International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

Scale-Aware and Boundary-Enhanced Semantic Segmentation

Wei Zhou

[1]North China Electric Power University, No. 2, Beinong Road, Changping District, Beijing, China
Email: *983897564[at]qq.com*

Abstract: ***Semantic segmentation is widely used in autonomous driving and medical imaging, but accurately segmenting objects with different scales and preserving object boundaries remain challenging. This paper proposes a semantic segmentation network that combines a Multi-Scale Feature Fusion (MSFF) module with a Boundary Enhancement Module (BEM). The MSFF module captures contextual information using multiple receptive fields to improve multi-scale object recognition, while the BEM exploits shallow edge features and spatial attention to enhance boundary reconstruction. Experiments on the PASCAL VOC 2012 and Cityscapes datasets demonstrate that the proposed method achieves competitive segmentation performance and improves boundary accuracy compared with representative baseline methods.***

Keywords: Semantic Segmentation, Multi-Scale Feature Fusion, Boundary Enhancement, Spatial Attention, Deep Learning, Computer Vision**.**

1. Introduction

As a cornerstone task within the realm of computer vision, semantic segmentation focuses on assigning a categorical label to each individual pixel, enabling pixel-level scene understanding. Distinct from image classification and bounding-box-based object detection, this paradigm demands that models not only recognize target categories but also meticulously delineate their spatial coordinates and geometric boundaries. Consequently, it plays a pivotal role in practical deployments such as autonomous navigation, remote sensing analysis, clinical imaging diagnostics, and intelligent surveillance systems.

With the rapid development of deep learning, approaches predicated on Convolutional Neural Networks (CNNs) have witnessed remarkable breakthroughs [1]. Conventional architectures predominantly adopt an encoder-decoder framework to execute feature extraction and spatial restoration. Within this topology, the encoding phase is tasked with assimilating high-level semantic abstractions, whereas the decoding phase undertakes the reconstruction of the target's spatial layout. Pioneering models—including Fully Convolutional Networks (FCNs) [2], U-Net [3], and the DeepLab [4] series—have continuously augmented segmentation efficacy through the integration of skip connections, multi-scale context modeling, and dilated convolutions.

Nevertheless, contemporary methodologies continue to encounter two fundamental bottlenecks. First, regarding targets exhibiting extreme scale variations, relying on single-tier features proves inadequate for simultaneously balancing local details and global semantics. Although shallow feature maps preserve abundant spatial configurations, their semantic representation capacity remains restricted. Conversely, deep features demonstrate potent abstraction capabilities but inevitably incur the loss of edge-related nuances. Consequently, depending exclusively on single-scale representations frequently fails to yield robust segmentation outcomes. Second, existing architectures are prone to classification ambiguity within target boundary vicinities. Because consecutive convolutional downsampling intrinsically depletes spatial resolution, peripheral regions frequently suffer from incomplete object contours, insufficient detail recovery, and inter-class aliasing. These deficiencies are particularly pronounced when segmenting miniature objects or parsing highly complex scenes, thereby degrading the overarching model performance.

To address these challenges, this paper introduces a novel semantic segmentation network that seamlessly couples multi-scale feature fusion with boundary enhancement. From the perspective of macroscopic scale perception, a Multi-Scale Feature Fusion (MSFF) module is designed at the apex of the feature encoding stage. By parallelizing multi-level dilated convolutions and global pooling operations, this module dynamically captures contextual dependencies across heterogeneous receptive fields, substantially augmenting the network's resilience to target scale fluctuations. In terms of microscopic detail remodeling, a Boundary Enhancement Module (BEM) is incorporated into the decoding phase. By leveraging spatial attention mechanisms, the BEM precisely extracts high-frequency edge priors from shallow features. These extracted cues subsequently serve as guiding signals to execute fine-grained spatial compensation on deep features, improving segmentation accuracy near object boundaries.

