Computers, Materials & Continua
DOI: 10.32604/cmc.2023.032757
Article

Tech Science Press

A Survey on Image Semantic Segmentation Using Deep Learning Techniques

Jieren Cheng[1,3], Hua Li[2,*], Dengbo Li[3], Shuai Hua[2] and Victor S. Sheng[4]

[1]School of Computer Science and Technology, Hainan University, Haikou, 570228, China
[2]School of Cyberspace Security (School of Cryptology), Hainan University, Haikou, 570228, China
[3]Hainan Blockchain Technology Engineering Research Center, Hainan University, Haikou, 570228, China
[4]Department of Computer Science Texas Tech University TX, 79409, USA
*Corresponding Author: Hua Li. Email: lihua32022@163.com
Received: 28 May 2022; Accepted: 12 July 2022

Abstract: Image semantic segmentation is an important branch of computer vision of a wide variety of practical applications such as medical image analysis, autonomous driving, virtual or augmented reality, etc. In recent years, due to the remarkable performance of transformer and multilayer perceptron (MLP) in computer vision, which is equivalent to convolutional neural network (CNN), there has been a substantial amount of image semantic segmentation works aimed at developing different types of deep learning architecture. This survey aims to provide a comprehensive overview of deep learning methods in the field of general image semantic segmentation. Firstly, the commonly used image segmentation datasets are listed. Next, extensive pioneering works are deeply studied from multiple perspectives (e.g., network structures, feature fusion methods, attention mechanisms), and are divided into four categories according to different network architectures: CNN-based architectures, transformer-based architectures, MLP-based architectures, and others. Furthermore, this paper presents some common evaluation metrics and compares the respective advantages and limitations of popular techniques both in terms of architectural design and their experimental value on the most widely used datasets. Finally, possible future research directions and challenges are discussed for the reference of other researchers.

Keywords: Deep learning; semantic segmentation; CNN; MLP; transformer

1 Introduction

Image semantic segmentation is a basic task in the field of computer vision. It can be regarded as a pixel level classification task, which achieves fine-grained reasoning by intensively predicting and inferring labels for each pixel, so that each pixel is labeled and divided into a specific category. Image semantic segmentation not only provides category prediction, but also provides spatial location information about these classes. In recent years, semantic segmentation has been applied more and more widely. It plays an important role in medical image analysis [1], automatic driving [2], virtual/augmented reality [3], video surveillance [4] and three-dimensional reconstruction [5].

Reviewing the development of semantic segmentation methods [6], Early methods were mostly based on mathematical methods, such as thresholding, k-means clustering, and conditional random fields. Then, with the great success of deep learning in various fields [7], researchers tried to use deep learning techniques for semantic segmentation task, and successfully designed the full convolution neural network (FCN) [8]. Since then, convolution neural network has swept the field and become the mainstream method. In the past two years, transformer has become popular in computer vision, and the application of MLP technology in this field has inspired researchers to explore more possibilities in the field of semantic segmentation.

With the rapid emergence of new semantic segmentation methods based on deep learning in recent years, many past reviews have some shortcomings. Although they have [9,10] introduced common datasets in the field of semantic segmentation and technical details of some classical methods, they lacked generalizations and descriptions of some new technologies (e.g., transformer- and MLP-based methods). It is well known that there is no extensive survey covering many types of semantic segmentation methods such as CNN-based, transformer-based and MLP-based.

The goal of this paper is to summarize and classify current deep learning methods in semantic segmentation to provide comprehensive information reference for scholars and practitioners. Inspired by the work of Zhao et al. [11], this paper compares and analyzes the image segmentation work of three main neural network architectures in deep learning technology, and proposes a new classification method, which is shown in Fig. 1: it is based on network architecture. Existing semantic segmentation methods are divided into four categories according to different network architectures: CNN-based architectures, transformer-based architectures, MLP-based architectures, and others.

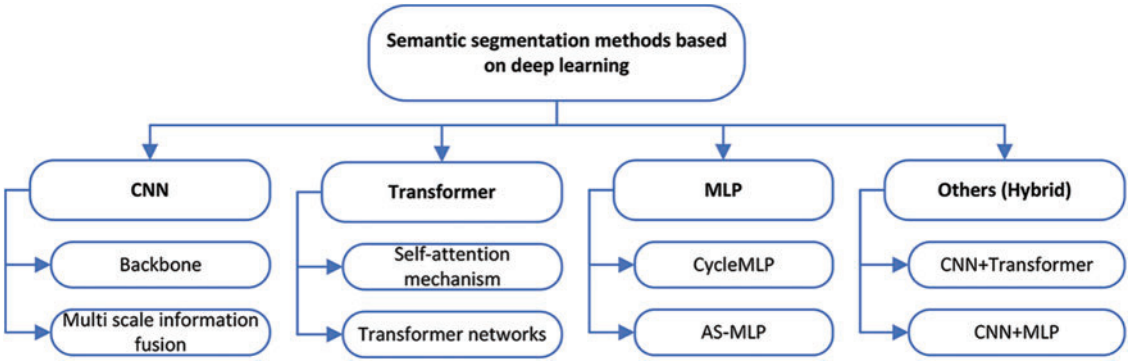

Figure 1: Taxonomy of methods on semantic segmentation

The key contributions of this paper include a systematic review of image semantic segmentation methods which covers the latest literature in the field of image semantic segmentation. Various deep learning algorithms used in image segmentation are described and divided into four categories according to different network architectures. The advantages and limitations of existing segmentation methods are compared and analyzed on popular benchmarks. The results of this study provide trends in semantic segmentation using deep learning, and challenges, and future research directions.

The remainder of this survey is organized as follows: Section 2 reviews some of the most popular image segmentation datasets and their characteristics. Section 3 is the main body of our survey. Section 4 summarizes some common metrics used in the performance evaluation of segmentation models, and then evaluates and analyzes the performance of the models. Section 5 discusses the main future

research directions and challenges in the field of image segmentation. Finally, Section 6 makes a summary.

