electronics

MDPI

Article

MFSNet: Enhancing Semantic Segmentation of Urban Scenes with a Multi-Scale Feature Shuffle Network

Xiaohong Qian [1], Chente Shu [1,2], Wuyin Jin [3], Yunxiang Yu [4] and Shengying Yang [1,3,*]

[1] Department of School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China
[2] Zhejiang Development & Planning Institute, Hangzhou 310030, China
[3] School of Mechanical and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, China
[4] Zhejiang Dingli Industrial Co., Ltd., Lishui 321400, China
* Correspondence: syyang@zust.edu.cn

Abstract: The complexity of urban scenes presents a challenge for semantic segmentation models. Existing models are constrained by factors such as the scale, color, and shape of urban objects, which limit their ability to achieve more accurate segmentation results. To address these limitations, this paper proposes a novel Multi-Scale Feature Shuffle NetWork (MFSNet), which is an improvement upon the existing Deeplabv3+ model. Specifically, MFSNet integrates a novel Pyramid Shuffle Module (PSM) to extract discriminative features and feature correlations, with the objective of improving the accuracy of classifying insignificant objects. Additionally, we propose an efficient feature aggregation module (EFAM) to effectively expand the receptive field and aggregate contextual information, which is integrated as a branch within the network architecture to mitigate the information loss resulting from downsampling operations. Moreover, in order to augment the precision of segmentation boundary delineation and object localization, we employ a progressive upsampling strategy for reinstating spatial information in the feature maps. The experimental results show that the proposed model achieves competitive performance, achieving 80.4% MIoU on the Pascal VOC 2012 dataset, 79.4% MIoU on the Cityscapes dataset, and 40.1% MIoU on the Coco-Stuff dataset.

Keywords: semantic segmentation; contextual information; pyramid pooling module; attention mechanism; multi-scale fusion

Citation: Qian, X.; Shu, C.; Jin, W.; Yu, Y.; Yang, S. MFSNet: Enhancing Semantic Segmentation of Urban Scenes with a Multi-Scale Feature Shuffle Network. *Electronics* **2024**, *13*, 12. https://doi.org/10.3390/electronics13010012

Academic Editor: Hüseyin Kusetogullari

Received: 19 October 2023
Revised: 27 November 2023
Accepted: 16 December 2023
Published: 19 December 2023

1. Introduction

Semantic segmentation is a crucial aspect of image processing tasks in the field of electronics, which is important for image comprehension; it aims to assign class labels to each individual pixel within an image and make a finer classification of the image. With the advancement of semantic segmentation techniques and the breakthroughs achieved through weakly supervised learning [1,2], semantic segmentation techniques find extensive applications in various domains of electronics, such as autonomous driving, where they enable accelerated signal processing through the utilization of semantic segmentation. In the process of segmenting images in urban scenes, several challenges arise related to complex spatial relationships, irregular layout positions, and multi-scale objects. The traditional method [3] for making predictions at a single scale is insufficient to achieve robust segmentation and address the problem of pixel-level object consistency across different scenes. Therefore, incorporating multi-scale information aggregation is necessary to enhance the performance of the model to accurately locate and detect objects.

The integration of multi-scale information has received attention in subsequent studies [4–8]. PSPNet [9] introduced a novel pooling structure known as the Pyramid Pooling Module (PPM), which aims to effectively integrate multi-scale contextual information, enabling the precise segmentation of objects of different scales in complex scenes. ForkNet [10] employed a novel multi-scale fusion approach to capture scale variations of objects by

integrating a Siamese feature pyramid network. LSTNet [11] introduces the Pyramid Texture Feature Extraction Module (PTFEM) to effectively extract multi-scale statistical texture features. The Atrous Spatial Pyramid Pooling (ASPP) method, introduced in [12], is designed to effectively enlarge the receptive field and fuse multi-scale information in semantic segmentation. However, due to the sparse downsampling effect caused by dilated convolutions, there is a potential risk of losing spatial details and fine-grained information, consequently leading to the issue of unclear segmentation boundaries in the resulting model [13]. Furthermore, this method also exhibits limitations when segmenting neighboring objects with similar attributes such as color and texture.

As the importance of attention mechanisms has been demonstrated in previous research literature [14], it has been widely applied in several tasks for semantic segmentation, such as [15], who combined [14] and multilayer perception decoders for image semantic segmentation. The objective is to extract more useful features by assigning higher weights to feature representations with significant information while suppressing the weights of less informative ones. OCNet [16] combined attention mechanisms with ASPP to extract contextual dependencies. GCNet [17] unified the SENet [18] into a framework to model the global context. However, these methods are susceptible to noise interference in urban scenes, which can result in the erroneous establishment of long-range pixel dependencies. Refs. [19–21] leveraged the Swin Transformer [22] to construct hierarchical feature maps and perform self-attention computations for semantic segmentation. Nevertheless, the approach of partitioning the feature maps into windows restricts the establishment of inter-window feature connections, which restricts the ability of the model to comprehensively capture contextual information. Additionally, the lack of effective integration of attention mechanisms in multi-scale feature extraction resulted in a weak ability to extract discriminative features in the case of indistinct urban scene targets. As a consequence, this limitation leads to a misclassification of pixels and discontinuous segmentation results.

To overcome these limitations, we propose a novel pyramid shuffle module (PSM) integrates channel and spatial attention mechanisms and utilizes channel shuffling operations on feature maps to extract informative feature representations. In addition, in order to mitigate the loss of spatial and fine-grained information caused by the downsampling process, this paper introduces an efficient feature aggregation module (EFAM) as a secondary branch of the model, which employs a multi-layer fusion strategy to jointly model and complement features of various characteristics and expand the receptive field. MFSNet enhances the segmentation accuracy of elongated objects or neighboring objects with similar features by establishing long-range dependencies between key pixels and utilizing feature reuse mechanisms.

The main contributions are summarized as follows:

- To mitigate the issue of segmentation errors in urban scenes resulting from the presence of neighboring objects with similar features such as texture or color. We introduce a Pyramid Shuffle Module (PSM), which improving the multi-scale feature representations and segmentation robustness of the network by facilitating channel interaction among multi-scale features and highlighting discriminative features.
- To achieve precise segmentation of object boundaries. We proposed an efficient feature aggregation module (EFAM), serving as a branch for network multi-layer feature fusion in the network, to compensate for feature loss caused by pyramid pooling and to facilitate network backpropagation.
- Extensive experimental results on Pascal VOC 2012, Cityscapes, and Coco-Stuff datasets reveal that the proposed method exhibits a good generalization ability.

2. Methods

In this section, we first introduce the overall structure of the model. Then, we provide a detailed explanation of the different modules used for constructing the network.

