Performance Comparison and Principle Analysis of Deep Learning-Based Models for Semantic Segmentation

Yuankai Su

School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin, 300072, China
jysyk_5772@tju.edu.cn

Abstract. Nowadays, the concept of artificial intelligence is widely popularized and attracting more and more attention. Computer vision is a hot field of artificial intelligence, and semantic segmentation in computer vision has been a worthwhile research direction in recent years. Not only does it play a crucial role in the current highly focused autonomous driving, but it also has many application scenarios, so it is necessary to study semantic segmentation, The breakthrough in semantic segmentation methods may greatly solve many practical problems currently in use. At present, the main-stream method is semantic segmentation based on deep learning. Compared to traditional machine learning based semantic segmentation methods, it has many advantages, and many excellent researchers are constantly optimizing methods and creating models. Therefore, many models worth learning have emerged during this period. This article will introduce eight semantic segmentation models based on deep learning, analyze the ideas, innovations, and contributions of these methods. After learning and understanding these methods, it may provide new ideas for one's own research and make useful contributions in this field.

Keywords: Semantic Segmentation, Deep Learning, Computer Vision.

1 Introduction

Semantic segmentation is a visual task, with the goal and role of dividing images into different semantic categories at the pixel level. Unlike traditional image classification tasks, semantic segmentation requires each pixel to be classified and labeled as belonging to a different semantic category, such as people, cars, road lines, obstacles, etc.

Semantic segmentation is a popular research direction of artificial intelligence, which has great help for society and humans, and has many positive impacts, such as automation and efficiency improvement, solving complex problems, promoting innovative development, liberating human time and labor.

Semantic segmentation has mature applications in many fields of today's society.In the world of autonomous driving, it is crucial. It is used to identify and segment

© The Author(s) 2024

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_65

different objects on the road, such as vehicles, pedestrians, bicycles, traffic signs, lane lines, etc., to help the auto drive system accurately perceive and understand the surrounding environment, and make correct and reasonable decisions during driving. Medical image analysis uses semantic segmentation in a broad variety of ways. It can help doctors accurately locate and segment lesion areas, such as tumors, organs, blood vessels, etc., thereby assisting in diagnosis, surgical planning, and treatment monitoring, greatly improving the efficiency of diagnosis. Semantic segmentation is widely used in land classification and land use research in remote sensing and satellite image analysis. It can divide surface areas into different categories, such as forests, farmland, cities, water bodies, etc., providing important information for urban planning, environmental monitoring, and resource management. Semantic segmentation is the key in video analysis and monitoring. It can be used for object detection, tracking, and behavior analysis in real-time videos, such as pedestrian tracking, abnormal event detection, and behavior recognition.

There are also some potential application areas for semantic segmentation, such as Assisted drone navigation, urban intelligent transportation, agricultural and crop monitoring, building maintenance and safety, image based retrieval and recommendation, environmental protection and ecological monitoring, etc.

The research on semantic segmentation could promote the advancement of computer vision, AR/VR, medical image analysis, autonomous driving, and other fields. It can provide more refined and accurate image understanding and perception capabilities, bringing more possibilities and benefits to various application scenarios. That's also why it has become one of the hottest research directions. This article will introduce some tested and effective models in the Cityscapes test dataset.

2 Method

Traditional machine learning algorithms and deep learning-based models are the two primary categories of semantic segmentation approaches.

2.1 Traditional Machine Learning-Based Semantic Segmentation

Graph-based methods: Traditional graph-based methods utilize graph representations to model the relationship between image pixels or superpixels. These methods often focus on achieving a global optimization by considering the connectivity and similarity between neighboring pixels or regions. Graph cuts and random walks are commonly used techniques in this category.

Markov Random Fields (MRFs): MRFs are probabilistic graphical models that capture the spatial dependencies between pixels or regions. They model the joint probability distribution of labels (semantic classes) and incorporate local and pairwise potentials to enforce label consistency. Inference algorithms, such as belief propagation or iterative decoding, are employed to estimate the optimal labeling.

Conditional Random Fields (CRFs): CRFs are an extension of MRFs that incorporate observed features, such as color, texture, or edge information, as additional inputs to improve segmentation accuracy. CRFs model the conditional probability distribution of the labels given the observed features, allowing for more informed and context-aware predictions.

Support Vector Machines (SVMs): SVMs are classical algorithms in classificaiton, including semantic segmentation. In this approach, features extracted from image patches or superpixels are fed into an SVM classifier, which learns to distinguish between different semantic classes. These classifiers can be applied to each pixel or region independently to obtain segmentation results.