The primary contributions of this work are summarized as follows:

- We propose a Multi-Scale Feature Fusion (MSFF) module that constructs a parallelized contextual capture mechanism encompassing diverse receptive fields. This effectively alleviates the representational homogeneity of deep features and significantly elevates the segmentation consistency for objects of disparate dimensions.
- We design a lightweight Boundary Enhancement Module (BEM) that utilizes shallow spatial features to activate high-frequency edge clues, achieving precise compensation and reconstruction of object contours at a negligible parameter cost, successfully ameliorating inter-class aliasing phenomena.

2. Related Work

2.1 CNN-Based Semantic Segmentation

Historically, semantic parsing heavily depended on hand-crafted descriptors and conventional machine learning paradigms, which struggled to extract robust semantics within intricate scenes. The advent of Fully Convolutional Networks (FCNs) [2] catalyzed a paradigm shift by replacing dense connected layers with convolutional operations to achieve end-to-end mappings from input images to dense predictive maps, thereby inaugurating the deep learning era for this domain. Subsequently, architectures dedicated to feature extraction and spatial restoration evolved exponentially: U-Net [3] incorporated skip connections to facilitate the preliminary cross-tier amalgamation of shallow spatial details and deep semantics; meanwhile, the DeepLab series [4] and PSPNet [5]leveraged Atrous Spatial Pyramid Pooling (ASPP) and Pyramid Pooling Modules (PPM), respectively, to drastically dilate the receptive field without compromising spatial resolution, profoundly elevating the parsing capacity for complex environments.

Despite these significant progress, traditional architectures exhibit rudimentary synergistic interactions between shallow and deep representations. Specifically, when confronted with severe object scale disparities and convoluted contours, these models frequently abstract away local spatial configurations, precipitating localized misclassifications and excessive edge smoothing. Consequently, the lossless and highly efficient utilization of multi-hierarchical features remains a pivotal focal point in contemporary research.

2.2 Multi-Scale Feature Fusion Mechanisms

The severe variance in object dimensions inherent to natural scenes necessitates robust multi-scale perceptual capabilities within segmentation models. Feature Pyramid Networks (FPN) [6] pioneered a top-down multi-scale aggregation paradigm, augmenting the model's detection and segmentation robustness across cross-scale targets. Building upon this foundation, subsequent methodologies evolved parallelized probing strategies featuring varying dilation rates [7], alongside dynamic feature recalibration techniques driven by attention mechanisms [8], empowering networks to adaptively zero in on critical scale-specific representations.

While extant fusion frameworks successfully amplify global semantic expression, they disproportionately prioritize the alignment and smoothing of high-level semantics, inadvertently neglecting the foundational role of low-level spatial structures in contour reconstruction. This semantic-heavy, structure-light fusion tendency results in the irreversible obliteration of details in miniature objects or slender structures. To counteract this, our research is dedicated to formulating a lightweight feature interaction mechanism that harmonizes macroscopic scale perception with microscopic detail fidelity.

2.3 Boundary-Enhanced Segmentation Methods

Image boundaries encapsulate highly discriminative geometric priors, and their delineation accuracy directly dictates the visual fidelity of the predictive outputs. To mitigate the edge blurring and inter-class spatial aliasing induced by successive network downsampling, recent investigations have pursued multi-dimensional strategies. A subset of the literature introduces auxiliary edge detection branches or synergistic loss formulations [9],[10], explicitly coercing the network to assimilate geometric morphologies and contour distributions. Alternative approaches resort to intricate fine-grained reconstruction networks or high-frequency information mining modules to execute secondary reinforcement on localized ambiguous zones [11].

3. Proposed Method

3.1 Overall Network Architecture

To address the deficiencies of existing methodologies in handling object scale variations and recovering boundary nuances, this paper proposes a novel semantic segmentation network that seamlessly integrates multi-scale feature fusion with boundary enhancement. The overarching architecture adheres to the classical encoder-decoder paradigm. Specifically, the network comprises three core components: a foundational feature extraction network (Encoder), a Multi-Scale Feature Fusion (MSFF) module, and a Boundary Enhancement Module (BEM).