2 Datasets

There are many datasets that can be used for semantic segmentation tasks. This paper introduces a total of ten representative general image segmentation datasets, including PASCAL visual object classes (VOC) [12], Cityscapes [13], Microsoft common objects in context (COCO) [14], ADE20K [15], CamBridge-driving labeled video database (CamVid) [16], COCO-stuff [17], Indian driving dataset (IDD) [18], Dark Zurich [19], adverse conditions dataset with correspondences (ACDC) [20], and PartImageNet [21]. According to different purposes of these datasets, they can be divided into generic, urban/Driving, generic-part, etc. Although there are related works [9,10] that have described datasets in detail, they suffer from the problem of partial content invalidation and lack of recent datasets. Therefore, several image semantic segmentation datasets are briefly summarized, and detailed information (such as their purpose, number of classes, training/validation/testing splits, and access hyperlink.) about the characteristics of each dataset are provided. Tab. 1 shows a summarized view of the above datasets, where the first five are the more popular datasets, and the last five are the most recent datasets. In addition, some segmentation models only select the classes of interest when training the model on the dataset, instead of using all classes. Therefore, the number in brackets in the class column is used to indicate the number of frequently used classes. The above summary is intended to facilitate readers to have a basic understanding of commonly used semantic segmentation data sets when reading this article. Readers can refer to the corresponding link address to query the detailed description of the relevant data set according to their own needs.

Table 1: Popular semantic segmentation datasets

Datasets and challenges	Purpose	Year	Classes	Resolution	Samples (train/val/test)
Pascal VOC [12]	Generic	2012	21	Variable	1,464/1,449/Private
Cityscapes [13]	Urban	2015	30(8)	2048 × 1024	2,975/500/1,525
Microsoft COCO [14]	Generic	2014	80+	Variable	82,783/40,504/81,434
ADE20K [15]	Generic	2017	150	Variable	20,210/2,000/3,000
Camvid [16]	Urban/Driving	2009	32	960 × 720	701/None/None
COCO-stuff [17]	Generic	2018	172	Variable	118,000/5,000/20,000
IDD [18]	Urban/Driving	2019	34	Variable	6,993/981/2,029
Dark Zurich [19]	Urban/Driving	2019	19	1920 × 1080	8,377/50/151
ACDC [20]	Driving	2021	30(19)	2048 × 1024	1,600/406/2,000
PartImageNet [21]	Generic-Part	2021	158	Variable	16,540/2,957/4,598

3 Deep Learning-based Segmentation Methods

In recent years, with the rapid development of neural networks, image semantic segmentation methods based on deep learning have entered a new stage of development. Model architectures for semantic segmentation are becoming more and more diverse, and the basic modules used to build different architectures are shown in Fig. 2, where PE denotes positional encoding. This section

divides segmentation methods into four categories according to different architectures: CNN-based, transformer-based, MLP-based, and others, also introduces the typical segmentation methods based on these architectures in detail.

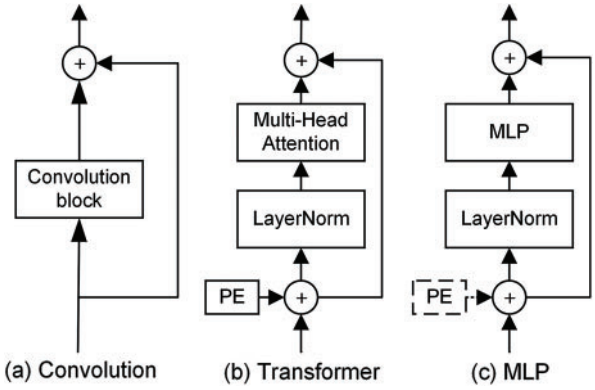

Figure 2: Three basic modules used to build different architectures

3.1 CNN-based Architectures

3.1.1 Backbone

After a long-term evolution and development, backbone networks in the field of image semantic segmentation has given birth to different types, such as large-scale classic backbone and its variants, lightweight backbone, encoder-decoder backbone and multi-scale backbone. Next, the backbone networks will be introduced based on the above four types.

Classical backbone and its variants: Some deep networks (e.g., VGGNet [22], ResNet [23]) have made great contributions to the field of Semantic segmentation and have laid a solid foundation for subsequent development. The subsequent backbones [24,25] are mostly combined with the design of the previous backbones, making full use of the idea of residual connection and grouping to improve the network, and extracting more spatial information without adding a lot of parameters to improve the performance of the network.

Lightweight backbone: The lightweight backbone is suitable for terminal devices that lack computing resources. They have few parameters and fast inference speed, which are suitable for real-time semantic segmentation. The two more famous series of backbones are MobileNet [26] and ShuffleNet [27]. The MobileNet proposed the deep separable convolution instead of the standard convolution. The ShuffleNet [27] designed a ShuffleNet unit, which used pointwise group convolution to replace the original pointwise convolution and added channel shuffle to strengthen the connection between groups. They all greatly reduce the computational overhead while maintaining accuracy.

Encoder-decoder backbone: Encoder-decoder network architecture [28,29] mainly includes encoder and decoder. The encoder generates a down sampling feature map, and the decoder up samples the feature map to match the input resolution. Usually, the input of each encoder layer is also bypassed to the decoder of the same feature map scale to help recover the missing spatial information.

Multi-branch backbone: Dual branch networks [30–32] a generally divide the network structure into spatial detail branches and depth feature branches, and then fuse the information of the two branches to reduce the loss of detail information. In addition,Sun et al. [33] proposed a high-resolution network (HRNet), which designed a multi-branch structure. They considered that the

down sampling operation of both branches will lose part of the spatial information and reduce the network performance. Therefore, the HRNet started from a high-resolution subnet and gradually down sampled to form a subnet from high to low resolution. Each subnet is connected in parallel and continuously integrates information, all branches are aggregated to directly affect the output.

3.1.2 Multi Scale Information Fusion

In the CNN-based architecture, the deep network has strong representation ability of semantic information, and the shallow network contains rich spatial detail information. Fully integrating the deep semantic information and shallow spatial information can effectively improve the accuracy of the network. In recent years, the work related to multi-scale information fusion has emerged one after another. Some works [34,35] constructed a long jump connection branch, which connects shallow features to deep features to reduce the loss of spatial information. Huang et al. [36] proposed the feature-aligned pyramid network (FaPN) for dense image prediction. The FaPN improved the feature pyramid network through feature alignment module and feature selection module, emphasized low-level features with rich spatial detail information, and solved the problem of prediction and classification errors caused by misalignment of context information in the process of feature fusion.

