EAI Endorsed Transactions
on e-Learning

Research Article EAI.EU

A Review of Real-Time Semantic Segmentation Methods for 2D Data in the Context of Deep Learning

Meng Gao[1,*], Haifeng Sima[1,2]

[1]School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, 454000, China
[2]School of Software, Henan Polytechnic University, Jiaozuo, 454000, China

Abstract

Semantic segmentation is a key research topic in the field of computer vision, aiming to assign each pixel to the corresponding category based on the semantic information in the image. This technology has significant application value in fields such as virtual reality and autonomous driving. With the rapid development of deep learning, particularly with the advent of FCN, image semantic segmentation has made substantial progress. Fully supervised learning, which trains deep learning models using labeled data, has demonstrated excellent performance in semantic segmentation tasks. This paper provides a comprehensive discussion and analysis of fully supervised semantic segmentation algorithms for 2D data in deep learning. First, it introduces the concept of semantic segmentation, its development, and its application scenarios. Next, it systematically reviews and categorizes current real-time semantic segmentation algorithms, analyzing the characteristics and limitations of each. Additionally, this paper presents a complete evaluation framework for real-time semantic segmentation, including relevant datasets and evaluation metrics. Based on this foundation, it identifies several challenges currently facing the field and suggests potential directions for future research. Through this summary and analysis, the paper aims to provide valuable insights for researchers conducting studies on image semantic segmentation.

Keywords: Image Semantic Segmentation, Deep Learning, Fully Supervised Learning, 2D Data

Received on 12 01 2025; accepted on 20 02 2025; published on 25 02 2025

doi:10.4108/eetel.8433

1. Introduction

The research on computer vision and all related technologies has made great progress. Image classification, object detection, and semantic segmentation are currently three prominent research directions in the field. As a pixel-level perception task in computer vision, semantic segmentation aims to assign each pixel in an input image to its corresponding category label. Compared with target detection and image classification, semantic segmentation provides more fine-grained information and plays an important role in practical applications. In particular, it plays a vital role in diverse fields such as medical image processing, robot vision, remote sensing image classification, augmented reality, image compression and transmission, autonomous driving vision, and intelligent video analysis. Its applications are essential for advancing these domains. Fully supervised semantic segmentation methods rely on large amounts of labeled data and learn the pixel-level classification task through deep neural networks.

The development of semantic segmentation can be traced back to traditional image processing and pattern recognition methods, which relied on hand-designed feature extraction and segmentation algorithms. These encompass a diverse array of region-based and boundary-based algorithms, including the OTSU method for optimal thresholding [1], watershed for image segmentation based on gradient magnitude [2], region growing for pixel aggregation based on similarity [3], active contours for boundary detection

*Corresponding author. Email: gaomeng@home.hpu.edu.cn

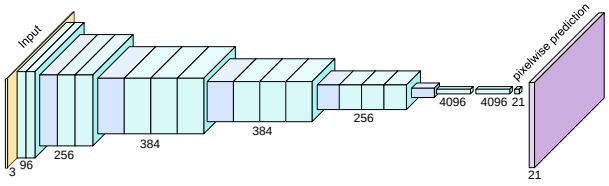

Figure 1. Diagram of FCN network structure

through energy minimization [4], graph cutting for optimizing segmentation boundaries [5], conditional random fields for contextual pixel labeling [6] and Markov random fields for modeling spatial dependencies in image data [7]. However, these traditional methods exhibit significant limitations when dealing with complex scenes and large-scale data. The advent of deep learning has greatly enhanced the capabilities of semantic segmentation algorithms, driving significant progress in their associated applications. In particular, the introduction of FCN [8] marked a breakthrough in semantic segmentation, enabling end-to-end segmentation models and ushering in a new era for semantic segmentation research. The architecture of the FCN is illustrated in Figure 1.

Inspired by FCN, CNN [9] is gradually becoming a new paradigm for image segmentation algorithms. Researchers have developed various improved models to further enhance the performance of semantic segmentation. For example, U-Net [10] achieves remarkable performance in medical image segmentation by merging high-resolution details features with low-resolution semantic information through its encoder-decoder architecture enhanced with Skip Connections. SegNet [11] achieves a more accurate upsampling of semantic information in the decoding stage by introducing Max-Pooling Indices. To capture richer contextual information, a variety of methods have been proposed to enhance the model's ability to utilize multi-scale features. Since semantic segmentation requires the restoration of detailed information lost during down-sampling, multi-scale features, and contextual information are crucial for achieving high segmentation accuracy. As a result, numerous segmentation models have introduced diverse approaches to effectively capture and utilize rich contextual information. For example, the spatial pyramid pooling (ASPP) module in the DeepLab family [12–15] efficiently fuses multi-scale contextual information through parallel null convolution operations. The Pyramid Pooling Module (PPM) of PSPNet [16] enhances the global semantic information of a scene by combining global and local features. Feature Pyramid Network (FPN) [17] fuses features of different resolutions layer by layer in a bottom-up feature pyramid structure, thus improving the fine-grained segmentation of target boundaries.

