2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

FDLNet: Boosting Real-time Semantic Segmentation by Image-size Convolution via Frequency Domain Learning

Qingqing Yan[1*], Shu Li[1*], Chengju Liu [1,3✉], Ming Liu[2] and Qijun Chen[1]

Abstract— **This paper proposes a novel real-time semantic segmentation network via frequency domain learning, called FDLNet, which revisits the segmentation task from two critical perspectives: spatial structure description and multilevel feature fusion. We first devise an image-size convolution (IS-Conv) as a global frequency-domain learning operator to capture long-range dependency in a single shot. To model spatial structure information, we construct the global structure representation path (GSRP) based on IS-Conv, which learns a unified edge-region representation with affordable complexity. For efficient and lightweight multi-level feature fusion, we propose the factorized stereoscopic attention (FSA) module, which alleviates semantic confusion and reduces feature redundancy by introducing level-wise attention before channel and spatial attention. Combining the above modules, we propose a concise semantic segmentation framework named FDLNet. We experimentally demonstrate the effectiveness and superiority of the proposed method. FDLNet achieves *state-of-the-art* performance on the Cityscapes, which reports 76.32% mIoU at 150+ FPS and 79.0% mIoU at 41+ FPS. The code is available at https://github.com/qyan0131/FDLNet.**

I. INTRODUCTION

Real-time semantic segmentation aims to assign dense labels to all pixels in the image at a low-cost and has drawn increasing interest due to its fast-growing practical application demands. Previous real-time segmentation methods [1]–[5] have achieved promising performances on various benchmarks [6], [7]. However, as a high-resolution prediction task, real-time semantic segmentation still faces the speed vs. accuracy contradiction caused by the challenges of high-resolution input/output and complex scale variations.

To this end, in contrast to existing methods that directly acquire features from the image domain, we extend the conventional feature space to the frequency domain (FD) and investigate the corresponding feature representation methods. In this paper, we propose a novel CNN-based segmentation architecture coined as FDLNet in view of FD learning. Specifically, it is shown in Fig. 1 that FDLNet adopts a pre-trained model as the backbone and improves the segmentation performance in the following two aspects.

First, given the paradox between the detailed structural representation and the respectable computational cost in predicting high-resolution output, modern methods [8]–[10] commonly sacrifice spatial information to achieve real-time inference, which leads to broken performance. Since the dramatic progress of the two-stream structure proposed by BiSeNet [1], a series of recent works [4], [11], [12] have constructed various extra spatial paths to offer low-level structural descriptions, like exact edges for segmentation models, which improves localization accuracy. However, the existing spatial path design usually biases real-time inference to employ shallow convolution layers with limited receptive fields and fails to capture long-range spatial interactions. This is often manifested in their failure to extract contours for complex texture regions, poor robustness to environmental noise such as shadows and halation, and ineffective foreground extraction in complex backgrounds, which results in terrible position guidance for segmentation results.

Drawn inspiration from the spatial frequency model for natural images, one crucial idea is that lower frequencies often describe smoothly changing structures, like areas with similar colors, while higher frequencies typically depict drastically changing structures such as edges, textures, and noise. Therefore, we propose to employ fast Fourier transform (FFT) [13] to convert the structure description problem in the image domain into a spectral selection problem in the FD. In particular, we first construct an efficient image-size convolution (IS-Conv) and present adaptive global kernel generation and efficient frequency feature learning algorithms for global perception in a single shot.

Besides, considering the high computational cost induced by direct operations on large-scale inputs, we leverage a space-to-depth (S2D) conversion to losslessly decrease the input resolution. Then, with the promotion of IS-Conv, we build a global structure representation path (GSRP) which consists of only one S2D layer, an IS-Conv layer, and one transition layer. The method is empirically demonstrated to be competent to learn a unified edge-region representation in the FD at a low cost for the global structure description.