These conventional machine learning techniques for semantic segmentation are still in use today and have made significant contributions to the area in terms of insights and solutions. Convolutional neural networks (CNNs) have, however, become the preferred method for semantic segmentation with the introduction of deep learning because of their capacity to automatically learn hierarchical features and capture complicated patterns in images [1, 2].

The Cityscapes test dataset contains a large number of deep learning-based models with outstanding semantic segmentation performance.

2.2 Densely Connected Neural Architecture Search (DCNAS)

DCNAS aims to solve the problem of expanding search space as much as possible within limited computing resource providers, avoiding the use of small proxy tasks, and directly searching for the optimal multi-scale network structure on large datasets. Specifically, by the hierarchical organization of learnable weights, a densely connected search space can encompass a large number of mainstream network designs. In addition, the sampling strategy includes two levels: path and channel, and a fusion module and a hybrid layer are designed to reduce the consumption of computing resources. This agentless search scheme can reduce the differences between search and training environments.

DCNAS can select appropriate paths throughout the entire set to derive the final network structure, while aggregating multi-scale contextual semantic information. In order to compensate for the differences between the proxy task and the target task, a fusion module was designed by relaxing the discrete structure into a continuous structure. During the search process, path level and channel level sampling strategies were adopted to reduce memory requirements. On this basis, SGD can be used to perform a search process without agents, selecting the optimal one from all candidate models.

In addition, considering that the NAS method requires angle measurement (such as ranking each model), if high-performance models are found, it can serve as a sign of training convergence. However, there is currently relatively little research on the relative ranking of search stage models. DCNAS has studied a method that can represent the accuracy of the network structure search phase and the correlation between fine-tuning independent models.

The optimal network structure for directly searching for multi-scale representations of visual information on large-scale target datasets. It designed a tightly connected search space that can encompass the design of existing models. On large-scale segmented datasets, proxy models are not used during the search process, and the optimal model can be found from candidate models [3].

2.3 High-Resolution Network with Object-Contextual Representations

High-Resolution Network Version 2 (HRNetV2) architecture with the Object-Contextual Representations (OCR) module is named HRNetV2+OCR. The model is based on the following ideas and designs:

HRNet is a traditional CNNs used for semantic segmentation often suffer from the loss of resolution due to multiple downsampling operations. HRNet addresses this issue by maintaining high-resolution features alongside low-resolution features in multiple branches and integrating them through a progressive fusion process. This construction helps preserve fine-grained details and recover resolution loss.

The OCR module is designed to capture pixel-level object context information to enhance the understanding of object boundaries and contextual relationships in semantic segmentation tasks. It introduces both object boundary attention and context attention mechanisms to model contextual dependencies within the feature maps. This improves the accuracy of object boundaries and semantic segmentation results.

The HRNetV2+OCR model incorporates multi-scale fusion techniques to combine features from different scales. It fuses features across various resolution levels, allowing the model to be aware of local details and global context, thus improving semantic segmentation performance.

The HRNetV2+OCR model aims to address the challenges of resolution loss and context modeling in semantic segmentation. By leveraging the HRNet architecture and incorporating the OCR module, the model achieves higher accuracy and preserves fine details. Additionally, the multi-scale fusion technique enhances the model's robustness and generalization ability by integrating information across different scales.

In summary, HRNetV2+OCR combines the strengths of HRNet's high-resolution features and OCR's object context modeling to obvtain state-of-the-art performance in semantic segmentation.

Overall, the multi-scale fusion technique in the HRNetV2+OCR model enables better contextual understanding, robustness to scale variations, preservation of fine details, hierarchical feature representation, and improved segmentation performance. These advantages contribute to the model's ability to performs good in semantic segmentation tasks [4].

2.4 Panoptic-DeepLab

The Panoptic-DeepLab model is based on the DeepLab framework and incorporates several new designs to improve the accuracy of panoptic segmentation and semantic segmentation tasks.

The Panoptic-DeepLab model extends the DeepLab framework, a well-known architecture for semantic segmentation, which leverages a fully CRF for refinement and dilated convolution to achieve the information from multi-scale context.

The Panoptic-DeepLab model adopts the DeepLabv3+ variant, which further improves upon DeepLab. DeepLabv3+ replaces the original encoder network with a more powerful feature extraction backbone called Xception, which enhances the model's ability to capture fine-grained details and semantic information.