2.1. Overview

As shown in Figure 1, the Multi-scale Feature Shuffle Network (MSFNet) used for urban scenes was constructed under a generic encoder–decoder framework. The encoder performed downsampling on the input image to capture rich semantic information. The learned high-level features were then decoded and reconstructed through the decoder for pixel-level semantic prediction. Inspired by the aforementioned limitations of the existing ASPP method, MSFNet introduced PSM and an EFAM to process the feature maps extracted by ResNet in the encoder. Specifically, PSM integrates a channel shuffle attention mechanism, leveraging parallel strategies to construct spatial and channel attention. This mechanism effectively utilizes the multi-scale context information extracted by the pyramid pooling module, suppressing noise and accentuating discriminative semantic regions. Additionally, in combination with EFAM in the sub-branch, the model expands the receptive field while complementing the output features of the main branch.

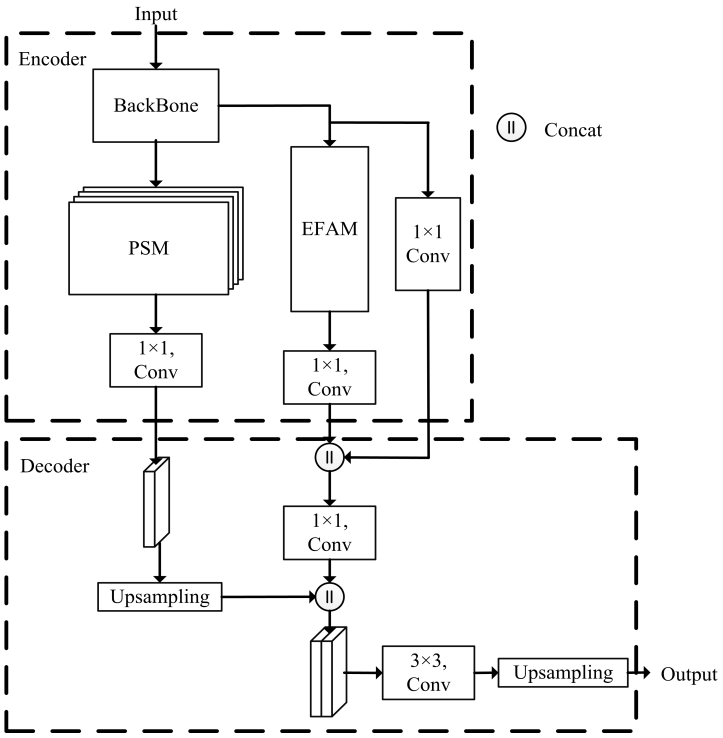


Figure 1. Overview of our proposed MFSNet for semantic segmentation. "PSM" denotes the pyramid shuffle module. "EFAM" denotes the efficient feature aggregation module.

During the decoding stage, MFSNet divides the upsampling process into two parts. Firstly, the feature maps output by the PSM were upsampled by a factor of 4. Subsequently, they were complemented by fusing them with the output feature maps from EFAM. This approach effectively avoided the loss of fine-grained information caused by high magnification upsampling. It is important to note that upsampling was not about fully restoring the image resolution, but rather smoothly recovering a feature map of a credible initial size using a computational formula. The refined features were further processed by a 3 × 3 convolutional layer, followed by another 4-fold upsampling. Finally, the features were mapped to classes, allowing the class mapping to be rescaled back to the input resolution. This model was constructed using a multi-strategy fusion approach, integrating functionalities such as multi-scale fusion, discriminative feature highlighting, and feature reuse. These functionalities ensured the extraction of high-level semantic information and the restoration of high-resolution details during the training process.

2.2. Pyramid Shuffle Model

PSM incorporates dilated convolutions in pyramid pooling, enabling the model to flexibly adjust the receptive field size and fuse multi-scale object feature information during feature extraction. As depicted in Figure 2, to enhance the model's comprehension of images, the approach employed parallel dilated convolution layers with dilation rates of 1, 12, 24, and 36, along with a global average pooling layer. After these operations, the model effectively captured local multi-scale information while integrating the global semantic information of the image. Subsequently, the resulting feature map M was obtained. Inspired by SANet [23], this module divided the feature map M, which had dimensions of $W \times H$ and C channels, into G groups based on the channel dimension. This process yielded a collection of feature maps, referred to as X, $X = [X_1, \ldots, X_G]$, $X_i \in \mathbb{R}^{W \times H \times C/G}$. Then, each sub-feature X_i was further split into two branches based on the channel dimension, which yielded two distinct feature sets, denoted as $X_{i1}, X_{i2} \in \mathbb{R}^{W \times H \times C/2G}$. These two branches integrated the channel attention mechanism and spatial attention mechanism to capture discriminative features. Specifically, the channel attention branch incorporated the global average pooling (GAP) to compute the average value of each channel in the feature map, generating global feature information denoted as $y_{i1} \in \mathbb{R}^{1 \times 1 \times C/2G}$. Subsequently, y_{i1} was enhanced using function F_c by $Wy + b$, where W and b are parameters can be updated by network training. Unlike SANet, PSM adopts the Hardsigmoid H_σ activation to expedite the convergence speed of the model and extract channel dependencies. The calculation process is as follows:

$$\alpha_{i1} = H_\sigma(W_1 y_{i1} + b_1) \times X_{i1} \tag{1}$$

where $W_1, b_1 \in \mathbb{R}^{1 \times 1 \times C/2G}$ adjusts the weights of each channel and highlights more significant features. This mechanism helps mitigate the interference from redundant information in the feature representation.

