 6777–6786, https://doi.org/10.1109/ICCV.2019.00688.

[96] 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: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 2502–2511, https://doi.org/10.1109/CVPR.2019.00261.

[97] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, Y. Yang, Macro-micro adversarial network for human parsing, in: 2018 European Conference on Computer Vision (ECCV), Springer, 2018, pp. 424–440, https://doi.org/10.1007/978-3-030-01240-3_26.

[98] X. Lv, Z. Liu, J. Xin, N. Zheng, A novel approach for detecting road based on two-stream fusion fully convolutional network, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2018, pp. 1464–1469, https://doi.org/10.1109/IVS.2018.8500551.

[99] N. Ma, X. Zhang, H. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: 2018 European Conference on Computer Vision (ECCV), Springer, 2018, pp. 116–131.

[100] Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE. p. 3226–3229. DOI: 10.1109/IGARSS.2017.8127684..

[101] Mazzini, D., 2018. Guided upsampling network for real-time semantic segmentation. arXiv preprint arXiv:1807.07466..

[102] McEver, R.A., Manjunath, B.S., 2020. Pcams: Weakly supervised semantic segmentation using point supervision. arXiv preprint arXiv:2007.05615v1..

[103] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, H. Hajishirzi, Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation, in: 2018 European Conference on Computer Vision (ECCV), Springer, 2018, pp. 552–568, https://doi.org/10.1007/978-3-030-01249-6_34.

[104] M. Menze, A. Geiger, Object scene flow for autonomous vehicles, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 3061–3070, https://doi.org/10.1109/CVPR.2015.7298925.

[105] J. Metzen, M. Kumar, T. Brox, V. Fischer, Universal adversarial perturbations against semantic image segmentation, IEEE International Conference on Computer Vision (ICCV) 2017 (2017) 2774–2783.

[106] U. Michieli, M. Biasetton, G. Agresti, P. Zanuttigh, Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation, IEEE Transactions on Intelligent Vehicles 5 (2020) 508–518, https://doi.org/10.1109/TIV.2020.2980671.

[107] M. Måns Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, F. Kahl, A cross-season correspondence dataset for robust semantic segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 9524–9534, https://doi.org/10.1109/CVPR.2019.00976.

[108] Y. Mo, Z. Hou, X. Chang, J. Liang, C. Chen, J. Huan, Structural feature representation and fusion of behavior recognition oriented human spatial cooperative motion. Beijing Hangkong Hangtian Daxue Xuebao/Journal of Beijing University of Aeronautics and Astronautics 45 (2019) 2495–2505, https://doi.org/10.13700/j.bh.1001-5965.2019.0373.

[109] K. Mopuri, A. Ganeshan, R. Babu, Generalizable data-free objective for crafting universal adversarial perturbations, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019) 2452–2465, https://doi.org/10.1109/TPAMI.2018.2861800.

[110] Mukherjee, A., Das, S.D., Ghosh, J., Chowdhury, A.S., Saha, S.K., 2019. Fast geometric surface based segmentation of point cloud from lidar data, in: 2019 Pattern Recognition and Machine Intelligence (PReMI), Springer. pp. 415–423. DOI: 10.1007/978-3-030-34869-4_45..

[111] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, K. Kim, Image to image translation for domain adaptation, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 4500–4509, https://doi.org/10.1109/CVPR.2018.00473.

[112] G. Neuhold, T. Ollmann, S.R. Bulò, P. Kontschieder, The mapillary vistas dataset for semantic understanding of street scenes, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 5000–5009, https://doi.org/10.1109/ICCV.2017.534.

[113] Noh, H., Hong, S., Han, B., Learning deconvolution network for semantic segmentation, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 1520–1528. DOI: 10.1109/ICCV.2015.178..

[114] M. Oršic, I. Krešo, P. Bevandic, S. Šegvic, In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 12607–12616, https://doi.org/10.1109/CVPR.2019.01289.

[115] G. Papandreou, L. Chen, K.P. Murphy, A.L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, pp. 1742–1750, https://doi.org/10.1109/ICCV.2015.203.

[116] Paszke, A., Chaurasia, A., Kim, S., Culurciello, E., 2016. Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147..

[117] Pathak, D., Krähenbühl, P., Darrell, T., 2015. Constrained convolutional neural networks for weakly supervised segmentation, in: ICCV 2015, IEEE. pp. 1796–1804. DOI: 10.1109/ICCV.2015.209..

[118] Pathak, D., Shelhamer, E., Long, J., Darrell, T., 2014. Fully convolutional multi-class multiple instance learning. arXiv preprint arXiv:1412.7144v4..

