Feature Fusion in Deep-Learning Semantic Image Segmentation: A Survey

Jie Yuan, Zhaoyi Shi(✉), and Shuo Chen

Minzu University of China, Beijing 100000, China
wzzhaoyi@outlook.com

Abstract. Semantic image segmentation is a necessary research and application direction for intelligent systems. Many researchers have tried to design advanced feature fusion to extract beneficial information from different feature maps selectively. However, there is no published review currently that focuses on feature fusion for semantic image segmentation. Therefore, we seek to compile related works and analyze the trends and challenges of feature fusion. In this paper, we introduce feature fusion modules based on different semantic image segmentation models. Then, we analyze typical and state-of-the-art approaches in terms of several effective from fusion. Third, we comprehensively present fusion strategies. Finally, we summarize the challenges as well as the development trend of feature fusion. This survey infers that although significant developments have been obtained, there is still plenty of room for improvement of feature fusion. Interpretability in deep-learning segmentation and the application of novel mechanisms have been important directions for future exploration.

Keywords: Feature fusion · Deep learning · Semantic segmentation

1 Introduction

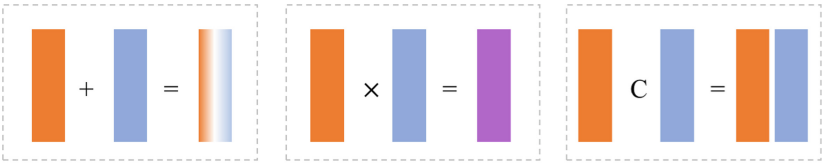
The main objects of intelligent system research are mathematical models with uncertainty, high nonlinearity, and complex scenes [1]. The neural network is an essential subfield of the intelligent system. Neural network control systems have better intelligence and robustness and can handle high-dimensional, nonlinear, and strongly coupled control problems. Computer vision is one of the most important research directions.

Semantic Image segmentation is a hot branch of computer vision. It tries to understand the class of each pixel in an image semantically (e.g., identifying whether the target is a bicycle, a motorcycle, or some other type). In recent years, semantic segmentation based on deep convolutional neural networks has gained tremendous attention and development, playing an essential role in several visual understanding systems, such as autonomous driving [2].

Accurate classification and fine-grained boundaries rely on both semantic and detailed information [3]. However, it is difficult to obtain them on the same feature map. The shallow layer of the network retains more details, while the deep layer mainly

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
S. Paiva et al. (Eds.): SmartCity360° 2021, LNICST 442, pp. 284–292, 2022.
https://doi.org/10.1007/978-3-031-06371-8_18

extracts contextual information. Therefore, much work has focused on feature fusion modules for multi-scale features. A part of the work extracts multi-scale features from different levels of the backbone and then fuses them, such as FCN [4], PSPNet [5], Deeplab [6], U-Net [17]. Some work introduces parallel branching to preserve multi-scale features simultaneously during inference and merge them finally, like ICNet [8], BiSeNet [9], Fast-SCNN [10], STDC [11]. Other work applies complex mechanisms to carefully filter and adjust the weight of multi-scale features [12]. For example, Segmenter [13] introduces a transformer-based model for semantic segmentation.



(a) elementwise add (b) elementwise multiply (c) concatenation

Fig. 1. Basic fusion operation

This paper reviews the feature fusion for semantic segmentation. Section 2 details typical feature fusion methods for different models and shows common defects and improvements. In Sect. 3, we analyze various fusion mechanisms with different architectures. Section 4 provides a comprehensive account of the feature fusion strategy. Section 5 analyzes the challenges and trends based on the above. Finally, the whole paper is concluded.

2 Feature Fusion in Various Models

In at least seven years, thousands of models have emerged in the field of semantic segmentation. They can be classified into several types according to contributions and structures. One thing is common, a suitable mechanism to merge high-level features with low-level high-resolution feature maps.

FCN [4] first introduces a fully convolutional network for semantic image segmentation for inputting an arbitrary size image and the output of the corresponding resolution segmentation map. The model applies to skip connections to fuse feature maps from shallow layers and uses elementwise add to merge them. It is not fast enough for the real-time task and cannot easily convert to 3D images. For feature fusion, the work does not sufficiently consider global contextual information. ParseNet [14] extracts additional global context by adding global average pooling for the feature map to overcome local confusion and smooth segmentation.

The encoder-decoder structure continuously reduces the feature map size. It increases the number of channels in the first half of the model while upsampling the feature map in the second half to achieve encoding and resolution recovery [15]. Initially, the encoder-decoder architecture serves the primary purpose of image compression and noise removal. The input is a picture encoded by downsampling to obtain a string of

features smaller than the original image, corresponding to squeeze, and then a decoder that ideally restores the image to its original state.