In addition to improving the performance of the model by fusing deep and shallow information, global semantic information can provide clues to segment category distribution and the robustness of the model can be improved by fusing the global and local features. Zhao et al. [37] developed the pyramid scene parsing network (PSPNet) to better learn the multi-scale context representation of a scene. The PSPNet proposed a pyramid pooling module (PPM), which extracted the global information of different sub-regions through four pooling modules with different down sampling scales, and then sampled on each branch to restore the resolution and concatenated the original feature map to fully integrate the local and global information. Zhang et al. [38] proposed an EncNet model, which designed a context encoding module to capture global semantic information and calculated the scaling factor of the feature graph based on the coding information to highlight the information categories that need to be emphasized. Finally, the EncNet used semantic coding loss (SE loss) to force the network to understand the global information.

To expand the receptive field and obtain rich semantic information, repeated maximum pooling and down sampling operations will be carried out, which will lead to the decline of feature map resolution and the loss of spatial information. Dilated convolutions can enlarge the receptive field without additional computational cost, so it has been very popular in the field of real-time segmentation, and many works are based on this technology to improve the performance of the model. Some of most important works include the DeepLap family proposed by Chen et al. [39,40] and the densely connected atrous spatial pyramid pooling (DenseASPP) proposed by Yang et al. [41]. They all used dilated convolution to replace the original down sampling method and expanded the receptive field to obtain more context information without increasing the number of parameters and calculation.

3.2 Transformer-based Architectures

Transformer is a deep neural network based on self-attention. It was first used in the field of natural language processing [42], then the model had been continuously improved [43] and had achieved excellent performance in a wide range of language tasks such as machine translation, text classification, and question answering. The great achievements of transformer in the field of natural language processing have greatly encouraged researchers to explore the role of transformer in the field of computer vision. In recent two years, transformer structure and its variants have been successfully

applied to visual tasks such as image classification [44], image captioning [45], object detection [46], and segmentation [47]. The Google team also analyzed the training of vision transformer and provided an effective guidance for future research on visual transformers [48]. This section mainly introduces self-attention mechanisms and transformer networks (a specific form of self-attention) in the field of image segmentation.

3.2.1 Self-attention Mechanisms

The self-attention mechanism has played a key role in delivering efficient and accurate computer vision systems [49]. The essence of visual attention mechanism is to imitate the way of human visual signal processing, quickly scan the global image, filter out important information, invest more attention resources and suppress other useless information, thereby greatly improving the efficiency and accuracy of visual information processing. Attention mechanisms can be divided into channel attention, spatial attention and mixed attention according to their function.

Channel attention, such as squeeze-and-excitation networks (SENet) [50]. The SENet generated feature weights for each feature channel to represent the importance of the channel, and then multiplied the feature weights with the original feature map to enhance the beneficial feature channels and suppress the useless feature channels.

Spatial attention emphasizes the correlation between pixels, such as Non-local Neural Networks [51] and criss-cross network (CCNet) [52]. They captured long-distance dependence through non-local operation, calculated the correlation between pixels through a variety of coding methods, and made full use of non-local information to enhance the representation ability of pixels. Wang et al. [51] proposed a differentiable non-local operation, which computed the response at a position as a weighted sum of the features at all positions in the feature map to capture long-range dependencies both in both space and time in a feed-forward fashion. To reduce computational overhead, Huang et al. [52] proposed the criss-cross attention module that for each pixel position generated a sparse attention map only on the criss-cross path.

Mixed attention enhances feature expression from two perspectives at the same time. For example, the dual attention network (DANet) [53] and the dual relation-aware attention network (DRANet) [54] not only emphasized the information expression between channels, but also emphasized the importance of local information in channels, and they fully integrated the advantages of spatial attention and channel attention. Similarly, Sagar et al. [55] proposed a dual multi-scale attention network (DMSANet), which first divided the input features into multiple groups to capture features at different scales, then fused the multi-scale features and performs channel and spatial dimension enhancement in parallel. Furthermore, considering that complex attention mechanisms are not suitable for lightweight models, there are many works from a lightweight perspective. For example, the Channelized Axial Attention (CAA) [56], the Axial-DeepLab [57] and Coordinate Attention [58] proposed the position-sensitive axis-attention, which reformulated the 2D self-attention mechanism into two 1D directional axial attention to capture full contextual information with high computational efficiency.

3.2.2 Transformer Networks

Vision transformer networks refer to how transformers can "altogether" replace the work of standard convolutions in deep neural networks on large-scale computer vision datasets. They first transform the original image into a series of image "patches", which are then fed into the original

transformer model for training. Next, we will introduce the transformer networks specifically designed to solve the image semantic segmentation task.

Zheng et al. [59] first performed semantic segmentation based on transformer and constructed a segmentation transformer network (SETR) to extract global semantic information. The input and output in transformer need to be serialized, so the SETR first divided the two-dimensional images into (H/16) $*$ (W/16) patches and converted them into a one-dimensional sequence with a length of H $*$ W/256, and then learned the specific coding of each patch through position coding to retain spatial information. Then, the sequence is input into the transformer encoder composed of multi-head self-attention (MSA) and multi-layer perceptron (MLP) modules to learn features. In addition, to effectively evaluate the effect of the encoder, the SETR designed three different decoders: naive up-sampling (naive), progressive up-sampling (pup) and multi-level feature aggregation (MLA). Inspired by SETR, Strudel et al. [60] designed a pure transformer model named Segmenter to apply to semantic segmentation tasks. Each layer of the encoder in the Segmenter models the global context information, then the mask transformer decodes the output of the encoder and class embedding, decodes the encoded sequence features into three-dimensional segmentation feature map, and then up-samples the feature map to obtain the final image segmentation map. Based on the advantage that transformer can build global dependencies in images, segmentation transformer [61] combined transformer with object contextual representations to represent the OCR pipeline to enhance feature expression.

