Urban Street View Semantic Segmentation Based on Multi-scale Network and Ghost Attention Head

Fei Zetao(23020241154391)[1*], Xu Wenjing(23020241154463)[2*], Zhang Xiaoran(23020241154474)[3*]

[*] Informatics Class

Abstract

Semantic segmentation models currently struggle with urban street view image segmentation due to an underutilization of multi-scale feature information and the excessive computational costs associated with generating redundant information, which significantly hinders the performance of segmentation models. Furthermore, considering the innate spatial particularities of urban street view data, it is imperative for segmentation models to place a greater emphasis on spatial information. To address this, we integrate multi-scale convolutions and Ghost Attention Heads into the segmentation framework. Multi-scale convolutions will be incorporated into the encoder to expand the network's receptive field, thereby capturing contextual information more effectively. A Ghost Attention Head is designed during the decoding process. This module, by employing efficient and cost-effective operations along with a separable attention mechanism, guides the network to focus more on spatial information. This approach not only significantly reduces computational expenses but also addresses the issue of neglecting spatial information. Experimental results on CamVid show a 9.7% mIoU improvement (71.32%) over FCN, with ablation studies validating the effectiveness of the module.

Introduction

Semantic segmentation is a pivotal domain within computer vision that entails pixel-level prediction, assigning each pixel in an image to its corresponding class or category. Urban scene segmentation, as a foundational yet challenging task within semantic segmentation, aims to dissect and deeply analyze scene objects into distinct regions associated with semantic category information. In recent years, algorithms for urban street view segmentation have been widely applied to autonomous driving tasks in the field of computer vision, achieving significant outcomes.

Image segmentation algorithms are broadly categorized into conventional and deep learning-based approaches. Traditional methods, rooted in mathematical principles, rely on features like texture and shape for prediction. Key techniques include threshold-based and edge detection methods. Threshold segmentation leverages pixel color or grayscale values, using specific thresholds to classify pixels (Kapur, Sahoo, and Wong 1985), but struggles in scenarios with similar or intertwined grayscale levels, such as in autonomous driving. Edge detection identifies boundaries to segment images (Canny 1986), performing well with distinct grayscale variations and low noise but showing limitations in more complex settings.

The advent of deep learning has revolutionized image segmentation, driven by the powerful feature representation of Convolutional Neural Networks (CNNs). The Fully Convolutional Network (FCN) (Long, Shelhamer, and Darrell 2015), a seminal approach, replaces fully connected layers with convolutional ones and employs transposed convolution to restore feature map size, enabling direct segmentation predictions. However, FCN lacks spatial consistency and fails to fully leverage pixel relationships, leading to coarse results. U-Net (Ronneberger, Fischer, and Brox 2015) addresses these limitations with a U-shaped architecture, using upsampling instead of pooling and introducing skip connections between encoding and decoding layers to recover and reuse features.

Semantic segmentation network architectures are typically classified into encoder-decoder structures and multi-scale feature-based networks. Encoder-decoder models, such as U-Net and SegNet (Badrinarayanan, Kendall, and Cipolla 2017), use a backbone for feature extraction and a decoder for segmentation, but often fail to fully utilize multi-scale features from the backbone, limiting contextual integration and performance. Multi-scale networks enhance contextual understanding by processing features of varying scales through convolutions and pooling. However, they often overlook the varying semantic importance of features across scales, amplifying noise and introducing redundancy. Additionally, many networks upscale images using deconvolution or interpolation without addressing noise in low-level features or the varying importance of cross-layer features, which impacts detail recovery. Although PSPNet (Zhao et al. 2017) employs pyramid pooling to capture multi-scale information, it neglects spatial locality, further constraining segmentation accuracy.

In summary, the contributions of this paper are as follows:

- **Multi-Scale Dilated Convolution:** Integrated into the backbone, it expands the receptive field for better contextual feature extraction and captures multi-level semantic information.

- **Ghost Attention Head Decoder:** Reduces computational cost by minimizing redundancy and, with separable attention, enhances spatial focus for refined segmentation.