SegNet [16] adds a series of novel shortcuts from the encoder to the decoder shown in Fig. 2. The decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform nonlinear upsampling. This eliminates the need for learning to upsample.

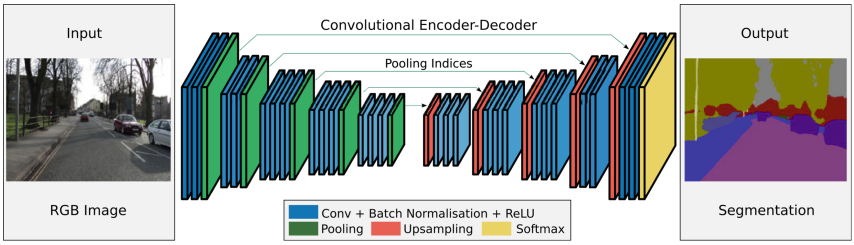


Fig. 2. A Typical Model with Encoder-Decoder Structure from [16]

U-Net [17] achieved excellent performance in medical image segmentation. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. The fusion method applies cropping, concatenation, and convolution in turn to achieve size consistency. The complexity of medical images is very high compared to ordinary photos, with an extensive greyscale range and unclear boundaries. U-Net combines low-resolution information in downsampling to provide a basis for object class recognition and high-resolution information in upsampling to provide a basis for accurate segmentation. It also fills in the underlying data with the skip connection to improve segmentation accuracy.

As medical images are relatively difficult to acquire and the amount of data is small, it is easy to over-fit the model if it has too many parameters. In contrast, the U-Net model has fewer parameters, making it more suitable than FCN.

Feature Pyramid Network (FPN) [18] is one of the most notable works using multi-scale analysis for object detection and then applied to segmentation. The model uses a pyramid parsing module to harvest different sub-region representations, followed by upsampling and concatenation layers (or other different feature fusing methods) to form the final feature representation [19, 20].

Multiple downsampling or upsampling may lead to the loss of high-resolution representations. Some models contain several branches due to a lack of high-resolution representations. HRNet [21] connect the multi-resolution streams in parallel. The connection fusion unit includes stridden convolution or bilinear upsampling followed by 1 $\times$ 1convolution, and elementwise add to merge feature maps finally in Fig. 3.

Li *et al.* [22] study a dynamic routing method to alleviate the scale variance in semantic representation. After the model is trained, the activation of fusion connections is data-dependent, adapting to the scale distribution of each image.

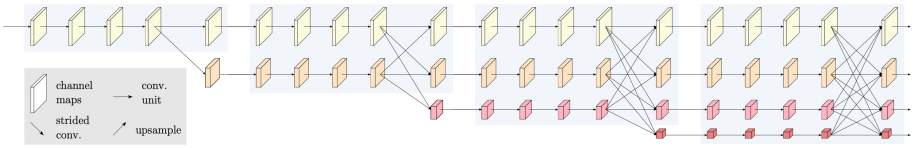


Fig. 3. An example of a high-resolution network [21] with parallel branches

3 Fusion Mechanisms

In this section, we classify feature fusion according to their effectiveness, architecture, and the methods they introduce and present some typical representative work. Table 1 provides the comparison with several representative methods on the Cityscapes test set in terms of frames per second and mIoU class.

3.1 Direct Fusion

For different scales, the direct fusion is to merge the feature maps after unifying the dimensions—no additional connections except for final fusion. Typical merge operations include elementwise add, elementwise multiplication, and concatenation. Channel consistency mainly uses convolution operations, such as 1×1 convolution.

FastSCNN [10] uses bilinear interpolation and convolution to resample feature maps. After that, elementwise add is applied to merge. Although a simple structure is not good enough for a high-accurate segmentation map, it receives a low cost of computing resources.

FasterSeg [23] present an automatically designed semantic segmentation network discovered from a multi-branch search space. The head module aggregates these outputs by direct fusion.

3.2 Multi-level Fusion

Most semantic segmentation backbones output multi-scale feature maps. Some work contains multi-level fusion, where feature maps are merged two by two according to an order. Detailed information can be progressively introduced into deeper features level by level, enabling more delicate boundaries.

FCN [4] chooses three feature maps with different levels and simply merges them sequentially from shallow layer to deep layer by elementwise add. In this way, it generates a segmentation map with the same dimension as the input image.