Considering the key role of multi-scale features in CNN based segmentation model, researchers began to explore methods to combine the advantages of transformer and encode multi-scale information. For example, Xie et al. [62] proposed a SegFormer model, which designed a hierarchical transformer encoder and lightweight MLP decoder. Hierarchical transformer encoder will generate multi-level features, and the output feature size decreased layer by layer. Large-scale feature map provided coarse-grained information, and small-scale feature map provided fine-grained information. All-MLP decoder aggregated different levels of features and combined global attention and local attention to obtain more powerful information representation. The model achieved excellent segmentation performance. Liu et al. [63] proposed a mobile window scheme, which limited the self-attention calculation to non-overlapping local windows, constructed a hierarchical network architecture named Swin Transformer, merged patch blocks layer by layer to expand the perception range, and then encoded information of different scales to adapt to multi-scale tasks in vision. Chen et al. [64] proposed a vision transformer adapter (ViT-Adapter), which first used the spatial prior module to model the local spatial contexts of the input image, then injected the captured prior features into the patches encoded by the vision transformer backbone (VIT) [65], and finally extracted hierarchical features from the output of the block through the multi-scale feature extractor.

Besides using multi-scale information, effectively utilizing different distance attention mechanisms to enhance features can also improve the performance of the model. For example, the cross-scale transformer (CrossFormer) [66] proposed a long and short distance attention (LSDA), which not only paid attention to the dependency between adjacent embedding, but also emphasized the dependency between embedding far away from each other and retained the embedding characteristics of small-scale and large-scale while reducing the cost. Yang et al. [67] proposed a focal self-attention and constructed a transformer architecture, named as Focal Transformer. It constructed fine-grained attention in local scope and coarse-grained attention in global scope. At the same time, it effectively captured the visual dependence of short-range and long-range, reduced the amount of calculation and improved the performance of the model.

3.3 MLP-based Architectures

Multi-layer perceptron (MLP) is a neural network with forward structure. It has a simple structure and strong adaptive ability, and it is widely used in the fields of natural language processing [68] and computer vision [69]. Although CNN-based and transformer-based networks are the mainstream choices in the field of computer vision, researchers still try to build the network architecture completely using MLP to explore more possibilities for visual network architecture. Recent studies have shown that the MLP-based architecture abandoning convolution and self-attention is simple in design, and its performance in many visual tasks is comparable to the CNN-based and Transformer-based architectures. This section will introduce the specific application of MLP in semantic segmentation task.

Chen et al. [70] proposed an MLP-like architecture called CycleMLP for dense prediction. Compared with the previous MLP-based architectures, the CycleMLP can flexibly processed input images of various scales and is easy to migrate to downstream tasks. Compared with the transformer based on architecture, CycleMLP has considerable performance and fewer parameters. It designed Cycle Fully-Connected Layer and used Cycle FC to replace the Spatial MLP in the MLP-Mixer architecture for optimization. It also defined the concept of pseudo-kernel, which changed the original fixed sampling point into a pseudo-kernel window centered on the original sampling point and with a certain receptive field. Combined with the advantages of channel FC and spatial FC, it not only maintained the linear computational complexity of the image, but also expanded the receptive field and fully aggregates the context information.

To capture the correlation of local features, Lian et al. [71] proposed an axial shift strategy and designed an axial shifted MLP (AS-MLP), which was conducive to the interaction of spatial information. The AS-MLP designed AS-MLP block, which extracted features from a single spatial direction through horizontal shift and vertical shift respectively, and then mapped the features to a linear layer through channel projection, and fused features from both directions to effectively extract local information. As-MLP extracted local features through a simple axial shifted strategy, which effectively improved the model performance and reduced the computational complexity.

3.4 Others

In [11], a unified framework called SPACH has been developed to fairly compare the performance of CNN, transformer and MLP. The experimental results show that under the same pre-training conditions, all three architectures can perform the classification task well. CNN and transformer are complementary, CNN-based structure has the best generalization ability, and transformer-based structure has the largest model capacity.

Obviously, each type of architecture has its own advantages, so future researches do not need to stick to a single architecture. Researchers can integrate the advantages of multiple architectures to achieve more efficient performance according to actual task requirements. This section will mainly introduce the related research of hybrid architecture.

The convolution operation in CNN-based architectures has low computational cost, but it has the limitation that it is unable to model long-term dependencies. In contrast, transformer has global attention mechanism, but low-level details are insufficient. Therefore, part of the work combines CNN and transformer to make use of their advantages to achieve a more efficient architecture. The following methods all combine CNN and transformer. The nnFormer [72] built an interleaving architecture based on self-attention and convolution to achieve a better combination of transformer and CNN. Zhang et al. [73] proposed an architecture called TransFuse which integrated transformer and CNN. The TransFuse consists of parallel transformer branches and CNN branches, in which transformer

branches capture global information and build remote dependencies, while CNN branches capture rich local information. Guo et al. [74] proposed a method to segment objects with transformers (SOTR), which extracted shallow features through feature pyramid and captured remote context dependencies based on parallel two branch transformer.

In addition to the work of combining CNN and transformer, researchers also try to combine CNN and MLP to build an architecture. MLP-based architectures can encode feature information well, but most of them have fixed dimension input, which is difficult to adapt to downstream tasks (such as Object detection, semantic segmentation), and has large amount of calculation and limited performance; Convolutional neural network can greatly reduce the amount of network calculation and flexibly adapt to different inputs. Therefore, the integration of CNN and MLP can build a lighter, phased and high-performance architecture. Li et al. [75] proposed a hierarchical convolutional MLPs (ConvMLP) for vision, which mainly included convolution stage and Conv-MLP stage. Tokenizer includes convolution, normalization, activation function and maximum pooling operation to extract initial features. The convolution stage is responsible for enhancing the spatial connection. In the Conv-MLP stage, convolution is used to increase the interaction ability of adjacent information in the process of patch merging and down sampling, and a depth wise convolution layer is embedded between the two MLP blocks to further promote the blending of adjacent information and effectively improve the performance of the model.

4 Evaluation of Methods

In this section, it first introduces several common metrics for model performance evaluation and then analyzes the performance of the segmentation model based on the evaluation metrics.

