How deep learning is empowering semantic segmentation

Traditional and deep learning techniques for semantic segmentation: A comparison

Uroosa Sehar1 · Muhammad Luqman Naseem2

Received: 3 June 2021 / Revised: 28 January 2022 / Accepted: 9 March 2022 / Published online: 6 April 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Semantic segmentation involves extracting meaningful information from images or input from a video or recording frames. It is the way to perform the extraction by checking pixels by pixel using a classification approach. It gives us more accurate and fine details from the data we need for further evaluation. Formerly, we had a few techniques based on some unsupervised learning perspectives or some conventional ways to do some image processing tasks. With the advent of time, techniques are improving, and we now have more improved and efficient methods for segmentation. Image segmentation is slightly simpler than semantic segmentation because of the technical perspective as semantic segmentation is pixels based. After that, the detected part based on the label will be masked and refer to the masked objects based on the classes we have defined with a relevant class name and the designated color. In this paper, we have reviewed almost all the supervised and unsupervised learning algorithms from scratch to advanced and more efficient algorithms that have been done for semantic segmentation. As far as deep learning is concerned, we have many techniques already developed until now. We have studied around 120 papers in this research area. We have concluded how deep learning is helping in solving the critical issues of semantic segmentation and gives us more efficient results. We have reviewed and comprehensively studied different surveys on semantic segmentation, specifically using deep learning.

Keywords Deep learning algorithm · Pattern recognition · Semantic segmentation · Image segments · Object detection

Uroosa Sehar
1505253990@qq.com

Muhammad Luqman Naseem
mluqmannaseem@outlook.com

1 University of Engineering and Technology, Taxila, Pakistan
2 Northeastern University, Shenyang, China
1 Introduction

Semantic segmentation is a pixel-to-pixel classification based on label data [54, 55]. It is the basic technique to get some meaningful information to extract the required [7, 53, 92]. After getting the information, you can mention that certain specific objects belong to specific classes [9, 59, 81]. It is the most challenging field nowadays. It has many more diverse applications in every field like education, medicine, weather prediction, climatic change prediction, and so on [4]. Deep learning-based segmentation methods have been widely used in other vision tasks such as video object segmentation [53–55, 83]. The experimental results have proved that the dataset training and preliminary changes, i.e., pre-processing, affect the results. You can say before training; we can make our data feasible to perform some operations altogether. The datasets we will take also depend on the type of operation you need to perform [46]. The related work in this area is more technical and challenging [38, 105]. Many surveys have been conducted to have a thorough overview of semantic segmentation [2]. The challenges and the new approaches have led this particular field, more worth and attention. Recent research and development show that the perfect classification can lead to better semantic segmentation results [39, 58, 113]. Subsequently, many methods are deriving day by day; we will go through all the basic details, and after that, we will see how deep learning algorithms will help us get the most efficient result [96]. Region-based segmentation, graph-based segmentation, image segmentation [26, 117], instance segmentation [56], semantic segmentation all had the same basic but different procedures. Figure 1 describes the moving trend for segmentation in recent years. Figure 2 will show you all the state-of-the-art techniques that can be used for semantic segmentation. The related work in this field using deep learning is progressing every day.

1.1 Key challenges

Some algorithms can perform on a specific dataset we give as input; it will not provide the same results on other datasets [102]. The primary reason behind it is that the different dataset does not perform the same operations before getting to the training phase and the testing phase; the second reason is that while going through the machine learning process,
we imagine that our whole data set does not have any ambiguity and in result, it will yield the best result which is more efficient and accurate [122]. Sometimes the dataset has few samples, like maybe less than 1000 images as a whole. In that case, our model will be under-fitted for such types of problems [90, 106, 107]. Model overfitting and the underfitting problem often occur when we have many or few samples [16]. We can now imagine that an algorithm is not equally performed in all dataset types, so in this paper, we will see how we can change something and improve efficiency [37]. For example, if you have a satellite image dataset [116], that means images from satellites of any region need to do some predictive analysis. You need to evaluate the climatic change in a specific area [121]. For that, you will get all the training datasets from already captured pictures taken from any satellite. Now you will examine all the images you have already accepted as a map. Here, you can use any map like Apple map, Baidu map, Google map, and so on [76]. After getting all the pictures, you can say on the images dataset for that specific area and apply some deep learning algorithms to evaluate semantic segments.

1.2 Classification

Classifying the objects that belong to the forest region [91], i.e., the tree is your required masked area. Now, you can easily highlight the regions of the area where we have trees. Our next step is to mention the label across that particular area [70, 71]. In this way, you can easily see which part of the country has more forest and less. Due to the forest, we are facing some climatic changes. Hence, you have noticed that using a semantic segmentation technique, and you can quickly get a prediction about a specific period, also like when this area grows more trees and has good weather and so on. Semantic segmentation has many more examples applicable in any place, like in the medical field for automatic diagnosis of Schizophrenia, where we can use CNN-LSTM models and the EEG Signals [22, 69, 75].
1.3 Object detection

Object detection [10] is also another essential part of semantic segmentation. After classification, you will make some edges and then detect the object [42]. Another helpful example is detecting some diseases by taking any medical training dataset [72]. The most recent research in this particular area is to detect COVID-19 [21, 73] when you have a sample of lungs, all the images for the infected lungs you can train. After that, you can guess and evaluate the differences you can see from infected lungs’ images from healthy lungs [105]. The most popular is the chest X-Ray scan. Here in this example, you can take all the data set and then evaluate which image had a particular sort of disease, and then you can spot the infected area and then mask that specific area. After that, the difference between the input lung image and the healthy lung image will have the calculation for the evaluation based on some critical analysis [62]. Therefore, you can see that using a good and accurate algorithm can give you more efficient results. You can easily detect the climatic change in a specific area in weather prediction cases. Here, the problem you can encounter is getting the primary data set and all the behavioral changes with time. Before getting all the data set and images, you will need to analyze before making your dataset. So, in this field, you can say getting all the data is also a critical step in dealing with or applying some deep learning algorithms [40]. Our purpose is to deliver a core understanding of all these concepts. We have first used supervised learning algorithms [37], and after that, we have seen how it will not give us the most efficient result, and we will jump to the most efficient algorithm. The algorithm that we have tested in the beginning uses some simple images or saves some Lighter images. We have used some already available libraries to get the results, and after that, we tried the more complicated algorithms.