Related Works

Multi-scale

To capture multi-scale contextual information, PSPNet (Zhao et al. 2017) employs average pooling of varying sizes to fuse features at different scales, effectively extracting context. However, successive pooling and downsampling significantly reduce image resolution. DeepLabv2 (Chen et al. 2017) addresses this by using dilated convolution to expand the receptive field without increasing parameters, integrating richer context. Its ASPP module employs parallel dilated convolutions with different rates to capture multi-scale information, improving robustness in multi-class segmentation and mitigating resolution loss. However, its high computational cost and slow inference limit its suitability for real-time applications. Inception networks (Szegedy et al. 2015; Ioffe 2015; Szegedy et al. 2017) leverage parallel branches with varied convolutional kernels to aggregate features but at the expense of increased complexity. DFANet (Li et al. 2019), in contrast, introduces a multi-tiered connectivity structure to optimize receptive fields across scales while maintaining efficiency.

BiSeNetV2 (Yu et al. 2021) employs a dual-channel architecture for efficient real-time semantic segmentation. Its detail branch encodes rich spatial information, while the semantic branch uses rapid downsampling to enhance feature representation and expand the receptive field. A bilateral guidance aggregation layer fuses these complementary features, achieving a balance between speed and accuracy. HyperSeg (Nirkin, Wolf, and Hassner 2021), on the other hand, adopts a nested U-shaped architecture to capture multi-scale semantic information. However, both approaches are optimized for low-resolution street view images, with significant inference speed degradation when processing high-resolution inputs.

Spatial Information

In terms of spatial information, standard convolutional operations do not explicitly consider spatial information and interactions, focusing primarily on feature extraction. In contrast, the SENet (Hu, Shen, and Sun 2018)introduces dynamic weighting to address this limitation. It first compresses global spatial information and learns feature importance across channels. The excitation component then dynamically allocates weights, enabling enhanced spatial interactions and more effective feature representation.

The Convolutional Block Attention Module (CBAM) (Woo et al. 2018) refines feature maps using separate channel and spatial attention modules. To capture pixel relationships, a self-attention mechanism (Vaswani 2017) is introduced, enabling the extraction of contextual information between pixels. This mechanism models spatial dependencies by calculating the relevance between a given Query and different Keys, and then computing weight coefficients for the corresponding Values. The attention Value is derived as a weighted average based on these coefficients.

Dual Attention Network (DANet) (Fu et al. 2019) employs two parallel attention modules to capture spatial and channel dependencies, thereby achieving interrelation. The spatial attention module obtains feature information at each location through spatial feature weighting, while the channel attention module correlates relevant features on the channel dimension. Eventually, the outputs of these two modules are fused to obtain enhanced feature representations.

Method

The overview of network structure

Figure 1 shows the overall structure of the network, and the overall network in this paper is the structure of Encoder-Decoder. The specific descriptive information of each part of this network is as follows:

1) The network employs ResNet-50 (He et al. 2016) as the feature extractor, with its architecture serving as the encoder. As input passes through each BasicBlock, a feature map is produced with a spatial resolution that is reduced by a factor of 2. After traversing a total of 4 BasicBlocks, the final feature map, with a resolution of 1/16 of the original input size, is obtained as the output of the backbone network.

2) To capture richer contextual information and facilitate the fusion of features across varying receptive fields and subregions for improved feature representation, Zhao et al. introduced a pyramid pooling module (PPM) in PSPNet (Zhao et al. 2017). In contrast, this paper proposes a novel multi-scale convolution approach, which is directly integrated into the ResNet-50 backbone. This method inserts three sets of multi-scale convolutions sequentially between different layers of ResNet-50. The proposed design enables the encoder to effectively fuse and extract contextual information during the feature extraction process, thereby enhancing the overall feature representation.

3) In the head of the network, to leverage the feature map for extracting additional information with minimal computational overhead, we incorporate the Ghost module (Han et al. 2020) within Seghead. This approach utilizes a more efficient method than traditional convolutions to generate redundant features. By combining a "small amount of traditional convolutional computation" with a "lightweight redundant feature generator" ,the overall computational burden of the network is reduced while maintaining its predictive accuracy.