PSPNet [18] uses the pyramid pooling module to gather context information providing an effective global contextual prior for pixel-level scene parsing. The pyramid pooling module can collect levels of information more representative than global pooling [14].

3.3 Multi-layer Mesh Fusion

Some work tries a more flexible direction of multi-scale feature fusion. The connection between multi-scale branches contains high-level to low-resolution directions and transports high-level features to low-level feature maps.

FRRN [24] employs a two-stream system, where full-resolution information is carried in one stream and context information in the other pooling stream. Full-resolution residual units have both context and detail information through the network with a bidirectional information flow. This results in a network that successively combines and computes features at two resolutions.

DCNAS [25] propose a novel neural architecture search framework. The search space contains cross-level connections. The fusion module can aggregate semantic features from preceding fusion modules and attach transformed semantic features to succeeding ones.

3.4 Weighted Fusion

Different levels of feature maps have different contributions to the generation of fine segmentation maps. Therefore, finding the fusion weights between features is a key direction for improving the semantic image segmentation accuracy. If A and B are the feature maps to be merged, direct fusion and weighted fusion can be following:

$$Direct\ fusion:\ F_d(A,\ B)\ =\ A\ +\ B \tag{1}$$

$$Weighted\ fusion:\ F_w(A,\ B)\ =\ \alpha * A\ +\ \beta * B \tag{2}$$

where α, β represent the weights learned from the feature maps to be fused.

BiSeNet [9] proposes a specific feature fusion module to fuse different levels of the features. It first concatenates the multi-level features of two branches' multi-level features, pools the concatenated feature to a feature vector, and computes a weight vector, like SENet [26]. This weight vector can re-weight the features, which amounts to feature selection and combination by multiplication. With attention mechanism, BPNet [27] introduces a context aggregation module to filter information that learns pixel-wise unary attention to emphasize small patterns and pairwise attention for long-range information dependency modeling.

DeepLabv3 [28] propose to augment ASPP with image-level features, similar to [5][14], to incorporate global context information properly. Li *et al*. [12] infer that simply directly combining multi-level features suffers from the semantic gap. And they propose Gated Fully Fusion (GFF) to fuse features selectively.

3.5 Graphical Models

Several methods introduce probabilistic graphical models for more accurate context, such as Conditional Random Fields (CRFs) and Markov Random Field (MRFs).

Chen *et al*. [29] show that responses at the final layer of DCNNs are not sufficiently localized for accurate object segmentation. Due to the very invariance properties, DCNNs are good at high-level tasks. This work combines the responses at the final DCNN layer with CRFs. Liu *et al*. [30] address semantic image segmentation by incorporating rich information into MRFs, including high-order relations and a mixture of label contexts.

Table 1. Performance of segmentation models on the Cityscapes test set

Model	Fusion mechanism	mIoU	FPS
DeepLab-MSc-CRF [29]	Graphical model	61.6	4.9
DPN [30]		66.8	-
Fast-SCNN [10]	Direct fusion	68.0	123.5
FasterSeg [23]		71.5	163.9
FCN [4]	Multi-level fusion	65.3	2.0
PSPNet [5]		78.4	-
FRRN [24]	Multi-layer mesh fusion	71.8	2.1
HRNetV2 [21]		81.6	-
DCNAS [25]		83.6	-
STDC2-50 [11]	Weighted fusion	71.9	**250.4**
BiSeNet-ResNet-18 [9]		74.7	65.5
DeepLabV3 [28]		81.3	-
GFF [12]		82.3	-
HRNet-OCR [33]		**85.1**	-

4 Fusion Strategy

The final stage of feature fusion generally uses one of three essential fusion operations, including element-wise addition, element-wise multiplication, and concatenation shown in Fig. 1. The addition has low computational complexity and is easy to compute. Multiplication increases the training difficulty [28]. Concatenation does not require a consistent number of channels and is more flexible. But it requires post-convolution, such as 1×1 convolution, to merge and filter redundant channels.

Nie *et al*. [27] find that either add or multiplication is not sufficient for feature fusion. They propose a feature fusion block that first adds and multiplies the characteristic graphs separately and then brings these two signals together.

For feature forms, methods usually pre-process the original image and detach the pixel space of the image with poor correlation. Traditional approaches tend to transform an image into a feature vector, while convolutional neural network extract information based on the feature map.

Transformer models have revolutionized Neuro-Linguistic Programming [31]. Recently, there has been some novel work for the usage of transformer structures in semantic image segmentation [32]. They formulate the problem of semantic segmentation as a sequence-to-sequence problem and use a transformer architecture. This work split the image into patches and treated linear patch embedding as input tokens for the transformer encoder, translating the image map to the sequence.