2 Datasets for semantic segmentation

This research area is becoming popular day by day. We have tried the algorithms on CamVid [113], Pascal VOC [76], and COCO dataset [100]. The dataset that we have used throughout is CamVid. Most of the time, it works for any technique [57]. While doing masking, we have used the COCO dataset. Cambridge-driving Labeled Video Dataset consists of 367 training pairs, 101 validation pairs, 233 test pairs with 32 semantic classes. It is supposed to be the most advanced dataset for real-time semantic segmentation [63]. The other dataset that we have experimented on is the COCO dataset. COCO, or we can refer to Microsoft Common Object in Context dataset, is a large image dataset explicitly designed for image processing tasks like object detection, person keypoint detection, caption generation, segmentation, and many other prevalent problems these days. It has around 80 classes and has more detailed objects inside the images. Moreover, we have also used some small datasets like balloons dataset, shapes [19] (to detect rectangle, circle, triangle, etc.), nucleus [74, 105] (for medical relevant field). The purpose of using these datasets is to check how algorithmic network architecture can work on other datasets with the same accuracy [123, 124] and loss.

Many other semantic segmentation datasets like Mapillary Vistas [61] contain around 25000 high-resolution images with the 66 defined semantic classes. Moreover, it also has instance-specific labels for the 37 classes. Mainly, the classes belong to the road scenes same as in the cityscapes dataset, but it has more annotations than cityscapes. Moreover, the Semantic KITTI is also used as the outclass dataset to understand the semantics of scenes [5]. It is based on the KITTI Vision with all the sequences or frames with overall angles (Table 1).
### Table 1  State of the art

| Algorithm implemented | Dataset |
|-----------------------|---------|
| Mask R-CNN            | CityScapes, CoCo, Balloons, Nucleus, Shapes |
| Faster R-CNN          | CoCo |
| CNN                   | Cifar-10, Mnist Dataset |
| ESSN                  | CamVid |
| UNet                  | IJB-A, Camvid |
| VGG-UNet for          | |
| Aerial segmentation   | Semantic drone dataset |

BDD is called Berkeley Deep Drive Dataset as the detailed dataset for urban scenes for autonomous driving [33, 85, 91]. This dataset can perform the complete tasks, including object detection, multi-object tracking, semantic segmentation, instance segmentation, segmentation tracking, and lane detection. For the semantic segmentation, BDD has 19 classes, and samples are not so practical for urban scenes semantic segmentation. Wildash 2 is also a primary dataset for semantic segmentation, but it has limited material, i.e., training and testing samples, to fulfill the algorithm’s requirements. So, it is advisable to prefer the other highly organized and well-managed datasets [110].

Wildpass is considered as a panoramic semantic segmentation taken from the cities. It has two versions with the same datasets. One is WildPass and other is WildPass2K. The first one contains 500 panoramas from 25 cities, and WildPASS2K contains 2000 labeled panoramas taken from 40 cities. Although the variation is acceptable, this dataset is not recommended to deal with the complex scenarios while dealing with urban scenes [104].

### 3  Algorithms: Classical VS. Deep learning

#### 3.1  Grab-Cut algorithm: Region-based segmentation

Grab-cut is an image-based segmentation [67] that is important in getting the object based on the defined area or region. Here, we can also extract the image information based on dividing the image into two regions: background and foreground; after that, we can make a segment. In the region-based algorithm, we have used Grab-cut [62] using open CV in Implementation using Python; the steps are under below, 1. Take in an input image. 2. We have taken foreground and background separately so that we can input them in our grab cut algorithm function. 3. After that, the defined rectangle will map on the foreground image, showing the segment separately.

#### 3.1.1  Results

This experiment result depends on the rectangle size that we have already made in Step 3. So, you cannot define this algorithm as efficient because if you change the area argument value, it will not give us the same result (Fig. 3).
3.2 K-map image segmentation

This algorithm is mainly designed to get the required segment of an image using a K map. It will get the necessary mass based on the neighborhood, and after that, we can get the required section of the image based on RGB values. The images we will use as input are also not high-resolution images. This algorithm can quickly work only on low-resolution images and easily get all the segments based on colors and neighborhoods (Table 2). The main steps of this algorithm are explained below

- First, take input of the image; you can take any image you want.
- Apply the Gaussian filter to let smoothen the image.
- Then, we have created a graph for building segments.
- Segment graph by merging the gay neighborhood, which is in this case we have used four or eight you can also say based on neighborhood or similarity.
- After those segments have been created, and it will generate an output image with all the details mapped on the plot.
- Then, it will return a graph with edges of the input image.
- Two segment graphs we have used the threshold to give us the result output image as a segmented image.

3.2.1 Results

This algorithm is efficient and can get all the segments after the image based on colors. The only drawback that we have evaluated is that when you take a high-resolution image, this algorithm becomes slower and also causes a delay. There is an ambiguity of classes. As we have not defined this particular algorithm, we can see that the road that should be colored with one color is also segmented into so many parts. Furthermore, the person in the original image must mask as one object, the same as the car beside the person. Meeting and coping these all things in mind, we refer to the deep learning approach in which we will see how classes can categorize and read by the machine (Fig. 4).

3.3 Deep learning approaches

The algorithm that we have tried is based on the up-down settings of the convolutional blocks and found the semantic segments as a color map. Currently, State-of-the-art methods
Table 2  CAMVID dataset image and segmentation result

| Image resolution | Number of components | Running time |
|------------------|----------------------|--------------|
| 500, 281 (RGB)   | 15                   | 12.26(s)     |
| 1244,422 (RGBA)  | 20                   | 41.4(s)      |
| 960,720 (RGB)    | 63                   | 53.06        |
| 3455,4320 (RGB)  | 890                  | 1889(s)      |

include many approaches to deal with semantic segmentation problems. The encoder-decoder-based, i.e., Fully transformer-based network models, are very popular, and they also give us promising result [80]. Modern research has applied the fully transformer-based architectures, and some adopted the CNN-based semantic segmentation model. Moreover, the hybrid network models are also a practical approach to solve these problems [95].

For semi-supervised segmentation, consistency regularization is a popular topic. Consistent predictions and intermediate characteristics are enforced by this method. Input perturbation techniques randomly augment the input pictures and apply a consistency constraint between the predictions of enhanced images, such that the decision function is in the low-density zone. Multiple decoders are used in a feature perturbation technique to ensure that the outputs of the decoder are consistent. Furthermore, the GCT technique further executes network perturbation by employing two segmentation networks with the same structure but started differently and ensured consistency between the perturbed networks [104].

Weakly Supervised Semantic Segmentation or WSSS with image-level labels has made significant progress in developing pseudo masks and training a segmentation network [43]. Recently, WSSS approaches have relied on CAMs to locate objects by identifying picture pixels helpful in classifying them. It does not mean that CAMs do not create helpful pseudo masks; they only accentuate the most discriminative portions of an object. There has been a great deal of work invested into finding a solution to this problem. For this purpose, they are employing tactics such as region-erasing, region monitoring, and regional growth to complete the CAMs. Other approaches use an iterative process to develop the CAM. Using random walks, PSA and IRN, for example, suggest spreading local replies across surrounding semantic entities. As previously stated, this problem stems from a lack of coordination between categorization and segmentation. Many academics have noticed this and
are investigating ways to decrease the gap using extra supervision, such as CAM consistency, cumulative feature maps, cross-image semantics, sub-categories, saliency maps, and multi-level feature maps requirements. These strategies are straightforward, yet they provide positive results [20].