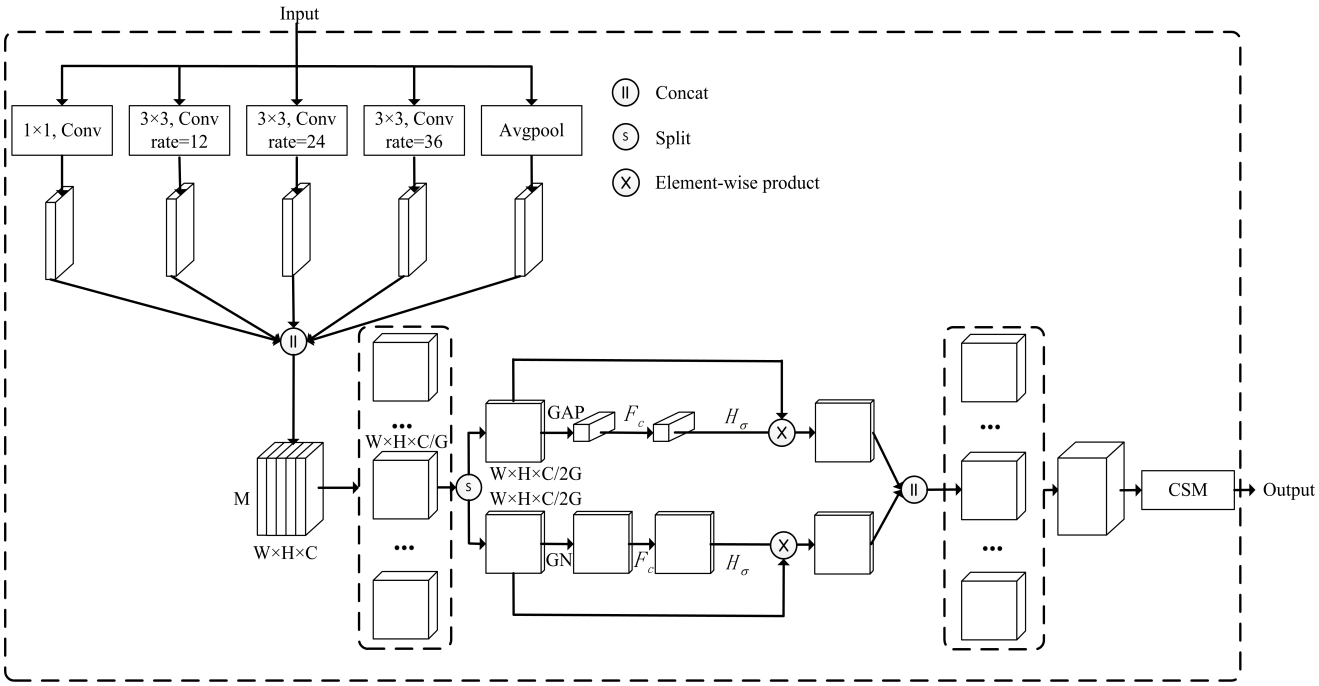


Figure 2. Pyramid shuffle module.

The spatial attention branch, in contrast, was utilized to focus on regions within the feature map that contained rich semantic information. It enhanced the capacity of the model to perceive local details and accurately locate objects in space. Firstly, $y_{i2} \in \mathbb{R}^{1 \times 1 \times C/2G}$ is obtained by GroupNorm using mean and variance calculations. Subsequently,

the spatial attention vector is computed to enhance the feature representation of y_{i2}. Lastly, the Hardsigmoid activation is applied to accelerate the convergence speed of the model:

$$\alpha_{i2} = H_\sigma(W_2 y_{i2} + b_2) \times X_{i2} \tag{2}$$

where $W_2, b_2 \in \mathbb{R}^{1\times1\times C/2G}$ are parameters that can also be updated by network learning. Subsequently, the sub-features α_{i1} and α_{i2} are combined along the channel dimension to yield the feature $\alpha_i \in \mathbb{R}^{W\times H\times C/G}$.

The G sub-features are aggregated and subjected to channel shuffling operations to facilitate inter-channel information interaction, as depicted in Figure 3. The input feature maps are split into three groups based on the channel dimension. Secondly, a dimension reshape is applied to compose a new matrix, followed by a transpose operation and flattening to accomplish channel shuffling. This operation effectively enhances the feature representation capabilities of the multi-scale and contextual information extracted through pyramid pooling. By enabling mutual influence among feature channels with similar semantics, it suppresses interference from redundant information and mitigates classification errors during segmentation, ultimately improving the robustness of the segmentation process.

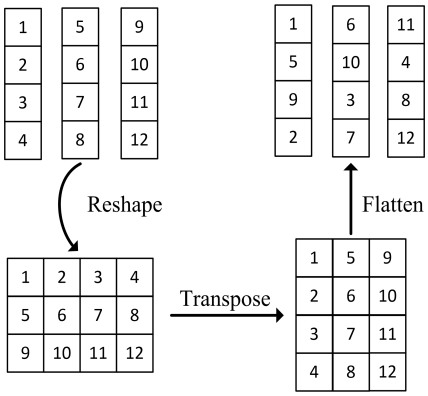


Figure 3. Channel shuffle module.

2.3. Efficient Feature Aggregation Module

In previous studies on semantic segmentation [24,25], it has been demonstrated that expanding the receptive field is beneficial for improving the performance of semantic segmentation models. EFAM has been presented as a feature reuse branch in the network. It effectively mitigates the loss of spatial and detail information both internally and externally, leading to a notable enhancement in the richness of the extracted semantic information, as illustrated in Figure 4.

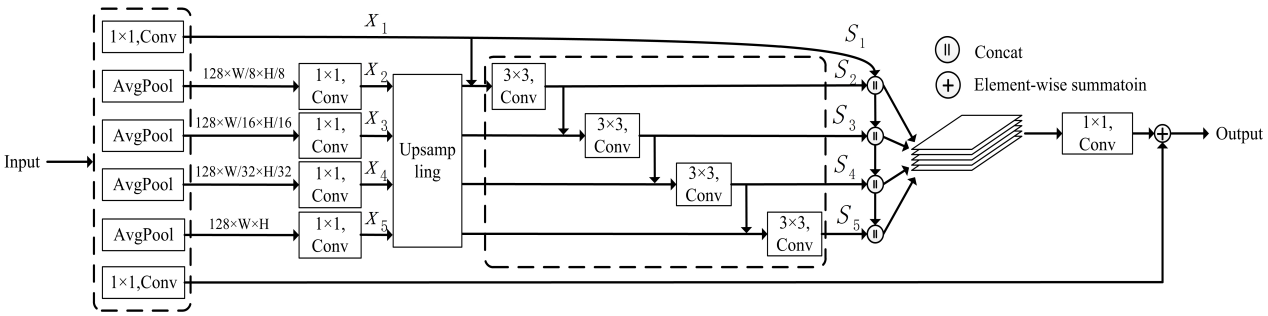


Figure 4. Efficient feature aggregation module.

EFAM receives the feature maps of the input image at a 1/4 resolution from the backbone. Subsequently, it undergoes multiple branches of average pooling to extract feature maps at resolutions of 1/8, 1/16, and 1/32. Additionally, a global average pooling

operation is performed to integrate spatial information, generating image-level information, which is subsequently used for upsampling. Inspired by the hierarchical residual-like connections in Res2Net [26], we introduced a multi-branch network architecture based on 3×3 convolutions to successively fuse multi-scale contextual information. The output feature maps at each scale s_i are expressed by Equation (3):

$$s_i = \begin{cases} x_i & , i = 1 \\ C_{3\times3}(Up(x_i) + x_{i-1}), & 1 < i \leq n \end{cases} \quad (3)$$

where $C_{3\times3}$ represents a 3×3 convolution and Up denotes bilinear interpolation upsampling. With the exception of x_1 and s_1, which are directly mapped without any operations, each input x_i undergoes an addition operation with its corresponding x_{i-1}, followed by a 3×3 convolution. Subsequently, the feature maps obtained from each branch are concatenated and then subjected to a 1×1 convolutional layer for dimensionality reduction. In the end, the output feature maps will be obtained by performing element-wise summation operations.