In recent years, the emergence of self-attention mechanisms and Transformer [18] architectures provide new solutions for semantic segmentation tasks. For instance, Vision Transformer (ViT) [19] splits an image into fixed-size patches and employs Multi-Head Self-Attention (MHSA) to effectively model global features. SETR [20] applies the Transformer for the first time to semantic segmentation tasks. SegFormer [21], a pioneering semantic segmentation framework, achieves high efficiency and advanced performance by merging a hierarchical Transformer encoder with a streamlined multilayer perceptron (MLP) decoder. It advanced performance with both local and global feature representation capabilities. MCTformer+ [22] achieves accurate category-specific target localization through multi-class Token and contrasting class Token modules, which substantially improves the performance of weakly supervised semantic segmentation. These Transformer-based models achieved significant performance gains on multiple semantic segmentation datasets, exhibiting stronger global feature modeling capabilities compared to traditional CNNs.

Although these approaches bring significant segmentation accuracy improvements, they are also accompanied by higher computational costs, especially in the self-attention mechanism and Transformer networks. The self-attentive mechanism is able to effectively capture long-range contexts by virtue of its global modeling capability, but its computational complexity is square to the image resolution, significantly increasing the inference latency. This delay is unacceptable for application scenarios that require real-time performance. Some heavy semantic segmentation networks may incur second-level latency, which can pose significant limitations in real-world deployments. To address these challenges, real-time semantic segmentation networks have emerged. These networks typically refer to models capable of performing inference at a rate of 30 frames per second or higher on a specified device. They meet the minimum frame rate requirement for perceiving smooth video flow by the human eye. However, to achieve this on resource-constrained mobile and edge devices, the model needs to accommodate both rich spatial detail information and multi-scale contextual information. On the one hand, preserving high-resolution underlying feature maps is crucial for obtaining spatial detail information, but this can significantly increase the computational cost. On the other hand, the capture and fusion of multi-scale contexts requires the design of complex modules, which can similarly increase the inference latency. Therefore, A key challenge in real-time segmentation is balancing computational efficiency with the retention of rich spatial information while effectively capturing multi-scale context, making this a crucial area of research.

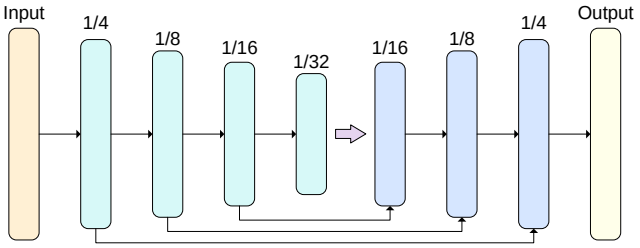

Figure 2. Illustration of single branch semantic segmentation network structure

2. Real-time Semantic Segmentation Methods

With the continuous progress of deep learning technology, real-time semantic segmentation has become an essential research focus in the field of computer vision, attracting increasing attention from researchers. This chapter systematically organizes and summarizes the current real-time semantic segmentation algorithms, offering an in-depth analysis from two key perspectives: network structure and fundamental framework. By examining these aspects, the chapter aims to provide insights into the development trends and challenges in real-time segmentation. Furthermore, Table 1 presents a comprehensive comparison of the performance of several widely used real-time semantic segmentation networks on the Cityscapes dataset.

2.1. Network structure

Single branch network. Single-branch real-time semantic segmentation network accomplishes feature extraction and semantic segmentation tasks through a unified network structure. Due to its simplicity and low computational cost, it is widely applied in scenarios demanding high real-time performance. As illustrated in Figure 2, a single-branch network typically follows an Encoder-Decoder (ED) architecture. The encoder progressively downsamples the image to extract semantic features. The decoder utilizes an upsampling method to recover the spatial resolution and generate segmentation results consistent with the original image size.

To improve real-time performance, ENet [23], one of the earliest semantic segmentation networks designed for real-time applications, achieves a substantial increase in inference speed by incorporating techniques such as factorized convolution and a reduced network depth. Its network structure is similar to SegNet, but its performance is optimized by early downsampling and a lightweight decoder. ERFNet [24] is further optimized on the basis of ENet and proposes Factorized Residual Modules, which reduce the number of parameters while maintaining good feature representation. Compared to ENet, ERFNet has significantly improved segmentation