4.1 Evaluation Metrics

The performance evaluation of segmentation model needs a unified evaluation metrics. Next, some common metrics (speed, accuracy, memory footprint) used to evaluate the performance of segmentation models will be outlined.

Speed is a very important metrics. The inference speed (IS) represents the amount of forward pass time in millisecond (ms) that a network takes process an image and usually measured by frames per second (FPS). A fast inference speed is conducive to the landing application of image segmentation methods. Therefore, it is very meaningful to understand the time required in the process of model reasoning.

Accuracy in this paper mainly refers to the accuracy of the model, The Mean Intersection over Union (MIoU) is usually used as the accuracy of semantic segmentation. Assuming k is the number of classes, MIoU can be defined as Eq. (1):

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

where P_{ij} refers to the number of pixels inferred from category i as category j, P_{ii} refers to the true positive. P_{ij} and P_{ji} refer to the false positive and false negative, respectively.

Memory footprint is another important factor in segmentation methods. Models with large parameters and complex calculations may not be applicable to some edge devices (such as unmanned aerial vehicles (UAVs), autopilot cars, and robots) with less memory than high-performance servers.

Therefore, memory constraints need to be considered in model design. It may be very useful to fully describe the peak and average memory occupation during model operation.

4.2 Performance Analysis

For devices with limited computing power and storage space in industrial applications, the accuracy of the model is no longer the only concern, fast and lightweight models are also very important. Therefore, a comprehensive evaluation of model performance will help readers fully understand the advantages of different segmentation methods. This section will evaluate the performance of the methods described in Section 3.

Tab. 2 summarizes the performance of the main segmentation models based on deep learning on different datasets (e.g., Pascal VOC 2012 test, Cityscapes test, ADE20K Val). In the table, prams refer to network parameters. IS is measured on a NVIDIA 1080Ti GPU cards. Methods marked with "$*$" represents the IS measured on other GPU cards. The last letter of the value in the column of IS represents the input of different resolutions, in which the corresponding resolutions of a, b, c and d are 2048×1024, 1536×768, 1024×512, 713×713. "-" in table, when referring to the backbone or other evaluation metrics, respectively means that the prior works have not used the backbone network or have not published the corresponding value.

Table 2: Performance comparison of segmentation methods on different datasets

Type	Method	Backbone	Performance				
			IS (FPS)	Params (M)	Accuracy on different dataset		
					Cityscapes test set	Pascal VOC test set	ADE20K Val set
CNN	FCN [8]	VGG-16	–	134.5	65.3	62.2	29.36
	ResNest [25]	ResNest-200	–	–	83.3	–	48.36
	HFEN* [28]	–	112a	1.1	69.5	–	–
	LAANet* [29]	–	95.8c	0.67	73.6	–	–
	BiSeNetV2 [30]	–	65.5b	49	74.7	–	–
	STDC2-Seg75 [31]	STDC2	97b	–	76.8	–	–
	DDRNet [32]	DDRNet-39	22.7a	32.3	80.4	–	–
	HRNet [33]	HRNet-w48	–	65.9	81.6	–	42.99
	CIFReNet [34]	MobileNetV2	34.5c	1.9	70.9	–	–
	MIFNet [35]	MobileNetV2	35.7c	2.0	72.2	–	–
	FaPN [36]	ResNet-18	78.1b	–	75.0	–	–
	PSPNet* [37]	ResNet-101	0.8d	250.8	78.4	85.4	43.29
	EncNet [38]	ResNet-101	–	–	76.9	82.9	44.65
	DeepLab-CRF [39]	ResNet-101	–	44.0	70.4	79.7	–
	DeepLabV3+ [40]	Xception-JFT	–	–	82.1	89.0	–
	DenseASPP [41]	DenseNet121	–	28.6	76.2	–	–
Trans-former	CCNet [52]	ResNet-101	–	–	81.3	–	45.2
	DANet [53]	ResNet-101	–	–	81.5	82.6	–

(Continued)

Table 2: Continued

Type	Method	Backbone	Performance				
			IS (FPS)	Params (M)	Accuracy on different dataset		
					Cityscapes test set	Pascal VOC test set	ADE20K Val set
	DRANet [54]	ResNet-101	–	–	82.9	–	–
	CAA [56]	ResNet-101	–	–	82.6	–	–
	SETR [59]	SETR-PUP	–	–	81.64	–	53.5
	Segmenter [60]	ViT-L	–	–	–	–	53.63
	OCR [61]	HRNetV2-W48	–	–	83.0	84.5	45.66
	SegFormer [62]	SegFormer-B5	–	84.7	84.0	–	51.8
	Swin-L [63]	UperNet	–	234.0	–	–	53.5
	ViT-Adaptor-L [64]	UPerNet	–	363.8	85.2	–	60.5
	CrossFormer [65]	UPerNet	–	125.5	–	–	51.4
	Focal-L [66]	UperNet	–	240.0	–	–	55.4
MLP	CycleMLP [70]	CycleMLP-B5	–	79.4	–	–	45.6
	AS-MLP [71]	AS-MLP-B	–	121.0	–	–	49.5
Others	Conv-MLP [75]	ConvMLP-L	–	46.3	–	–	40.0

From the perspective of segmentation accuracy, it is not difficult to find that transformer-based models perform well on multiple benchmarks (e.g., Cityscapes test, ADE20K Val) compared to CNN and MLP-based models. If the goal is to improve network accuracy without paying attention to model size and computational cost, choosing to use transformer to design a segmentation network is a good choice. The MLP-based networks are comparable in accuracy to large CNN-based segmentation networks on the ADE20K Val dataset. From the perspective of network size and inference speed, CNN-based methods have absolute advantages.