Second, aggregating multilevel features is crucial to capturing multiscale context for addressing the complex scale variations. In mainstream semantic segmentation algorithms, the pyramid-style feature fusion modules [14], [15] are often exploited to enrich the feature space but leads to a dramatic increase in computational cost. Meanwhile, the direct fusion of multilevel features is prone to semantic confusion due to the huge semantic gap between shallow and deep features. To address the problem, some pioneering approaches reweigh the channel features [16] or spatial pixels [17], [18] by leveraging attention mechanisms to model channel-wise or space-wise feature interactions. Though these methods seem

* These authors contributed equally.
✉ C. Liu is the corresponding author (liuchengju@tongji.edu.cn).

This work is supported by the National Natural Science Foundation of China under Grants (62173248, 62073245), Suzhou Key Industry Technological Innovation-Core Technology R&D Program (SGC2021035), and Shanghai Science and Technology Innovation Action Plan (22511104900).

[1] Tongji University, Shanghai, 201804, China.
[2] Hong Kong University of Science and Technology, Hong Kong.
[3] Tongji Artificial Intelligence (Suzhou) Research Institute, Suzhou, 215000, China.

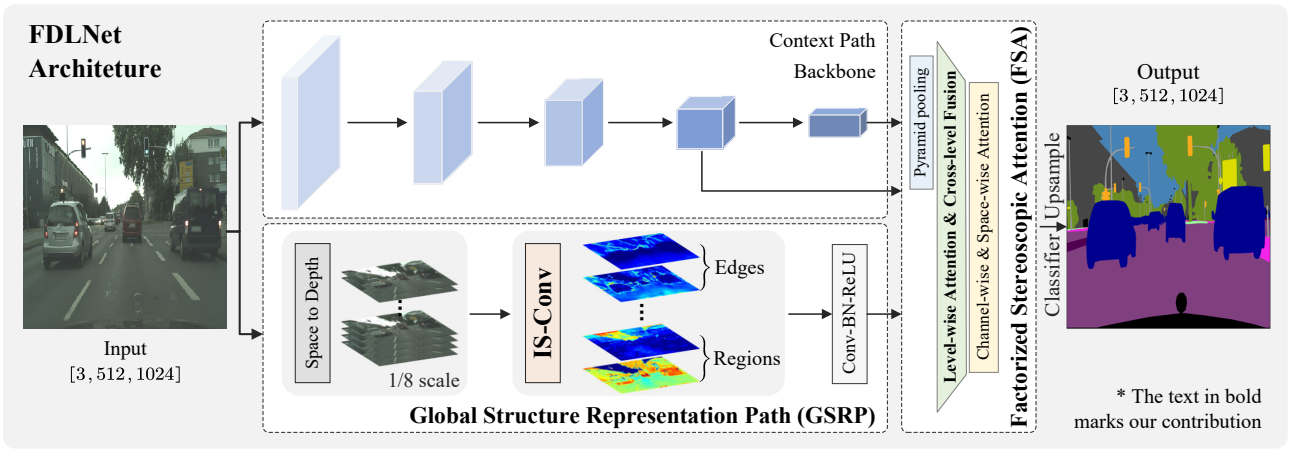


Fig. 1. The proposed FDLNet framework. The context path contains a backbone for high-level semantic features. The GSRP is proposed for global structure description via image-size convolution (IS-Conv). The FSA is proposed to generate feature pyramids and fuse multilevel features for pixel classification.

effective, they usually ignore the level-wise redundancy and suffer from a huge computation burden due to the simultaneous attention modeling across all pixels.

To tackle this, we propose a factorized stereoscopic attention (FSA) in three decoupled dimensions: level-wise, channel-wise, and space-wise. It emphasizes level-wise interaction modeling and highly reduces feature redundancy at a low cost. FSA first employs level-wise attention to estimate the contribution of each level to each pixel in the fused feature space and achieves fast integration of multilevel features. Then it constructs inter-channel dependencies to select key features by channel-wise attention and builds spatial context associations by space-wise attention.

Based on the above modules, we present the FDLNet framework and conduct experiments on the Cityscapes [6]. We have analyzed and validated the effectiveness of GSRP and FSA. The experiments demonstrate that FDLNet achieves state-of-the-art performance, which reports 76.32% mIoU at 150+ FPS and 79.0% at 41+ FPS on the test set.

