Real-time Semantic Segmentation with Parallel Multiple Views Feature Augmentation

Jian-Jun Qiao
Southwest Jiaotong University
Chengdu, China
qjjai56@gmail.com

Zhi-Qi Cheng
Carnegie Mellon University
Pittsburgh, United States
zhiqic@cs.cmu.edu

Xiao Wu
Southwest Jiaotong University
Chengdu, China
wuxiaohk@swjtu.edu.cn

Wei Li*
Southwest Jiaotong University
Chengdu, China
liwei@swjtu.edu.cn

Ji Zhang
Southwest Jiaotong University
Chengdu, China
jizhang901@gmail.com

ABSTRACT

Real-time semantic segmentation is essential for many practical applications, which utilizes attention-based feature aggregation into lightweight structures to improve accuracy and efficiency. However, existing attention-based methods ignore 1) high-level and low-level feature augmentation guided by spatial information, and 2) low-level feature augmentation guided by semantic context, so that feature gaps between multi-level features and noise of low-level spatial details still exist. To address these problems, a new real-time semantic segmentation network, called MvFSeg, is proposed. In MvFSeg, parallel convolution with multiple depths is designed as a context head to generate and integrate multi-view features with larger receptive fields. Moreover, MvFSeg designs multiple views feature augmentation strategies that exploit spatial and semantic guidance for shallow and deep feature augmentation in an inter-layer and intra-layer manner. These strategies eliminate feature gaps between multi-level features, filter out the noise of spatial details, and provide spatial and semantic guidance for multi-level features. By combining multi-view features and augmented features from the lightweight networks with progressive dense aggregation structures, MvFSeg effectively captures invariance at various scales and generates high-quality segmentation results. Experiments conducted on Cityscapes and CamVid benchmark show that MvFSeg outperforms the existing state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies → Image segmentation**.

KEYWORDS

Real-time Semantic Segmentation, Multi-view Features, Feature Augmentation, Attention Mechanism, Feature Aggregation

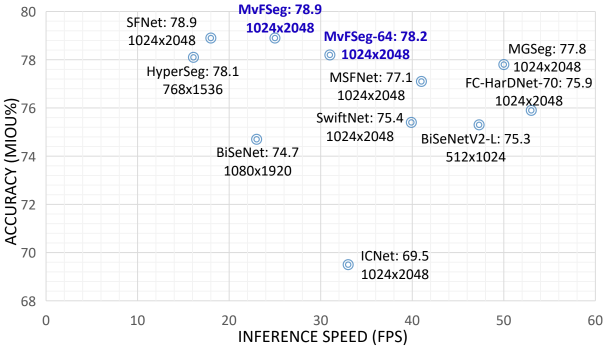

Figure 1: Speed-accuracy comparison on the Cityscapes test set. The proposed MvFSeg achieves the highest accuracy and possesses real-time inference speed.

ACM Reference Format:
Jian-Jun Qiao, Zhi-Qi Cheng, Xiao Wu, Wei Li, and Ji Zhang. 2022. Real-time Semantic Segmentation with Parallel Multiple Views Feature Augmentation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547786

1 INTRODUCTION

Semantic segmentation is a crucial task closely related to vision and multimedia applications [6–10, 16, 35, 44–46], which detects and delineates each object of interest appearing in an image by assigning each pixel a class label [12, 24, 26, 29, 42]. With the development of deep learning technologies, semantic segmentation achieves great progress. However, objects of interest in the real world are often difficult to be identified due to complex backgrounds, inconsistent scales, and diverse visual appearances, which bring great difficulties to semantic segmentation, especially for real-time segmentation. To achieve effective real-time performance, many approaches have been explored to restrict the input image size [47], migrate lightweight network [42] and prune the redundant channels [32] to improve efficiency. Unfortunately, these lightweight designs borrowed from other tasks such as image classification suffer from limitations in segmentation and lead to dramatic performance degradation due to insufficient task-specific designs [12]. It remains a challenging task to develop an effective method for real-time semantic segmentation.

*Corresponding author: Wei Li

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547786

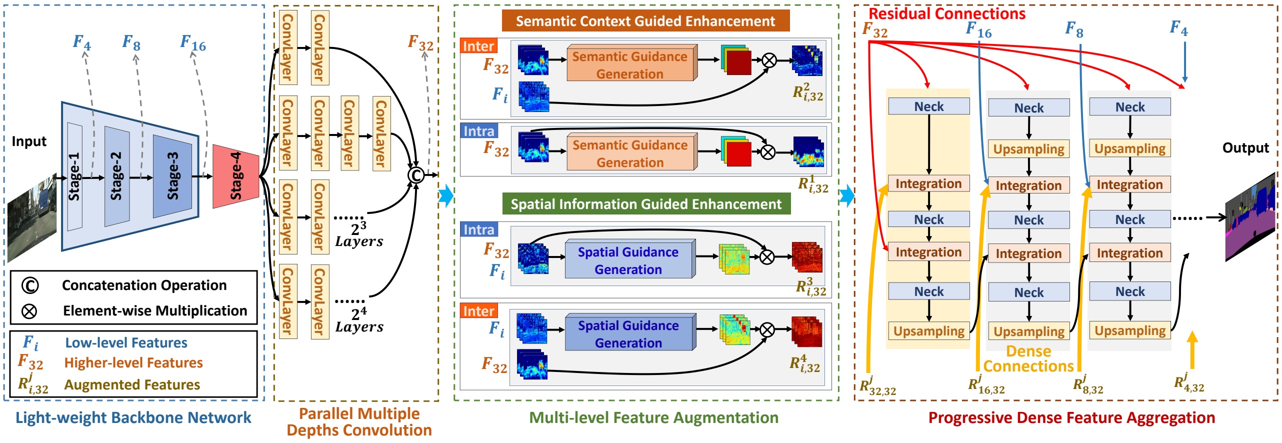
Input
F_4
F_8
F_{16}
Stage-1
Stage-2
Stage-3
Stage-4
© Concatenation Operation
⊗ Element-wise Multiplication
F_i Low-level Features
F_{32} Higher-level Features
$R^j_{i,32}$ Augmented Features
Light-weight Backbone Network
ConvLayer
F_{32}
2^3 Layers
2^4 Layers
Parallel Multiple Depths Convolution
Semantic Context Guided Enhancement
Inter
Intra
Semantic Guidance Generation
$R^2_{i,32}$
$R^1_{i,32}$
Spatial Information Guided Enhancement
Spatial Guidance Generation
$R^3_{i,32}$
$R^4_{i,32}$
Multi-level Feature Augmentation
Residual Connections
Neck
Integration
Upsampling
Dense Connections
Output
$R^j_{32,32}$
$R^j_{16,32}$
$R^j_{8,32}$
$R^j_{4,32}$
Progressive Dense Feature Aggregation

Figure 2: The framework of the proposed MvFSeg is composed of the lightweight backbone network, parallel multiple depths convolution (PMDC), multi-level feature augmentation (MFA), and progressive dense feature aggregation (PDFA).