This strategy of splitting and reconnecting greatly enhances the capacity of the module to extract both the global and local information, substantially improving feature extraction and processing in the semantic segmentation tasks. Furthermore, ref. [27] provides a detailed discussion on the combination of traditional convolution, normalization, and activation, presenting a comprehensive pre-activation design. In the construction of EFAM, a similar sequential combination of BN-ReLu-Conv was employed, which effectively reduced the model overfitting and enhanced model generalization.

3. Experimental Results

3.1. Datasets

The Pascal VOC 2012 [28] benchmark contains 20 foreground object classes and 1 background class. The original dataset consisted of 1464 pixel-level annotated images for training, 1449 for validation, and 1456 for testing. In addition, the dataset was augmented with extra annotations provided by [29], resulting in 10,582 training augmented images that were divided into 21 classes.

Cityscapes [30] is one of the more well-known scene semantic segmentation datasets, focusing on the analysis of urban street scenes. It consists of 5000 high-quality pixel-level finely annotated images collected from 50 cities, which are divided into 2975 images for training, 500 images for validation, and 1525 images for testing, with a total of 19 classes. In addition, we did not use its extra 20,000 coarse labeled images during training.

Coco-Stuff [31] comprises a comprehensive collection of 10,000 annotated images, with 9000 images allocated for training and 1000 images for testing. Compared with Cityscapes and Pascal VOC 2012, the Coco-Stuff dataset is more challenging due to its more complex classes, includes 80 thing classes, 91 stuff classes, and 1 class labeled as 'unlabeled'.

The experimental design in this paper aligns with the approach in [32–37]. We conducted experiments on the validation sets of the Pascal VOC 2012 and Cityscapes datasets to thoroughly analyze the improvements and contributions of the proposed model. Furthermore, we further validated the effectiveness of MFSNet on the COCO-Stuff test set, demonstrating its performance in semantic segmentation tasks. The experiments employed Mean Intersection over Union (MIoU) and Mean Accuracy (MAcc) as the evaluation criteria for assessing the segmentation performance. A higher MIoU and MAcc value suggested a more precise image segmentation.

3.2. Train Setting

The training and validation of this experiment were conducted on the Ubuntu 20.0.4.01 environment, using the PyTorch framework version 1.8.1. The training was performed on a GeForce RTX 3090 with 24GB of memory. All of the experiments were conducted with exactly the same data augmentation. Specifically, Pascal VOC 2012 was cropped to a 440×440 resolution, Cityscapes was cropped to 769×769 resolution and the Coco-Stuff

was cropped to a 380 × 380 resolution. We trained the datasets with cross-entropy loss function and a stochastic gradient descent (SGD) optimizer, and optimized the network by adopting the poly learning rate Equation (4). The initialization of parameters W_1 and W_2 were set to 1, and b_1 and b_2 were set to 0.

$$I = I_{init} \times (1 - \frac{epoch}{epoch\_max})^{power} \quad (4)$$

where *epoch_max* represents the maximum number of epochs, I represents the learning rate, I_{init} represents the initial learning rate, and *power* is set to 0.9.

3.3. Ablation Study

To validate the feasibility and effectiveness of this model, this study evaluated the contribution of each module to improving the overall accuracy through a series of experiments. In order to ensure fairness in the results, ResNet101 was uniformly chosen as the backbone for this model during the experiments. Considering that the PSM in our model was improved based on ASPP in Deeplabv3+, we incorporated ASPP into the experiments for the comparative analysis.

In Table 1, "Non-EFAM" refers to the absence of the EFAM as the feature reuse branch, while "EFAM" represents the adoption of EFAM as the feature reuse branch. By analyzing the table, it becomes apparent that both introduced modules significantly enhanced the segmentation performance. PSM introduced in this model achieved 79.7% and 80.4% MIoU, as well as 88.7% and 89.3% MAcc, in the Non-EFAM and EFAM cases, respectively. Compared with ASPP, PSM achieved a superior 1.3% and 1.0% MIoU, as well as 1.5% and 1.0% MAcc. Notably, when ASPP was replaced with PSM, this resulted in a minimal increase in model parameters, while significantly improving the performance of model. Furthermore, the addition of EFAM as the feature reuse branch led to a slight increase in 0.8M in the model parameters. However, this trade-off yielded a substantial improvement in segmentation accuracy.

Table 1. Ablation analysis of the Pascal VOC 2012 val set.

Method	BackBone	ASPP	PSM	Params. (M)	MIoU (%)	MAcc (%)
Non-EFAM	ResNet101	✓		59.3	78.4	87.2
Non-EFAM	ResNet101		✓	59.3	79.7	88.7
EFAM	ResNet101	✓		60.1	79.4	88.3
EFAM	ResNet101		✓	60.1	80.4	89.3

In addition, Figure 5 illustrates the effectiveness of each module in a more intuitive manner through the MIoU curve. The curve clearly demonstrates that the inclusion of each module significantly enhanced the accuracy of the segmentation model. Analyzing the curve trends revealed that the ASPP gradually tendedtowards overfitting after 80 training epochs. Conversely, the MIoU curve of MFSNet consistently exhibited a steady upward trend without any indications of overfitting, although the rate of improvement gradually slowed down.

As the proposed method in this study was an improvement based on ASPP, we further compared it with Deeplabv3+ in terms of precision, recall, and F1 value to validate its effectiveness. As shown in Table 2, MFSNet achieved precision, recall, and F1 values that were 1.4%, 2.7%, and 2.1% higher than Deeplabv3+, respectively, indicating that our method exhibited a higher accuracy. Additionally, to ensure comprehensive experimental results, we further computed the false negative rate (FNR) for MFSNet, which resulted in 10.7%.