To sum up, the contributions of this paper are as follows:

1) We put forward the idea of learning global frequency features in the FD and introduced an image-size convolution algorithm (IS-Conv), which enables the capture of global interactions at a low cost in a single shot.
2) We build the global structure representation path (GSRP) based on IS-Conv for global structure description by learning a unified edge-region representation in the FD.
3) We propose the factorized stereoscopic attention (FSA) module for efficient feature fusion by decoupling it into the previously under-appreciated level-wise, channel-wise, and space-wise attentions.
4) We present the FDLNet framework based on GSRP and FSA, which achieves state-of-the-art performance.

II. Related Work

A. Application of FD transformations in vision tasks

The Fourier transform has been widely used in classical digital image processing [19]. With the rapid development of CNNs in vision, some works [20], [21] proposed to use the convolution theorem for convolution acceleration with FFT. In recent years, there has been some works that introduce frequency information into deep learning networks and has achieved good results on various tasks [22], [23]. DCTNet [24] learns on the discrete cosine transform results and downsamples images, which avoids accuracy loss by excluding trivial frequencies. FFC [25] proposed to replace part of the convolution with an Local Fourier Unit and operate convolutions in the FD. FFC combines image-domain convolution and frequency-domain convolution which is used only for global dependency construction. More recently, GFNet [26] models the network by learning in the FD as a global filter. However, the FD kernel size of GFNet is set in advance and cannot be changed dynamically. In this paper, we propose to implement an image-size convolution with adaptive kernel size by FD learning, which acts as a more flexible global filter with fewer learnable parameters.

B. Real-time semantic segmentation

Over the years, the interest in real-time semantic segmentation has rapidly intensified due to their fast-growing application demands. Some networks use efficient convolutional operations [27]–[30] or apply a lightweight backbone for real-time inference [4], [31]–[33]. Many networks are designed with efficient modules to deal with structural information and multilevel features for precise segmentation.

1) Dealing with structural information: Some encoder-decoder structures [9], [34], [35] have employed skip connections to combine low-level features with refined high-level representations for detail restoration. Other multi-path fusion methods [1], [2], [11], [36], [37] adopt different branches to encode both context and structural information to generate high-accuracy predictions. MSFNet [38] proposed a strategy of multiplying spatial fusion and class boundary supervision.

However, these methods model the structural information only at a shallow level with limit receptive fields, which brings redundancy and noise and even leads to semantic confusion. In this paper, we transform the spatial structure

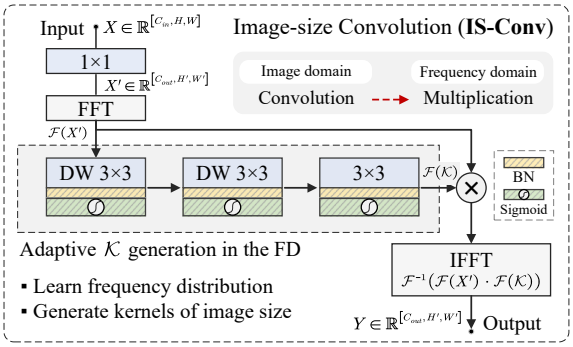


Fig. 2. The IS-Conv structure. *DW* is the depthwise separable convolution.

description problem into a frequency selection problem in the FD and provide high-quality structural information by constructing edge-region consistent descriptions.

2) Handling multilevel features: Modern approaches [3], [10], [39] use dilated convolutions with different dilation rates in cascade or parallel to obtain multiscale features. PSPNet [14] encoded context information by Pyramid Pooling Module. DeepLab [15] employed ASPP with global average pooling to obtain multiscale information. DFANet [2] adopted deep feature aggregation to fully use features of different levels. In addition, since SENet [16], many works [17], [18], [40]–[42] have established attention mechanisms in channels, spaces and achieved excellent performance.

However, the level-wise association and redundancy are often ignored, resulting in information overload and extensive computation during feature fusion. We thus introduce level-wise attention and design a factorized attention module to achieve efficient and cost-effective feature fusion.

III. Proposed Method

The proposed FDLNet framework is depicted in Fig. 1, which consists of two paths and a multilevel feature fusion module. We describe them in detail below.