The Panoptic-DeepLab model introduces Panoptic Feature Pyramid Networks (PANet), a novel design that enables efficient fusion of features from different scales. PANet incorporates a top-down pathway and lateral connections to aggregate features from multi-scale and generate high-resolution feature maps. This design helps in accurately capturing objects of various sizes and handling the scale variations present in the Cityscapes dataset.

The Panoptic-DeepLab model includes a Context Encoding Module (CEM) to capture global context information. The CEM module utilizes global average pooling to aggregate contextual information from the entire image and then incorporates it into the feature representations. This helps the model to better understand the overall scene context and improve the segmentation performance, especially for objects with complex structures and large spatial extent.

The new designs of Panoptic-DeepLab aim to address the challenges of accurately segmenting objects and scenes in the Cityscapes test dataset. By combining the strengths with modern models, the model achieves outstanding performance. It accurately labels each pixel with a semantic class and assigns an instance ID to each object, providing a comprehensive understanding of the scene's content. This makes the Panoptic-DeepLab model suitable for a range of applications, such as autonomous driving, scene understanding and urban planning [5].

2.5 EfficientPS

The initial panoramic segmentation method was to simultaneously perform instance segmentation and semantic segmentation, and then combine the predicted results of the two in the post-processing step. It can be imagined that this method has high computational overhead, information redundancy, and differences in predictions for each network, making it difficult to combine. Although bottom-up sequential approaches or top-down shared network components have been successfully used in recent methods to solve this challenge, these approaches still have issues with insufficient computational accuracy and efficiency.

EfficientPS is related to EfficientNet. It includes an improved EfficientNet backbone network and dual FPN, semantic segmentation head, instance segmentation head, and finally a panoramic fusion module. Design features:

The shared backbone network using mobile inverted bottleneck units is improved from EfficientNet, and its biggest innovation is the use of composite scaling to evenly expand all dimensions of the network (input image size, network width, depth, etc.) in the scaling strategy.

The author discovered the 2-way Feature Pyramid Network because the single path information flow in the regular FPN has limits in terms of aggregating multi-scale features. Therefore, a new bidirectional FPN was proposed, which can achieve dual path flow of information and significantly improve the panoramic segmentation quality of foreground classes while maintaining little change in runtime.

In the semantic segmentation header, separable convolutions are used to better capture fine features and long-range contextual information, achieving better target boundary refinement.

In the instance segmentation header, Mask RCNN is used, and separable convolutions and iABN synchronization layers are used to enhance it.

When integrating semantic segmentation and instance segmentation results to generate panoramic segmentation output, the author proposes a new panoramic fusion module that can adaptively adjust their fusion based on the confidence level of the mask of semantic and instance heads. In addition, the conjugate integration of specific foreground class instances and background classes forms the final output result.

EfficientPS's design objective is to outperform earlier state-of-the-art models while achieving excellent computing efficiency [6].

2.6 Vision Transformer Adapter (ViT-Adapter)-L

Weak prior assumptions cause the plain Vision Transformer (ViT) to perform worse on dense predictions. Researchers have developed a solution to this problem called the ViT-Adapter, which enables standard ViT to perform as well as vision-specific transformers. The framework's main component is a straightforward ViT, which can learn potent representations from massive amounts of multimodal input. A pre-training-free adaptor is utilized to include the image-related inductive biases into the model prior to transfer to downstream tasks, making the model appropriate for these tasks.

When compared to transformers designed specifically for vision, the ordinary ViT has clear flaws in dense predictions. Plain ViTs struggle to compete with vision-specific transformers due to slower convergence and inferior performance caused by the absence of image-related prior information. This work seeks to create an adapter to narrow the performance gap between the ordinary ViT and vision-specific backbones for dense prediction tasks, drawing inspiration from adapters in the NLP sector.

The ViT-Adapter, an additional network that doesn't need pre-training and that successfully adapts the basic ViT to downstream heavy prediction tasks without altering its original design, is suggested by researchers as a means to doing this. In order to introduce the vision-specific inductive biases into the plain ViT, researchers developed three customized modules for ViT-Adapter, including a spatial prior module for extracting the local semantics (spatial prior) from input images,a multi-scale feature extractor to rebuild the multi-scale features needed for dense prediction tasks, as well as a spatial feature injector to incorporate spatial prior into the ViT.This study investigates a new paradigm for incorporating the plain ViT with vision-specific inductive biases. ViT benefits from multi-modal pre-training in addition to conventional ImageNet pre-training in achieving performance comparable to current transformer variations. To inject the image prior without changing the ViT architecture, two feature interaction operations and a spatial prior module were designed. For dense prediction problems, they can reorganize fine-grained multi-scale characteristics and fill in the gaps in local knowledge. testing the ViT-Adapter against a variety of difficult benchmarks.Under the fair pre-training technique, the models regularly perform better than the prior arts [7].