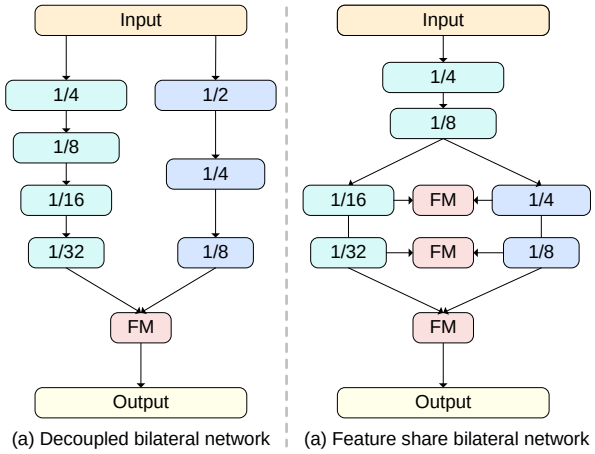

Figure 3. Illustration of the structure of Two-branch semantic segmentation network

accuracy in complex scenes. Its residual connection design ensures the stability of the gradient, which helps to train the deep model and performs well in the edge detail segmentation task. In contrast, SCTNet [25] combines Transformer with traditional single-branch CNN structure. The Transformer branch is introduced in the training phase to capture long-range contextual information while the lightweight single-branch CNN is retained in the inference phase, resulting in efficient real-time inference.

Although the single-scale lightweight decoder significantly improves inference speed, its low-resolution single-scale feature input struggles to recover image details. This results in limited accuracy advantages for this type of real-time segmentation network.

Two-branch network. As a pixel-level dense prediction task, semantic segmentation necessitates capturing both global contextual information and detailed local spatial features at the same time. Two-branch networks provide an effective solution for this purpose. Such networks usually consist of two branches, one focusing on extracting high-resolution spatial information and preserving the detailed features of the image. The other emphasizes learning from low-resolution data to extract high-level semantic features. To effectively integrate the features of the two branches, a specific fusion strategy is usually employed. As shown in Fig. 3, the two-branch real-time semantic segmentation network is broadly categorized into a decoupled bilateral network and a feature-sharing bilateral network.

1) Decoupled two-branch network. The decoupled two-branch network captures this spatial and contextual information separately through independent branches and performs feature fusion at the end of the network, as shown in Figure 3(a). BiSeNet [26] proposed this network structure for the first time. Its

Table 1. Performance comparison of different real-time semantic segmentation networks on the Cityscapes dataset. "–" indicates that no data is provided.

Type	Method	Resolution	Params (M)	FPS	Val mIoU (%)	Test mIoU (%)
Single Branch network	ENet	1024 × 512	0.4	76.9	-	58.3
	ERFNet	640 × 360	-	-	71.3	70.33
	SCTNet	1024 × 512	4.6	160.3	72.8	-
Two-branch network	BiSeNet	1536 × 768	49.0	65.5	74.8	74.7
	BiSeNet V2	1024 × 512	-	47.3	75.8	75.3
	STDC	1536 × 768	22.2	97.0	77.0	76.8
	Fast-SCNN	2048 × 1024	1.1	123.5	68.6	68.0
	DDRNet	2048 × 1024	20.1	37.1	79.5	79.4
	RTFormer	2048 × 1024	16.8	39.1	79.3	-
	SeaFormer	1024 × 512	-	-	77.7	77.5
Single Branch network	ICNet	2048 × 1024	26.5	30.3	-	69.5
	ESPNet	1024 × 512	0.4	113	-	60.3
	DFANet	1024 × 1024	7.8	100.0	-	71.3
	PIDNet	2048 × 1024	36.9	31.1	80.9	80.6

proposed spatial path uses a shallower network structure to maintain high-resolution information and capture the underlying detail information. The contextual path branch extracts high-level semantic information through a deep structure. To effectively integrate features from both branches, BiSeNet designs an FFM module, which realizes efficient feature fusion with low computational overhead. Based on the framework of BiSeNet, BiSeNet V2 [27] was optimized in various aspects. The introduction of global average pooling in the context path enhances the network's sensory field and semantic information extraction ability. In addition, BiSeNet V2 is designed with a Bilateral Guided Aggregation Layer to realize the bidirectional fusion of the features of the two branches. To solve the problem of large computation of spatial path branching and lack of underlying supervision, the STDC network proposed by Fan et al [28] constructs a multi-scale feature map by adjusting the depth and number of convolutional kernels. The detail recovery ability of spatial path branching is enhanced by adding a detailed header, which greatly reduces the computation amount.