While we consider the context-based mechanism, OCNet (Object Context Network for Scene Parsing) would be a better option to select as the baseline. Logically, it contains the Resnet-FCN and objects Context Module. After the classifier, it again upsamples the image to parse the scene and provide the mask [109]. It has many variations like Base-OC, Pyramid-OC, ASP-OC (Atrous Spatial Pooling). Besides OCNet, we can have significantly matured network models like RFNET or ACNET that use asymmetric convolution blocks to strengthen the kernel structure. This network also helps us save extra computation time. Moreover, SETR (Segmentation Transformer) is the latest network architecture for the transformer-based mechanism that challenges the excellent mIoU of 50.28% for the ADE20K dataset and 55.83% for Pascal Context, and also give us promising results on the Cityscapes dataset [36, 77]. There are other latest transformer-based semantic segmentation models, i.e., Trans4Trans(Transformer for Transparent Object Segmentation) and SegFormer(Semantic Segmentation with Transformers) that are significantly less computational network architecture that can give us multi-scale features [99, 114]. SegFormer minimizes the effect of using complex decoders. Technically, it adds the learned features from all layers and the maximized and enriched representation. [99] also re-scale the basic approach and found very well-noted and robust results for up to 84.0% while experimenting on the Cityscapes dataset.

Omnisupervised learning framework is also designed for efficient CNNs, which adds different data sources. In this way, it will improve the reliability in unseen areas. So, the traditional CNN uses an unsupervised framework to take advantage of both labeled and unlabeled panoramas [103]. Now, researchers plan to take a panoramic panoptic segmentation approach to better scene understanding.

The traditional Semantic segmentation is based on RGB images, which is not a reliable way to deal with complex outdoor scenarios. Polarization sensing can be adopted as an efficient approach for dealing with these issues. By getting the information from the optical sensors, we can get the exact information regardless of what materials we are incorporating [47, 98].

There is another challenging aspect of semantic segmentation, i.e., 3D Segmentation in computer vision, that can be applied in autonomous driving, medical image analysis, and robotics. Usually, it applies to the handcrafted features while considering the shortcomings of the 2D-based segmentation [31]. Several models are famous in 3D Semantic segmentation, including ShapeNet, PSB, COSeg, ScanNet, etc. [60],

Another fascinating aspect of semantic segmentation is mapping the associated pixels in night scenes where the illumination concept is negligible. In this way, the real objects are not fully seen, and network models need to look down into deep details. The re-weighting methodology has been adopted before to cope with the false predictions [94]. The publicly available datasets can be used for that subject. These scenarios can be compared with the medical images, where images are almost grey, and the features are not correctly shown. The dataset we have used for this is the CamVid dataset, i.e., the dataset contains images of all around 960x720 dimensions, with approximately. The network architecture we will be following is based on core UNET. We will have 32 semantic classes; each class has a defined color (RGB) defined by the CamVid. The classes have been categorized by the characteristic of a particular object in our input. Mainly, moving objects has animals, pedestrians, children,
rolling carts, bicyclists, motorcycles, cars, trucks, trains. Moreover, we have Road (shoulder, lane marking drivable, non-drivable), Ceiling (sky, tunnel, archway), and Fixed Objects (Building, wall, Tree, Fence, sidewalk, parking block, pole, traffic cone, bridge, Sign, Text, Traffic lights) respectively. The exact labeled image is shown in Figs. 5 and 6.

### 3.3.1 Network architecture

This model was developed by Olaf Ronneberger [66]. They developed a model for doing image segmentation on biomedical data. We have tried the UNET model on the CamVid dataset. Taking the input dataset image of size 960x720 and all the images are in png format. It is recommended to use images in png format if you are preprocessing data or making your dataset in certain environmental conditions and camera quality because image in other formats is not so feasible for doing all types of operations that we want to perform while performing deep neural network operations. The model used for this type of segmentation is by upsampling the image matrix by using a convolutional block [116]. While using the max-pooling layer, the image changes from high resolution to low resolution. Here we can get the extract part of the image, i.e., a feature required to be extracted. Moreover, here we can see that the image size decreases, but the depth, context, or receptive field enhances [30]. In return, we can see how size increases, resulting in losing the information that where our actual matrix value was where place. To solve this problem, we have another step for decoding the information that was downsampled before, and then it will pass to the transposed convolutional network to upsample it. During downsampling, we compute the parameters for the transpose convolution such that the image’s height and breadth are doubled, but the

![Fig. 5 Comparison for All state-of-the-art network architectures class IoU and Category IoU using Cityscapes dataset](image-url)
number of channels is halved. We will get the required dimensions with the exact information that will increase the accuracy in return. Results in Fig. 9 show results for the tested algorithm on CamVid Dataset.
Table 3  Experimental results after Applying VGG-UNET on Aerial Dataset  Experimental results for CamVid and Drone dataset using VGG-UNet

| Epochs | Camvid dataset          | Aerial drone dataset      |
|--------|-------------------------|---------------------------|
|        | Train Acc | Train loss | Train Acc | Train loss |
| 0      | 0.6453002 | 1.38478502 | 0.5847   | 1.5115     |
| 1      | 0.7320985 | 0.94604099 | 0.6521   | 1.1718     |
| 2      | 0.7695483 | 0.81193501 | 0.6796   | 1.0628     |
| 3      | 0.7769584 | 0.77720939 | 0.7054   | 0.9806     |
| 4      | 0.8051298 | 0.68544247 | 0.7066   | 0.9753     |
| 5      | 0.8133175 | 0.65179542 | 0.7287   | 0.9028     |

UNet has also been applied on the Aerial or Drone dataset [59] and with VGG [48] as the network backbone. The drone dataset is the freely available dataset. It has 23 semantic classes. Figure 7 shows the segmentation results after applying UNET to this dataset. UNet outperforms in most situations. The Table 3 shows the accuracy and loss values for both datasets in 5 epochs (Figs. 8, 9, 10, 11 and 12).