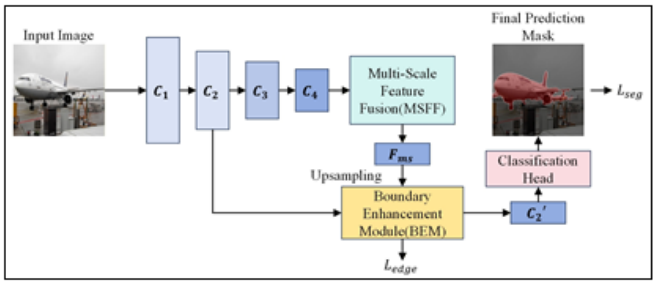

Figure 1: The overall architecture of the proposed semantic segmentation network.

During the encoding phase, given an input image $I \in \mathrm{R}^{H\times W\times 3}$, we employ a mainstream Convolutional Neural Network (e.g., the ResNet series [12]) as the backbone to extract hierarchical representations. As the network deepens, the spatial resolution of the feature maps progressively decreases while their semantic richness intensifies, sequentially yielding four stages of feature maps denoted as $\{C_1, C_2\}$, C_3, $C_4\}$. Among these, C_1 and C_2 function as shallow features, encapsulating abundant spatial details and local structural priors; conversely, C_3 and C_4 serve as deep features, possessing formidable global semantic representation capabilities.

In the decoding phase, the deep feature C_4 is initially fed into the MSFF module, which generates an aggregated representation highly robust to scale fluctuations by capturing contextual dependencies across parallel, heterogeneous receptive fields. Subsequently, during the progressive upsampling process, this aggregated feature is fused with the shallow feature C_1 that has been refined by the BEM. The BEM facilitates the meticulous reconstruction of spatial details by mining high-frequency edge cues inherently embedded within the shallow representations. Ultimately, the fused feature maps are forwarded to a Classification Head,

outputting dense pixel-wise semantic predictions with the resolution restored to H×W.

3.2 Multi-Scale Feature Fusion Module

In intricate natural scenarios, the dimensional disparities among objects are frequently pronounced, rendering single-tier features insufficient for providing consistent and accurate classification criteria. To circumvent the limitations of low spatial resolution and homogeneous contextual information inherent to deep features, we construct the Multi-Scale Feature Fusion (MSFF) module. Deployed at the apex of the backbone network, the MSFF module accepts the feature map C_4 as its input. To dilate the model's receptive field without incurring prohibitive computational overhead, we adopt a multi-branch parallelized feature extraction strategy, comprising a global context branch, a multi-scale receptive field branch, and a local detail preservation branch.

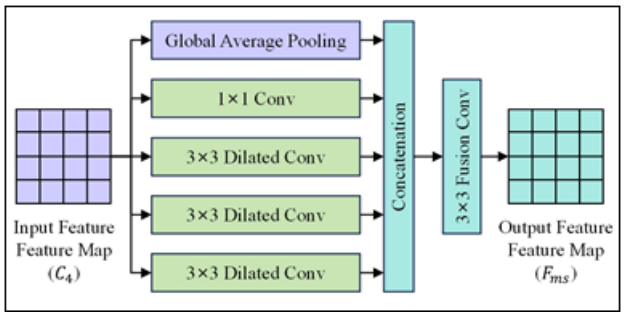

Figure 2: Detailed structure of the Multi-Scale Feature Fusion (MSFF) module.