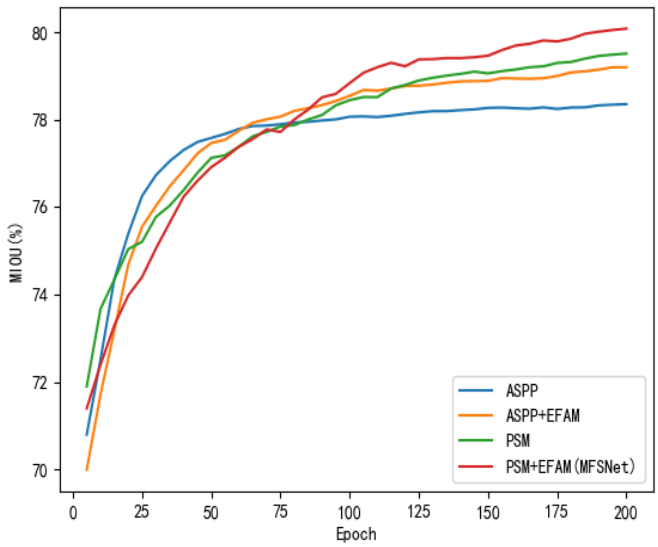


Figure 5. Curve of MIoU for ablation experiments on the Pascal VOC 2012 val set.

Table 2. Ablation analysis of Deeplabv3+ and MFSNet.

Method	BackBone	Precision (M)	Recall (%)	F1 Score (%)
Deeplabv3+	ResNet101	86.5	86.6	86.5
MFSNet	ResNet101	87.9	89.3	88.6

We conducted comparative ablation experiments on the visual results of ASPP, PSM, and the entire MFSNet model, as depicted in Figure 6. A comparative analysis of the second and third rows clearly demonstrated that MFSNet outperformed the ASPP method, exhibiting a superior segmentation efficacy. An analysis of the first and fourth rows clearly demonstrated the superior performance of our model compared with the existing ASPP method. The incorporation of PSM effectively enhanced the discriminative features of "person" and "motorcycle", successfully mitigating the issue observed in existing methods where neighboring objects with similar features were incorrectly identified as the same object. Moreover, observations from the first and forth rows revealed that the model efficiently integrated multi-scale information, establishing long-range dependencies among critical pixels. As a result, it accurately segmented "person" within the "car" and produced more complete outlines for smaller objects. Comparing column (d) with column (e), it becomes evident that our model exhibited a notable enhancement in robustness by leveraging EFAM as a feature reuse branch to recover spatial information. This improvement effectively reduced the occurrence of discontinuity during the segmentation process. The visualizations clearly demonstrate that the introduced modules in our model significantly improved the accuracy of object segmentation.

We investigated the impact of the pre-activation structure in EFAM in terms of accuracy. Observing Table 3 reveals that when utilizing the shallow ResNet50 as the backbone, the influence of post-activation and pre-activation was relatively minor, with a marginal difference of approximately 0.1% in MIoU. However, as the depth of the backbone increased, the disparity became more pronounced, with the pre-activation scheme surpassing the post-activate counterpart by 0.2% MIoU. These results indicate the efficacy of incorporating the pre-activation structure in our network, as it contributed to an improvement in the model performance.

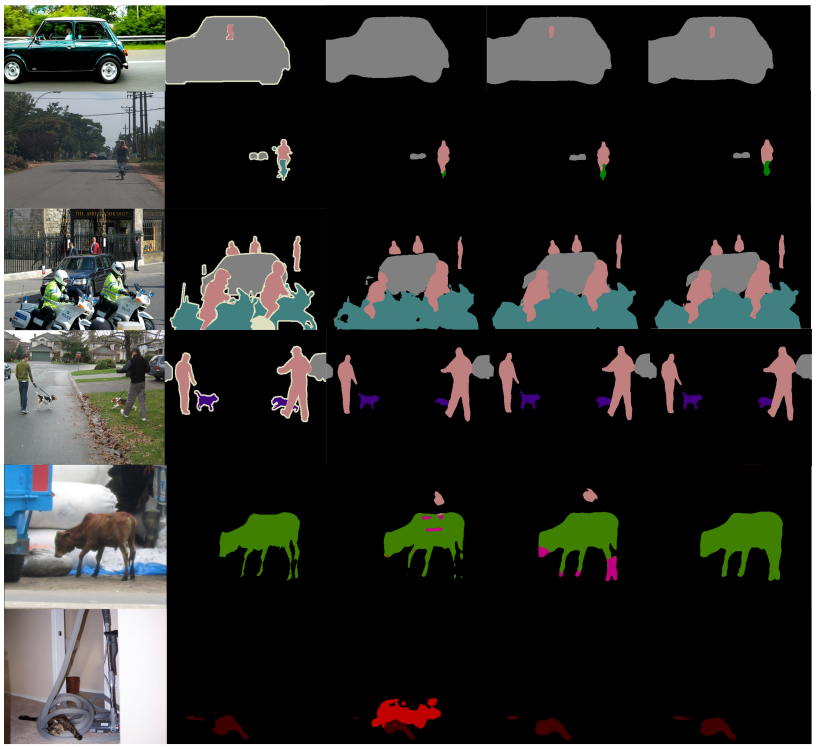


Figure 6. Visual comparison of segmentation results for the ablation experiments.

Table 3. Ablation study for pre-activate structure.

BackBone	Post-Activation	Pre-Activation	MIoU (%)
ResNet50	✓		78.8
ResNet50		✓	78.9
ResNet101	✓		80.2
ResNet101		✓	80.4

PSM is an improvement based on Shuffle Attention (SA), as it is theoretically reasoned that SA can efficiently combine the channel and spatial attention through its multi-branch structure, which enables the extraction of the discriminative feature information, while increasing the feature diversity, effectively enhancing the robustness of the segmentation model. Experimental comparisons with other attention mechanisms, as illustrated in Table 4, unequivocally demonstrate that our proposed method yielded the most substantial performance improvement. Specifically, under the condition of a comparable parameter count of 60.1M to the ECA and SA methods, our approach outperformed CBAM, SE, ECA, and SA by 0.5%, 0.6%, 0.5%, and 0.4% in terms of MIoU, respectively.

Table 4. Ablation study for attention mechanisms.

Method	Params. (M)	MIoU (%)
ASPP + CBAM [38]	60.4	79.9
ASPP + SE [18]	60.4	79.8
ASPP + ECA [39]	60.1	79.9
ASPP + SA [23]	60.1	80.0
PSM	60.1	80.4

3.4. Results on Pascal VOC 2012

This experiment evaluated the proposed approach on the validation set of the Pascal VOC 2012 dataset, comparing it with other methods. To improve the training efficiency, the input image resolutions were cropped during both the training and evaluation processes. As presented in Table 5, MFSNet achieved 78.9% and 80.4% MIoU on ResNet50 and ResNet101, respectively, while also achieving 87.8% and 89.3% MAcc. Furthermore, the study [40] found that larger image sizes during the training and validation resulted in improved segmentation accuracy, as indicated by previous research. It is noteworthy that the image sizes utilized in our method were comparatively smaller than those employed by other state-of-the-art models. Nevertheless, MFSNet still managed to outperform these advanced methods, achieving the highest MIoU and MAcc, which further validated the superiority of our method. Additionally, even when the backbone was changed to ResNet50, MFSNet demonstrated a superior performance compared with the majority of existing models.