For the purpose of real-time processing, many semantic segmentation approaches usually adopt lightweight backbones. Unfortunately, due to the limited parameters for learning complex relations, the pixels of the final prediction map get insufficient perceptual regions and thus often achieve suboptimal performance. Typically, FCN [26] network with a lightweight backbone has insufficient receptive fields on large objects, but has oversized receptive fields on small objects. This will lead to incomplete segmentation of large objects and coarse segmentation of small objects. To alleviate such condition, different strategies are adopted, including attention mechanisms [12, 42] and aggregation structures [23, 37] for feature augmentation. However, existing attention-based strategies in real-time methods [12, 42] focus on the semantic context guided high-level feature augmentation, while the semantic context guided low-level feature augmentation tends to be ignored. Moreover, spatial information guided high-level and low-level feature augmentation is not investigated in real-time segmentation. Accordingly, feature gaps among multi-level features and noise of spatial details in shallow features still exist. Therefore, the investigation of multi-view feature augmentation is limited and the aggregation structure usually receives features with insufficient spatial information and semantic context, resulting in limited fusion results.

To alleviate the problem of lack of sufficient spatial information and semantic context, this paper proposes a new real-time semantic segmentation network, MvFSeg, which fully utilizes hierarchical spatial information and high-level semantic context for multi-view feature enhancement and aggregation. As can be seen in Figure 2, it consists of four components: 1) light-weight backbone network, 2) parallel multiple depths convolution (PMDC), 3) multi-level feature augmentation (MFA), and 4) progressive dense feature aggregation (PDFA). The structure of MvFSeg is described as follows:

- First, four stages of features are generated with the lightweight backbone, where the first three stages correspond to low-level features that are used for localization, while the last stage represents the high-level features that are exploited for classification.
- Second, the high-level features from stage 4 are fed to PMDC to integrate long-range context and generate higher-level features (F_{32}), which constructs multiple convolution paths with different depths to capture features with different receptive fields. By combining features from these different paths, multi-scale receptive fields are integrated.
- Third, multi-level features (F_4, F_8, F_{16} and F_{32}) are fed to semantic and spatial augmentation branches to capture and utilize semantic context and spatial information, which potentially boosts the capability of classification and localization. In order to integrate these complemented semantic context and spatial information for multi-level feature augmentation, inter-layer and intra-layer feature augmentation strategies are designed. The inter-layer feature augmentation bridges feature gaps between higher-level and low-level features by capturing long-range semantic and spatial guidance and conducting cross-layer feature augmentation. The intra-layer feature augmentation acquires short-range semantic and spatial guidance from shallow and deep features, which refines features from the same layer.
- Finally, these augmented features ($R^j_{i,32}$) as well as the multi-level features (F_4, F_8, F_{16} and F_{32}) are passed to sequential fusion blocks for effective dense feature aggregation, which aims to integrate multi-view features. In addition, higher-level features are connected to each fusion block in a residual manner to preserve higher-level semantics. Moreover, each fusion block is composed of stacked compression and integration structure, which guarantees efficiency. Therefore, high-resolution segmentation maps are efficiently reconstructed with a progressive increase in semantics and spatial information.

MvFSeg integrates hierarchical spatial information and high-level semantic context from features of all layers, which acts as a general framework, so that it can be applied to different backbone networks for real-time semantic segmentation. To verify the efficiency and effectiveness of MvFSeg, extensive experiments are conducted on two datasets and the state-of-the-art performance is obtained, as shown in Figure 1.

The contributions are summarized as follows:

- A real-time segmentation approach called MvFSeg is proposed, which is designed with parallel multiple views feature augmentation structure to generate high-quality features that contain sufficient spatial information and semantic context, and is a general framework for different lightweight backbone networks.

- Parallel multiple depths convolution is designed as context head, which constructs multiple convolutional learning paths with different depths to capture long-range information and generates higher-level features with multi-scale and larger receptive fields.
- Multi-level feature augmentation with inter-layer and intra-layer feature augmentation strategies is designed by integrating the spatial information and semantic context of the network to bridge feature gaps among cross-layer features and conduct intra-layer feature self-reinforcement, which improves semantic classification ability and spatial localization capability.
- Progressive dense feature aggregation is proposed to combine augmented and multi-view features in an efficient progressive manner for stepwise high-resolution segmentation map reconstruction, with the effective residual structure to preserve semantics and dense connections for multi-view feature aggregation.
- MvFSeg achieves the state-of-the-art performance on the Cityscapes [11] and CamVid [2] datasets in terms of accuracy and efficiency with the well-designed structure.

2 RELATED WORKS

2.1 Semantic Segmentation

With the prosperity of deep learning, convolutional neural networks [18, 33, 34, 38–40] are applied to semantic segmentation. FCN [26] first adopts fully convolutional networks for image segmentation and hits a new peak on segmentation task. Based on FCN, methods of different designs [4, 31] are proposed and achieve impressive progress. Approach like UNet [36] is constructed in U-shape structure, which bridges the information transmission gap between the encoder and decoder with skip-connections [26]. Deeplab series [4, 5, 25] perform high-resolution feature map learning by conducting dilated convolution operation and exploit multi-scale information with astrous spatial pyramid pooling (ASPP). PSPNet [48] devises pyramid pooling module (PPM) to model multi-scale contexts. In [21], a depth-adaptive network consisting of adaptive perception neurons and in-layer multi-scale neurons is proposed to adjust the receptive field. In [14], the depth information of training images is treated as the privileged information to mine the hard pixels in semantic segmentation. Recently, transformer [49] is developed for scene parsing and makes impressive efficacy. SETR [49] treats semantic segmentation as a sequence-to-sequence prediction task and models global context in every layer of the transformer to achieve powerful performance.

2.2 Real-time Semantic Segmentation

In pursuit of real-time performance for real-world applications, more and more attention is drawn to fast segmentation [1, 24, 28, 32, 47]. ENet [32] drops the redundant channels for light-weight segmentation structure. SegNet [1] constructs U-shape structure with small networks and pooling indices reused for segmentation. ICNet [47] provides original and decreased resolution as inputs to generate spatial details and high-level context for fast and effective segmentation. ESPNet [28] decomposes standard convolution into point-wise convolution and spatial pyramid of dilated convolutions to save computation. SFNet [24] learns semantic context between feature maps of adjacent levels and propagates high-level features to high-resolution features effectively. HyperSeg [30] devises dynamic patch-wise convolution with weights that vary both per input and per spatial location, under the nested U-shape structure.