The performance achieved by these methods stems from the unique advantages of their respective architectural designs. The convolution operation extracts image features through a fixed-size convolution kernel, and only needs to learn the parameters of the fixed-size window without encoding global information. Therefore, CNN-based networks are more lightweight and have faster inference speed. However, convolution also comes with the disadvantage that it cannot capture long-range dependencies such as the relationship between arbitrary pixels in an image. Furthermore, convolution filter weights remain fixed after training, and thus cannot dynamically adapt to variation to the input. Currently, methods to capture global information mainly include expanding the receptive field and embedding non-local self-attention mechanism in the networks. Self-attention mechanism is an integral part of the transformer, unlike convolution operations, its weights are dynamically calculated and can capture long-range dependencies. Therefore, the transformer-based networks have achieved SOTA results on multiple datasets, due to its unique advantages of encoding global information and having general modeling capabilities. However, transformer-based models have complex structures, modeling global information requires huge computational overhead, and training models requires high costs (enough computing resources, large-scale datasets). MLP-based segmentation methods are new attempts to semantic segmentation architecture. Its existence indicates that convolution operation

and self-attention mechanism are not necessary conditions for good performance and provides more ideas for future development.

In summary, since the emergence of FCN, deep learning-based semantic segmentation methods have made significant progress in both accuracy and speed, their MIoU scores have increased by nearly 30% on multiple datasets. Among all segmentation models, the CNN-based models still occupy the majority. Transformer based methods achieve SOTA in terms of accuracy. At present, as new networks from different camps continue to improve the accuracy on image segmentation benchmarks, no conclusion can be made as which structure among CNN, Transformer, and MLP performs the best or is most suitable for semantic segmentation tasks. In general, each architecture has its own advantages. With the development of deep learning technology, future semantic segmentation models will integrate the advantages of multiple architectures to achieve better performance.

5 Future Directions and Challenges

According to the research we reviewed, there is no doubt that image segmentation based on deep learning has developed rapidly and made great achievements. Image segmentation is not only the foundation of scene understanding, but also the focus of future research. Next, some major challenges in this field are summarized and future research directions are given.

5.1 Self-Supervised and Unsupervised Learning

Image semantic segmentation tasks rely on high-quality labeled data. However, labeling data is a time-consuming and labor-intensive task. To alleviate the dependence of semantic segmentation tasks on labeled data, self-supervised [76] and unsupervised learning [77] can be utilized. These technologies have great research value in image segmentation and can effectively alleviate the problem of poor image analysis results in some fields due to the lack of image segmentation datasets. Based on the idea of transfer learning, the general image segmentation model is trained on the upstream tasks, and then fine-tuned on the target dataset to perform new tasks in the target domain. With the help of self-supervised learning, a large amount of unlabeled data is used to capture the detailed information in the image to guide the model training, to reduce the dependence on labeled data in the model training process.

5.2 3D Point-cloud Segmentation

A lot of image segmentation works focus on two-dimensional image segmentation and cannot be directly applied to video sequences. However, visual recognition and scene understanding in the real world are usually based on video sequences. Point-cloud segmentation has a wide range of applications in the fields of autopilot, robot, 3D reconstruction, building modeling and so on. At present, point cloud segmentation faces many challenges, such as how to deal with 3D disordered and unstructured point-cloud data, how to construct an image segmentation model based on the original point cloud to reduce the loss of effective information in the process of information processing, and how to use deep learning method to segment different individuals in complex scenes, which are worthy of research and exploration.

5.3 Real-time Segmentation

In many applications, high accuracy is no longer the only pursuit, and it is also critical that the model can infer at near real-time speeds (it needs to reach the speed of at least 25 frames per second for general-purpose cameras). This is very useful for computer vision systems deployed in devices

such as self-driving cars, robot dogs, and UAVs. Although large-scale transformer networks and MLP-based networks can achieve high accuracy, they have intensive power and computational requirements, and high resource costs affect their deployment on devices. Currently, most models either have high segmentation accuracy but long inference time, or fast inference but low segmentation accuracy, and do not achieve a good balance between speed and accuracy. Therefore, future work needs to pay more attention to real-time constraints and efficient hardware designs, continuously improve model performance, simplify the network, and find a balance between accuracy and running time to promote the implementation and application of relevant technologies.

5.4 Hybrid Architecture Segmentation

Existing methods are often limited to a single network architecture and cannot fully integrate the advantages of different architectures. Each architecture type has its advantages and disadvantages. CNN-based architectures have low computational overhead, but they cannot capture long-distance features. In the future, it can be compensated by expanding the receptive field by using large convolution kernels, atrous convolution and attention mechanisms. Transformer-based architectures and MLP-based architectures can encode global dependencies, but they have high parameter complexity, which leads to high training and inference costs. The future work can build an efficient visual network by reasonably compressing the network architecture or using network architecture search methods. In the future, in addition to optimizing different architectures separately, researchers can also fully integrate the advantages of multiple architectures to build an effective hybrid architecture. Although there have been some related researches on hybrid architectures, there is still a lot of room for improvement. In addition, some useful experiences in the development of CNN-based architectures (e.g., constructing multi-scale information and fusing shallow information) can be applied to transformer- and MLP-based architectures to optimize feature extraction.

6 Conclusion

The research of semantic segmentation methods based on deep learning has made great progress in recent years. This survey first presents some commonly used image segmentation datasets and later reviews pioneering methods in the field of general image semantic segmentation. Furthermore, these methods are divided into 4 types according to their different architectures and highlights: CNN-based, transformer-based, MLP-based and others. To discover and utilize the power of different types of architectures, existing methods are compared and analyzed based on evaluation metrics (such as model size, inference speed, segmentation accuracy), and the key strengths and limitations of different types of architectures are reported. In generally, CNN-based methods have lighter models and faster inference speed, transformer-based methods can encode global information, MLP-based models are simple in design and do not require convolution operations and self-attention mechanisms. At present, there is no conclusion can be made as which structure among CNN, Transformer, and MLP performs the best or is most suitable for semantic segmentation tasks. Finally, possible research directions and challenges are specifically elaborated. We believe that combining the advantages of multiple deep learning architectures to design high-precision and high-efficiency networks is a key point to be explored to solve the current bottleneck, which is also the future scope of our present work.