2) **Feature-sharing two-branch network.** These networks reduce computational redundancy by sharing feature maps at an early stage and dividing them into spatial and contextual branches at a later stage, whose structure is schematically shown in Figure 3(b). Fast-SCNN [29] adopts this design by fusing the early convolution of these spatial and contextual branches into a single branch for learning downsampling. It is then deeply divided into contextual and spatial branches and fuses dual-branch features to reduce computational redundancy. DDRNet proposed by Pan et al [30] further explores the feature-sharing structure by using a shallow layer to share a single branch and dividing it into dual branches at the deeper layer. DDRNet boosts the sensory field and improves the segmentation accuracy with the Bilateral Fusion Module and Deep Aggregation Pooled Pyramid Module while maintaining a low inference latency. RTFormer [31] improves DDRNet by replacing the dyadic branch with the RTFormer module. RTFormer is also a dual-resolution module where the low-resolution branch uses a linear self-attention mechanism and the high-resolution branch uses cross-resolution attention. Its attention module is similar to external attention but optimized for GPU operations, and reduces inference latency by moving the head-splitting operation to the activation function to preserve the full large matrix. On the other hand, SeaFormer [32] proposes a hybrid Transformer-CNN two-branch real-time semantic segmentation network architecture. Its overall structure adopts a feature-sharing dual-branching framework similar to DDRNet but uses a one-sided fusion module to fuse only the context branch features into the spatial branch. By combining the advantages of Transformer and CNN, SeaFormer perfectly balances speed and accuracy on mobile.

By modeling the two-way decoupling of images, the two-branch network can efficiently capture the spatial details and long-range contextual information required for semantic segmentation and improve segmentation efficiency. However, its extra branches and inter-branch interactions also bring extra computational cost and inference delay.

Multi-branch network. Multi-branch network is an efficient architecture designed for real-time semantic segmentation, which extracts different types of features through multiple branches to segment multi-scale and complex scenes accurately. The core idea behind this

design is to leverage the distinct structural characteristics of each branch, effectively integrate multi-resolution features, and achieve an optimal trade-off between computational efficiency and segmentation accuracy. ICNet [33] designs a unique multi-branch network architecture by adopting the characteristics of low-resolution images that are faster in inference and high-resolution images that have higher prediction accuracy. ICNet employs a cascaded feature fusion unit to integrate multi-resolution features, progressively refining the prediction details and enhancing segmentation quality.

Building on this, ESPNet [34] introduces an efficient spatial pyramid (ESP) structure. It first applies the 1×1 convolution to downscale the feature map dimensions and then expands the sensory field by null convolution to capture broader contextual information. In addition, ESPNet solves the grid effect problem of null convolution by layer-by-layer image fusion technique and improves the network operation efficiency by using channel splicing on the basis of multiple resolution image inputs. DFANet [35] adopts the structure of additional subnetworks to refine the high-level features by fusing the multiscale features. Unlike ICNet and ESPNet, DFANet re-inputs the multiscale features output from the subnetwork into the subnetwork for processing. To avoid the lack of spatial details in large sensory fields and small-scale features leading to accuracy degradation, DFANet uses high-resolution feature refinement layer by layer in the decoder stage. Large-scale features with rich details are eventually recovered, which leads to a notable improvement in segmentation accuracy, particularly in complex scenes.

PIDNet [36] constructs a three-branch network from the PID control algorithm. It improves the ICNet branching structure and expands the DDRNet branching structure, and its structure is shown in Figure 4. At the initial stage of network processing, a single-branch convolutional network downsamples the input image to 1/8 of its original resolution. Then three branches are used to parse the details (I branch), context (P branch), and boundary information (D branch) respectively. PIDNet uses the Boundary-Attention-Guided and Pixel-Attention-Guided fusion modules to fuse the feature information of the three branches. For supervised training, besides using ground truth labels for overall supervision, boundary labels are also used to supervise the P and D branches to improve its feature extraction ability. By achieving an optimal trade-off between inference speed and accuracy, PIDNet stands out as a representative multi-branch network.

Although multi-branch networks have obvious advantages in terms of accuracy, their complex architectural design also brings high computational overhead and inference delay. How to improve computational efficiency while maintaining segmentation accuracy remains a key research focus for the development of multi-branch networks in the future.

2.2. Foundation framework

In deep learning, real-time semantic segmentation methods can be broadly classified into three types based on their foundational architectures: CNN-based frameworks, Transformer-based frameworks, and hybrid frameworks that integrate both CNNs and Transformers. The algorithms included in these three frameworks are shown in Figure 5. With the long-term development of the CNN framework in image processing, real-time semantic segmentation methods based on CNN have been widely researched and applied. The CNN framework is still the mainstream choice for real-time semantic segmentation tasks due to its linear complexity in image resolution and good optimization of hardware acceleration. CNN frameworks are structurally diverse and can realize single-branching, two-branching, multi-branching, and other network architectures. These different architectures can be adapted to meet different real-time and accuracy requirements by adjusting the network depth, branch design, and module configuration. On the other hand, Transformer-based frameworks have attracted significant interest due to their capability to model long-range dependencies and capture global context, both essential for semantic segmentation. At the heart of the Transformer architecture is the self-attention mechanism, which can be formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Where Q, K, and V denote the query, key, and value matrices, respectively, while d_k represents the dimensionality of the keys. This mechanism enables the model to dynamically assess the importance of different regions in the input image. However, the self-attention operation has a quadratic complexity of $O(n^2)$ concerning the number of input tokens n. This will lead to significant computational overhead, especially for high-resolution images. This limitation has spurred research into efficient variants of Transformers, such as sparse attention mechanisms and hierarchical structures, to make them more suitable for real-time applications. Nevertheless, some Transformer-based real-time semantic segmentation methods, such as SegFormer and AFFormer [37], have made breakthroughs in recent years. By lightweighting the Transformer module and adopting a single-branch network architecture, these methods greatly reduce the inference latency and can effectively maintain real-time segmentation accuracy while ensuring real-time performance. Therefore, although there are relatively few real-time semantic