2.3 Attention and Feature Aggregation

Attention strategy [12, 17, 42] and feature aggregation [23, 37] are usually adopted for efficient and effective segmentation. DANet [13] adaptively generates and integrates local features and global dependencies for fine-grained features, with dual-attention mechanism. But the inference speed of DANet is non-real-time. CCNet [20] proposes criss-cross attention to harvest the fine-grained contextual information of all the pixels in an efficient way. BiSeNet [42] takes multi-branch structure to generate deep and shallow features for attention refinement and feature aggregation. STDC-Seg [12] refines deep features with channel attention [19]. MGSeg [17] designs light-weight channel attention to extract dominant visual characteristics and fuses multi-granularity features. DFANet [23] efficiently incorporates high-level context into encoded features with multiple times, achieving promising trade-off efficacy between speed and accuracy. MSFNet [37] fuses all feature maps of different scales to enlarge the receptive field and recover spatial information.

3 PARALLEL MULTIPLE VIEWS FEATURE AUGMENTATION

3.1 Framework

The framework of MvFSeg is illustrated in Figure 2, which consists of four components: 1) light-weight backbone network, 2) parallel multiple depths convolution (PMDC), 3) multi-level feature augmentation (MFA) and 4) progressive dense feature aggregation (PDFA). First, to preserve high efficiency in real-time processing, the lightweight backbone network, e.g., ResNet18 [18] is chosen as the backbone network for 4-stage hierarchical shallow and deep features. Second, high-level features are fed to multiple convolution paths with different depths to get features with different and larger receptive fields. By integrating features from these different convolution paths, higher-level features with multi-view receptive fields are generated. Third, low-level features (F_4, F_8 and F_{16}) and higher-level features (F_{32}) from PMDC are fed to MFA for inter-layer and intra-layer feature augmentation. In inter-layer feature augmentation, low-level and higher-level features are combined in pairs to produce cross-layer spatial and semantic guidance and conduct coarse-to-fine semantic context and spatial information guided feature augmentation. In intra-layer augmentation, shallow and deep features are independently sent to semantic and spatial augmentation branches for intra-layer feature self-reinforcement. By providing spatial information and semantic context to features at all levels, MFA generates augmented features ($R^j_{i,32}$). Fourth, these complemented features, including the original multi-level features (F_4, F_8, F_{16} and F_{32}) and the refined features ($R^j_{i,32}$) are combined with effective dense connections and residual structure to integrate multi-view features. Specifically, they are connected to each aggregation block of PDFA for effective dense feature aggregation. In addition, higher-level features (F_{32}) are fed to each fusion step of PDFA in a residual manner to preserve semantic context. With PDFA, high-quality aggregation results are progressively generated,

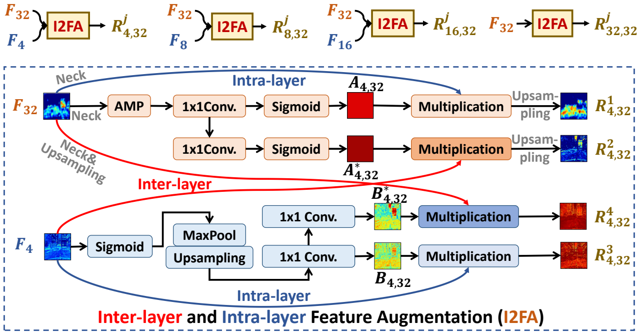

Figure 3: The structure of multi-level feature augmentation.

with step-by-step reconstruction of feature map resolution and stepwise increase of semantic context as well as spatial information.

3.2 Parallel Multiple Depths Convolution

The outputs of light-weight backbone networks have small receptive fields due to the depth of the network, resulting in limited performance. To solve this problem, parallel multiple depths convolution (PMDC) is proposed. In PMDC, extra convolution path is built to generate higher-level features with larger receptive fields. In addition, to generate higher-level features with multi-view receptive fields, the single convolution path is expanded to multiple paths. These convolution paths have different depths, which aim to generate features with multi-scale receptive fields. By combining features from these paths, higher-level features with multi-view receptive fields are generated.

As can be seen in Figure 2, the high-level features from stage 4 of backbone network are fed to PMDC. To reduce the parameters and preserve efficiency, a Neck operation is conducted to compress the channel number, which is a 1×1 convolution. The compressed features are fed to multiple convolution paths with different depths for higher-level features with multi-scale receptive fields. Finally, the output of these convolution paths as well as the compressed high-level features are fused with a concatenation operation and a Neck operation to integrate and compress features with multi-scale receptive fields. Therefore, higher-level features with multi-view receptive fields are generated. The process is formulated as follows:

$$F'_{32} = Neck(OF_{32}) \tag{1}$$

$$F_{32} = Neck(Cat(F'_{32}, \theta_1(F'_{32}), \theta_2(F'_{32}), \theta_3(F'_{32}), \theta_4(F'_{32}))) \tag{2}$$

where $\theta_z(F'_{32})$ denotes the output of the z_{th} convolution path of PMDC. OF_{32} and F'_{32} refer to the high-level features from stage 4 of the light-weight backbone and the compressed features, respectively. *Cat* denotes concatenation operation.

3.3 Multi-level Feature Augmentation

In previous methods, feature gaps existing in features from different levels (e.g., F_4 and F_{32}) restrict the fusion quality. In addition, the noise of spatial details in low-level features damages the feature representation ability. To diminish the feature gaps, including semantic gaps and spatial gaps of different features, e.g., F_4 and F_{32}, MFA is proposed to augment multi-view features. The inter-layer feature augmentation strategy is designed to handle these complementary multi-level features and make multi-view feature augmentation. With MFA, both shallow and deep features obtain sufficient semantic context and spatial information, which eliminates the gaps among multi-level features. The second problem is that the noise in low-level features will reduce the effectiveness of feature fusion. To alleviate this problem, a feature selection structure is designed in MFA, which can select important spatial information and filter out the noise of spatial details with a max-pooling operation and convolutional learning layers. Max pooling is used to obtain salient features that contain abundant spatial information while convolutional learning layers are exploited to learn pixel-level relations among these salient features.

As can be seen in Figure 3, MFA is composed of four inter-layer and intra-layer feature augmentation (I2FA) components. For each I2FA component, paired higher-level and low-level features, e.g., F_{32} and F_4 are exploited to produce short-range and long-range semantic guidance $A_{i,32}$ and $A^*_{i,32}$, as well as short-range and long-range spatial guidance $B_{i,32}$ and $B^*_{i,32}$. The semantic guidance from higher-level features is used to provide semantic context to both shallow and deep features. With this, the network pays more attention to the identification of categories. The spatial guidance from different layers aims to provide spatial information to both higher-level and low-level features. In this way, the network can focus on local details such as boundaries and small objects. With the spatial and semantic guidance, MFA devises inter-layer and intra-layer feature augmentation strategies to conduct a coarse-to-fine process and produce the augmented features $R^j_{i,32}$, $j \in \{1, 2, 3, 4\}$. With the I2FA structure, semantic and spatial gaps among multi-level features are eliminated by providing semantic context and spatial information at all levels.