A. Image-size convolution(IS-Conv)

It is well known that the demand for capturing global context often calls for a feature extraction process with a sizeable receptive field. In practice, the optimal kernel size for acquiring full receptive fields is consistent with the input image size in a single shot, whose corresponding convolution is referred to as the image-size convolution. However, image-size convolution faces the following dilemmas: (1) enormous computational overhead especially for large-scale inputs, (2) dynamic convolutional kernel size due to variable image sizes, (3) massive training parameters. To tackle this, we propose the IS-Conv operator via FD learning.

1) Structure: The proposed IS-Conv is constructed as Fig. 2. For the ease of description, we use $\otimes$ for convolution operator and $X \in \mathbb{R}^{[C_{in},H,W]}$, $Y \in \mathbb{R}^{[C_{out},H',W']}$ and $\mathcal{K}$ for input, desired output and image-size kernel, respectively. Before performing image-size convolution, X is transformed to X' for desired output shape, and then the image-size convolution is calculated as Eq. (1), where ϕ is the transform function. According to the convolution theorem as Eq. (2), we rewrite Eq. (1) as Eq. (3), where $\mathcal{F}\{\cdot\}$ denotes the FFT, $\mathcal{F}^{-1}\{\cdot\}$ stands for the inverse FFT (IFFT), and $\mathcal{F}(X')$ and $\mathcal{F}(\mathcal{K})$ are homomorphic matrices.

$$X' = \phi(X), \quad Y = X' \otimes \mathcal{K} \tag{1}$$

$$\mathcal{F}\{f \otimes g\} = \mathcal{F}\{f\} \cdot \mathcal{F}\{g\} \tag{2}$$

$$Y = X' \otimes \mathcal{K} = \mathcal{F}^{-1}(\mathcal{F}(X') \cdot \mathcal{F}(\mathcal{K})) \tag{3}$$

Based on Eq. (3), IS-Conv performs the following steps:

(1) Transform the input to the desired output shape by a 1×1 Conv transition layer with weights $W_{transition}$ as Eq. (4), and obtain the spectrum $\mathcal{F}(X')$ using FFT.

$$X' = \phi(X) = X \otimes W_{transition} \tag{4}$$

(2) Generate $\mathcal{K}$ in the FD ($\mathcal{F}(\mathcal{K})$) conditioned on the spectrum of $\mathcal{F}(X')$ using a series of convolutions as Eq. (5),

$$\mathcal{F}(\mathcal{K}) = \mathcal{F}(X') \otimes W_1 \otimes \cdots \otimes W_N, \tag{5}$$

where W_i $(i \in N)$ represents the convolution. In this paper, we set $N = 3$ and the detailed convolution setups are shown in Fig. 2.

(3) Multiply $\mathcal{F}(X')$ with $\mathcal{F}(\mathcal{K})$ and conduct IFFT to get the final output just as Eq. (3).

The generated $\mathcal{K}$ can be adaptive to various image sizes, and there are very few learnable parameters, which is friendly and flexible to both training and inference.

2) Design rational and analysis:

a) FD learning principle and its effectiveness: Convolution considers the association of local pixels in the image domain, which restricts the receptive field of a single operation to the kernel size. In contrast, the convolution learning on the spectrum considers the frequency correlation in the FD and extracts the frequency combinations with specific patterns. Since all pixels of the image are involved in the computation of the spectrum, the equivalent receptive field in the image domain is no longer limited by the convolution kernel size. Besides, frequency better highlights the essential properties of a particular pattern in an image, e.g. zero frequency for the global average of the feature map, low-frequency for uniform color regions, and high-frequency for edges and complex textures. Also, FD learning can remove ultra-high frequency noise and improve the feature robustness. Thanks to that, the spectrum of the IS-Conv kernel, $\mathcal{F}(\mathcal{K})$, actually acts as a weight mask in FD generated by learning the spectrum. The mask then realizes the frequency selection by multiplying it with the input spectrum, which enables IS-Conv with global perception when the multiplication is converted to the image domain.