[119] A. Patil, S. Malla, H. Gang, Y. Chen, The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 9552–9557, https://doi.org/10.1109/ICRA.2019.8793925.

[120] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters - improve semantic segmentation by global convolutional network, CVPR IEEE (2017) 1743–1751, https://doi.org/10.1109/CVPR.2017.189.

[121] Q.H. Pham, P. Sevestre, R.S. Pahwa, H. Zhan, C. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, J. Lin, A*3d dataset: Towards autonomous driving in challenging environments, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 2267–2273, https://doi.org/10.1109/ICRA40945.2020.9197385.

[122] P.O. Pinheiro, R. Collobert, From image level to pixel-level labeling with convolutional networks, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2015, pp. 1713–1721, https://doi.org/10.1109/CVPR.2015.7298780.

[123] F. Pizzati, Charette, R.d., Zaccaria, M., Cerri, P., Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2020, pp. 2979–2987, https://doi.org/10.1109/WACV45572.2020.9093540.

[124] Poudel, R.P.K., Bonde, U., Liwicki, S., Zach, C., 2018. Contextnet: Exploring context and detail for semantic segmentation in real-time. arXiv preprint arXiv:1805.04554..

[125] Poudel, R.P.K., Liwicki, S., Cipolla, R., 2019. Fast-scnn: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502..

[126] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, T. Huang, Weakly supervised scene parsing with point-based distance metric learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8843–8850, https://doi.org/10.1609/aaai.v33i01.33018843.

[127] C. Qin, L. Wang, Y. Zhang, Y. Fu, Generatively inferential co-training for unsupervised domain adaptation, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 1055–1064, https://doi.org/10.1109/ICCVW.2019.00135.

[128] F. Remondino, Heritage recording and 3d modeling with photogrammetry and 3d scanning, Remote Sensing 3 (2011) 1104–1138, https://doi.org/10.3390/rs3061104.

[129] E. Romera, J.M. Àlvarez, L.M. Bergasa, R. Arroyo, Erfnet: Efficient residual factorized convnet for real-time semantic segmentation, IEEE Transactions on Intelligent Transportation Systems 19 (2018) 263–272, https://doi.org/10.1109/TITS.2017.2750080.

[130] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[131] H. Roth, L. Lu, A. Farag, H. Shin, J. Liu, E. Turkbey, R. Summers, Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 556–564.

[132] Saito, K., Ushiku, Y., Harada, T., Saenko, K., 2017. Adversarial dropout regularization. arXiv preprint arXiv:1711.01575v3..

[133] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 3723–3732, https://doi.org/10.1109/CVPR.2018.00392.

[134] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, Mobilenetv 2: Inverted residuals and linear bottlenecks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 4510–4520.

[135] Schneider, L., Jasch, M., Fröhlich, B., Weber, T., Franke, U., Pollefeys, M., Rätsch, M., 2017. Multimodal neural networks: Rgb-d for semantic segmentation and object detection, in: SCIA..

[136] L. Shang, D. Huang, J. Du, C. Zheng, Palmprint recognition using fastica algorithm and radial basis probabilistic neural network, Neurocomputing 69 (2006) 1782–1786.

[137] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE transactions on pattern analysis and machine intelligence 39 (2017) 640–651, https://doi.org/10.1109/TPAMI.2016.2572683.

[138] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: 2012 European Conference on Computer Vision, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 746–760.

[139] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR (2015), abs/1409.1556.

[140] X. Song, P. Wang, D. Zhou, R. Zhu, C. Guan, Y. Dai, H. Su, H. Li, R. Yang, Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 5447–5457, https://doi.org/10.1109/CVPR.2019.00560.

[141] Spadotto, T., Toldo, M., Michieli, U., Zanuttigh, P., 2020. Unsupervised domain adaptation with multiple domain discriminators and adaptive self-training. arXiv preprint arXiv:2004.12724v1..

[142] L. Sun, K. Yang, X. Hu, W. Hu, K. Wang, Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images, IEEE Robotics and Automation Letters 5 (2020) 5558–5565, https://doi.org/10.1109/LRA.2020.3007457.

[143] Y. Sun, W. Zuo, M. Liu, Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes, IEEE Robotics and Automation Letters 4 (2019) 2576–2583, https://doi.org/10.1109/LRA.2019.2904733.

[144] Y. Sun, W. Zuo, P. Yun, H. Wang, M. Liu, Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion, IEEE Transactions on Automation Science and Engineering 18 (2021) 1000–1011, https://doi.org/10.1109/tase.2020.2993143.