Multi-level feature augmentation with semantic guidance. Inspired by the previous attention mechanism [19], semantic guidance is generated with channel attention. In the attention structure, the Neck operation aims to reduce channel number by a 1×1 convolution layer. Adaptive max pooling layer is used to generate high response values, which refer to activated semantics. The first convolutional learning block generates short-range semantic guidance $A_{i,32}$ to conduct feature self-reinforcement, while the extra learning block (ExtraLB) produces long-range semantic guidance $A^*_{i,32}$ to eliminate semantic gaps among cross-layer features. Then sigmoid operation models interdependencies between channels and generates semantic guidance. Different from the previous methods, the designed structure is more efficient with Neck. In addition, the generated short-range and long-range semantic guidance will not only be exploited for higher-level feature refinement, but also for low-level feature augmentation. With the semantic guidance, two operations are conducted to augment features: 1) enhancing the representation ability of higher-level features with short-range semantic guidance $A_{i,32}$, and 2) guiding low-level features for classification with semantic information provided by long-range semantic guidance $A^*_{i,32}$. The formulations are written as follows:

$$R^1_{i,32} = F_{32} \odot A_{i,32}; R^2_{i,32} = F_i \odot A^*_{i,32} \tag{3}$$

where ⊙ denotes element-wise multiplication.

Multi-level feature augmentation with spatial guidance. To get spatial guidance, the sigmoid activation is adopted to generate pixel-level spatial details. The max-pooling and upsampling

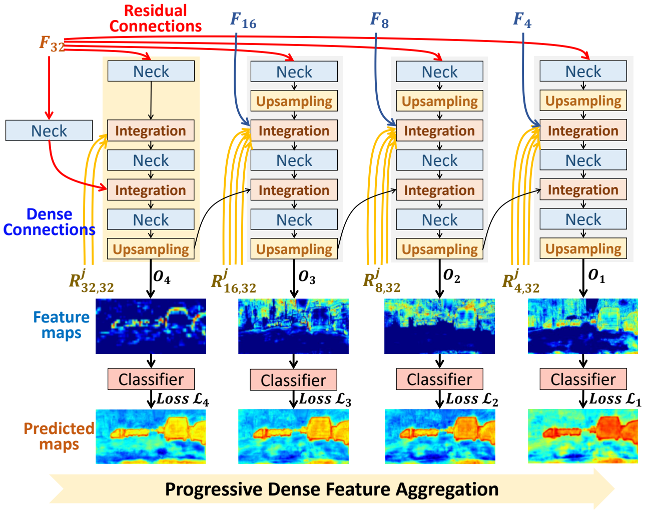

Figure 4: Progressive dense feature aggregation structure.

operations are utilized to select important spatial structures, filter out the noise of spatial details and restore the map resolution. The first learning block is 1×1 convolution followed by batch normalization that learns spatial dependencies of pixels to get short-range spatial guidance $B_{i,32}$. When generating spatial information as guidance of higher-level features, an extra convolutional learning block (ExtraLB) is used to construct a long learning path, which generates long-range spatial guidance $B^*_{i,32}$ to eliminate spatial gaps in multi-level features. With the spatial guidance, two operations are conducted to augment features: 1) improving the representation ability of low-level features with short-range spatial guidance $B_{i,32}$, and 2) providing spatial information to higher-level features for localization with long-range spatial guidance $B^*_{i,32}$. The augmentation process is formulated as follows:

$$R^3_{i,32} = F_i \odot B_{i,32}; R^4_{i,32} = F_{32} \odot B^*_{i,32} \tag{4}$$

3.4 Progressive Dense Feature Aggregation

To effectively and efficiently combine the augmented features as well as the original features and generate high-quality prediction map with multi-view features, progressive dense feature aggregation (PDFA) is designed. It consists of sequential fusion blocks with dense and residual connections, which effectively integrates these complemented features. In this way, the multi-level spatial information and semantic context are progressively integrated along the fusion path. Consequently, high-resolution maps are reconstructed in progressive manner with stepwise increase of semantic context, spatial information and images resolution. With the stacked channel compression and integration structure, the computational cost is saved.

Compared to directly combining all features, PDFA has three advantages. First, PDFA makes the network trainable. There are many features, including 14 refined features (brown lines in Figure 4) and 8 original features (blue and red lines in Figure 4). Besides, these features need to be resized to 1/4 of the original image size for fusion. Therefore, directly combining these features will produce huge parameters and the network will be extremely difficult to train. Second, the computational cost is greatly reduced. PDFA is divided into different steps, and at the front and end of each fusion step, channel compression is conducted to reduce parameters and computation. Third, the accuracy of combination result is further improved. PDFA designs dense and residual connection structure, which progressively propagates the previous fusion results O_k, the original phased features F_i, as well as the next fine-grained feature group $R^j_{i,32}$ to the following fusion step. In this way, multi-view features are integrated, which is of great importance in complex scene segmentation.

The structure of PDFA is illustrated in Figure 4, which combines the original higher-level and low-level features F_i, as well as the refined features $R^j_{i,32}$ with a 4-step structure to generate fusion results $O_k, k \in \{1, 2, 3, 4\}$. The fusion results O_k are fed into classifiers that are composed of Conv-BN-ReLU-Conv for prediction. The integration structure in Figure 4 denotes the operation of concatenation. Neck is used for channel compression, with which the number of output channels is reduced.

3.5 Network Efficiency

Efficiency is indispensable for real-time semantic segmentation. In MvFSeg, a series of lightweight strategies such as adopting lightweight backbone networks and learning low-resolution feature map are adopted to achieve high efficiency. Light-weight backbone networks such as ResNet18 [18], GhostNet [15] are exploited in MvFSeg for feature extraction. These efficient lightweight backbone networks are important to achieve real-time performance. In addition, the resolution of multi-level features generated by backbone network is smaller than the original image size, which dramatically reduces the computational complexity.

In MvFSeg structure, most of the operations such as 1×1 convolution and concatenation in PMDC, MFA and PDFA are efficient, which guarantees efficiency. Moreover, except for these lightweight operations, the efficiency of MvFSeg can be controlled by a hyperparameter. This hyperparameter is exploited to control the channel number in PMDC, MFA, PDFA, and four values are set for it: O, 32, 64, 128. When it is set to O, PMDC and PDFA use 128 as the basic channel number for series operations. MFA adopts the original channel number of features from different layers for inter-layer and intra-layer feature augmentation. When it is set to 32, 64, or 128, all PMDC, MFA and PDFA utilize the same channel number of 32, 64 or 128 as the basic channel number.