b) Computational efficiency: For comparison, the number of parameters for a single convolution and an IS-Conv with the same receptive field are given by Eq. (6) and (7), respectively. The computational complexity of FFT and IFFT is $\mathcal{O}(H'W'C_{out}\log_2(H'W'))$, and the total complexity of IS-Conv is $\mathcal{O}(HWC_{in}C_{out} + H'W'C_{out}\log_2(H'W') + H'W'C_{out}^2 + H'W'C_{out})$. Taking the image size as the

dominant variable, the IS-Conv complexity can be simplified to $\mathcal{O}(HW\log_2(HW))$. Comparing the global convolution with a complexity of $\mathcal{O}(HWH'W'C_{in}C_{out})$ which can be simplified to $\mathcal{O}(H^2W^2)$, the proposed IS-Conv effectively reduces the computational complexity, especially for large-scale image inputs.

$$Params_{conv} = C_{in}C_{out}HW, \tag{6}$$

$$\begin{cases} Params_{1\times1} = C_{in}C_{out} \\ Params_{W_{1,2,3}} = 18C_{out} + 9C_{out}^2 \\ Params_{IS-Conv} = Params_{1\times1} + Params_{W_{1,2,3}} \end{cases} \tag{7}$$

In summary, IS-Conv realizes image-size convolution via image-frequency transformation and FD learning, which cuts off numerous computation and parameters, and enables global representation in a single shot.

B. Global structure representation path(GSRP)

Regions and edges are equally important for semantic segmentation, which we refer to collectively as spatial structure information. Since various spatial structure features are often represented separately at different levels in successive convolutional models, existing methods using shallow layers in spatial path often fail to describe completely spatial structure and long-range dependencies. Benefit from the effectiveness of IS-Conv, we recommend transforming the spatial structure description problem into a spectral selection problem in the FD. So we construct global structure representation path (GSRP) based on IS-Conv, which realizes the edge-region consistency representation and global dependence capture.

1) Structure: The GSRP structure is depicted in the lower middle part of Fig. 1, which consists of an S2D defined in our previous work [5], an IS-Conv and a transition layer. Due to the huge calculation and high spatial redundancy associated with large-scale input, we first use the S2D to losslessly reduce the original input to 1/8 scale. Then, we apply one IS-Conv to extract rich edge and region information. Finally, a 1×1 convolution is used to re-combine different variants and control the number of output channels.

2) Design rational and analysis: As mentioned earlier, whether edges or regions, they are a specific pattern in the FD, which can be captured through proper learning. To prove this, we make a simple demonstration as shown in Fig. 3. We perform FFT on the input images and then directly remove the 30% low-frequency component or 30% high-frequency one, respectively, and then we correspondingly obtain (b) the edge maps, and (c) the region maps. It follows that a unified edge-region representation can be achieved by a proper frequency selection in the FD. In other words, it is reasonable and effective to convert the structural description problem into a frequency selection problem in the FD.

The computational cost of GSRP mainly exists in IS-Conv and the transition layer. Since Sect. III-A.2.b demonstrates the high efficiency of IS-Conv and the transition layer is a 1×1 conv on a low-resolution image with low computational cost, GSRP is naturally quite efficient and cost-effective.

To summarize, GSRP yields global structure representations with affordable complexity, providing rich and precise location information for the segmentation model.

(a) Original Image (b) Edge (70% high frequency) (c) Region (30% low frequency)

Fig. 3. Example of extracting edges and regions in the FD. (a) is the input image, (b) shows the edge extraction by keeping 70% high frequency component, while (c) is the region obtained by removing the high frequency.

C. Factorized stereoscopic attention(FSA)

Multiscale context and multi-level semantics have proven to be critical to segmentation performance. As mentioned earlier, modern methods for the feature fusion usually suffer from the huge semantic gaps among different levels, and the information overload problem caused by feature redundancy. To handle the problem, we propose the FSA to model attention mechanisms with three decoupled dimensions: level-wise, channel-wise, and space-wise, which achieves lightweight but powerful context aggregation and multiscale fusion simultaneously. Notably, FSA introduces level-wise attention, which was being overlooked by previous methods, and enables fast multi-level feature merging.