2.7 InverseForm

With the help of an inverse-transformation network, researchers provide a unique boundary-aware loss term for semantic segmentation. This plug-in loss term enhances the cross-entropy loss in capturing boundary transitions and provides segmentation backbone models with a consistent and considerable performance gain without increasing their size and computational complexity. They include their loss function into the training phase of several backbone networks in single-task and multi-task contexts and examine the quantitative and qualitative results on three benchmarks for indoor and outdoor segmentation.

The crossentropy loss ignores the distance between the target boundaries and the pixels and instead concentrates on predicted and actual pixel label changes.Localized spatial alterations like translation, rotation, or scaling between the expected and goal bounds cannot be accurately assessed by it. To get over this restriction, researchers have added the boundary distance-based InverseForm measure to the well-known segmentation loss functions. In order to forecast the distance between border maps, researchers build an inverse transformation network that can efficiently learn the degree of parametric modifications between small geographical regions. Using any backbone model, this measure enables us to significantly and consistently increase semantic segmentation without growing the network's inference size or computational complexity.

The contributions to the work are given below. It first recommended enhancing semantic segmentation by employing InverseForm, a boundary distance-based method. In order to achieve more precise segmentation findings, we show that our specially created measure can capture the spatial boundary transforms noticeably better than cross-entropy based measures. Second, the technique is very adaptable and can be easily integrated into any current segmentation model without incurring any additional inference costs, regardless of the backbone architecture preference. Due to its plug-and-play capability, it has no effect on the network's fundamental architecture. It is adaptable and compatible with frameworks for multi-task learning. Thirdly, in different tasks, the boundary-aware-segmentation approach achieves outstanding performance compared to others [8].

2.8 Hierarchically Supervised Semantic Segmentation (HS3)-Fuse

By varying task difficulty, HS3, a training approach, learns meaningful representations of the intermediate layers of a neural network. To ensure high performance with moderated computational cost, researchers use unique class clusters as the supervision of inter mediate layers. They develop HS3-Fuse, a fusion framework, to merge the hierarchical features generated by these layers. This provides rich semantic contexts and further enhances the segmentation process. The HS3 system greatly outperforms deep supervision, according to numerous tests, and does not impose any extra inference costs.

Finding the appropriate learning activity for each intermediate layer that needs supervision is the aim of HS3. By grouping semantic labels into a set with fewer classes and therefore lower complexity, we are able to complete these segmentation tasks. The relevant sub-network, or the portion of the network up to the present layer, is particularly supervised with a reduced set of classes to meet its learning capacity. To estimate the quantity of class clusters, we offer a tenet-based training technique with two steps. This method performs automatic hierarchical class categorization using confusion matrices that were created after training a deep supervision baseline. The second (and last) step of training is when hierarchical supervision is used.

The work's contributions include the following:

Introducing HS3, a unique hierarchical supervision method that enables the supervised intermediate layers to learn with the tasks with moderate complexity, for the training of semantic segmentation networks. With no additional inference costs, this improves the feature learning of the intermediate layers.

Creating HS3-Fuse, a cutting-edge framework, to fully leverage the hierarchical features produced by the intermediate supervised layers. To improve the overall segmentation performance, the integrated features are fed into the output layer with suitable and helpful hierarchical semantic context [9].

2.9 InternImage-H

InternImage is a brand-new, massive CNN-based foundation model that researchers have developed. It can profit from growing parameters and training data like ViTs. InternImage uses deformable convolution as its primary operator, in contrast to more contemporary CNNs that prioritize big dense kernels. As a result, our model has the adaptive spatial aggregation required for tasks like detection and segmentation in addition to the large effective receptive field needed for upstream tasks like detection and segmentation. As a result, the recommended InternImage decreases the extreme inductive bias of traditional CNNs and makes it possible to learn stronger and more accurate patterns with large-scale parameters from enormous data, such as ViTs.

They focus on creating a CNN-based foundation model in this work so that it can effectively handle large number of parameters and data.