The global context branch acquires the macroscopic prior of the entire image via Global Average Pooling, followed by a 1×1 dimensionality reduction convolution and bilinear upsampling to the original scale. The multi-scale receptive field branch employs three parallel 3×3 dilated convolutions with diverse dilation rates (e.g., configured to 6, 12, and 18) to capture localized contextual dependencies across varying spatial extents. The local detail preservation branch utilizes a standard 1×1 convolution to retain the intrinsic deep feature information. The outputs from all branches are subsequently concatenated along the channel dimension and fed into a 3×3 fusion convolutional layer to facilitate cross-channel interaction and smoothing, yielding the multi-scale aggregated feature F_{ms}. This representation successfully couples local nuances with global semantics, thereby fortifying the network's recognition capacity for targets of disparate dimensions.

3.3 Boundary Enhancement Module

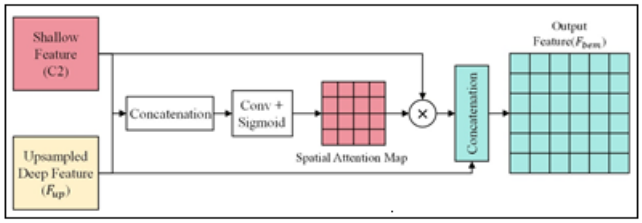

Figure 3: Detailed structure of the Boundary Enhancement Module (BEM).

Following successive convolutional and downsampling operations, object boundaries encapsulated within deep features frequently manifest blurring and aliasing phenomena. Although shallow features (e.g., C_2) retain higher-resolution spatial structures, directly concatenating them with deep features across hierarchical levels often introduces substantial background noise. To this end, we design the Boundary Enhancement Module (BEM) to extract high-frequency edge activation responses from shallow features, thereby guiding the upsampling trajectory of the decoder. The BEM takes the shallow feature C_2 and the deep guiding feature (i.e., F_{up}, derived by upsampling F_{ms} to align with the spatial resolution of C_2) as inputs.

Specifically, we initially concatenate C_2 and F_{up}, and subsequently deploy a lightweight convolutional module-comprising a 3×3 convolution (with 256 filters), a Batch Normalization layer, and a 1×1 projection layer—to learn a spatial attention map, denoted as $A_{boundary} \in \mathrm{R}^{H^{'}\times W^{'}\times 1}$.

Constrained by a Sigmoid activation function, this attention map is normalized to a range of [0,1], serving to accentuate high-frequency peripheral regions while suppressing responses in homogeneous, flat areas. Thereafter, we perform an element-wise multiplication between this attention map and the raw shallow feature C_2 to generate the boundary-enhanced feature C_2

$$C_2^{'} = C_2 \otimes A_{boundary} \quad (1)$$

where $\otimes$ denotes the Hadamard (element-wise) product. The boundary-enhanced feature C_2 is fused with the deep feature F_{up} and forwarded to the subsequent decoding layers. Through this fine-grained spatial compensation, the network manages to drastically ameliorate the structural integrity of object contours without incurring conspicuous inflation in parameter count.

3.4 Loss Function

To execute a joint optimization of semantic classification and boundary restoration, the overarching objective function of the proposed framework comprises two distinct components: a primary semantic segmentation loss and an auxiliary boundary constraint.

We employ the standard Cross-Entropy (CE) loss as the principal segmentation loss, denoted as L_{seg}, to supervise the pixel-level categorical predictions. Furthermore, to compel the BEM to effectively assimilate and project geometric boundaries, we extract the boundary predictive map from the BEM and compute the Binary Cross-Entropy (BCE) loss, denoted as L_{edge}, against the generated ground truth edge labels. Specifically, these binary edge labels are derived online by applying a morphological gradient operation (with a 3×3 kernel) to the discrete semantic segmentation masks. The total loss function, L_{total}, is formulated as:

$$L_{total} = L_{seg} + \lambda L_{edge} \quad (2)$$

where λ functions as a hyperparameter to orchestrate the equilibrium between the two loss terms. This joint

supervisory strategy not only accelerates the convergence trajectory of the network but also intrinsically coerces the model- from a gradient backpropagation perspective- to allocate disproportionate attention to peripheral zones.