3.6 Loss Function

In MvFSeg, features from the progressive dense feature aggregation (PDFA) are exploited for loss computation and optimization. PDFA has four blocks and the output $O_k, k \in \{1, 2, 3, 4\}$ of each block will be used to calculate the cross entropy loss with the ground truth. Online hard example mining (OHEM) [37] is adopted to select the hard pixels having large loss values. With OHEM, the errors of the selected pixels are back-propagated in the top-V positions. In this way, losses of the four blocks are obtained. The summation of the four losses is the final loss that would be optimized. The loss function [17] is formulated as follows:

$$\mathcal{L} = -\frac{1}{V} \sum_{k=1}^{4} \sum_{l=1}^{L} y_l \cdot \mathbb{I}(p^{y_l}_{l,k} < t_V) \cdot log p^{y_l}_{l,k} \tag{5}$$

where $\mathcal{L}$ denotes the final loss that is computed from feature $O_k, k \in \{1, 2, 3, 4\}$ and label y. L denotes the number of pixels in O_k. $p_{l,k}^{y_l}$ is the predicted posterior probability of pixel l in O_k. The threshold t_V is set to select the pixels with the highest top-V losses. $\mathbb{I}(x) = 1$ if x is true, and 0 otherwise.

4 EXPERIMENTS

4.1 Datasets

Cityscapes [11] is a scene parsing benchmark for urban roads with image resolution of 1024×2048. The training, validation and test set are 2975, 500 and 1525 images with fine annotations, respectively. To fairly compare MvFSeg with the previous works [42, 47], 19 classes of Cityscapes are used for segmentation.

CamVid [2] contains images of urban roads and the size of every image is 720×960. The training, test, and validation set are 367, 223, and 101 images, respectively. Similarly, MvFSeg follows the previous works [23, 42, 47] and utilizes 11 classes of CamVid for segmentation.

4.2 Evaluation Metrics

Mean Intersection over Union (mIoU) [26] is widely used in semantic segmentation [4, 13, 26, 42, 43] for accuracy index, which calculates the average ratio of the intersection over union of ground truth and the predicted pixel regions. **Frames Per Second** (FPS) and **Parameters of Model** (PM) are used for efficiency evaluation.

4.3 Implementation Details

MvFSeg is constructed by PyTorch, an open-source framework of deep learning. Stochastic Gradient Descent (SGD) [22] is adopted as optimizer with momentum set to 0.9, and weight decay set to 5e-4. Following the pioneering works [4, 13, 42, 43], MvFSeg adopts the "poly" learning rate strategy $lr = lr_{base} \times (1 - \frac{C_{iter}}{T_{iter}})^{power}$, where lr is current learning rate, lr_{base} is the base learning rate and $power = 0.9$. C_{iter} and T_{iter} are the current iteration number and the total iteration number, respectively. MvFSeg takes light-weight network such as ResNet18 [18] as the backbone. For max-pooling operation in MFA structure, the kernel size, stride and padding are set to 3, 2, and 1, respectively. The final augmented outputs O_1, O_2 and O_3 are utilized to compute auxiliary cross entropy loss. A single RTX 2080Ti GPU is adopted to test the inference speed of MvFSeg. The "profile" library of Python is adopted to compute the model parameters.

Cityscapes and CamVid training details: The base learning rate is 0.01. Images of Cityscapes are randomly cropped to 1024×1024 for training with 80,000 iterations and batch size 24. For CamVid, images are randomly cropped to 720×960, with iteration number and batch size set to 8,000 and 32, respectively. Data augmentation of random left-right flipping and random resizing with the scale range of [0.75, 2.0] are utilized in the training process. For Cityscapes and CamVid test set evaluation, the training and validation set are combined for model training.

4.4 Comparison with State-of-the-Art Methods

In this subsection, experiments are conducted on the Cityscapes and CamVid datasets to evaluate the effectiveness of the proposed method. The experiment results of the state-of-the-art methods and MvFSeg are listed in Table 1. "-" denotes the corresponding result is not provided by the listed method. For a fair comparison, the specific input size of the image related to speed is listed. The highest and second highest mIoU are highlighted. These methods only use single image scale for evaluation. For the Cityscapes dataset, methods trained with fine labels and using ImageNet for pretraining are listed for fair comparison. For CamVid test set, the accuracy of MvFSeg with ImageNet and Cityscapes pretraining are listed, respectively.

Table 1: Performance comparison with the state-of-the-art methods. CiV and CiT denote Cityscapes validation set and test set, respectively. CaT refers to CamVid test set.

Method	InputSize	PM	FPS	mIoU%		
				CiV	CiT	CaT
ENet [32]	360×640	0.4	135.4	-	58.3	51.3 / -
ICNet [47]	1024×2048	26.5	33	-	69.5	67.1 / -
DFANet [23]	1024×1024	7.8	100	71.9	71.3	64.7 / -
BiSeNet [42]	1080×1920	49.0	23	74.8	74.7	68.7 / -
SwiftNet [29]	1024×2048	11.8	39.9	75.5	75.4	- / -
BiSeNetV2-L [41]	512×1024	12.9	47.3	75.8	75.3	73.2 / 78.5
FC-HarDNet-70 [3]	1024×2048	16.1	53	-	75.9	67.7 / -
MSFNet [37]	1024×2048	-	41	77.2	77.1	75.4 / -
STDC2-Seg75 [12]	768×1536	-	97	77.0	76.8	73.9 / -
MGSeg [17]	1024×2048	13.3	50	-	77.8	72.7 / -
SFNet [24]	1024×2048	12.9	18	78.7	**78.9**	73.8 / -
HyperSeg [30]	768×1536	10.2	16.1	78.2	78.1	**78.4** / -
MvFSeg (GhostNet [15])	1024×2048	8.5	43	76.0	76.3	72.7 / 76.6
MvFSeg (MuxNet-m [27])	1024×2048	7.6	42	77.5	76.7	72.8 / 77.0
MvFSeg (MuxNet-l [27])	1024×2048	8.0	32	77.0	77.5	73.2 / 77.1
MvFSeg (ResNet18 [18])	1024×2048	19.6	32	78.0	77.5	75.6 / **79.2**
MvFSeg (ResNet18-D [18])	1024×2048	19.0	24	78.8	78.4	75.2 / 78.4
MvFSeg (ResNet34 [18])	1024×2048	28.7	25	**79.4**	**78.9**	**76.2** / **80.1**
MvFSeg-64 (ResNet18 [18])	1024×2048	12.9	45	77.9	77.0	75.2 / 78.3
MvFSeg-64 (ResNet34 [18])	1024×2048	23.0	31	**79.3**	78.2	75.4 / 79.1

Among the previous methods, SFNet [24] achieves the highest accuracy on Cityscapes, which refines high-level features with semantic flow [24]. But it ignores shallow feature refinement, resulting in poor ability in segmentation of small objects and has limited accuracy on CamVid, which has a large portion of small objects. Without Cityscapes pretraining, HyperSeg [30] achieves the highest accuracy on CamVid with a nested UNet [36] to draw higher level context features. But the large encoder-decoder structure limits the efficiency. Methods like MGSeg [17], STDC2-Seg75 [12] and MSFNet [37] achieve promising trade-off efficacy between speed and accuracy. But without sufficient spatial information and semantic context, they have limitations in segmentation of complex scenes.