4) When utilizing the semantic information from the encoder, it is crucial to account for the presence of noise in the features and the variability of the information across different feature maps. To focus more spatially on relevant feature information in specific regions of the extracted feature map, while minimizing the influence of noise, a separable spatial attention mechanism is employed in Seghead. The feature maps are fed into the attention module, which consists of three deep separable convolutions, each with a kernel size of 3 and a stride of 2, to generate the spatial attention map. This map encodes spatial locations that warrant greater attention and is then resized to the original dimensions via lin-

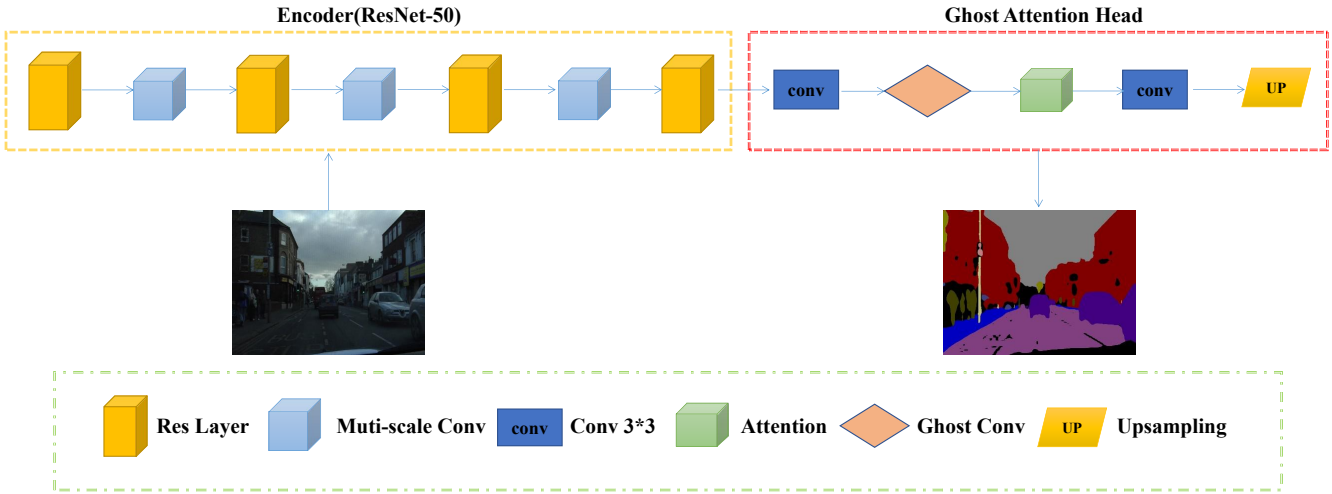


Figure 1: The overview of our network architecture.

ear interpolation upsampling. The final output is obtained by applying a Sigmoid activation to the attention map.

5) Given the integration of the Ghost module and the Spatial Attention module in the head of the network, this component is collectively referred to as the **Ghost Attention Head**.

Multi-scale convolution operator

Since the ResNet-50 backbone lacks the capability to capture the relationships between different categories in the global scene when addressing image segmentation tasks, this paper aims to improve the feature extraction performance of ResNet-50 by incorporating multi-scale image information and leveraging features of varying dimensions. These enriched features are then utilized for the segmentation task. To achieve this, we propose a multi-scale atrous convolution operator, which is integrated into different layers of the ResNet-50 backbone. This addition enables the network to effectively capture multi-scale feature information, thereby enhancing the model's expressive power.

The three multi-scale convolution operators, denoted as [MsConv2, MsConv3, MsConv4], are positioned prior to layer2, layer3, and layer4, respectively. The number of grouped convolutions for each operator is set to [2, 3, 4], with the sequence of group sizes being [1, 4, 8, 16]. For each operator, the input channel size of the convolution kernels remains constant, while the output channels are reduced to 1/N of the original number of convolutions, where N represents the number of convolutions in the respective operator. The final output is obtained by concatenating the results from all N convolutions. This design ensures the extraction of richer contextual features and more effective fusion of information, while simultaneously limiting the increase in the number of parameters, thereby improving network accuracy.