1) Structure: The FSA sketch shown in Fig. 4 is designed with the following operations:

(1) Pyramid pooling for the 1/32 scale output for multiscale context. To accommodate non-square inputs, the pooling parameters are set as shown in Fig. 4.
(2) Channel adjustment and scale alignment by 1×1 convolution and interpolation. The FSA has 7 sets of features F^i, $i \in [1, 7]$ with the same size and output channels.
(3) Level-wise attention and cross-level fusion: (a) With the structure marked as *level-wise attention* in Fig. 4, we calculate the weight of F^i for each spatial location and obtain the score map S^i. The physical meaning of S^i is the degree of contribution of F^i to each pixel on the fused output. (b) Compute fusion output F_{fuse} by weighting and summing all F^i with S^i as weights.
(4) Apply commonly used channel-wise [16] and space-wise [17] attentions on F_{fuse} to obtain the final FSA's output.

2) Design rational and analysis: We quantify the redundancy between feature maps by Pearson correlation coefficient. By calculating the correlation of 1/8, 1/16, 1/32 scale feature maps of pre-trained ResNet-18, we find that 14.9% of the feature maps have a correlation higher than 0.3, which indicates a moderate or stronger correlation. Typical examples are shown in Fig. 5. In terms of context aggregation, we plot the correlation matrix between the pooling pyramids according to the PSPNet [14] as depicted on the right of Fig. 5, which shows extremely high correlations. Therefore, we propose to reweigh the contribution of each level to the fusion output at each space location, and obtain the fused features directly by weighting and summing across all levels.

In conclusion, FSA couples three attention mechanisms

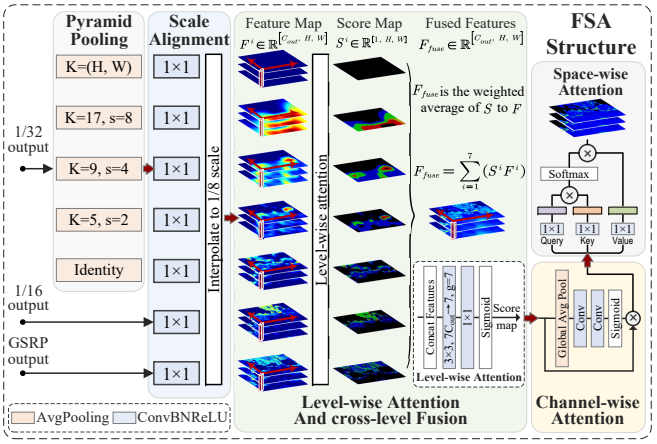


Fig. 4. The FSA structure. The FSA accepts input from the context path and GSRP, and outputs fused features.

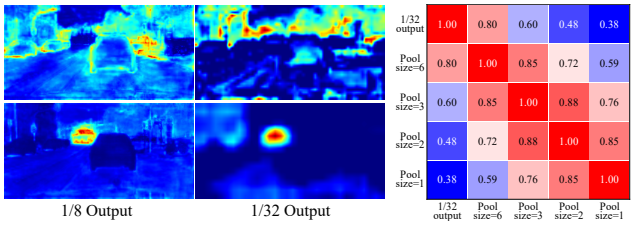


Fig. 5. Feature redundancy demonstration. LEFT: Correlated feature maps from different scales. RIGHT: Correlation matrix of pyramid poolings.

from level-wise, channel-wise, and space-wise and achieves lightweight but powerful multi-level feature aggregation.

D. FDLNet Architecture

Recalling Fig. 1, FDLNet employs a two-path structure. The context path contains a backbone which extracts high-level semantics, while the proposed GSRP describes the global structure which provides accurate location guidance for high-resolution output. Finally the two paths are fed into FSA for efficient multiscale and multilevel feature fusion, and the classifier outputs prediction results. FDLNet is a plug-and-play framework for flexible model designs using a variety of backbone networks, and its components GSRP and FSA can be easily applied to other models as well.

IV. Experiments

In this section, we first introduce the implementation details, and then perform ablation study to further analyze and demonstrate the effectiveness and superiority of our method. Finally we present the comparison with other methods on the Cityscapes dataset [6].