Funding Statement: This work was supported by the Major science and technology project of Hainan Province (Grant No. ZDKJ2020012), National Natural Science Foundation of China (Grant No. 62162024 and 62162022), Key Projects in Hainan Province (Grant ZDYF2021GXJS003 and Grant ZDYF2020040), Graduate Innovation Project (Grant No. Qhys2021-187).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Naqvi, D. Hussain and W. Loh, "Artificial intelligence-based semantic segmentation of ocular regions for biometrics and healthcare applications," *Computers, Materials & Continua*, vol. 66, no. 1, pp. 715–732, 2021.

[2] X. Tang, W. Tu, K. Li and J. Cheng, "DFFNet: An IoT-perceptive dual feature fusion network for general real-time semantic segmentation," *Information Sciences*, vol. 565, pp. 326–343, 2021.

[3] T. Leonardo, P. Piazzolla, F. Porpiglia and E. Vezzetti, "Real-time deep learning semantic segmentation during intra-operative surgery for 3D augmented reality assistance," *International Journal of Computer Assisted Radiology and Surgery*, vol. 16, no. 9, pp. 1435–1445, 2021.

[4] S. Nedevschi, "Weakly supervised semantic segmentation learning on UAV video sequences," in *Proc. of 29th European Signal Processing Conf. (EUSIPCO)*, Dublin, Ireland, pp. 731–735, 2021.

[5] T. B. Zhu, D. Wang, Y. H. Li and W. J. Dong, "Three-dimensional image reconstruction for virtual talent training scene," *Traitement du Signal*, vol. 38, no. 6, pp. 1719–1726, 2021.

[6] S. Mahajan and A. K. Pandit, "Image segmentation and optimization techniques: A short overview," *Medicon Engineering Themes*, vol. 2, no. 2, pp. 47–49, 2022.

[7] J. Cheng, Y. Yang, X. Tang, N. Xiong, Y. Zhang *et al.,* "Generative adversarial networks: A literature review," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 12, pp. 4625–4647, 2020.

[8] E. Shelhamer, J. Long and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[9] S. Minaee, Y. Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz *et al.,* "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 1–20, 2021.

[10] F. Cao and Q. Bao, "A survey on image semantic segmentation methods with convolutional neural network," in *Proc. CISCE*, Kuala Lumpur, Malaysia, pp. 458–462, 2020.

[11] Y. Zhao, G. Wang, C. Tang, C. Luo, W. Zeng *et al.,* "A battle of network structures: An empirical study of CNN, transformer, and MLP," arXiv preprint, arXiv:2108.13002, 2021.

[12] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler *et al.,* "The cityscapes dataset for semantic urban scene understanding," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 3213–3223, 2016.

[14] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona *et al.,* "Microsoft COCO: Common objects in context," in *Proc. of European Conf. Computer Vision (ECCV)*, Zurich, Switzerland, vol. 8693, pp. 740–755, 2014.

[15] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso *et al.,* "Scene parsing through ADE20K dataset," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5122–5130, 2017.

[16] G. J. Brostow, J. Fauqueur and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[17] H. Caesar, J. R. R. Uijlings and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 1209–1218, 2018.

[18] G. Varma, A. Subramanian, A. M. Namboodiri, M. Chandraker and C. V. Jawahar, "IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *Proc. of IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Waikoloa Village, HI, USA, pp. 1743–1751, 2019.

[19] C. Sakaridis, D. X. Dai and L. V. Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea, pp. 374–7383, 2019.

[20] C. Sakaridis, D. X. Dai and L. V. Gool, "ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 10745–10755, 2021.
[21] J. He, S. Yang, S. K. Yang, A. Kortylewski, X. D. Yuan *et al.,* "PartImageNet: A large, high-quality dataset of parts," arXiv preprint, arXiv: 2112.00933, 2021.
[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of Int. Conf. on Learning Representations (ICLR)*, San Diego, CA, USA, pp. 1–14, 2015.
[23] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.
[24] X. Li, W. Wang, X. Hu and J. Yang, "Selective kernel networks," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 510–519, 2019.
[25] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang *et al.,* "ResNeSt: Split-attention networks," arXiv preprint, arXiv:2004.08955, 2004.
[26] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 4510–4520, 2018.
[27] N. Ma, X. Zhang, H. Zheng and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. of European Conf. Computer Vision (ECCV)*, Munich, Germany, pp. 122–138, 2018.
[28] M. Liu and Y. Zhang, "A hierarchical feature extraction network for fast scene segmentation," *Sensors*, vol. 21, no. 22, pp. 7730–7747, 2021.
[29] X. Zhang, B. Du, Z. Wu and T. wan, "LAANet: Lightweight attention-guided asymmetric network for real-time semantic segmentation," *Neural Computing & Applications*, vol. 34, no. 5, pp. 3573–3587, 2022.
[30] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen *et al.,* "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3051–3068, 2021.
[31] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai *et al.,* "Rethinking BiSeNet for real-time semantic segmentation," in *Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 9716–9725, 2021.
[32] Y. Hong, H. Pan, W. Sun and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," arXiv preprint, arXiv: 2101.06085, 2021.
[33] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao *et al.,* "High-resolution representations for labeling pixels and regions," arXiv preprint, arXiv:1904.04514, 2019.
[34] B. Jiang, W. Tu, C. Yang and J. Yuan, "Context-integrated and feature-refined network for lightweight object parsing," *IEEE Transactions on Image Processing*, vol. 29, pp. 5079–5093, 2020.
[35] J. Cheng, X. Peng, X. Tang, W. Tu and W. Xu, "MIFNet: A lightweight multiscale information fusion network," 2021. [Online]. Available https://doi.org/10.1002/int.22804.
[36] S. Huang, Z. Lu, R. Cheng and C. He, "FAPN: Feature-aligned pyramid network for dense image prediction," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 844–853, 2021.
[37] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid scene parsing network," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 6230–6623, 2017.
[38] H. Zhang, K. J. Dana, J. Shi, Z. Zhang, X. Wang *et al.,* "Context encoding for semantic segmentation," in *Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7151–7160, 2018.
[39] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and 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*, vol. 40, no. 4, pp. 834–848, 2018.
[40] L. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. of European Conf. Computer Vision (ECCV)*, Munich, Germany, pp. 833–851, 2018.