3.4 Masking: Instance segmentation using object detection

Another famous and more useful algorithm for datasets like VOC2012 [49], Camvid, and Coco datasets using Mask RCNN is the extension for Fast RCNN, mainly for object detection. The experiment was done using resnet101 as the backbone. Resnet-101 [39] is the CNN network which is 101 deep layers. What you have to do is to load a pre-trained network trained on more than a million images from the ImageNet dataset with 1.5 million images with the 1000 image classes. All input images are of size 224*224. Resnet pre-trained on the ImageNet dataset can be further evaluated using the Fast RCNN (Figs. 13, 14, 15 and 16). After that, it can validate all the datasets containing classes that contain object classes similar to it. For example, the COCO dataset configured pre-trained weight file can validate...
the COCO dataset. Moreover, when we tested it with VOC, it showed a good efficiency of around 98-99% in every object class. After detecting the objects that belong to different classes, it will mask that instance. Eventually, it proved fruitful with class-wise object detection on every image while testing the CamVid dataset.

4 Experimental conditions

The experiments are performed under good hardware specifications. For conventional algorithms and Mask-RCNN experiment configurable to 2.2GHz dual-core Intel Core i7, Turbo Boost up to 3.2GHz, with 4MB shared L3 cache. UNET experiment was done under-speed
boost with NVidia Tesla P100 GPUs. Selecting the system or hardware for semantic segmentation algorithms’ customization and performance analysis is also a key aspect [113]. Efficiency affects the results for segmentation. It is due to the reason that we have non-local information in semantic segmentation. Sometimes, we focus on the feature map level rather than the image level. However, we have already lost spatial information while focusing on the last feature map. We must save the residuals from managing the dependency between the pixels far away within the image.

5 Evaluation metrics

The performance evaluation can vary from problem to problem when making the deep learning neural network. Mainly, traditional methods like KNN [81], Decision tree with Boosting
Fig. 13  Masking on CoCo dataset
Fig. 14  (a) Results for Mask-RCNN applied on COCO dataset

[85], SVM [53], conditional random fields, or any statistical-based approach use accuracy or precision as performance evaluation metrics. As far as deep learning is concerned, we have more performance metrics for Classification, Object Detection, and Semantic Segmentation [89].

- Normally, we evaluate the final performance based on the accuracy (mIOU - mean intersection over Union). mIoU we can get by comparing the ground truth values with the output map after passing our image into the derived model.
- The second measure is usually the time it takes to process the image in the CNN network. In terms of FPS, we can also refer to it as Latency. The third one is the Network Parameters that we have used to learn by the derived network.
- The storage space we will need to save all the network parameters. This is also known as network size.
- The computational cost for GPU that we are using. Sometimes, we use it as the execution time for the frequency of GPU. The more the value is, the more efficient our GPU will work.
- Utilization of the hardware resources that mainly deal with CPU, GPU, and memory.
- The amount of power that our system mainly consumed.
- Memory bandwidth that we are using. It refers to the ratio of a number of bytes to a several transferred bytes from memory (sometimes it is shared and sometimes not shared).
- Training metrics also matter. Which environment, IDE, or libraries for deep learning are we using.

Fig. 15  (b) Results for Mask-RCNN applied on VOC2012
6 Performance evaluation

We have done experiments on the algorithms and advanced deep learning algorithms and concluded that traditional algorithms are not working fine for real-time images. It does not have the meaning for each pixel. A model-like UNET, SegNet, Deeplabv3+ is the right choice to be selected now [51]. The reason is they can give you exact pixel information without losing the pixel values. Like in encoder-decoder structure, i.e., UNET, SegNet model, we use skip connection in decoder side to regain the lead that we have lost in performing Max-Pooling operation on the encoder side. Losing information refers to the pixel values while we are doing feature extraction. Deeplab V3 has three more updated versions with improvements. UNet also has improved versions, namely Residuals U-Net and Fully Dense UNet. UNet refers to helping detect low-level features. It has initially been proposed for medical image segmentation. Residual UNet was introduced to improve the performance of UNet architecture. Further, a series of residual blocks are stacked together that benefits in terms of degradation problems with the help of skip connections, as same as in UNet, which helps to propagate the low-level features. Moreover, while modifying the UNet architecture using dense blocks, Dense UNet was introduced. It helps to improve the artifact while allowing each layer to learn the features at various spatial scales. We show in Table 4 the comparative data of JPANet composed of three different lightweight backbone networks and other models on the camvid test set. JPANet can not only achieve 67.45% mIoU but also obtains 294FPS once we input 360 × 480 low-resolution images. The data in Table 4 another time proves the effectiveness of the JPANet model. Figure 5 shows the visual comparison effect of JPANet on the CamVid test set (Figs. 17 and 18).

Eventually, UNET is easily applied in every field, especially in Biomedical (for medical image datasets) and Industry 4.0 related problems, like detecting the defects for Hot-Rolled Steel Strips, Surface, or Road Defects [79]. Mask-RCNN also has advanced in recent years like Mask- Scoring RCNN. For solving real-time scene understanding, Mask RCNN would be a better choice. The performance also matters. In some dense networks like Yolo V5 or Fully Dense UNET, the network parameters are abundant. While selecting a network model, you must consider the lightweight architectures to be applied to real-time applications and fast in computation.

We can see from Table 5 that JPANet achieved the very best scores in 18 of the 19 classification categories. It’s because JPANet emphasizes the importance of shallow spatial information. The development of JPANet on small object samples is the most blatant. For instance, the JPANet accuracy on the traffic signal and traffic sign are 24.6% and 19.8%
Table 4  Comprehensive performance comparison on the CamVid test set

| Method       | Input size | Params(M) | FPS | mIoU(%) |
|--------------|------------|-----------|-----|---------|
| JPANet-S [124] | 360 × 480  | 2.30      | 434 | 63.80   |
| JPANet-M [124] | 360 × 480  | 3.05      | 256 | 68.29   |
| JPANet-G [124] | 360 × 480  | 3.49      | 294 | 67.45   |
| SegNet [3]    | 360 × 480  | 29.5      | 46  | 55.6    |
| FSSNet [111]  | 360 × 480  | 0.2       | 179 | 58.6    |
| EDANet [52]   | 360 × 480  | 0.68      | 163 | 66.4    |
| ERFNet [68]   | 360 × 480  | 2.06      | 164 | 63.7    |
| DFANet [45]   | 720 × 960  | 7.8       | 120 | 64.7    |
| RPNet [13]    | 360 × 480  | 1.89      | 149 | 64.8    |
| BiSeNet [108] | 720 × 960  | 5.8       | –   | 65.6    |
| SwiftNet [64] | 720 × 960  | 12.9      | –   | 65.7    |
| ENet [65]     | 360 × 480  | 0.36      | 227 | 51.3    |
| DABNet [44]   | 360 × 480  | 0.81      | 117 | 66.4    |
| ICNet [118]   | 720 × 960  | 26.5      | 27.8| 67.1    |
| ESPNet [89]   | 512×1024   | –         | –   | 60.3    |
| VGG-16-Unet [124] | 512×512   | 2.1       | –   | 68.9    |
| EFSNet [35]   | 360 × 480  | 0.17      | 332 | 61.1    |

above ESPNet, respectively. Besides, JPANet also pays attention to extracting multi-scale semantic information. Thus JPANet also improves the segmentation results of huge targets to a particular extent. For instance, the accuracy of JPANet on the sidewalk and car is 1.7%, and 1.2% above the state-of-the-art ERFNet, respectively (Table 5).