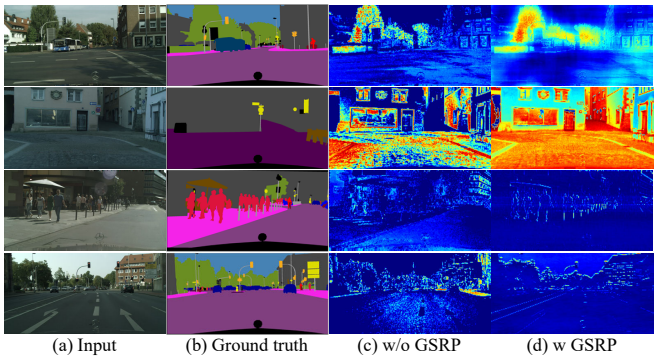
A. Implementation Details

We conducted experiments on PyTorch 1.9.0. The environment is NVIDIA RTX2080Ti, CUDA 11.3 and CuDNN v8.2. For training, we used mini-batch stochastic gradient descent (SGD) with momentum 0.9, batch size 12 and weight decay 5e-4. Our model was trained 500 epochs with standard cross-entropy loss. We applied the poly learning rate policy

TABLE I

Ablation study results. GFLOPs are calculated with input shape [3,512,1024]

GSRP	FSA	#Param.	GFLOPs	mIoU	Inf.Time
-	-	13.48M	29.64G	74.78%	8.82ms
✓	-	13.50M	29.70G	75.65%	8.82ms
-	✓	11.89M	22.15G	75.77%	**6.64ms**
✓	✓	11.92M	22.21G	**76.61%**	**6.65ms**



(a) Input (b) Ground truth (c) w/o GSRP (d) w GSRP

Fig. 6. Some typical examples that demonstrate the effectiveness of GSRP.

with power 0.9 and initial learning rate 0.01. For evaluation metrics, both mean classwise intersection-over-union (mIoU) and frames per second (FPS) were used. For data processing, we used mean subtraction and normalization and applied some data augmentation techniques: random resizing from 0.75 to 1.5, crop, and horizontal flip.

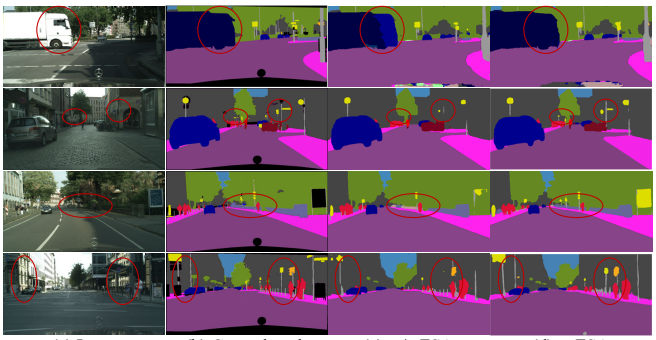
B. Ablation study

We evaluated the proposed GSRP and FSA on the validation set of the Cityscapes [6]. For fair comparison, we chose the popular BiSeNet [1] structure as the baseline. We verified the effectiveness of each module by replacing the spatial path (denoted as SP) in BiSeNet with GSRP and the Feature Fusion Module (FFM) with FSA. Table I shows the result, where the first line is the BiSeNet and the last one is FDLNet. Marked with a ✓ sign indicates the replacement. Below we analyze the role of each module in detail.

1) Effectiveness of the GSRP: After replacing the original SP with GSRP, the model achieved a 0.9% accuracy improvement. Although the number of parameters and computation increases slightly due to the global kernel generation in IS-Conv, GSRP barely hurts the real-time performance. We compare the cases with and without GSRP in Fig. 6. The GSRP's effectiveness are reflected in two aspects:

(1) The first two lines demonstrate the ability to highlight regions. The original SP tends to extract too many edges at the texture and thus results in discontinuities in the region owing to the lack of long-range dependence. In contrast, GSRP is able to extract almost all regions of tree and road, respectively.

(2) The last two lines shows the edge extraction capability. Due to the shallow layers and low-level features, SP sometime extracts more noise than edges. However,



(a) Input (b) Ground truth (c) w/o FSA (d) w FSA

Fig. 7. Some examples of FSA's role on improving network performance.