Ghost module

The ResNet-50 backbone generates many similar feature map pairs during the feature extraction process on the input image. Redundancy in feature maps has been identified as a key factor contributing to the success of deep neural networks [13], and as such, we choose to leverage, rather than eliminate, these redundant feature maps. Building on this, this paper incorporates the Ghost module into Seghead, based on the ResNet-50 backbone. The Ghost module employs a series of linear transformations and cheap operations to generate multiple similar feature maps, which can be extracted from the original feature maps at minimal computational cost. This approach reduces the model's computational complexity while preserving its ability to maintain high performance in terms of network similarity. The specific architecture is illustrated in Figure 2.

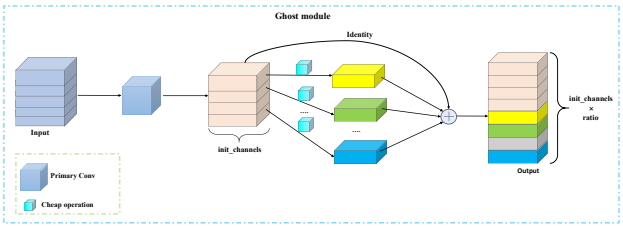


Figure 2: The details of Ghost module.

The input feature map is first passed through a 1×1 standard convolution, followed by feature extraction via a depthwise separable convolution with a 3×3 kernel. The depthwise separable convolution is an efficient linear operation. In this paper, the number of output channels is set to $init\_channels \times ratio$, where the standard convolution generates $init_channels$ feature maps. Each of these feature maps is then processed through a linear operation, producing $init\_channels \times (ratio - 1)$ feature maps. Finally, the feature maps obtained from the linear operation are concatenated with those generated by the standard convolution, re-

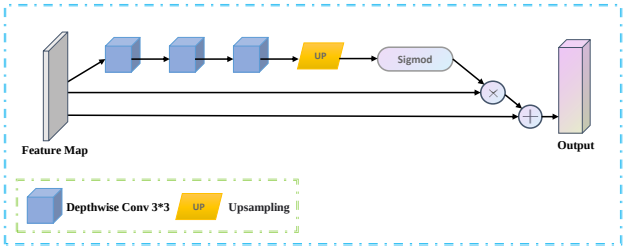


Figure 3: The details of attention module.

sulting in the final output. This output represents the final result of the Ghost module.

Spatial attention module

When viewing an image, a person does not perceive the entire picture at once, but rather focuses on key or focal regions. The same principle applies to semantic segmentation. The backbone network, however, does not account for the varying importance of different spatial locations during feature extraction, and the multiple convolutional operations lead to a loss of spatial information. To address this, we introduce a spatial attention module in Seghead, as shown in Figure 3, to generate a spatial weight matrix. This enables the network to focus more effectively on specific regions of the image, thereby refining the feature representations. The implementation steps are as follows:

The input feature map $P \in \mathbb{R}^{C\times H\times W}$ is downsampled using three 3×3 depthwise separable convolution operations to extract features, yielding the feature map $D(P)$. This process is computationally more efficient than standard convolution. Subsequently, a linear interpolation upsampling operation restores the feature map to its original spatial dimensions $H \times W$, which is then processed through a Sigmoid function to obtain the final spatial attention map $S(A) \in \mathbb{R}^{H\times W}$. This is illustrated in Formula 1.

$$S(A) = \sigma(Upsample(D(P))) \tag{1}$$

where $\sigma()$ represents the Sigmoid operation, and UpSample denotes the upsampling operation. The spatial attention map $S(A) \in \mathbb{R}^{H\times W}$ is then element-wise multiplied with the original input feature map $P \in \mathbb{R}^{C\times H\times W}$ to produce the spatially reweighted feature map $Z \in \mathbb{R}^{C\times H\times W}$, as described in Formula 2.

$$Z = S(A) \times P \tag{2}$$

Finally, the spatially reweighted features $Z \in \mathbb{R}^{C\times H\times W}$ are fused with the original input feature map $P \in \mathbb{R}^{C\times H\times W}$ to produce the module's final output L, as defined in Formula 3.