Table 5. Comparison of PASCAL VOC 2012 val set results with other models.

Method	Backbone	Resolution	MIoU (%)	MAcc (%)
CCNet [41]	ResNet101	512 × 512	77.8	85.1
GCNet [17]	ResNet101	512 × 512	77.8	86.0
PSPNet [9]	ResNet101	512 × 512	78.5	87.0
PSANet [42]	ResNet101	512 × 512	77.7	85.0
DeepLabv3+ [12]	ResNet101	769 × 769	78.4	86.0
OCRNet [43]	HRNetV2p-W48	512 × 512	77.1	85.9
MARS [44]	ResNet101	-	77.7	-
WASS-SAM [2]	ResNet101	-	77.2	-
MFSNet (Ours)	ResNet50	440 × 440	78.9	87.8
MFSNet (Ours)	ResNet101	440 × 440	80.4	89.3

We conducted comparative evaluations to validate the superior segmentation results of MFSNet against DeepLabv3+ and PSPNet, as shown in Figure 7. The findings reveal that MFSNet exhibited significant advantages in terms of both distinguishing visually similar categories, such as "motorcycle" and "person", and maintaining the overall integrity of the segmented objects. While DeepLabv3+ efficiently captured multi-scale contextual information through the integration of ASPP, it failed to address the issue of feature information loss caused by sparse downsampling and lacked the ability to capture critical pixel-level feature dependencies, resulting in incomplete segmentation of complex objects or neighboring objects with similar features. This deficiency was evident in the first row where the segmentation result exhibited discontinuity. Similarly, PSPNet, despite incorporating a pyramid pooling module to extract contextual information by performing multiple pooling operations at different scales, suffered from the loss of fine-grained feature details due to repeated downsampling. Consequently, the model was prone to noise interference during segmentation and exhibited imprecise object edge delineation, similar to DeepLabv3+. In contrast, MFSNet effectively mitigated feature information loss through the deep aggregation of global and local feature information, while establishing long-range dependencies between the pixels to extract discriminative features. This approach significantly enhanced segmentation accuracy. Comparing the third and fourth rows further demonstrates that MFSNet surpassed DeepLabv3+ and PSPNet in segmenting edges and small-scale objects.

3.5. Results on Cityscapes

To further demonstrate the good generalizability of our semantic segmentation model in urban scenes, we conducted experiments on the Cityscapes dataset, as presented in Table 6. Our method with ResNet50 as a backbone achieved 78.5% MIoU and 87.4% MAcc, while it achieved 79.4% MIoU and 88.3% MAcc with ResNet101. A comprehensive evalua-

tion of both MIoU and MAcc metrics revealed that our model outperformed DeepLabv3+ by a slight margin of 0.4% in terms of MIoU, while surpassing it by a noteworthy 1.6% in terms of MAcc. The experimental results on both datasets unequivocally demonstrated the outstanding performance of our proposed network, underscoring its substantial generalizability in urban scenes. While the MIoU of our method on the Cityscapes dataset exhibited a marginal improvement of only 0.4% over DeepLabv3+, it outperformed DeepLabv3+ by a substantial margin of 2.0% in terms of MIoU on the Pascal VOC 2012 dataset.

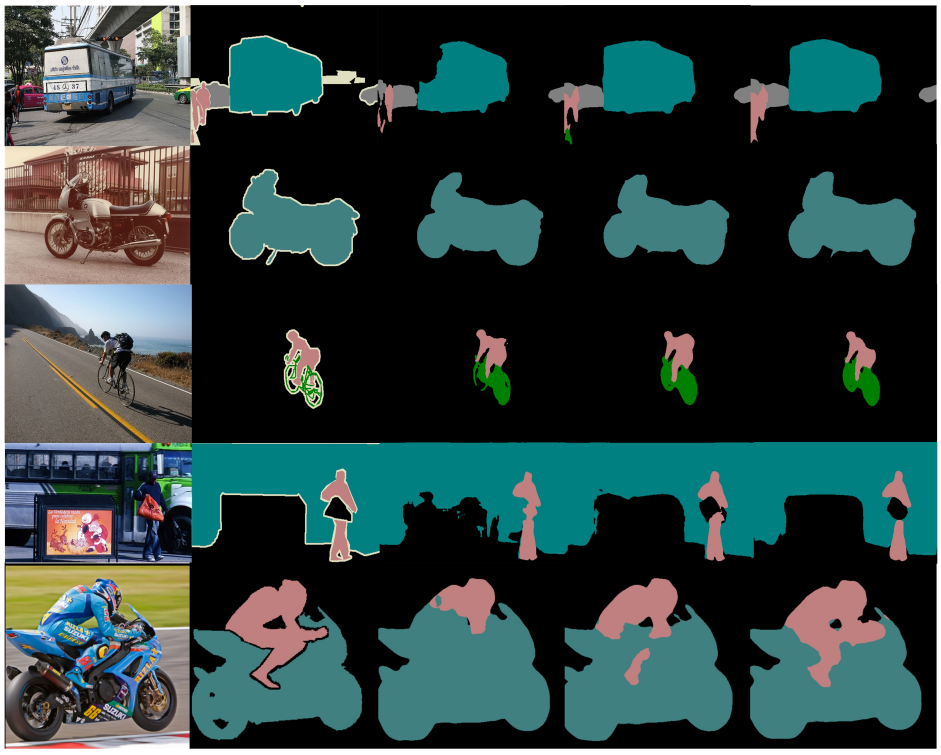


Figure 7. Visual comparison of the segmentation results with the other methods on PASCAL VOC 2012.

Table 6. Comparison of the Cityscapes val set results with other models.