![Fig. 17](image-url)  Comparison for All state-of-the art network architectures based on mIoU (in %) using Camvid dataset
7 Conclusion and future challenges

In this paper, a comprehensive overview of semantic segmentation algorithms and their grouping has been discussed. Some classical and some machine learning algorithms are compared and examined. Furthermore, by adding some parameters, we have calculated the efficiency of the model. Semantic segmentation is becoming a hot topic in every field, whether we study in industrial projects or the medical fields. Most of the time helps us in examining the critical details about the particular application after the valuable data preprocessing. The Future aspects of this research area can be studied as below:

- An annotation problem is also a big challenge for the dataset, which has very few samples for training, as in segmenting the surface defect datasets, some medical diagnostic or in soft robots where the field is quite new and publicly available datasets are not as much.
- Few-shot segmentation can be used to solve the less annotated dataset problems. The same problem can also be tackled by using the data augmentation technique.
- Small objects detection and segmentation is also an essential aspect for semantic segmentation that most researchers want to solve these days. Small is not tiny, but it is not as straightforward as nearer objects. Typically aerial images and long distant scenes are the main subjects to examine the classification or segmentation of the small object.
- Weather conditions could be a significant point of discussion in urban scene datasets for semantic segmentation.
- Moreover, light effect and control are crucial aspects that need to be addressed.
- Computational loaded segmentation approaches are also becoming quite negligible because of the robust and less-computational network architectures introduced daily like Trans4Trans or SegFormer. They have already taken the place of traditional encoder-decoder-based network architectures.
Table 5 Experiment on the Cityscapes dataset for each class one by one

| Method      | Road | Sidewalk | Building | wall | Fence | Pole | Traffic light | Traffic sign | Vegetation | Terrain | Sky | Person | Rider | Car | Truck | Bus | Train | Motorcycle | Bicycle | Class IoU | Category IoU |
|-------------|------|----------|----------|------|-------|------|----------------|--------------|-------------|----------|-----|--------|-------|------|-------|-----|-------|-------------|---------|-----------|--------------|
| JPANet [124]| 98.1 | 82.7     | 90.6     | 47.1 | 48.2  | 53.1 | 60.2          | 66.1         | 91.6        | 69.2      | 94.5 | 578.6  | 60.5  | 94.0 | 64.2  | 75.9| 66.5  | 63.7         | 65.9    | 71.6     | 86.7         |
| EFSNet [35] | 96.6 | 74.9     | 86.4     | 37.5 | 39.6  | 48.0 | 49.8          | 55.1         | 89.7        | 64.9      | 92.8 | 70.3  | 51.5  | 90.2 | 43.0  | 49.7| 41.6  | 41.5         | 53.5    | 61.9     | 82.8         |
| SQ [82]     | 96.9 | 75.4     | 87.8     | 31.6 | 35.7  | 50.9 | 52.0          | 61.7         | 90.9        | 65.8      | 93.0 | 73.8  | 42.6  | 91.5 | 18.8  | 41.2| 33.3  | 34.0         | 59.9    | 59.8     | 84.0         |
| CGNet [93]  | 95.5 | 78.7     | 88.1     | 40.0 | 43.0  | 54.1 | 59.8          | 63.9         | 89.6        | 67.6      | 92.9 | 74.9  | 54.9  | 90.2 | 44.1  | 59.5| 25.2  | 47.3         | 60.2    | 64.8     | 85.7         |
| ERFNet [68] | 97.7 | 81.0     | 89.8     | 42.5 | 48.0  | 56.3 | 59.8          | 65.3         | 91.4        | 68.2      | 94.2 | 76.8  | 57.1  | 92.8 | 50.8  | 60.1| 51.8  | 47.3         | 61.7    | 68.0     | 86.5         |
| ESPNet [89] | 97.0 | 77.5     | 76.2     | 35.0 | 36.1  | 45.0 | 35.6          | 46.3         | 90.8        | 63.2      | 92.6 | 66.7  | 40.9  | 92.3 | 38.1  | 52.5| 50.1  | 41.8         | 57.2    | 60.3     | 82.2         |
| ENet [65]   | 96.3 | 74.2     | 75.0     | 32.2 | 33.2  | 43.4 | 34.1          | 44.0         | 88.6        | 61.4      | 90.6 | 65.5  | 38.4  | 90.6 | 36.9  | 50.5| 48.1  | 38.8         | 55.4    | 58.3     | 80.4         |
| SegFormer-B5 [99] | 90.5 | 81.9     | 46.4     | 37.5 | 49.6  | 38.0 | 29.8          | 45.1         | 80.7        | 64.9      | 90.8 | 74.3  | 51.5  | 90.2 | 43.0  | 49.7| 41.6  | 41.5         | 83.5    | 82.4     | 84.0         |
So, to cover the whole understanding in every specific field and to understand the fundamental challenges, we must have a clear understanding of how we extract features, whether it is for detecting big objects or for detecting the smaller object in images due to the variation in distance, or lightening conditions.