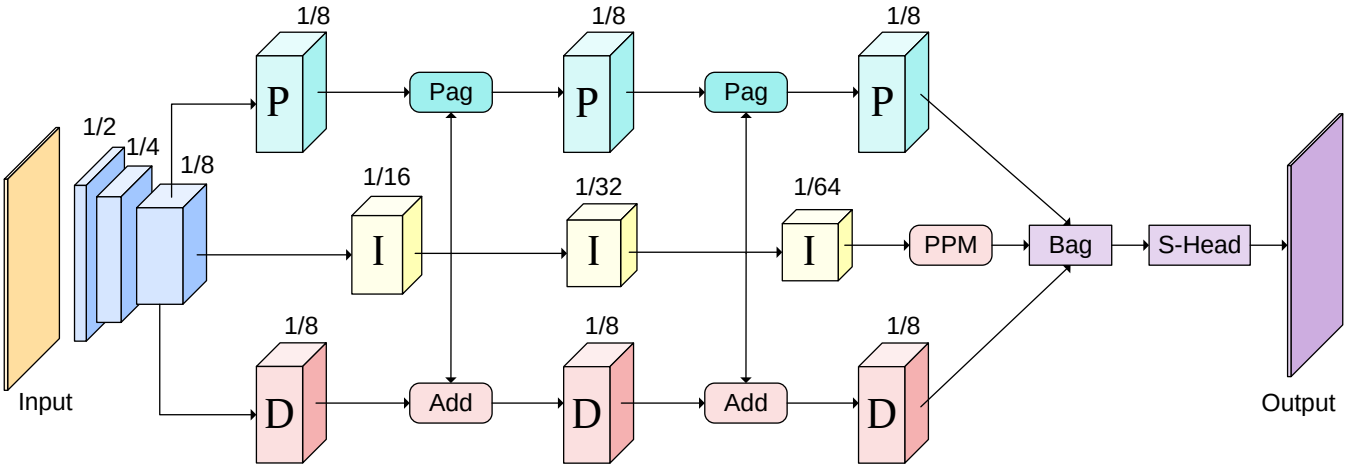

Figure 4. Overall architecture of PIDNet

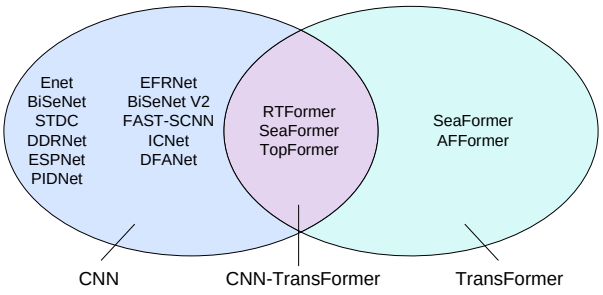

Figure 5. Algorithms included in the three frameworks

segmentation methods based on Transformer, with the development of the technology, its application potential is still huge, especially when dealing with complex scenes with long-distance dependencies.

Real-time semantic segmentation methods based on CNN-Transformer hybrid frameworks aim to fully utilize the respective advantages of CNN and Transformer to achieve more efficient and accurate semantic segmentation. Such approaches typically use the Transformer module in the deeper layers of the network to capture global contextual information and reserve the CNN in the shallow or middle layers to extract local features. For example, Topformer [38] is a representative method based on this framework. It achieves a balance between speed and accuracy by embedding the Transformer in the deeper layers to capture long-range dependencies. Meanwhile, the CNN remains in the shallow layers to preserve local details. Another common design follows a two-branch structure. Here, the CNN extracts spatial features, while the Transformer models long-range context. RTFormer and SeaFormer exemplify this design, which, by combining the Transformer as a context branch with the spatial branch of the CNN, achieves efficient processing of semantic segmentation tasks.

The CNN-Transformer hybrid framework can significantly improve real-time semantic segmentation performance by integrating the local modeling capability of CNN and the advantage of the Transformer in capturing long-range contextual information. More importantly, this framework avoids the problem of high inference latency that the Transformer may incur when processing high-resolution images, thus ensuring the efficiency of real-time inference. Therefore, the algorithm based on the hybrid CNN-Transformer framework is not only able to achieve high segmentation accuracy in most complex scenarios but also meets the real-time requirements, showing great potential and advantages in practical applications.