5 Challenges and Opportunities

Learning-based feature fusion modules have achieved excellent performance. However, research needs to investigate their underlying mechanisms further. For example, can the module implement the current functionality in a more compact structure? Can the contribution of multi-scale features be explained in an exact and easy-to-understand form? The interpretability of deep neural networks can help people achieve more efficient designs, approaching truly intelligent systems. A compelling fusion structure can create a more effective flow of semantic information and benefit gradient propagation.

For high availability, the feature fusion module needs low computational complexity, high memory efficiency in some tasks, such as autopilot. The application of neural architecture search helps researchers automatically generate application-competitive modules with the balance of precision and real-time. The boom in the mobile internet has created a massive demand for security, traffic scene awareness, and the need for portable models. Moreover, the design of models with higher energy efficiency contributes to maintaining a low-carbon smart city.

6 Conclusion

Feature fusion as a filtering and merging module for context and detail information plays a massive role in segmentation. Applying the model to specific scenarios and making it more efficient has been a key topic. In this paper, we provide a detailed overview of the feature fusion modules. We introduce representative fusion mechanisms and the essential fusion strategies. Finally, we infer the current challenges and future trends. We hope to provide readers with some helpful, modest guidance.

References

1. Antsaklis, P.J., Passino, K.M.: An Introduction to Intelligent and Autonomous Control. Kluwer Academic Publishers (1993)
2. Forsyth, D., Ponce, J.: Computer Vision: A Modern Approach. Prentice hall (2011)
3. Minaee, S., Boykov, Y.Y., Porikli, F., et al.: Image segmentation using deep learning: A survey. IEEE Trans. Pattern Anal. Mach. Intell. (2021)
4. Long, J., Shelhamer, E., Darrell, T.:: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
5. Zhao, H., Shi, J., Qi, X., et al.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
6. Chen, L.C., Papandreou, G., Kokkinos, I., et al.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
7. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

8. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: ICNet for real-time semantic segmentation on high-resolution images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 418–434. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_25
9. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: BiSeNet: bilateral segmentation network for real-time semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11217, pp. 334–349. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_20
10. Poudel, R.P.K., Liwicki, S., Cipolla, R.: Fast-SCNN: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502 (2019)
11. Fan, M., Lai, S., Huang, J., et al.: Rethinking BiSeNet for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9716–9725 (2021)
12. Li, X., Zhao, H., Han, L., et al.: Gated fully fusion for semantic segmentation. Proc. AAAI Conf. Artif. Intell. **34**(07), 11418–11425 (2020)
13. Strudel, R., Garcia, R., Laptev, I., et al.: Segmenter: Transformer for Semantic Segmentation. arXiv preprint arXiv:2105.05633 (2021)
14. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
15. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. IN: Proceedings of the IEEE International Conference on Computer Vision, pp.1520–1528 (2015)
16. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241. Springer, Cham (2015)
18. Lin, T.Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. IN: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
19. Zhao H, Shi J, Qi X, et al.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
20. He, J., Deng, Z., Zhou, L., Wang, Y., Qiao, Y.: Adaptive pyramid context network for semantic segmentation. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 7519–7528 (2019)
21. Wang, J., Sun, K., Cheng, T., et al.: Deep high-resolution representation learning for visual recognition. In: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, (2020)
22. Li, Y., Song, L., Chen, Y., et al.: Learning dynamic routing for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8553–8562 (2020)
23. Chen, W., Gong, X., Liu, X., et al.: FasterSeg: Searching for faster real-time semantic segmentation. In: Proceedings of the International Conference on Learning Representations, (2019)
24. Pohlen, T., Hermans, A., Mathias, M., et al.: Full-resolution residual networks for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4151–4160 (2017)
25. Zhang, X., Xu, H., Mo, H., et al.: Dcnas: Densely connected neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13956–13967 (2021)

26. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
27. Nie, D., Xue, J., Ren, X.: Bidirectional pyramid networks for semantic segmentation. In: Proceedings of the Asian Conference on Computer Vision (2020)
28. Chen, L.C., Papandreou, G., Schroff, F., et al.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
29. Liang-Chieh, C., Papandreou, G., Kokkinos, I., et al.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: Proceedings of the International Conference on Learning Representations (2015)
30. Liu, Z., Li, X., Luo, P., et al.: Semantic image segmentation via deep parsing network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1377–1385 (2015)
31. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-XL: Attentive language models beyond a fixed-length context. In: Proceedings of the ACL (2019)
32. Zheng, S., Lu, J., Zhao, H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
33. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020)