MvFSeg achieves the highest mIoU on Cityscapes validation and test set, and the second highest (without Cityscapes pretraining) mIoU on CamVid test set, with the well-designed parallel multiple views feature augmentation structure. Despite SFNet achieves the highest accuracy (78.9% mIoU) on Cityscapes test set, its inference speed is slow. Moreover, SFNet adopts PPM (Pyramid Pooling Module) [48] as context head to achieve high result on Cityscapes dataset. Without PPM, the accuracy of SFNet is only 77.2% mIoU on Cityscapes validation set. The proposed MvFSeg only adopts public light-weight backbone network and all of the other parts are designed by MvFSeg itself. HyperSeg achieves the highest accuracy on CamVid test set, but its inference speed is very slow. With 768×1536 resolution, it only achieves 16 FPS. While the proposed MvFSeg achieves the second highest accuracy on CamVid test set and 25 FPS on larger image resolution (1024×2048). Compared

Table 2: Performance of PDFA, MFA and PMDC.

Method	PDFA	MFA	PMDC	mIoU%	FPS
Baseline				68.9	**75 / 508**
Baseline	✓			73.2	39 / 310
Baseline	✓	✓		76.7	34 / 230
MvFSeg	✓	✓	✓	**78.0**	32 / 221

to MvFSeg, MvFSeg-64 achieves better trade-off efficacy between speed and accuracy, which has high accuracy on both Cityscapes and CamVid datasets, and possesses faster speed. This is mainly because MvFSeg-64 conducts the parallel multiple views feature augmentation with smaller but effective channel number. The outstanding performance proves the effectiveness and efficiency of MvFSeg. Since MvFSeg fully exploits all level features from light-weight backbone networks to construct high-quality segmentation map with multi-view features, the deficiencies of light-weight backbone networks are addressed. Therefore, MvFSeg acts as a general framework for different light-weight backbone networks. The competitive experiment results of MvFSeg with six backbone networks, including GhostNet [15], MuxNet-m [27], MuxNet-l [27], ResNet18 [18], ResNet18-D [18] and ResNet34 [18], prove that the proposed method is appropriate for different light-weight backbone networks and achieves efficient and effective results.

4.5 Ablation Studies

Effect of MFA, PDFA and PMDC. To verify the effect of the proposed PDFA, MFA and PMDC, experiments are conducted with different settings on Cityscapes dataset. FPS of MvFSeg is tested on both 1024×2048 and 360×640 image resolution. The baseline is a FCN [26] model with ResNet18 [18] backbone. As listed in Table 2, the proposed method improves performance remarkably. With PDFA adopted, the original multi-level features F_4, F_8, F_{16} and F_{32} are exploited for effective multi-level feature aggregation. Compared with the baseline model, PDFA improves the accuracy by 4.3%. With MFA, MvFSeg eliminates feature gaps among features from different layers and passes spatial information as well as semantic context to all level features, which improves the feature quality and achieves more precise results with 7.8% mIoU improvement to the baseline model. By adopting PMDC as context head to integrate long-range context and generate higher-level features with multi-view receptive fields, the accuracy is further improved.

Effect of MvFSeg with different feature levels. In this subsection, different number of feature levels are selected to verify the effect of MvFSeg. The results on Cityscapes can be seen in Table 3. To verify the effect of original multi-level features, PMDC module that can generate higher-level features is not used. Therefore, F_{32} in Table 3 is the same to OF_{32} that from the stage 4 of the light-weight backbone network. The baseline model only exploits features of the highest level (F_{32}) for segmentation. When more feature levels are combined with MvFSeg structure to generate multi-view augmented features, the performance is significantly improved. Such result indicates the proposed feature augmentation structure is effective. Correspondingly, the consumption of computing resources is also increased, but affordable.

Table 3: Performance of different feature levels.

Method	F_{32}	F_{16}	F_8	F_4	mIoU%	FPS
Baseline	✓				68.9	**75 / 508**
Baseline+PDFA+MFA	✓				71.1	69 / 438
Baseline+PDFA+MFA	✓	✓			74.1	60 / 397
Baseline+PDFA+MFA	✓	✓	✓		75.6	49 / 311
Baseline+PDFA+MFA	✓	✓	✓	✓	**76.7**	34 / 230

Table 4: Performance of MFA.

Method	MFA				mIoU%	FPS
	Spatial	Semantic	Intra	Inter		
Baseline+PDFA					73.2	**39 / 310**
Baseline+PDFA	✓		✓	✓	73.8	36 / 251
Baseline+PDFA		✓	✓	✓	76.5	38 / 281
Baseline+PDFA	✓	✓	✓		75.4	36 / 250
Baseline+PDFA	✓	✓		✓	75.1	36 / 257
Baseline+PDFA	✓	✓	✓	✓	**76.7**	34 / 230

Table 5: Performance of PMDC.

Method	PMDC			mIoU%	FPS
	2^4	2^0-2^3	2^1-2^4		
Baseline+PDFA+MFA				76.7	**34 / 230**
Baseline+PDFA+MFA	✓			77.3	33 / 224
Baseline+PDFA+MFA		✓		77.3	33 / 225
Baseline+PDFA+MFA			✓	**78.0**	32 / 221

Table 6: MvFSeg with different context heads.

Method	Context Head			mIoU%	FPS
	PPM [48]	ASPP [5]	PMDC		
MvFSeg				76.7	**34 / 230**
MvFSeg	✓			76.9	33 / 226
MvFSeg		✓		77.8	32 / 223
MvFSeg			✓	**78.0**	32 / 221

Effect of MFA. The effects of spatial and semantic guidance are reported in Table 4, and both of them boost the accuracy significantly, especially the semantic guidance that provides high-level semantics to all level features. The effects of inter-layer and intra-layer augmentation are verified, respectively, with accuracy of 75.1% and 75.4% mIoU. When spatial guidance and semantic guidance are exploited with inter-layer and intra-layer feature augmentation strategies, sufficient spatial information and semantics are integrated to generate refined features, which boost the performance and achieve 76.7% mIoU on the Cityscapes validation set.

Effect of PMDC. To verify the performance of parallel multiple depths convolution, experiments of single path and multiple paths, as well as multiple paths with different depths are conducted. As can be seen in Table 5, compared to single path of 2^4 layers, multiple paths of 2^1 to 2^4 layers achieve better performance by integrating multi-view features. Multiple paths of 2^1 to 2^4 layers can produce larger receptive fields compared to multiple paths of 2^0 to 2^3 layers, thus better result is achieved.