3. Evaluation system for semantic segmentation

3.1. Relevant 2D datasets

Data sets are a prerequisite for algorithmic research, and some research institutions, large companies, and competition programs around the world have open-sourced quite a number of large-scale data sets, which have greatly promoted the development of related fields. The research on image semantic segmentation mainly focuses on 2D images, and the more commonly used 2D datasets are as follows:

Cityscapes [39]: a large-scale urban streetscape semantic understanding dataset with 5000 finely annotated images and 20000 roughly annotated images. Among them, 2975 of the fine annotated images are used for training, 500 for validation, and 1525 for testing. All images have a resolution of 2048×1024 pixels and cover 30 different classes, 19 of which are used for semantic segmentation. Real-time semantic segmentation methods usually use only finely

annotated images, and only a few methods use coarse data for further enhancement of network performance. It covers a variety of complex urban street scenes with high image resolution, high annotation quality, and a wide range of spatial and temporal spans, making it one of the most commonly used datasets in the field of real-time segmentation. The dataset is available at https://www.cityscapes-dataset.com.

CamVid [40, 41]: a densely labeled autopilot dataset containing 701 images of vehicle driving viewpoints from a 10-minute driving shot sequence in Cambridge, UK. It includes 367 training images, 101 validation images, and 233 test images. The resolutions are all 960×720 pixels and contain 32 categories, 11 of which are used for semantic segmentation. This dataset is weakly spatio-temporally diverse and small, and networks for CamVid often require Cityscapes pre-training. The dataset can be downloaded at mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid.

ADE20K [42, 43]: one of the most comprehensive datasets available for scene parsing and semantic segmentation, containing 22210 images. Among them, 20,210 images are used for training and 2,000 for validation. Its images are of different sizes and cover a variety of scenes, including indoor, outdoor, nature, and urban, etc. These images are annotated at the pixel level and cover 150 categories of objects, including common object categories as well as some rare categories. The problem of having rich semantic categories and long-tailed distributions is a challenge for lightweight real-time semantic segmentation networks. The dataset is hosted on the website https://groups.csail.mit.edu/vision/datasets/ADE20K.

COCOStuff-10K [44]: COCOStuff-10K is obtained by extending the large-scale scene understanding dataset COCO with numerous "stuff" categories. It is a set of 10,000 complex images with artificial dense labeling in COCOStuff, of which 9,000 are used for training and 1,000 for testing. It contains about 182 categories, including 91 thing categories and 91 stuff categories, but 11 of the thing categories are not labeled, so only 171 categories are used. This is also a challenging dataset for real-time semantic segmentation because it has more complex categories and more variable scenarios. The datasets can be downloaded at https://github.com/eulersantana/cocostuff.

PASCAL VOC 2012 [45]: this dataset is one of the important benchmark datasets in computer vision and is widely used for image classification, target detection, and semantic segmentation tasks. It contains 20 different semantic categories and 1 background category with 5717 training images, and 5823 validation images Each image is labeled at a fine pixel level to ensure high-quality semantic segmentation labels. The dataset is available at host.robots.ox.ac.uk/pascal/VOC/voc2012.

3.2. Evaluation indicators

Mean Intersection over Union (mIoU) is a widely metric for evaluating semantic segmentation. IoU measures the degree of overlap between predicted and ground truth masks by calculating the ratio of their intersection and the merge of two sets of pixels. MIoU is the average of the IoUs across all classes in the dataset, and it is a measure of the overall accuracy of the segmentation model. The calculation formula is as follows:

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{2}$$

$$mIoU = \frac{1}{C}\sum_{i=1}^{c} IoU_i \tag{3}$$

Where TP_i denotes the number of pixels correctly classified as class i. FP_i represents the pixels mistakenly predicted as class i. FN_i refers to the pixels belonging to class i but not correctly predicted. C is the total number of classes.

Frames per second (FPS) is the number of image frames per second that a network model can process, the size of which is affected by the performance of a specific device. Real-time segmentation requires the network's FPS to be greater than or equal to 30 frames/s. FPS is a visual measure of the network's inference speed and throughput when comparing across devices.

Inference Latency is the inverse of FPS, indicating the time required by the network model to process one frame of an image. It is usually used more in semantic segmentation models on mobile. GPU devices have more arithmetic power and are accelerated with TensorRT for faster inference. Many real-time segmentation networks have an FPS of well over 30 frames per second on GPUs, and it is not intuitive to compare latency at this point. While the mobile end CPU arithmetic resources are limited, the delay is large and the FPS is small, using the delay can be more intuitive to compare the inference speed of the network.

Number of parameters refers to the total number of parameters that need to be trained and learned in the model training process. The number of parameters is usually used to measure the size of the model. On edge devices with limited resources, the number of parameters is a key factor to consider.

4. Problems and directions for future research

1) Labeling Difficulty: In semantic segmentation, labeling each pixel in an image with an accurate semantic category is a heavy and time-consuming task. For example, complex organs and lesion regions in medical images require specialized knowledge for fine annotation, while the annotation standards of different annotators may differ, leading to consistency problems.