[145] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[146] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015 (2015) 1–9, https://doi.org/10.1109/CVPR.2015.7298594.

[147] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 (2016) 2818–2826, https://doi.org/10.1109/CVPR.2016.308.

[148] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, C. Schroers, Normalized cut loss for weakly-supervised cnn segmentation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 1818–1827, https://doi.org/10.1109/CVPR.2018.00195.

[149] M. Tang, F. Perazzi, A. Djelouah, I. Ayed, C. Schroers, Y. Boykov, On regularized losses for weakly-supervised cnn segmentation, in: 2018 European Conference on Computer Vision (ECCV), 2018, https://doi.org/10.1007/978-3-030-01270-0_31.

[150] M. Toldo, A. Maracani, U. Michieli, P. Zanuttigh, Unsupervised domain adaptation in semantic segmentation: A review, Technologies 8 (2020) 35, https://doi.org/10.3390/TECHNOLOGIES8020035.

[151] M. Toldo, A. Maracani, U. Michieli, P. Zanuttigh, Unsupervised domain adaptation in semantic segmentation: A review, Technologies 8 (2020) 35, https://doi.org/10.3390/TECHNOLOGIES8020035.

[152] M. Toldo, U. Michieli, G. Agresti, P. Zanuttigh, Unsupervised domain adaptation for mobile semantic segmentation based on cycle consistency and feature alignment, Image and Vision Computing 95 (2020), https://doi.org/10.1016/j.imavis.2020.103889 103889.

[153] X. Tong, G. Xia, Q. Lu, H. Shen, S. Li, S. You, L. Zhang, Learning transferable deep models for land-use classification with high-resolution remote sensing images, ArXiv (2018), abs/1807.05713.

[154] A. Torralba, R. Fergus, W.T. Freeman, 80 million tiny images: a large data set for nonparametric object and scene recognition, IEEE transactions on pattern analysis and machine intelligence 30 (2008) 1958–1970, https://doi.org/10.1109/TPAMI.2008.128.

[155] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, M. Chandraker, Learning to adapt structured output space for semantic segmentation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 7472–7481, https://doi.org/10.1109/CVPR.2018.00780.

[156] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, 2016, arXiv preprint arXiv:1607.08022v3.

[157] J. Van Brummelen, M. O'Brien, D. Gruyer, H. Najjaran, Autonomous vehicle perception: The technology of today and tomorrow, Transportation Research Part C: Emerging Technologies 89 (2018) 384–406, https://doi.org/10.1016/J.TRC.2018.02.012.

[158] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, C.V. Jawahar, Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1743–1751, https://doi.org/10.1109/WACV.2019.00190.

[159] P. Vernaza, M. Chandraker, Learning random-walk label propagation for weakly-supervised semantic segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2953–2961, https://doi.org/10.1109/CVPR.2017.315.

[160] Vijay Badrinarayanan, R.C. Alex Kendall, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE transactions on pattern analysis and machine intelligence 39 (2017) 2481–2495, https://doi.org/10.1109/TPAMI.2016.2644615.

[161] R.P.D. Vivacqua, M. Bertozzi, P. Cerri, F.N. Martins, R.F. Vassallo, Self-localization based on visual lane marking maps: An accurate low-cost approach for autonomous driving, IEEE Transactions on Intelligent Transportation Systems 19 (2018) 582–597, https://doi.org/10.1109/TITS.2017.2752461.

[162] T. Vu, H. Jain, M. Bucher, M. Cord, P.P. Pérez, Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 2512–2521, https://doi.org/10.1109/CVPR.2019.00262.

[163] T. Vu, H. Jain, M. Bucher, M. Cord, P.P. Pérez, Dada: Depth-aware domain adaptation in semantic segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 7363–7372, https://doi.org/10.1109/ICCV.2019.00746.

[164] Wang, B., Qi, G., Tang, S., Zhang, T., Wei, Y., Li, L., Zhang, Y., 2019a. Boundary perception guidance: A scribble-supervised semantic segmentation approach, in: Proceedings of the Twenty-Eighth International Joint

Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization. pp. 3663–3669. 10.24963/ijcai.2019/508..
[165] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2016) 2010–2023, https://doi.org/10.1109/TPAMI.2015.2505311.
[166] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153, https://doi.org/10.1016/j.neucom.2018.05.083.
[167] Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G., Understanding convolution for semantic segmentation, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. p. 1451–1460. DOI: 10.1109/WACV.2018.00163..
[168] X. Wang, D. Huang, A novel density-based clustering framework by using level set method, IEEE Transactions on Knowledge and Data Engineering 21 (2009) 1515–1531, https://doi.org/10.1109/TKDE.2009.21.
[169] X. Wang, D. Huang, H. Xu, An efficient local chan-vese model for image segmentation, Pattern Recognition 43 (2010) 603–618, https://doi.org/10.1016/j.patcog.2009.08.002.
[170] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, L.J. Latecki, Lednet: A lightweight encoder-decoder network for real-time semantic segmentation, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1860–1864.
[171] K. Watanabe, K. Saito, Y. Ushiku, T. Harada, Multichannel semantic segmentation with unsupervised domain adaptation, in: 2018 European Conference on Computer Vision Workshops (ECCVW), Springer, 2018, pp. 600–616, https://doi.org/10.1007/978-3-030-11021-5_37.
[172] J. Wei, G. Lin, K.H. Yap, T.Y. Hung, L. Xie, Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 4383–4392, https://doi.org/10.1109/cvpr42600.2020.00444.
[173] B. Wu, A. Wan, X. Yue, K. Keutzer, Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 1887–1893, https://doi.org/10.1109/ICRA.2018.8462926.
[174] Z. Wu, X. Han, Y. Lin, M.G. Uzunbas, T. Goldstein, S.N. Lim, L.S. Davis, Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation, in: 2018 European Conference on Computer Vision (ECCV), Springer, 2018, pp. 535–552, https://doi.org/10.1007/978-3-030-01228-1_32.
[175] Z. Wu, C. Shen, A. van den Hengel, Real-time semantic image segmentation via spatial sparsity, ArXiv (2017), abs/1712.00213.
[176] Z. Wu, X. Wang, J. Gonzalez, T. Goldstein, L. Davis, Ace: Adapting to changing environments for semantic segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 2121–2130, https://doi.org/10.1109/ICCV.2019.00221.
[177] F. Wulff, B. Schäufele, O. Sawade, D. Becker, B. Henke, I. Radusch, Early fusion of camera and lidar for robust road detection based on u-net fcn, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2018, pp. 1426–1431, https://doi.org/10.1109/IVS.2018.8500549.
[178] W. Xia, C. Domokos, J. Dong, L. Cheong, S. Yan, Semantic segmentation without annotating segments, in: 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, 2013, pp. 2176–2183, https://doi.org/10.1109/ICCV.2013.271.
[179] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, A. Yuille, Adversarial examples for semantic segmentation and object detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 1378–1387, https://doi.org/10.1109/ICCV.2017.153.
[180] J. Xie, M. Kiefel, M. Sun, A. Geiger, Semantic instance annotation of street scenes by 3d to 2d label transfer, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 3688–3697, https://doi.org/10.1109/CVPR.2016.401.
[181] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017 (2017) 5987–5995, https://doi.org/10.1109/CVPR.2017.634.
[182] J. Xue, J. Fang, T. Li, B. Zhang, P. Zhang, Z. Ye, J. Dou, Blvd: Building a large-scale 5d semantics benchmark for autonomous driving, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 6685–6691, https://doi.org/10.1109/ICRA.2019.8793523.
[183] Yan, Z., Sun, L., Krajník, T., Ruichek, Y., 2019. Eu long-term dataset with multiple sensors for autonomous driving. arXiv preprint arXiv:1909.03330v3..
[184] Yang, X., Wu, Y., Zhao, J., Liu, F., 2020a. Dense dual-path network for real-time semantic segmentation. arXiv preprint arXiv:2010.10778..
[185] Y. Yang, D. Lao, G. Sundaramoorthi, S. Soatto, in: Phase consistent ecological domain adaptation, in 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 9008–9017, https://doi.org/10.1109/cvpr42600.2020.00903.
[186] Y. Yang, S. Soatto, Fda: Fourier domain adaptation for semantic segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020, pp. 4084–4094, https://doi.org/10.1109/cvpr42600.2020.00414.
[187] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, S. Chennupati, M. Uricar, S. Milz, M. Simon, K. Amende, C. Witt, H. Rashed, S. Nayak, S. Mansoor, P. Varley, X. Perrotton, D. Odea, P. Pérez, Woodscape: A multi-task, in: multi-camera fisheye dataset for autonomous driving, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2019, pp. 9307–9317, https://doi.org/10.1109/ICCV.2019.00940.
[188] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: 2018 European Conference on Computer Vision (ECCV), Springer, 2018, pp. 325–341, https://doi.org/10.1007/978-3-030-01261-8_20.
[189] Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122v3..
[190] Zhang, D., Zhang, H., Tang, J., Hua, X., Sun, Q., 2020. Causal intervention for weakly-supervised semantic segmentation. arXiv preprint arXiv:2009.12547v2..
[191] Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A., Context encoding for semantic segmentation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. p. 7151–7160. DOI: 10.1109/CVPR.2018.00747..
[192] Y. Zhang, Z. Qiu, T. Yao, D. Liu, T. Mei, Fully convolutional adaptation networks for semantic segmentation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 6810–6818, https://doi.org/10.1109/CVPR.2018.00712.
[193] Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J., Icnet for real-time semantic segmentation on high-resolution images, in: 2018 European Conference on Computer Vision (ECCV), Springer. p. 418–434. DOI: 10.1007/978-3-030-01219-9_25..
[194] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 6230–6239, https://doi.org/10.1109/CVPR.2017.660.
[195] Z. Zhao, H. Glotin, Z. Xie, J. Gao, X. Wu, Cooperative sparse representation in two opposite directions for semi-supervised image annotation, IEEE Transactions on Image Processing 21 (2012) 4218–4231, https://doi.org/10.1109/TIP.2012.2197631.
[196] Z. Zhao, D. Huang, B. Sun, Human face recognition based on multi-features using neural networks committee, Pattern Recognition Letters 25 (2004) 1351–1358, https://doi.org/10.1016/j.patrec.2004.05.008.
[197] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5122–5130, https://doi.org/10.1109/CVPR.2017.544.
[198] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ade20k dataset, International Journal of Computer Vision 127 (2018) 302–321, https://doi.org/10.1007/s11263-018-1140-0.
[199] Zhou, Q., Feng, Z., Cheng, G., Tan, X., Shi, J., Ma, L., 2020a. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. arXiv preprint arXiv:2004.08878v1..
[200] Q. Zhou, Y. Wang, J. Liu, X. Jin, L.J. Latecki, An open-source project for real-time image semantic segmentation, Science China Information Sciences 62 (2019), https://doi.org/10.1007/s11432-019-2685-1.
[201] Z. Zhou, M. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Transactions on Medical Imaging 39 (2020) 1856–1867, https://doi.org/10.1109/TMI.2019.2959609.
[202] F. Zhu, L. Zhu, Y. Yang, in: Sim-real joint reinforcement transfer for 3d indoor navigation, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 11380–11389, https://doi.org/10.1109/CVPR.2019.01165.
[203] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2242–2251, https://doi.org/10.1109/ICCV.2017.244.
[204] X. Zhu, H. Zhou, C. Yang, J. Shi, D. Lin, Penalizing top performers: Conservative loss for semantic segmentation adaptation, in: 2018 European Conference on Computer Vision (ECCV), Springer, 2018, pp. 587–603, https://doi.org/10.1007/978-3-030-01234-2_35.
[205] J. Zhuang, J. Yang, L. Gu, N. Dvornek, Shelfnet for fast semantic segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 847–856, https://doi.org/10.1109/ICCVW.2019.00113.