MvFSeg with different context heads. To compare the performance of parallel multiple depths convolution (PMDC) with other context heads, experiments of MvFSeg with PMDC and Atrous Spatial Pyramid Pooling (ASPP) [5] as well as Pyramid Pooling Module (PPM) [48] are conducted. PPM and ASPP are famous and effective context heads. Specifically, the output of the light-weight backbone network is fed into PMDC, ASPP or PPM, and the output of the context head is adopted to provide higher-level semantics for features from the first three stages and features from the context head itself. The results of MvFSeg with different context heads are reported in Tabel 6. As the reported results, MvFSeg with PPM and ASPP achieves improvement on accuracy and is efficient. This mainly because both PPM and ASPP can capture multi-scale information for precise segmentation. MvFSeg with PMDC achieves the highest accuracy and also has real-time performance. Such result indicates PMDC is more powerful than PPM and ASPP in lightweight backbone networks, by generating multi-scale receptive fields and solving the deficiencies of lightweight backbone networks.

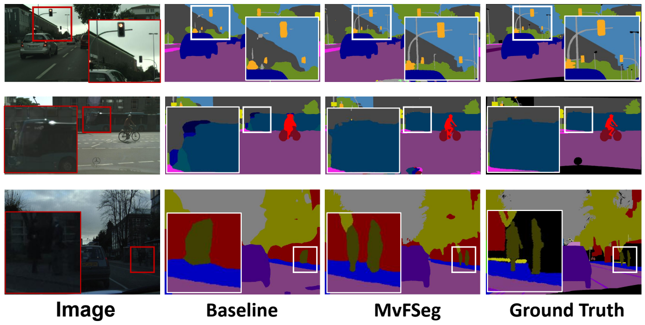

Figure 5: Visualization of the segmentation results from Cityscapes (first two rows) and CamVid (the last row).

Table 7: MvFSeg Efficiency.

Method	Channel Number				mIoU%	FPS	PM
	32	64	128	O			
Baseline					68.9	**75 / 508**	11.7
MvFSeg-32	✓				75.6	54 / 396	11.8
MvFSeg-64		✓			77.9	45 / 322	12.9
MvFSeg-128			✓		77.8	30 / 207	17.3
MvFSeg-O				✓	**78.0**	32 / 221	19.6

4.6 Network Parameters and Efficiency

The comparison of model parameters is listed in Table 1. As can be seen from this table, ENet [32] reduces the image resolution at the beginning of the network, having only 0.4M parameters. Methods such as MGSeg [17], SFNet [24] and HyperSeg [30] have around 7 to 16M parameters with well-designed light-weight networks. ICNet [47] and BiSeNet [42] have more than 26M parameters due to their multi-branch structure, inducing additional computations. Compared to the state-of-the-art methods, MvFSeg with GhostNet [15], MuxNet-m [27] and MuxNet-l [27] has small number of parameters, around 8M, due to the small channel number and feature map size. Taking ResNet18 [18] and ResNet18-D [18] as backbone, the model parameters of MvFSeg is increased to 19M, because of larger channel number and feature map size. MvFSeg with ResNet34 [18] has 28.6M parameters since ResNet34 is the largest network among the listed backbones. The model parameters of MvFSeg-64 are also listed in Table 1 for comparison. In MvFSeg-O, multi-level features from the backbone network are augmented and fused without reducing the channel number. The multi-level features in MvFSeg-64 are augmented after reducing the channel number to 64. Taking ResNet18 as backbone, MvFSeg-O has about 19.6M parameters, while MvFSeg-64 is 12.9M. With ResNet34 as backbone, MvFSeg-O has 28.7M parameters and MvFSeg-64 is 23.0M.

To verify the effect of the hyperparameter that is used to control the efficiency, experiments of different settings are conducted. As can be seen in Table 7, when the hyperparameter is set to 64, it achieves the best trade-off efficacy between the efficiency and accuracy. When it is set to 32, MvFSeg achieves the highest speed and induces only 0.1M extra parameters compared to the baseline model, but the accuracy is dropped compared to other MvFSeg models. This is because compressing channel number to 32 greatly decreases the spatial information and semantic context in multi-view features.

4.7 Visualization Results

Figure 5 shows visualization results of MvFSeg (ResNet18). Compared to the baseline (FCN [26] with ResNet18), the proposed method achieves more precise segmentation. The first row denotes the proposed approach makes full use of spatial information and has the ability of predicting details. The second row proves the proposed feature augmentation and aggregation structure enlarges the receptive field and large objects can be effectively predicted. The third row indicates MvFSeg has robust pixel-level semantic classification ability on small objects such as person in dark scenes that with blurred boundaries. The visualization results demonstrate the effectiveness of the proposed method.

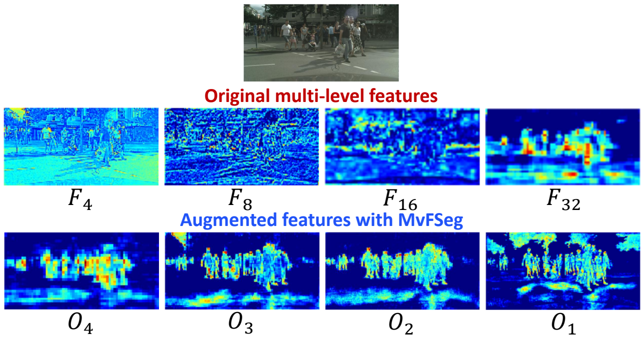

Figure 6: Visualization of heat maps from MvFSeg.

Figure 6 shows the heat maps of multi-level features before and after augmentation from MvFSeg (ResNet18). As illustrated, low-level features such as F_4 and F_8 extremely lack high-level semantics but have rich visual appearances such as edge and shape. While deep features (F_{32}) contain high-level semantics but are short of spatial information. With MvFSeg adopted, high-level and low-level features are integrated for the enhanced features, e.g., O_1 and O_2, which contain abundant spatial information for region localization and high-level semantics for category classification.

5 CONCLUSION

In this paper, a novel real-time semantic segmentation network called MvFSeg is proposed for real-world applications. It focuses on complemented high-level and low-level features integration for fine-grained multi-level feature generation, augmentation and aggregation, which aims to integrate multi-view features and achieve efficient and accurate performance. MvFSeg makes full use of all level features, by eliminating feature gaps among different features, filtering out noise of spatial details and refining all level features with spatial information and semantic context. Experiments on Cityscapes and CamVid demonstrate the effectiveness and efficiency of the proposed method, which achieves state-of-the-art performance on metrics of accuracy and efficiency. With the well-designed structure, MvFSeg can be adopted to different lightweight backbone networks for real-time semantic segmentation.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61772436 and 62001400), Sichuan Science and Technology Program (Grant No. 2020YJ0207 and 2021YJ0364), Foundation for Department of Transportation of Henan Province, China (2022-4-6 and 2019J-2-2), and Grant of Institute of Applied Physics and Computational Mathematics, Beijing (Grant No. HXO2020-118) and China Postdoctoral Science Foundation (Grant No. 2021M702713).

REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2015. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 12 (2015), 2481–2495.

[2] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. 2008. Segmentation and Recognition Using Structure from Motion Point Clouds. In *Proc. European Conference on Computer Vision.* 44–57.

[3] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. 2019. HarDNet: A Low Memory Traffic Network. In *Proc. IEEE International Conference on Computer Vision.* 3551–3560.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 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 (2017), 834–848.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proc. European Conference on Computer Vision.* 833–851.

[6] Zhi-Qi Cheng, Yang Liu, Xiao Wu, and Xian-Sheng Hua. 2016. Video Ecommerce: Towards Online Video Advertising. In *Proc. ACM international conference on Multimedia.* 1365–1374.

[7] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. 2017. Video ecommerce++: Toward large scale online video advertising. *IEEE transactions on multimedia* 19, 6 (2017), 1170–1183.

[8] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. 2017. Video2shop: Exact Matching Clothes in Videos to Online Shopping Images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 4048–4056.

[9] Zhi-Qi Cheng, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. 2017. On the selection of anchors and targets for video hyperlinking. In *Proc. ACM International Conference on Multimedia Retrieval.* 287–293.

[10] Wenqing Chu, Yao Liu, Chen Shen, Deng Cai, and Xian-Sheng Hua. 2018. Multi-Task Vehicle Detection With Region-of-Interest Voting. *IEEE Transactions on Image Processing* 27, 1 (2018), 432–441.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, and et al. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 3213–3223.

[12] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. 2021. Rethinking BiSeNet For Real-time Semantic Segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 9716–9725.

[13] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. 2019. Dual Attention Network for Scene Segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 3146–3154.

[14] Zhangxuan Gu, Li Niu, Haohua Zhao, and Liqing Zhang. 2020. Hard Pixel Mining for Depth Privileged Semantic Segmentation. *IEEE Transactions on Multimedia* 23 (2020), 3738–3751.

[15] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. 2020. GhostNet: More Features From Cheap Operations. In *Proc. European Conference on Computer Vision.* 1577–1586.

[16] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. 2022. Group Contextualization for Video Recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 928–938.

[17] Jun-Yan He, Shi-Hua Liang, Xiao Wu, Bo Zhao, and Lei Zhang. 2021. MGSeg: Multiple Granularity Based Real-time Semantic Segmentation Network. *IEEE Transactions on Image Processing* 30 (2021), 7200–7214.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

[19] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 7132–7141.

[20] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. CCNet: Criss-Cross Attention for Semantic Segmentation. In *Proc. IEEE International Conference on Computer Vision.* 603–612.

[21] Byeongkeun Kang, Yeejin Lee, and Truong Q. Nguyen. 2018. Depth-Adaptive Deep Neural Network for Semantic Segmentation. *IEEE Transactions on Multimedia* 20, 9 (2018), 2478–2490.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. Advances in Neural Information Processing Systems.* 1106–1114.

[23] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. 2019. DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 9522–9531.

[24] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. 2020. Semantic Flow for Fast and Accurate Scene Parsing. In *Proc. European Conference on Computer Vision.* 775–793.

[25] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan Yuille, and Fei-Fei Li. 2019. Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 82–92.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 3431–3440.

[27] Zhichao Lu, Kalyanmoy Deb, and Vishnu Naresh Boddeti. 2020. MUXConv: Information Multiplexing in Convolutional Neural Networks. In *Proc. European Conference on Computer Vision.* 12041–12050.

[28] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Shapiro, Linda G., and Hannaneh Hajishirzi. 2018. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. In *Proc. European Conference on Computer Vision.* 561–580.

[29] David Nilsson and Cristian Sminchisescu. 2018. Semantic Video Segmentation by Gated Recurrent Flow Propagation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 6819–6828.

[30] Yuval Nirkin, Lior Wolf, and Tal Hassner. 2021. HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 4061–4070.

[31] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. 2015. Learning Deconvolution Network for Semantic Segmentation. In *Proc. IEEE International Conference on Computer Vision.* 1520–1528.

[32] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. 2016. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv preprint arXiv:1606.02147* (2016).

[33] Xiaojiang Peng and Cordelia Schmid. 2016. Multi-region Two-Stream R-CNN for Action Detection. In *Proc. European Conference on Computer Vision.* 744–759.

[34] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. 2016. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. *Computer Vision and Image Understanding* 150 (2016), 109–125.

[35] Jian-Jun Qiao, Xiao Wu, Jun-Yan He, Wei Li, and Qiang Peng. 2022. SWNet: A Deep Learning Based Approach for Splashed Water Detection on Road. *IEEE Transactions on Intelligent Transportation Systems* 23, 4 (2022), 3012–3025.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proc. Medical Image Computing and Computer-Assisted Intervention.* 234–241.

[37] Haiyang Si, Zhiqiang Zhang, and Feng Lu. 2020. Real-Time Semantic Segmentation via Multiply Spatial Fusion Network. In *Proc. British Machine Vision Conference.*

[38] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. International Conference on Learning Representations.*

[39] Yuhang Wang, Jing Liu, Yong Li, Junjie Yan, and Hanqing Lu. 2016. Objectness-aware Semantic Segmentation. In *Proc. ACM International Conference on Multimedia.* 307–311.

[40] Peng Ying, Jin Liu, Hanqing Lu, and Songde Ma. 2015. Exclusive Constrained Discriminative Learning for Weakly-Supervised Semantic Segmentation. In *Proc. ACM International Conference on Multimedia.* 1251–1254.

[41] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. 2021. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *International Journal of Computer Vision* 129, 11 (2021), 3051–3068.

[42] Changqian Yu, Jingbo Wang, Chao Peng, and et al. 2018. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In *Proc. European Conference on Computer Vision.* 334–349.

[43] Changqian Yu, Jingbo Wang, Chao Peng, and et al. 2018. Learning a Discriminative Feature Network for Semantic Segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 1857–1866.

[44] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. 2021. Token shift transformer for video classification. In *Proc. ACM International Conference on Multimedia.* 917–925.

[45] Shanshan Zhang, Jian Yang, and Bernt Schiele. 2018. Occluded Pedestrian Detection Through Guided Attention in CNNs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 6995–7003.

[46] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. 2018. Multi-view Image Generation from a Single-view. In *Proc. ACM international conference on Multimedia.* 383–391.

[47] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. 2018. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *Proc. European Conference on Computer Vision.* 418–434.

[48] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid Scene Parsing Network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 6230–6239.

[49] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. 2021. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition.* 6881–6890.