**Declarations**

**Conflict of Interests** We have no conflict of interest.

**References**

1. Aakur SN, Sarkar S (2019) A perceptual prediction framework for self supervised event segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:1197–1206. https://doi.org/10.1109/CVPR.2019.00129
2. Agustsson E, Uijlings JR, Ferrari V (2019) Interactive full image segmentation by considering all regions jointly. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June(1):11614–11623. https://doi.org/10.1109/CVPR.2019.01189
3. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495
4. Batra A, Singh S, Pang G, Basu S, Jawahar CV, Paluri M (2019) Improved road connectivity by joint learning of orientation and segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:10377–10385. https://doi.org/10.1109/CVPR.2019.01063
5. Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C, Gall J (2019) SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In: Proceedings of the IEEE international conference on computer vision, 2019-October(iii), pp 9296–9306. https://doi.org/10.1109/ICCV.2019.00939
6. Benenson R, Popov S, Ferrari V (2019) Large-scale interactive object segmentation with human annotators. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:11692–11701. https://doi.org/10.1109/CVPR.2019.01197
7. Blokhinov YB, Gorbachev VA, Rakutin YO, Nikitin AD (2018) A real-time semantic segmentation algorithm for aerial imagery. Comput Opt 42(1):141–148. https://doi.org/10.18287/2412-6179-2018-42-1-141-148
8. Cao J, Pang Y, Li X (2019) Triply supervised decoder networks for joint detection and segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June(c):7384–7393. https://doi.org/10.1109/CVPR.2019.00757
9. Cerrone L, Zeilmann A, Hamprecht FA (2019) End-to-end learned random walker for seeded image segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:12551–12560. https://doi.org/10.1109/CVPR.2019.01284
10. Chang CY, Huang DA, Sui Y, Fei-Fei L, Niebles JC (2019) D3TW: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June(c):7384–7393. https://doi.org/10.1109/CVPR.2019.00757
11. Chang WL, Wang HP, Peng WH, Chiu WC (2019) All about structure: Adapting structural information across domains for boosting semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:1900–1909. https://doi.org/10.1109/CVPR.2019.00200
12. Chen PY, Liu AH, Liu YC, Wang YCF (2019) Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:2619–2627. https://doi.org/10.1109/CVPR.2019.00273
13. Chen X, Lou X, Bai L, Han J (2019) Residual pyramid learning for single-shot semantic segmentation. IEEE Trans Intell Transp Syst 21(7):2990–3000
14. Chen X, Williams BM, Vallabhaneni SR, Czanner G, Williams R, Zheng Y (2019) Learning active contour models for medical image segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:11624–11632. https://doi.org/10.1109/CVPR.2019.01190
15. Chen X, Yuan Y, Zeng G, Wang J (2021) Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. pp 2613–2622. https://doi.org/10.1109/cvpr46437.2021.00264
16. Chen K et al (2019) Hybrid task cascade for instance segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:4969–4978. https://doi.org/10.1109/CVPR.2019.00511
17. Cheng D, Liao R, Fidler S, Urtasun R (2019) Darnet: Deep active ray network for building segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:7423–7431. https://doi.org/10.1109/CVPR.2019.00761
18. Cholakkal H, Sun G, Shahbaz Khan F, Shao L (2019) Object counting and instance segmentation with image-level supervision. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:12389–12397. https://doi.org/10.1109/CVPR.2019.01268
19. Ding H, Jiang X, Shuai B, Liu A, Wang G (2019) Semantic correlation promoted shape-variant context for segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:8877–8886. 10.1109/CVPR.2019.00909
20. Du Y, Fu Z, Liu Q, Wang Y (2021) Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast. arXiv preprint arXiv:2110.07110
21. Fan R, Cheng MM, Hou Q, Mu TJ, Wang J, Hu SM (2020) S4net: Single stage salient-instance segmentation. Comput Vis Media 6(2):191–204. https://doi.org/10.1007/s41095-020-0173-9
22. Farha YA, Gall J (2019) MS-TCN: Multi-stage Temporal convolutional network for action segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:3570–3579. https://doi.org/10.1109/CVPR.2019.00369
23. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3141–3149. https://doi.org/10.1109/CVPR.2019.00326
24. Fu G et al (2019) A deep-learning-based approach for fast and robust steel surface defects classification. Opt Lasers Eng 121(May):397–405. https://doi.org/10.1016/j.optlaseng.2019.05.005
25. Ge Y, Zhang R, Wang X, Tang X, Luo P (2019) Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:5332–5340. https://doi.org/10.1109/CVPR.2019.00548
26. Griffin BA, Corso JJ (2019) BubbleNets: Learning to select the guidance frame in video object segmentation by deep sorting frames. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:8906–8915. https://doi.org/10.1109/CVPR.2019.00912
27. Gupta A, Dollar P, Girshick R (2019) Lvis: A dataset for large vocabulary instance segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:5351–5359. https://doi.org/10.1109/CVPR.2019.00550
28. He J, Deng Z, Zhou L, Wang Y, Qiao Y (2019) Adaptive pyramid context network for semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:7511–7520. https://doi.org/10.1109/CVPR.2019.00770
29. He K, Meng Q, Yan Y (2020) Song an end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Trans Instrum Meas 69(4):1493–1504. https://doi.org/10.1109/TIM.2019.2915404
30. He T, Shen C, Tian Z, Gong D, Sun C, Yan Y (2019) Knowledge adaptation for efficient semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:578–587. https://doi.org/10.1109/CVPR.2019.00067
31. He Y, Yu H, Liu X, Yang Z, Sun W, Wang Y, Fu Q, Zou Y, Mian A (2021) Deep learning based 3D segmentation: A survey. Proceedings Of 1(1):1–36. arXiv:2103.05423
32. Hou J, Dai A, Niblert M (2019) 3D-SIS: 3D semantic instance segmentation of RGB-D scans. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:4416–4425. https://doi.org/10.1109/CVPR.2019.00455
33. Hu Y, Lu, Han J, Schwing AG (2019) Sail-VOS: Semantic amodal instance level video object segmentation-A synthetic dataset and baselines. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:3100–3110. https://doi.org/10.1109/CVPR.2019.00322
34. Hu X, Jing L, Sehar U (2021) Joint pyramid attention network for real-time semantic segmentation of urban scenes. Appl Intell. https://doi.org/10.1007/s10489-021-02446-8
35. Hu X, Wang H (2020) Efficient fast semantic segmentation using continuous shuffle dilated convolutions. IEEE Access 8:70913–70924
36. Hu X, Yang K, Fei L, Wang K (2019) ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation. Proc - Int Conf Image Process ICIP 2019-September:1440–1444. https://doi.org/10.1109/ICIP.2019.8803025
37. Huang Z, Huang L, Gong Y, Huang C, Wang X (2019) Mask scoring r-CNN. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:6402–6411. https://doi.org/10.1109/CVPR.2019.00657
38. Hung W, Jampani V, Liu S, Molchanov P, Yang M, Kautz J SCOPS: Self-Supervised Co-Part Segmentation
39. Jain S, Wang X, Gonzalez JE (2019) Accel: a corrective fusion network for efficient semantic segmentation on video. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:8858–8867. https://doi.org/10.1109/CVPR.2019.00907

40. Jang WD, Kim CS (2019) Interactive image segmentation via backpropagating refinement scheme. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:5292–5301. https://doi.org/10.1109/CVPR.2019.00544

41. Jiao J, Wei Y, Jie Z, Shi H, Lau R, Huang TS (2019) Geometry-aware distillation for indoor semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:2864–2873. https://doi.org/10.1109/CVPR.2019.00298

42. Johander J, Danelljan M, Brissman E, Khan FS, Felsberg M (2019) A generative appearance model for end-to-end video object segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:8945–8954. https://doi.org/10.1109/CVPR.2019.00916

43. Lee J, Kim E, Lee S, Lee J, Yoon S (2019) Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:5262–5271. https://doi.org/10.1109/CVPR.2019.00541