In addition, it is difficult to obtain labeled data that comprehensively covers various scenes and categories, which limits the generalization ability of the model.

2) Category imbalance: In actual image data, the distribution of different semantic categories is usually uneven. Take the commonly used road scene images as an example, categories such as roads and buildings usually occupy most of the area, while categories like traffic signs and pedestrians are less frequent. This category imbalance causes the model to be biased towards learning common categories, leading to performance degradation when dealing with rare categories. For example, a model trained on a natural scene dataset may recognize common categories like trees better, but recognize rare animals less well.

3) Data Diversity and Complexity: Real-world image data is highly diverse and complex. Lighting, changes in scale, and occlusion phenomena all affect the appearance characteristics of objects. Strong direct light or low-light environments may lead to loss of detail information. Scale variations of objects also increase the difficulty of model recognition, especially in aerial images, where there are large-scale buildings as well as small vehicles and pedestrians. In addition, occlusion and overlap between objects, such as in crowded scenes, also increase the challenge of model segmentation accuracy.

4) Segmentation accuracy improvement bottleneck: Although existing semantic segmentation models have achieved better segmentation results in some specific scenes, it is still difficult to achieve ideal accuracy when dealing with some complex scenes or fine objects. For example, in medical image analysis, for tiny lesions or complex structures, the existing models may not be able to segment the details accurately, especially when the boundaries of tumor cells or neural tissues are fuzzy and the contrast is low.

5) Computational resource requirements and efficiency: Deep learning models, especially Transformer-based semantic segmentation models, demand substantial computational resources for high-resolution images. Their computational complexity grows in square steps with the increase of image resolution, posing challenges for deployment on resource-limited devices. In applications that require real-time decision-making like autonomous driving, slow inference can hinder rapid decision-making, compromising both efficiency and safety.

6) Generalization ability of the model: The performance of the model in different datasets and scenarios often varies, and the generalization ability needs to be improved. Even when a model that has been trained well on a specific dataset is applied to another dataset or a real-world scenario, performance degradation may occur. For example, a model trained on an urban street view dataset may not be able to accurately segment the target object when processing images in a rural or industrial environment. Many existing models lack generalization capabilities to adapt to diverse scenes and object classes.

7) Multi-modal data fusion challenges: With the development of technology, more and more application scenarios require fusion of RGB images, depth images, LiDAR and other multi-modal data for semantic segmentation. However, the differences in the characteristics of different modal data make data fusion complicated. For example, RGB images mainly provide color and texture information of objects, while depth images focus on distance information of objects. Existing methods are deficient in information integration, limiting the effectiveness of multimodal data fusion.

8) Balance between real-time and accuracy: In application scenarios with high real-time requirements such as automatic driving and video surveillance, how to ensure high segmentation accuracy while realizing fast inference is a current difficulty. Existing models usually require complex computational processes, leading to delayed inference, which may affect decision-making timeliness in rapidly changing environments.

9) Interpretability of models: Deep learning models have achieved good results in semantic segmentation, but they are often regarded as "black-box" models that lack a transparent decision-making process. This may pose a potential risk in applications that require high safety and reliability, such as medical diagnosis and autonomous driving. If a model's decision-making process is not interpretable, users may not be able to fully understand its output, which may affect the application and trust of the model in clinical and real-world scenarios.

References

[1] Otsu, N. (1979) A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1): 62–66.

[2] Meyer, F. and Beucher, S. (1990) Morphological segmentation. *Journal of Visual Communication and Image Representation* **1**(1): 21–46.

[3] Adams, R. and Bischof, L. (1994) Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(6): 641–647.

[4] Kass, M., Witkin, A. and Terzopoulos, D. (1988) Snakes: Active contour models. *International Journal of Computer Vision* **1**(4): 321–331.

[5] Boykov, Y. and Jolly, M.P. (2001) Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. *Proceedings of IEEE International Conference on Computer Vision (ICCV)* : 105–112.

[6] Lafferty, J., McCallum, A. and Pereira, F. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of*

the 18th International Conference on Machine Learning (ICML) : 282–289.

[7] Li, S.Z. (2009) *Markov random field modeling in image analysis* (Springer Science & Business Media).

[8] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 3431–3440.

[9] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25**.

[10] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W.M. and Frangi, A.F. [eds.] *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (Cham: Springer International Publishing): 234–241.

[11] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(12): 2481–2495.

[12] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K.P. and Yuille, A.L. (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR* **abs/1412.7062**. URL https://api.semanticscholar.org/CorpusID:1996665.

[13] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2018) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(4): 834–848.

[14] Chen, L., Papandreou, G., Schroff, F. and Adam, H. (2017) Rethinking atrous convolution for semantic image segmentation. *CoRR* **abs/1706.05587**. URL http://arxiv.org/abs/1706.05587. 1706.05587.