They begin with Flexible Convolution's variant Deformable Convolution (DCN). We use it in conjunction with numerous specific transformer-like block- and architecture-level designs to create the InternImage convolutional backbone network. InternImage's primary algorithm is a dynamic sparse convolution, whose sampling offsets may be adjusted to dynamically select the appropriate receptive fields (which could be long- or short-range) based on the input data. The conventional 3*3 convolution window also avoids the significant expenses and optimization problems associated with employing massive dense kernels.

This paper demonstrates that a worthwhile area of study for large-scale model research is convolutional models. In order to expand CNNs to large-scale settings, they additionally examine the customised basic block, stacking rules, and scaling strategies based on the operator. They also introduce long-range dependencies and adaptive spatial aggregation using an upgraded 3*3 DCN operator. These methods give the models access to large-scale parameters and data by effectively using the operator [10].

3 Result

Cityscapes is a widely used large-scale dataset for understanding urban scenes, aimed at promoting research and algorithm development in the field of computer vision [11]. This dataset contains high-resolution images from different cities, covering various scenes such as streets, vehicles, pedestrians, buildings, etc.

The majority of tasks involving the interpretation of urban scenes, such as semantic segmentation, instance segmentation, object detection, and road scene understanding, make use of the Cityscapes dataset. It offers a substantial amount of labeled data, including instance level annotations, accurate bounding box annotations, and pixel level semantic labels, which may be used to train and test different computer vision algorithms.

The goal of the Cityscapes dataset is to promote the development of urban scene understanding algorithms and provide a standard benchmark testing platform for researchers and developers. It has become one of the important datasets for evaluating and comparing many computer vision algorithms in urban scene understanding tasks.

Mean Intersection over Union (mIoU) is one of the indicators used to evaluate the performance of image semantic segmentation tasks. In image semantic segmentation, we hope to assign each pixel in the image to different categories, such as people, vehicles, roads, etc. MIoU is an indicator used to quantify the similarity between predicted results and real labels. All the above methods have been validated on Cityscapes and have a high mIoU index as demonstrated in Table 1.

Table 1. Result comparison.

Method	mloU
DCNAS	83.6%
HRNetV2 + OCR	83.7%
Panoptic-DeepLab	84.2%
EfficientPS	84.2%
ViT-Adapter-L	85.2%
InverseForm	85.6%
HS3-Fuse	85.8%
InternImage-H	86.1%

4 Conclusion

Semantic segmentation methods based on traditional machine learning methods are difficult to apply to real-time driving systems due to poor segmentation performance, low efficiency, and long segmentation time. Through research, this article found that with the rise of deep learning, semantic segmentation methods have advanced very quickly and have mature applications in many aspects. These models have made up for the problems existing in traditional machine learning solutions. The models summarized above have been validated and achieved good results, with their mIoU index above 83%. However, even so, there are still many technical difficulties that need to be overcome, such as the limitations of CNN, the limitations of annotated data, and the generalization ability of the model. Therefore, the semantic segmentation method based on deep learning is still a popular research direction with great potential for development and deserves more people to participate in the research.

References

1. Li, B., Shi, Y., Qi, Z., & Chen, Z. A survey on semantic segmentation. In 2018 IEEE International Conference on Data Mining Workshops, 1233-1240 (2018).
2. Yu, H., Yang, Z., Tan, L., Wang, Y., Sun, W., Sun, M., & Tang, Y. Methods and datasets on semantic segmentation: A review. Neurocomputing, 304, 82-103 (2018).
3. Zhang, X., Xu, H., Mo, H., Tan, J., Yang, C., Wang, L., & Ren, W. Dcnas: Densely connected neural architecture search for semantic image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 13956-13967 (2021).
4. Yuan, Y., Chen, X., & Wang, J. Object-contextual representations for semantic segmentation. In European Conference on Computer Vision, 16, 173-190 (2020).
5. Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H., & Chen, L. C. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 12475-12485 (2020).
6. Mohan, R., & Valada, A. Efficientps: Efficient panoptic segmentation. International Journal of Computer Vision, 129(5), 1551-1579 (2021).
7. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., & Qiao, Y. Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 (2022).
8. Borse, S., Wang, Y., Zhang, Y., & Porikli, F. Inverseform: A loss function for structured boundary-aware segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5901-5911 (2021).
9. Borse, S., Cai, H., Zhang, Y., & Porikli, F. Hs3: Learning with proper task complexity in hierarchically supervised semantic segmentation. arXiv preprint arXiv:2111.02333 (2021).
10. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14408-14419 (2023).

11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., et al. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3213-3223 (2016).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.