Method	Backbone	Resolution	MIoU (%)	MAcc (%)
GCNet [17]	ResNet101	512 × 1024	78.2	85.7
PSPNet [9]	ResNet101	512 × 1024	78.3	85.3
DMNNet [6]	ResNet101	769 × 769	77.6	86.4
CCNet [41]	ResNet101	769 × 769	76.9	84.9
DeepLabv3+ [12]	ResNet101	769 × 769	79.0	87.0
PointRend [45]	ResNet101	512 × 1024	78.3	85.7
APCNet [46]	ResNet101	769 × 769	77.9	87.1
PSANet [42]	ResNet101	769 × 769	78.4	87.4
Multiscale DEQ [47]	MDEQ-large	769 × 769	77.8	-
UOIFT [7]	UOIFT	-	78.0	-
STDC [8]	STDC2	512 × 1024	76.7	84.0
BiSeNetV2 [48]	BiSeNetV2	1024 × 1024	75.7	83.4
StreamDEQ [49]	MDEQ-iter8	768 × 768	78.2	-
EEA-NEt-C2 [50]	EEA-NEt-C2	320 × 320	76.8	-
MFSNet (Ours)	ResNet50	769 × 769	78.5	87.4
MFSNet (Ours)	ResNet101	769 × 769	79.4	88.3

Although our method on the Cityscapes dataset was only 0.4% MIoU higher than Deeplabv3+, in order to demonstrate the effectiveness of our model at extracting discriminative information and improving the segmentation of elongated objects, we compared the IoU scores of the original model and MFSNet on different categories of the Cityscapes dataset. As shown in Table 7, our model performed exceptionally well in all categories, except for "road", "sky", "truck", and "train", where the accuracy was lower compared with the original model. Specifically, our model achieved a significantly higher Intersection Over Uion (IoU) on elongated objects such as "pole" and "fence" compared with the original model, and also exhibitdc a higher accuracy for complex objects such as "traffic sign" and "rider".

Table 7. Comparison of different classes of IoU on the Cityscapes dataset.

Class	DeepLabv3+ (%)	MFSNet (%)
Road	98.5	98.2
Sidewalk	87.1	87.3
Building	92.8	93.2
Wall	51.3	51.6
Fence	62.4	63.2
Pole	66.5	69.1
Traffic light	70.6	73.1
Traffic sign	79.1	81.6
Vegetation	92.7	93.1
Terrain	64.1	64.5
Sky	95.0	94.2
Person	82.7	84.2
Rider	63.2	64.4
Car	95.7	95.9
Truck	86.1	83.2
Bus	89.1	89.6
Train	78.0	74.6
Motorcycle	68.9	69.1
Bicycle	78.1	79.5

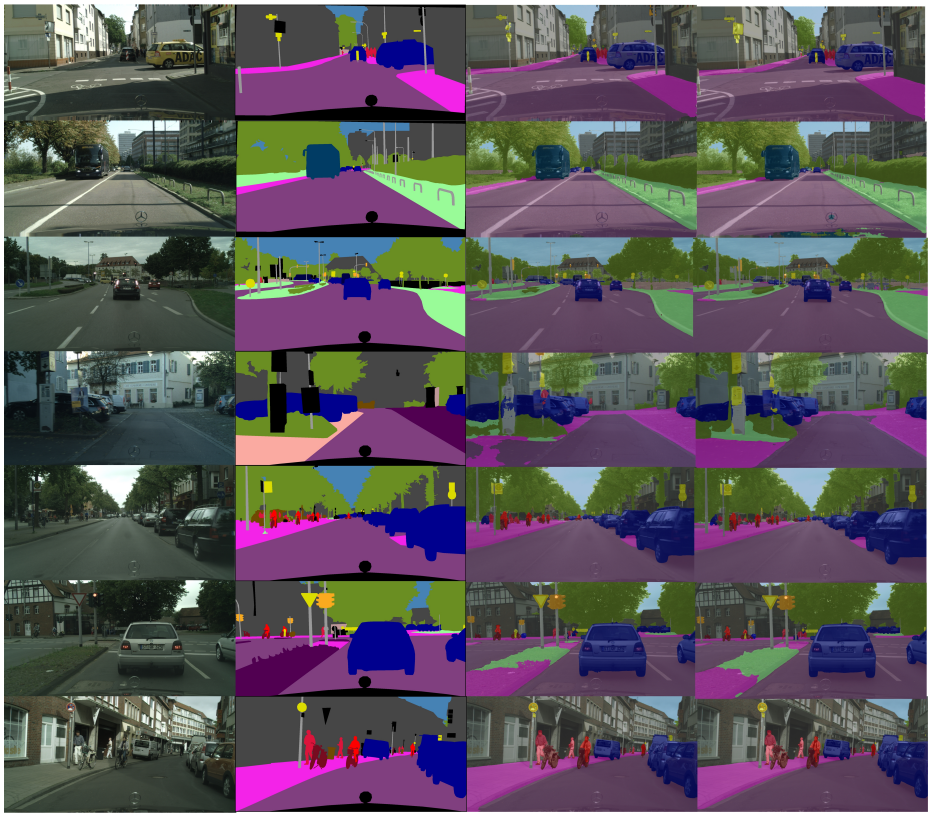
We performed a visual comparison of the results between MFSNet and DeepLabv3+ on the Cityscapes dataset to illustrate the segmentation performance of MFSNet in urban scenes, as shown in Figure 8. By analyzing the segmentation results of the first, fifth, and sixth rows, it is evident that MFSNet achieved more complete and continuous segmentation of object contours such as "pole" "traffic light", and "traffic sign" after establishing feature correlations. Additionally, by observing the results in the fifth, sixth, and seventh rows, MFSNet significantly outperformed DeepLabv3+ in the segmentation of complex objects like "rider" after incorporating multi-scale features fusion and extracting discriminative features from small-scale objects in the image.

3.6. Results on Coco-Stuff

Additionally, to ensure the completeness of our experiments, we conducted a comparative analysis of MFSNet with other models on the COCO-Stuff test set. By observing Table 8, we can see that our model still achieved competitive results. Particularly noteworthy is the performance improvement over Deeplabv3+, where our model achieved a 1.7% higher MIoU and a 1.4% higher MAcc.

Table 8. Comparison of the Coco-Stuff test set results with other models.

Method	Backbone	Resolution	MIoU (%)	MAcc (%)
PSPNet [9]	ResNet101	512 × 512	37.2	49.3
RefineNet [51]	ResNet101	513 × 513	33.6	-
DANet [52]	ResNet101	768 × 768	39.7	-
DeepLabv3	ResNet101	512 × 512	37.3	49.3
DeepLabv3+ [12]	ResNet101	512 × 512	38.4	50.2
OCRNet [43]	ResNet101	520 × 520	39.5	-
MFSNet (Ours)	ResNet101	380 × 380	40.1	51.6



(a) Image (b) Label (c)DeepLabv3+ (d) MFSNet

Figure 8. Visual comparison of segmentation results with other methods on Cityscapes.