$$L = Z + P \tag{3}$$

Experiment

Experimental design

Dataset: The proposed network model was evaluated using CamVid (Brostow, Fauqueur, and Cipolla 2009), a commonly used image dataset in the field of segmentation. Where the CamVid dataset provides 369 images for training, 100 images for validation and 232 images for testing, there are a total of 11 classes of objects (class = 12) in this dataset, where 0 is the background. **Evaluation metrics:** Using the mean intersection over union (mIoU), the mean accuracy (mAcc) for all categories. The formula of $mIoU$ is shown in Formula 4.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij} + \sum_{j=0}^{k}p_{ji} - p_{ii}} \tag{4}$$

where k represents the total number of categories in the dataset, i denotes the true class, j denotes the predicted class, and p_{ij} represents the number of pixels originally belonging to class i but predicted as class j. The metric $mloU$ indicates the calculated result. The formula for $mAcc$ is shown in Formula 5.

$$mAcc = \frac{\sum_{i=1}^{k} CA_i}{k} \tag{5}$$

where k denotes the number of categories in the dataset, CA_i denotes the classification accuracy of the i th category, and mAcc denotes the calculated result.

Network parameter settings: Our experiments were performed on Pytorch with graphics card 3090. first we cropped the image size to (480, 480) and batch size was set to 16. the base learning rate was 0.01, the weight decay was 0.0001, the optimizer used stochastic gradient descent (SGD), and the epoch was 120 in trraining.

Analysis of experimental results

1)Performance Analysis of the Ghost Module: To evaluate the effectiveness of the Ghost module and its impact on enhancing the overall network performance, this study removes the spatial attention module from the Seghead and conducts comparative experiments on the CamVid dataset. Specifically, the experiments compare the performance of the ResNet-50 backbone network with and without the Ghost module integrated as the Seghead. The results, as presented in Table 1 , indicate that the introduction of the Ghost module leads to a 0.29% improvement in mIoU. This demonstrates that the lightweight module, which generates redundant features through computationally inexpensive operations such as linear transformations, not only enables higher accuracy but also achieves more precise segmentation compared to the original baseline.

Baseline	Ghost Module	mIou(%)	mAcc(%)	allAcc(%)
ResNet-50	×	68.59	78.20	89.53
ResNet-50	√	**68.88**	**78.59**	**89.60**

Table 1: Performance of Ghost module on Camvid-test.

2)Performance Analysis of the Ghost Attention Head: Building on the previous findings, this study introduces the spatial attention module to further enhance the model and conducts comparative experiments. Specifically, starting

from the baseline, the Ghost module is first incorporated as the Seghead, followed by the addition of the spatial attention module to form the complete Ghost Attention Head. The experimental results are summarized in Table 2. The baseline also employs a pre-trained ResNet-50 as the backbone network, and the following three schemes are designed to verify the effectiveness of Ghost Attention Head:

- (a) Utilizing only the ResNet-50 backbone network;
- (b) Extending (a) by integrating the Ghost module as the Seghead;
- (c) Extending (a) by integrating the Ghost Attention Head as the Seghead.

Table 2 uses the Ghost Attention Head scheme compared to baseline and baseline+Ghost Module both improve the segmentation performance to some extent, so it can be concluded that the addition of spatial attention module on top of Ghost Module to assign weights to spatial location information features has more accurate segmentation performance than the first two schemes, where compared to baseline, mIoU improves by 0.73% and mAcc improves by 0.77%, compared with adding Ghost Module mIoU improves by 0.44% and mAcc improves by 0.38%, thus verifying the effectiveness of Ghost Attention Head.

Baseline	Ghost Module	Ghost Attention Head	mIou(%)	mAcc(%)	allAcc(%)
ResNet-50	×	×	68.59	78.20	89.53
ResNet-50	√	×	68.88	78.59	89.60
ResNet-50	×	√	**69.32**	**78.97**	**89.69**