44. Li G, Jiang S, Yun I, Kim J, Kim J (2020) Depth-wise Asymmetric bottleneck with Point-Wise aggregation decoder for Real-Time semantic segmentation in urban scenes. IEEE Access 8:27495–27506

45. Li H, Xiong P, Fan H, Sun J (2019) DFANEt: Deep feature aggregation for real-time semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:9514–9523. https://doi.org/10.1109/CVPR.2019.00975

46. Li Y, Yuan L, Vasconcelos N (2019) Bidirectional learning for domain adaptation of semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:6929–6938. https://doi.org/10.1109/CVPR.2019.00710

47. Li Y et al (2019) Attention-guided unified network for panoptic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:7019–7028. https://doi.org/10.1109/CVPR.2019.00719

48. Lin D et al (2019) Zigzagnet: Fusing top-down and bottom-up context for object segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:7482–7491. 10.1109/CVPR.2019.00767

49. Liu Y, Chen K, Liu C, Qin Z, Luo Z, Wang J (2019) Structured knowledge distillation for semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:2599–2608. https://doi.org/10.1109/CVPR.2019.00271

50. Liu H et al (2019) An end-to-end network for panoptic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:6165–6174. https://doi.org/10.1109/CVPR.2019.00633

51. Liu C et al (2019) Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:82–92. https://doi.org/10.1109/CVPR.2019.00017

52. Lo SY, Hang HM, Chan SW, Lin JJ (2019) Efficient dense mod- ules of asymmetric convolution for real-time semantic segmen- tation. In: Proceedings of the ACM Multimedia Asia, pp 1–6

53. Lu X, Wang W, Danelljan M, Zhou T, Shen J, Van Gool L (2020) Video object segmentation with episodic graph memory networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12348 LNCS:661–679. https://doi.org/10.1007/978-3-030-58580-8_39

54. Lu X, Wang W, Ma C, Shen J, Shao L, Porikli F (2019) (N.d.).lu see more kno more unsupervised video object segmentation with co-attention CVPR 2019 paper. Cvpr 1(d):3623–3632

55. Lu X, Wang W, Shen J, Crandall D, Luo J (2020) Zero-Shot Video object segmentation with Co-Attention siamese networks. IEEE Trans Pattern Anal Mach Intell 8828(c). https://doi.org/10.1109/ TPAMI.2020.3040258

56. Majumder S, Yao A (2019) Content-aware multi-level guidance for interactive instance segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:11594–11603. https://doi.org/10.1109/CVPR.2019.01187

57. Mans Larsson M, Stenborg E, Hammarstrand L, Pollefeys M, Attard T, Kahl F (2019) A cross-season correspondence dataset for robust semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:9524–9534. https://doi.org/10.1109/CVPR.2019.00976

58. Marin D, Tang M, Ben Ayed I, Boykov Y (2018) ADM for grid CRF loss in CNN segmentation, no. 1. Available: arXiv:1809.02322