[41] M. Yang, K. Yu, C. Zhang, Z. Li and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 3684–3692, 2018.
[42] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 1, pp. 4171–4186, 2019.
[43] B. Y. Xue, J. W. Yu, J. H. Xu, S. S. Liu, S. K. Hu *et al.,* "Bayesian transformer language models for speech recognition," in *Proc. ICASSP*, Toronto, ON, Canada, pp. 7378–7382, 2021.
[44] Z. W. Zhang, T. Li, X. Tang, X. Hu and Y. Peng, "CAEVT: Convolutional autoencoder meets lightweight vision transformer for hyperspectral image classification," *Sensors*, vol. 22, no. 10, pp. 3902–3923, 2021.
[45] Z. Deng, B. Zhou, P. He, J. Huang, O. Alfarraj *et al.,* "A position-aware transformer for image captioning," *Computers, Materials & Continua*, vol. 70, no. 1, pp. 2065–2081, 2022.
[46] Y. N. Dai, J. Y. Yu, D. Zhang, T. H. Hu and X. T. Zheng, "RODFormer: High-precision design for rotating object detection with transformers," *Sensors*, vol. 22, no. 7, pp. 2633–2646, 2022.
[47] Z. Y. Xu, W. C. Zhang, T. X. Zhang, Z. Yang and J. Y. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sensing*, vol. 13, no. 18, pp. 3585–3609, 2021.
[48] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit *et al.,* "How to train your vit? Data, augmentation, and regularization in vision transformers," arXiv preprint, arXiv:2106.10270, 2021.
[49] H. Ahmad, H. U. Khan, S. Ali, S. Ijaz, F. Wahid *et al.,* "Effective video summarization approach based on visual attention," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1427–1442, 2022.
[50] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7132–7141, 2018.
[51] X. Wang, R. B. Girshick, A. Gupta and K. He, "Non-local neural networks," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7794–7803, 2018.
[52] Z. L. Huang, X. G. Wang, L. C. Huang, C. Huang, Y. Wei *et al.,* "CcNet: Criss-cross attention for semantic segmentation," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea, pp. 603–612, 2019.
[53] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao *et al.,* "Dual attention network for scene segmentation," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 3146–3154, 2019.
[54] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao *et al.,* "Scene segmentation with dual relation-aware attention network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2547–2560, 2021.
[55] A. Sagar, "DMSANnet: Dual multi scale attention network," in *Proc. of Int. Conf. on Image Analysis and Processing (ICIAP)*, Lecce, Italy, pp. 633–645, 2022.
[56] Y. Huang, W. J. Jia, X. J. He, L. Liu, Y. X. Li *et al.,* "CAA: Channelized axial attention for semantic segmentation," arXiv preprint, arXiv:2101.07434, 2021.
[57] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille *et al.,* "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," in *Proc. of European Conf. Computer Vision (ECCV)*, Glasgow, UK, pp. 108–126, 2020.
[58] Q. Hou, D. Zhou and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 13713–13722, 2021.
[59] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo *et al.,* "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 6881–6890, 2021.
[60] R. Strudel, R. G. Pinel, I. Laptev and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 7242–7252, 2021.
[61] Y. Yuan, X. Chen and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. of European Conf. Computer Vision (ECCV)*, Glasgow, UK, pp. 173–190, 2020.
[62] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Alvarez *et al.,* "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. of NeurIPS*, Online Conference, Virtual Event, pp. 12077–12090, 2021.

[63] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei *et al.,* "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. of Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9992–10002, 2021.
[64] Z. Chen, Y. C. Duan, W. H. Wang, J. J. He, T. Lu *et al.,* "Vision transformer adapter for dense predictions," arXiv preprint, arXiv:2205.08534, 2022.
[65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai *et al.,* "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. of Int. Conf. on Learning Representations (ICLR)*, Virtual Event, Austria, pp. 1–21, 2021.
[66] W. Wang, L. Yao, L. Chen, D. Cai, X. He *et al.,* "CrossFormer: A versatile vision transformer hinging on cross-scale attention," arXiv preprint, arXiv:2108.00154, 2021.
[67] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao *et al.,* "Focal attention for long-range interactions in vision transformers," in *Proc. NeurIPS*, Online Conference, Virtual Event, pp. 30008–30022, 2021.
[68] J. Tae, H. Kim and Y. Lee, "MLP singer: Towards rapid parallel Korean singing voice synthesis," in *Proc. of IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Gold Coast, Australia, pp. 1–6, 2021.
[69] T. Yu, X. Li, Y. Cai, M. Sun and P. Li, "S2-MLP: Spatial-shift MLP architecture for vision," in *Proc. WACV*, Waikoloa, HI, USA, pp. 3615–3624, 2022.
[70] S. Chen, E. Xie, C. Ge, D. Liang and P. Luo, "CycleMLP: A MLP-like architecture for dense prediction," arXiv preprint, arXiv:2107.10224, 2021.
[71] D. Lian, Z. Yu, X. Sun and S. Gao, "AS-MLP: An axial shifted MLP architecture for vision," arXiv preprint, arXiv:2107.08391, 2021.
[72] H. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang *et al.,* "NnFormer: Interleaved transformer for volumetric segmentation," arXiv preprint, arXiv:2109.03201, 2021.
[73] Y. Zhang, H. Liu and Q. Hu, "Transfuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. MICCAI*, Strasbourg, France, pp. 14–24, 2021.
[74] R. Guo, D. Niu, L. Qu and Z. Li, "SOTR: Segmenting objects with transformers," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 7137–7146, 2021.
[75] J. Li, A. Hassani, S. Walton and H. Shi, "ConvMP: Hierarchical convolutional MLPs for vision," arXiv preprint, arXiv:2109.04454, 2021.
[76] M. S. Amac, A. Sencan, O. B. Baran, N. Ikizler-Cinbis and R. G. Cinbis, "MaskSplit: Self-supervised meta-learning for few-shot semantic segmentation," in *Proc. WACV*, Waikoloa, HI, USA, pp. 428–438, 2022.
[77] S. Kang and J. Choi, "Unsupervised semantic segmentation method of user interface component of games," *Intelligent Automation & Soft Computing*, vol. 31, no. 2, pp. 1089–1105, 2022.