[15] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*. URL https://api.semanticscholar.org/CorpusID:3638670.

[16] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017) Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 6230–6239. doi:10.1109/CVPR.2017.660.

[17] Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B. and Belongie, S.J. (2016) Feature pyramid networks for object detection. *CoRR* **abs/1612.03144**. URL http://arxiv.org/abs/1612.03144. 1612.03144.

[18] Vaswani, A. (2017) Attention is all you need. *Advances in Neural Information Processing Systems* .

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M. *et al.* (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR* **abs/2010.11929**. URL https://arxiv.org/abs/2010.11929. 2010.11929.

[20] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y. *et al.* (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 6877–6886. doi:10.1109/CVPR46437.2021.00681.

[21] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021) Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems* **34**: 12077–12090.

[22] Xu, L., Bennamoun, M., Boussaid, F., Laga, H., Ouyang, W. and Xu, D. (2024) Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(12): 8380–8395. doi:10.1109/TPAMI.2024.3404422.

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

[24] Romera, E., Álvarez, J.M., Bergasa, L.M. and Arroyo, R. (2018) Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems* **19**(1): 263–272. doi:10.1109/TITS.2017.2750080.

[25] Xu, Z., Wu, D., Yu, C., Chu, X., Sang, N. and Gao, C. (2024), Sctnet: Single-branch cnn with transformer semantic information for real-time segmentation. URL https://arxiv.org/abs/2312.17071. 2312.17071.

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

[27] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C. and Sang, N. (2020) Bisenet V2: bilateral network with guided aggregation for real-time semantic segmentation. *CoRR* **abs/2004.02147**. URL https://arxiv.org/abs/2004.02147. 2004.02147.

[28] Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J. and Wei, X. (2021) Rethinking bisenet for real-time semantic segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 9711–9720. doi:10.1109/CVPR46437.2021.00959.

[29] Poudel, R.P.K., Liwicki, S. and Cipolla, R. (2019) Fast-scnn: Fast semantic segmentation network. *CoRR* **abs/1902.04502**. URL http://arxiv.org/abs/1902.04502. 1902.04502.

[30] Pan, H., Hong, Y., Sun, W. and Jia, Y. (2023) Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems* **24**(3): 3448–3460. doi:10.1109/TITS.2022.3228042.

[31] Wang, J., Gou, C., Wu, Q., Feng, H., Han, J., Ding, E. and Wang, J. (2022), Rtformer: Efficient design for real-time semantic segmentation with transformer. URL https://arxiv.org/abs/2210.07124. 2210.07124.

[32] Wan, Q., Huang, Z., Lu, J., Yu, G. and Zhang, L. (2024), Seaformer++: Squeeze-enhanced axial transformer for mobile visual recognition. URL https://arxiv.org/abs/2301.13156. 2301.13156.

[33] Zhao, H., Qi, X., Shen, X., Shi, J. and Jia, J. (2018) Icnet for real-time semantic segmentation on high-resolution

images. In *Proceedings of the European conference on computer vision (ECCV)*: 405–420.

[34] Mehta, S., Rastegari, M., Caspi, A., Shapiro, L. and Hajishirzi, H. (2018) Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*: 552–568.

[35] Li, H., Xiong, P., Fan, H. and Sun, J. (2019) Dfanet: Deep feature aggregation for real-time semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*: 9514–9523. doi:10.1109/CVPR.2019.00975.

[36] Xu, J., Xiong, Z. and Bhattacharyya, S.P. (2023), Pidnet: A real-time semantic segmentation network inspired by pid controllers. URL https://arxiv.org/abs/2206.02066. 2206.02066.

[37] Dong, B., Wang, P. and Wang, F. (2023), Head-free lightweight semantic segmentation with linear transformer. URL https://arxiv.org/abs/2301.04648. 2301.04648.

[38] Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., Yu, G. *et al.* (2022), Topformer: Token pyramid transformer for mobile semantic segmentation. URL https://arxiv.org/abs/2204.05525. 2204.05525.

[39] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U. *et al.* (2016) The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 3213–3223.

[40] Brostow, G.J., Shotton, J., Fauqueur, J. and Cipolla, R. (2008) Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*: 44–57.

[41] Brostow, G.J., Fauqueur, J. and Cipolla, R. (2009) Semantic object classes in video: A high-definition ground truth database. *Pattern recognition letters* **30**(2): 88–97.

[42] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A. and Torralba, A. (2017) Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*: 5122–5130. doi:10.1109/CVPR.2017.544.

[43] Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A. and Torralba, A. (2018), Semantic understanding of scenes through the ade20k dataset. URL https://arxiv.org/abs/1608.05442. 1608.05442.

[44] Caesar, H., Uijlings, J. and Ferrari, V. (2018) Cocostuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on* (IEEE).

[45] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2010) The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* **88**(2): 303–338.