59. Mou L, Hua Y, Zhu XX (2019) A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:12408–12417. https://doi.org/10.1109/CVPR.2019.01270
60. Nekrasov V, Chen H, Shen C, Reid I (2019) Fast neural architecture search of compact semantic segmentation models via auxiliary cells. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:9118–9127. https://doi.org/10.1109/CVPR.2019.00934
61. Neuhold G, Ollmann T, Bulo SR, Kontschieder P (2017) The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision, 2017-October. pp 5000–5009
62. Ni T, Xie L, Zheng H, Fishman EK, Yuille AL (2019) Elastic boundary projection for 3D medical image segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:2104–2113. https://doi.org/10.1109/CVPR.2019.00221
63. Orsic M, Kreso I, Bevandic P, Segvic S (2019) In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:12599–12608. https://doi.org/10.1109/CVPR.2019.01289
64. Orsic M, Kreso I, Bevandic P, Segvic S (In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 12607–12616
65. Paszke A, Chaurasia A, Kim S, Culurciello E (2016) Enet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147
66. Porzi L, Bulo SR, Colovic A, Kontschieder P (2019) Seamless scene segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:8269–8278. https://doi.org/10.1109/CVPR.2019.00847
67. Ranjan A et al (2019) Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:12232–12241. https://doi.org/10.1109/CVPR.2019.01252
68. Romero E, Alvarez JM, Bergasa LM, Arroyo R (2017) Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Trans Intell Transp Syst 19(1):263–272
69. Sadeghi D et al (2021) An Overview on Artificial Intelligence Techniques for Diagnosis of Schizophrenia Based on Magnetic Resonance Imaging Modalities: Methods, Challenges, and Future Works, pp 1–74, [Online]. Available: http://arxiv.org/abs/2013.03081
70. Shen Y, Ji R, Wang Y, Wu Y, Cao L (2019) Cyclic guidance for weakly supervised joint detection and segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:697–707. https://doi.org/10.1109/CVPR.2019.00079
71. Shetty R, Schiele B, Fritz M (2019) Not using the car to see the sidewalk: Quantifying and controlling the effects of context in classification and segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:8210–8218. https://doi.org/10.1109/CVPR.2019.00841
72. Shoeibi A et al (2021) Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: a review. Comput Biol Med, vol 136. https://doi.org/10.1016/j.compbiomed.2021.104697
73. Shoeibi et al (2020) Automated Detection and Forecasting of COVID-19 using Deep Learning Techniques: A Review. [Online]. Available: http://arxiv.org/abs/2007.10785
74. Shoeibi A et al (2021) Applications of Epileptic Seizures Detection in Neuroimaging Modalities Using Deep Learning Techniques: Methods, Challenges, and Future Works. [Online]. Available: http://arxiv.org/abs/2105.14278
75. Shoeibi et al (2021) Automatic Diagnosis of Schizophrenia using EEG Signals and CNN-LSTM Models, pp 1–11. [Online]. Available: http://arxiv.org/abs/2109.01120
76. Song C, Huang Y, Ouyang W, Wang L (2019) Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. ProcIEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:3131–3140. https://doi.org/10.1109/CVPR.2019.00325
77. Sun L, Yang K, Hu X, Hu W, Wang K (2020) Real-time Fusion Network for RGB-D Semantic Segmentation Incorporating Unexpected Obstacle Detection for Road-driving Images. [Online]. Available: http://arxiv.org/abs/2002.10570
78. Sun R, Zhu X, Wu C, Huang C, Shi J, Ma L (2019) Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:4355–4364. https://doi.org/10.1109/CVPR.2019.00449
79. Tabernik D, Šela S, Skvara J, Skočaj D (2020) Segmentation-based deep-learning approach for surface-defect detection. J Intell Manuf 31(3):759–776. https://doi.org/10.1007/s10845-019-01476-x
80. Tian Z, He T, Shen C, Yan Y (2019) Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:3121–3130. https://doi.org/10.1109/CVPR.2019.00324
81. Tokunaga H, Teramoto Y, Yoshizawa A, Bise R (2019) Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:12589–12598. https://doi.org/10.1109/CVPR.2019.01288
82. Treml M, Arjona-Medina J, Unterthiner T, Durgesh R, Friedmann F, Schuberth P, Mayr A, Heusel M, Hofmarcher M, Widrich M, Nessler B, Hochreiter S (2016) Speeding up semantic segmentation for autonomous driving. In: MLITS NIPSWorkshop 2(7)
83. Ventura C, Bellver M, Girbau A, Salvador A, Marques F, Giro-I-Nieto X (2019) RVOS: End-to-end Recurrent network for video object segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:5272–5281. https://doi.org/10.1109/CVPR.2019.00542
84. Voigtlaender P, Chai Y, Schroff F, Adam H, Leibe B, Chen LC (2019) Feelvos: Fast end-to-end embedding learning for video object segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:9473–9482. https://doi.org/10.1109/CVPR.2019.00971.2696
85. Voigtlaender P et al (2019) Mots: Multi-object tracking and segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:7934–7943. https://doi.org/10.1109/CVPR.2019.00813
86. Vu TH, Jain H, Bucher M, Cord M, Perez P (2019) Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June(Mmd):2512–2521. https://doi.org/10.1109/CVPR.2019.00262
87. Wang F, Gu Y, Liu W, Yu Y, He S, Pan J (2019) Context-aware spatio-recurrent curvilinear structure segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:12640–12649. https://doi.org/10.1109/CVPR.2019.01293
88. Wang L, Huang Y, Hou Y, Zhang S, Shan J (2019) Graph attention convolution for point cloud semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:10288–10297. https://doi.org/10.1109/CVPR.2019.01054
89. Wang L, Li D, Zhu Y, Tian L, Shan Y (2020) Dual Super-Resolution Learning for Semantic Segmentation. pp 3773–3782. https://doi.org/10.1109/cvpr42600.2020.00383
90. Wang X, Liu S, Shen X, Shen C, Jia J (2019) Associatively segmenting instances and semantics in point clouds. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:4091–4100. https://doi.org/10.1109/CVPR.2019.00422
91. Wang M et al (2019) Example-guided style-consistent image synthesis from semantic labeling. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:1495–1504. https://doi.org/10.1109/CVPR.2019.00159
92. Wen Z, Zhao Q, Tong L (2020) CNN-based minor fabric defects detection. Int J Cloth Sci Technol. https://doi.org/10.1108/ICST-11-2019-0177
93. Wu T, Tang S, Zhang R, Gao J, Zhang Y (2020) Cgenet: A light-weight context guided network for semantic segmentation. IEEE Trans Image Process 30:1169–1179
94. Wu X, Wu Z, Guo H, Ju L, Wang S (2021) DANNet: A One-Stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation. pp 15769–15778. https://doi.org/10.1109/cvpr46437.2021.01551
95. Wu S, Wu T, Lin F, Tian S, Guo G (2021) Fully Transformer Networks for Semantic Image Segmentation 1(2):1–17. arXiv:2106.04108
96. Wu S, Wu T, Lin F, Tian S, Guo G (2021) Fully Transformer Networks for Semantic Image Segmentation 1(2):1–17. arXiv:2106.04108
97. Xian Y, Choudhury S, He Y, Schiele B, Akata Z (2019) Semantic projection network for zero-and few-label semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:8248–8257. https://doi.org/10.1109/CVPR.2019.00845
98. Xiang K, Yang K, Wang K (2021) Polarization-driven semantic segmentation via efficient attention-bridged fusion. Opt Express 29(4):4802. https://doi.org/10.1364/oe.416130
99. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P (2021) SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, pp 1–18. [Online]. Available: http://arxiv.org/abs/2105.15203
100. Xiong Y et al (2019) Upsnet: A unified panoptic segmentation network. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:8810–8818. https://doi.org/10.1109/CVPR.2019.00902
101. Xu S, Liu D, Bao L, Liu W, Zhou P (2019) MHP-VOS: Multiple Hypotheses propogation for video object segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:314–323. https://doi.org/10.1109/CVPR.2019.00040
102. Xu K, Wen L, Li G, Bo L, Huang Q (2019) Spatiotemporal CNN for video object segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:1379–1388. https://doi.org/10.1109/CVPR.2019.00147
103. Yang K, Hu X, Fang Y, Wang K, Stiefelhagen R (2020) Omnisupervised omnidirectional semantic segmentation. IEEE Trans Intell Transp Syst, no. September, pp 1–16. https://doi.org/10.1109/tits.2020.3023331

104. Yang K, Hu X, Stiefelhagen R (2021) Is Context-Aware CNN ready for the surroundings? panoramic semantic segmentation in the wild. IEEE Trans Image Process 30(December):1866–1881. https://doi.org/10.1109/TIP.2020.3048682

105. Ye L, Rochan M, Liu Z, Wang Y (2019) Cross-modal self-attention network for referring image segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:10494–10503. https://doi.org/10.1109/CVPR.2019.01075

106. Yi L, Zhao W, Wang H, Sung M, Guibas LJ (2019) GSPN: Generative Shape proposal network for 3D instance segmentation in point cloud. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:3942–3951. https://doi.org/10.1109/CVPR.2019.00407

107. Yu F, Liu K, Zhang Y, Zhu C, Xu K (2019) PartNet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:9483–9492. https://doi.org/10.1109/CVPR.2019.00972

108. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the european conference on computer vision (ECCV), pp 325–341

109. Yuan Y, Huang L, Guo J, Zhang C, Chen X, Wang J (2018) OCNet: Object Context Network for Scene Parsing. arXiv:1809.00916

110. Zendel O, Honauer K, Murschitz M, Steininger D, Domínguez GF (2018) Wilddash-creating hazard-aware benchmarks. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11210 LNCS:407–421. https://doi.org/10.1007/978-3-030-01231-1-25

111. Zhang X, Chen Z, Wu QMJ, Cai L, Lu D, Li X (2018) Fast semantic segmentation for scene perception. IEEE Trans Indus Inform 15(2):1183–1192

112. Zhou Y et al (2019) Collaborative learning of semi-supervised segmentation and classification for medical images. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:2074–2083. https://doi.org/10.1109/CVPR.2019.00218

113. Zhu Y et al (2020) Improving semantic segmentation via self-training. arXiv
124. Zhuang B, Shen C, Tan M, Liu L, Reid I (2019) Structured binary neural networks for accurate image classification and semantic segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019-June:413–422. https://doi.org/10.1109/CVPR.2019.00050

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.