Yujian Mo is a PhD candidate in the college of Electronic and Information engineering, Tongji University, Shanghai, China. Now he is mainly engaged in artificial intelligence, deep learning, image processing, pattern recognition and self-driving.

Yan Wu is a full professor and Doctoral Advisor in the College of Electronics and Information Engineering, Tongji University, Shanghai, China. She received her Ph. D. degree in traffic information engineering and control from Shanghai Tiedao University, China, in 1999. From 2000 to 2003, she had worked as a Postdoctoral Research Fellow in Department of Electric Engineering, Fudan University, China. She has published more than 140 papers on important national and international journals and conference proceedings. Now she is mainly engaged in deep learning, image processing, pattern recognition, and self-driving.

Xinneng Yang is a master degree candidate in the college of Electronic and Information engineering, Tongji University, Shanghai, China. Now he is mainly engaged in artificial intelligence, deep learning, image processing, pattern recognition and self-driving.

Feilin Liu is a master candidate in the college of Electronic and Information engineering, Tongji University, Shanghai, China. He received his B.Eng degree in software engineering from Wuhan University of Technology, China in 2019. Now he is mainly engaged in computer vision, deep learning, image processing, pattern recognition and self-driving.

Yujun Liao is a master degree candidate in the college of Electronic and Information engineering, Tongji University, Shanghai, China. He received his bachelor degree in computer science from Tongji University and Politecnico di Torino in 2020. Now he is mainly engaged in artificial intelligence, big data, deep learning, image processing, pattern recognition and self-driving.