4. Experiments

4.1 Datasets and Evaluation Metrics

To validate the effectiveness of the proposed method, we conduct comprehensive empirical evaluations on two mainstream public semantic segmentation benchmarks: PASCAL VOC 2012[13]and Cityscapes [14].

PASCAL VOC 2012: As a paradigmatic benchmark in the realm of semantic segmentation, this dataset encompasses 20 foreground object categories and one background class. The official split allocates 1,464 images for training, 1,449 for validation, and 1,456 for testing. In our experiments, we utilize the widely adopted augmented dataset, which expands the training set to 10,582 images.

Cityscapes: Concentrating exclusively on the comprehension of urban street scenes, this dataset provides 19 semantic categories for evaluation. It contains 5,000 finely annotated images (comprising 2,975 for training, 500 for validation, and 1,525 for testing). Characterized by severe object scale variations and a plethora of miniature targets, Cityscapes serves as an optimal touchstone to verify the efficacy of the proposed MSFF and BEM modules.

Regarding evaluation metrics, we adopt the Mean Intersection over Union (mIoU), the most prevalent quantitative criterion in semantic segmentation tasks.

4.2 Implementation Details

The proposed architecture is implemented utilizing the PyTorch deep learning framework. To accelerate convergence and bolster feature extraction capabilities, we adopt ResNet-50, pre-trained on ImageNet, as the foundational backbone. During the optimization phase, we employ the Stochastic Gradient Descent (SGD) optimizer, configuring the momentum to 0.9 and weight decay to 1×10^{-4}. The learning rate is modulated by the classical "poly" decay strategy, with the initial learning rate set to 0.01. Pertaining to data augmentation, the input images are subjected to random horizontal flipping, random scaling (within a ratio of 0.5 to 2.0), and random cropping (to 512×512 for VOC and 768×768 for Cityscapes). The model is trained on a single NVIDIA RTX 4090 GPU for 100 epochs with a batch size of 8. The balancing hyperparameter λ in the loss function is assigned a value of 0.5. This specific weight was determined empirically to ensure that the auxiliary boundary gradient regularizes the network effectively without overwhelming the primary semantic optimization trajectory.

4.3 Comparison with State-of-the-Art

To comprehensively validate the effectiveness and superiority of the proposed methodology, we compare our network against prevailing state-of-the-art architectures on both the PASCAL VOC 2012 and Cityscapes validation sets. To ensure an equitable comparison, all evaluations are conducted under a single-scale testing protocol, devoid of any auxiliary pre-training on external datasets. The comparative results are delineated in Table 1.

Table 1: Comparison with state-of-the-art methods on PASCAL VOC 2012 and Cityscapes.

Methods	*Backbone*	*PASCAL VOC 2012(mIOU)*	*Cityscapes (mIOU)*
SegNet	VGG-16	59.9%	-
ENet	Custom	-	58.3%
FCN-8s	VGG-16	62.2%	65.3%
ICNet	ResNet-50	-	69.5%
DeepLabV2	ResNet-50	68.7%	68.6%
PSPNet(Baseline)	ResNet-50	69.6%	70.2%
Ours	ResNet-50	71.5%	72.6%

The empirical outcomes demonstrate that the proposed method exhibits exceptional segmentation performance on both benchmarks, achieving mIoUs of 71.5% and 72.6%, respectively. Compared to nascent architectures such as FCN-8s and SegNet [15], as well as lightweight networks like ICNet [16], our method effectively overcomes the limitations of restricted receptive fields by virtue of the potent multi-scale contextual aggregation capacity of the MSFF module, realizing a substantial leap in accuracy (e.g., yielding a further 3.1% enhancement over the multi-scale cascade network ICNet on Cityscapes). Furthermore, when juxtaposed with classical multi-scale modeling paradigms such as DeepLabV2 and PSPNet, our method, relying merely on a ResNet-50 backbone, successfully surpasses DeepLabV2 (which utilizes a bulkier ResNet-101 backbone) and outperforms the equivalent PSPNet baseline. This robustly corroborates the superiority of the joint mechanism synergizing macroscopic scale perception (MSFF) with microscopic edge compensation (BEM). It indicates that the proposed network can more precisely elevate the classification quality of object contours without incurring a conspicuous inflation in computational overhead, striking a highly competitive equilibrium between scene parsing capacity and operational efficiency.