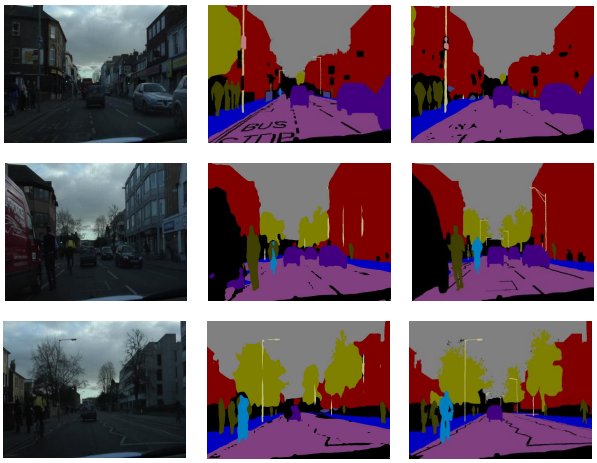
Table 2: Performance of Ghost Attention Head on Camvid-test.

3)Ablation experiments: In order to further verify the effectiveness of the proposed module, we conducted ablation experiments on this paper, as shown in Table 3. The experimental data show that the Ghost Attention Head and the multi-scale convolution operator proposed in this paper have different degrees of performance improvement compared with Baseline, which also proves the importance of incorporating Ghost Attention Head and multi-scale convolution into the model again.

Baseline	mIou(%)	mAcc(%)
baseline	68.59	78.20
baseline+GhostAttentionHead	69.32	78.97
baseline+GhostAttentionHead+Muti-conv	**71.32**	**79.23**

Table 3: The details performance of each component in our proposed.

4)Comparison experiments with other classical semantic models: In order to ensure the objectivity and universality of the views of this paper, our proposed model was compared with FCN, BiSeNetv2 and other algorithms on the CamVid dataset, and the experimental results are shown in Table 4. As can be seen from Table 4 the values of the models studied and proposed in this paper are 71.32% and 79.23% in mIoU and mAcc evaluation metrics, respectively, which is 9.7% and 5.58% improvement compared to FCN in mIoU and mAcc evaluation metrics, and 4.13% and 0.52% improvement compared to BiSeNetv2 in in mIoU and mAcc evaluation metrics, Compared with BisENetv1(Yu et al. 2018), DFANet(Li et al. 2019), EDANet(Lo et al. 2019) and DenseDecoder(Bilinski and Prisacariu 2018), the mIoU evaluation metrics have been improved to different degrees.



(a) image **(b) predict** **(c) label**

Figure 4: Fig. 4.The segmentation effect of this model on road streetscape.

Model	mIou(%)	mAcc(%)
FCN	61.62	73.65
BiSeNetv2	67.19	78.71
BiSeNetv1	68.70	-
DFANet	64.70	-
EDANet	66.40	-
DenseDecoder	70.90	-
Ours	**71.32**	**79.23**

Table 4: The details performance of each component in our proposed.

5)Model visualization on the CamVid dataset: In order to more intuitively show the segmentation accuracy of the model proposed in this paper, we also visualized the model on the CamVid dataset, and Figure 4 shows the prediction results of the model for some images in the test set. Where (a) is a random street view in Camvid, (b) is the prediction made by the model, and (c) is the label.

Conclusion

This paper presents a Multi-Scale Convolutional and Ghost Attention Head framework integrated into the ResNet-50 backbone. The multi-scale convolutional module enhances feature extraction by capturing contextual information at different scales. The Ghost Attention Head, with its Ghost and Separable Attention modules, improves segmentation performance by reducing computational costs and refining spatial features. Experimental results on the CamVid dataset confirm the effectiveness and efficiency of the proposed approach.

References

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.

Bilinski, P.; and Prisacariu, V. 2018. Dense decoder shortcut connections for single-pass semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6596–6605.

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.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. 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.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3146–3154.

Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; and Xu, C. 2020. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1580–1589.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Ioffe, S. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kapur, J. N.; Sahoo, P. K.; and Wong, A. K. 1985. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer vision, graphics, and image processing*, 29(3): 273–285.

Li, H.; Xiong, P.; Fan, H.; and Sun, J. 2019. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9522–9531.

Lo, S.-Y.; Hang, H.-M.; Chan, S.-W.; and Lin, J.-J. 2019. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, 1–6.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Nirkin, Y.; Wolf, L.; and Hassner, T. 2021. Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4061–4070.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; and Sang, N. 2021. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International journal of computer vision*, 129: 3051–3068.

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.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.