4. Conclusions

In this paper, we propose a multi-scale feature shuffle model (MFSNet) for semantic segmentation of urban scenes. By incorporating PSM and EFAM into our network architecture, we effectively mitigated issues related to information loss, weak feature correlations, and inadequate discriminative features during the feature extraction process. Specifically, our method introduced PSM, which leverages both multi-scale characteristics and attention mechanisms, enhancing the representation and consistency of the extracted features, and effectively reducing the impact of noise interference on model segmentation. Furthermore, EFAM combines pooling kernels of varying depths and sizes, which enables the aggregation of both local and global feature information, compensating for the loss of feature details caused by downsampling operations and promoting effective feature reuse. The ablation studies in the Pascal VOC 2012 dataset show the effectiveness of the proposed PSM and EFAM. The experimental results show that MFSNet achieves an outstanding performance on the Pascal VOC 2012, Cityscapes, and COCO-Stuff datasets, respectively.

Author Contributions: Conceptualization, X.Q., C.S. and S.Y.; methodology, C.S. and W.J.; formal analysis, X.Q.; validation, W.J., Y.Y. and X.Q.; visualization, X.Q.; writing—original draft, X.Q. and C.S.; writing—review and editing S.Y., Y.Y. and X.Q.; supervision, S.Y. and Y.Y.; funding acquisition, X.Q. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 12062009), Scientific Research Fund of Zhejiang Provincial Education Department (Y202352150, Y202352263).

Data Availability Statement: The code is available at: https://github.com/syyang2022/MFSNet.

Acknowledgments: We would like to acknowledge the valuable discussions with Qiang Xu about Pyramid Shuffle Module.

Conflicts of Interest: Author Yunxiang Yu was employed by the company Zhejiang Dingli Industrial Co. Ltd. This study is based on the publicly available datasets The Pascal VOC 2012, The Cityscapes and The Coco-Stuff and does not involve any company data. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Li, M.; Chen, D.; Liu, S. Weakly supervised segmentation loss based on graph cuts and superpixel algorithm. *Neural Process. Lett.* **2022**, *54*, 2339–2362. [CrossRef]
2. Sun, W.; Liu, Z.; Zhang, Y.; Zhong, Y.; Barnes, N. An Alternative to WSSS? An Empirical Study of the Segment Anything Model (SAM) on Weakly-Supervised Semantic Segmentation Problems. *arXiv* **2023**, arXiv:2305.01586.
3. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
4. Fu, J.; Liu, J.; Wang, Y.; Zhou, J.; Wang, C.; Lu, H. Stacked deconvolutional network for semantic segmentation. *IEEE Trans. Image Process.* **2019**. [CrossRef] [PubMed]
5. Hou, L.; Vicente, T.F.Y.; Hoai, M.; Samaras, D. Large scale shadow annotation and detection using lazy annotation and stacked CNNs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1337–1351. [CrossRef] [PubMed]
6. He, J.; Deng, Z.; Qiao, Y. Dynamic multi-scale filters for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3562–3572.
7. Bejar, H.H.; Guimaraes, S.J.F.; Miranda, P.A. Efficient hierarchical graph partitioning for image segmentation by optimum oriented cuts. *Pattern Recognit. Lett.* **2020**, *131*, 185–192. [CrossRef]
8. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking bisenet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9716–9725.
9. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
10. He, H.; Chen, Y.; Li, M.; Chen, Q. ForkNet: Strong semantic feature representation and subregion supervision for accurate remote sensing change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 2142–2153. [CrossRef]
11. Zhu, L.; Ji, D.; Zhu, S.; Gan, W.; Wu, W.; Yan, J. Learning statistical texture for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12537–12546.
12. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
13. Yang, Z. Semantic segmentation method based on improved DeeplabV3+. In Proceedings of the International Conference on Cloud Computing, Performance Computing, and Deep Learning (CCPCDL 2023), Huzhou, China, 17–19 February 2023; pp. 32–37.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
15. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
16. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Ocnet: Object context network for scene parsing. *arXiv* **2018**, arXiv:1809.00916.
17. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019; pp. 1971–1980 .
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
19. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.

20. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
21. Cui, L.; Jing, X.; Wang, Y.; Huan, Y.; Xu, Y.; Zhang, Q. Improved Swin Transformer-Based Semantic Segmentation of Postearthquake Dense Buildings in Urban Areas Using Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 369–385. [CrossRef]
22. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
23. Zhang, Q.-L.; Yang, Y.-B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.
24. Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; Liu, J. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Trans. Med Imaging* **2019**, *38*, 2281–2292. [CrossRef] [PubMed]
25. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv* **2018**, arXiv:1805.10180.
26. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part IV 14; pp. 630–645.
28. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
29. Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; Malik, J. Semantic contours from inverse detectors. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 991–998.
30. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
31. Caesar, H.; Uijlings, J.; Ferrari, V. Coco-stuff: Thing and stuff classes in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1209–1218.
32. Rashwan, A.; Du, X.; Yin, X.; Li, J. Dilated SpineNet for semantic segmentation. *arXiv* **2021**, arXiv:2103.12270.
33. Jin, Z.; Liu, B.; Chu, Q.; Yu, N. Isnet: Integrate image-level and semantic-level context for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7189–7198.
34. Zhou, Z.; Lei, Y.; Zhang, B.; Liu, L.; Liu, Y. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 11175–11185.
35. Zhou, T.; Wang, W.; Konukoglu, E.; Van Gool, L. Rethinking semantic segmentation: A prototype view. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2582–2593.
36. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
37. Li, L.; Zhou, T.; Wang, W.; Li, J.; Yang, Y. Deep hierarchical semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1246–1257.
38. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
39. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
40. Chu, X.; Chen, L.; Chen, C.; Lu, X. Improving image restoration by revisiting global information aggregation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 53–71.
41. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
42. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.
43. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VI 16; pp. 173–190.
44. Jo, S.; Yu, I.-J.; Kim, K. MARS: Model-agnostic Biased Object Removal without Additional Supervision for Weakly-Supervised Semantic Segmentation. *arXiv* **2023**, arXiv:2304.09913.
45. Kirillov, A.; Wu, Y.; He, K.; Girshick, R. Pointrend: Image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9799–9808.
46. He, J.; Deng, Z.; Zhou, L.; Wang, Y.; Qiao, Y. Adaptive pyramid context network for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7519–7528.

47. Bai, S.; Koltun, V.; Kolter, J.Z. Multiscale deep equilibrium models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5238–5250.
48. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
49. Ertenli, C.U.; Akbas, E.; Cinbis, R.G. Streaming Multiscale Deep Equilibrium Models. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 189–205.
50. Termritthikun, C.; Jamtsho, Y.; Ieamsaard, J.; Muneesawang, P.; Lee, I. EEEA-Net: An early exit evolutionary neural architecture search. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104397. [CrossRef]
51. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
52. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.