4.4 Ablation Study

To profoundly investigate the specific contributions of the individually proposed modules to the overarching model performance, we conduct an extensive suite of ablation studies on the PASCAL VOC 2012 dataset. We establish a foundational encoder-decoder network encompassing ResNet-50 as our baseline model. The results are delineated in Table 2.

Table 2: Ablation study of proposed modules on PASCAL VOC 2012.

Models	*mIOU*
Baseline	55.1%
Baseline + BEM	67.6%
Baseline + MSFF	70.1%
Ours	71.5%

The incorporation of the MSFF module into the baseline augments the mIoU metric by 15.0%. This substantiates that the parallelized dilated convolution branches effectively dilate the receptive field, resolving the "hole" prediction

dilemma inside large-scale objects. Independently integrating the BEM module alongside the auxiliary boundary loss into the baseline similarly yields a performance gain of 12.5%. By exploiting shallow spatial features, the BEM successfully rectifies the misclassification of local nuances. When both the MSFF and BEM are simultaneously deployed, the model achieves optimal performance. This signifies a synergistic complementarity between macroscopic scale adaptability and microscopic boundary refinement, the combination of which holistically ameliorates semantic segmentation quality.

4.5 Visualization Analysis

To intuitively illustrate the segmentation efficacy of the proposed method, we extract and compare the predictive outcomes of selected validation images against those of the baseline model.

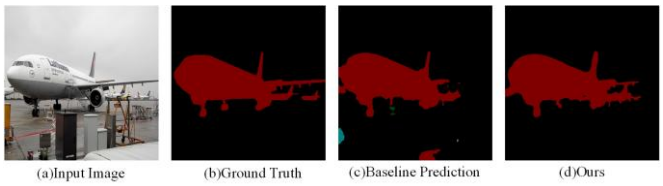

Figure 4: Qualitative visualization of the segmentation outcomes on the validation set. Columns from left to right correspond to: (a) Input Image, (b) Ground Truth, (c) Baseline Prediction, and (d) Ours.

5. Conclusions

This paper presented a semantic segmentation network integrating Multi-Scale Feature Fusion and Boundary Enhancement to improve segmentation performance for multi-scale objects and complex boundaries. Experimental results on the PASCAL VOC 2012 and Cityscapes datasets demonstrate that the proposed approach achieves improved segmentation accuracy compared with representative baseline methods while maintaining a relatively simple architecture. Future work will investigate lightweight model design, additional benchmark evaluation, and semi-supervised learning to further improve practical applicability.

References

[1] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(7), 3523-3542.

[2] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[3] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Cham: Springer international publishing.

[4] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & 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.

[5] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[6] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

[7] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[8] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3146-3154).

[9] Takikawa, T., Acuna, D., Jampani, V., & Fidler, S. (2019). Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5229-5238).

[10] Zhen, M., Wang, J., Zhou, L., Li, S., Shen, T., Shang, J., ... & Quan, L. (2020). Joint semantic segmentation and boundary detection using iterative pyramid contexts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13666-13675).

[11] Kirillov, A., Wu, Y., He, K., & Girshick, R. (2020). Pointrend: Image segmentation as rendering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9799-9808).

[12] Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: A survey. Applied sciences, 12(18), 8972.

[13] Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. International journal of computer vision, 111(1), 98-136.

[14] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).

[15] Badrinarayanan, V., Kendall, A., & 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.

[16] Zhao, H., Qi, X., Shen, X., Shi, J., & Jia, J. (2018). Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European conference on computer vision (ECCV) (pp. 405-420).