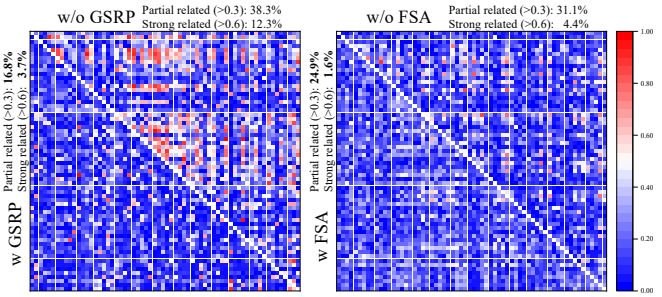


Fig. 8. Correlation matrix between feature maps. The closer to red, the higher the correlation.

GSRP can describe clear and explicit boundaries.

The results indicate that GSRP facilitates the learning of specific textures and boundaries in the global background and reduces noise interference. Both benefit from the edge-region consistency representation in the FD, which enables frequency feature learning and global structural descriptions.

2) Effectiveness of the FSA: The improvement brought by FSA is significant in both accuracy ($\sim 1\% \uparrow$) and inference time ($\sim 24.7\% \uparrow$). As illustrated in Fig. 7, we summarize its promotion effect in four aspects.

(1) Provides enough receptive field to improve the perception of large objects (1st row).
(2) Enhances tiny object detection through multilevel and multiscale learning (2nd row).
(3) Alleviates local misclassification via contextual aggregation (3rd row).
(4) Mitigates semantic confusion due to similar appearance and large background by coupling multidimensional attention mechanisms (4th row).

These results coincide with our initial design intention and fully prove the efficiency and effectiveness of our method.

3) Study on feature redundancy: To further reveal the superiority of GSRP and FSA, we traverse the validation set of Cityscapes and calculated the correlation matrix of GSRP/FSA outputs, as shown in Fig. 8. Intuitively, the feature correlation of the networks with GSRP/FSA is lower than those without these modules. To quantitatively describe the redundancy, we count the proportion of partially correlated (correlation>0.3) and strongly correlated (correlation>0.6) for each correlation matrix, as shown in the tick labels on Fig. 8. Both GSRP and FSA effectively reduce the redundancy of the output features. It can be concluded that GSRP and FSA are superior in improving the learning and generalization abilities of the network.

C. Comparison with state-of-the-art methods

We tested our networks on the Cityscapes [6] test set and compared our results with the state-of-the-art real-time methods for semantic segmentation in Table II. Among the methods listed, our FDLNet achieves the best trade-off between speed and accuracy. Compared to the CABiNet [43] with similar accuracy, FDLNet runs nearly 2x faster. And

TABLE II

COMPARISON ON CITYSCAPES TEST SET. *Py.* STANDS FOR TESTING ON PYTORCH PLATFORM, WHILE *TRT* IS ON TENSORRT.

Method	Input Size	GPU	mIoU (%)	FPS	
				Py.	TRT
DFANet A [2]	1024×1024	Titan X	71.3	100	
DFANet B [2]	1024×1024	Titan X	67.1	120	
BiSeNet v1 [1]	768×1536	Titan Xp	74.7	65.6	-
BiSeNet v2 [12]	512×1024	GTX 1080Ti	72.6	-	156
BiSeNet v2-L [12]	512×1024	GTX 1080Ti	75.3	-	47.3
SwiftNet-18 [44]	1024×2048	GTX 1080Ti	75.5	39.9	-
FANet [45]	1024×2048	Titan X	74.4	72	-
ShelfNet [46]	1024×2048	GTX 1080Ti	74.8	36.9	-
GAS [47]	769×1537	Titan Xp	71.8	108.4	-
HMSeg [48]	768×1536	GTX 1080Ti	74.3	83.2	-
SFNet [49]	512×1024	GTX 1080Ti	74.5	121	-
MSFNet [38]	1024×2048	RTX 2080Ti	77.1	41	-
CABiNet [43]	1024×2048	RTX 2080Ti	75.9	76.5	-
STDC1-Seg50 [4]	512×1024	GTX 1080Ti	71.9	-	250.4
STDC2-Seg50 [4]	512×1024	GTX 1080Ti	73.4	-	188.6
STDC1-Seg75 [4]	768×1536	GTX 1080Ti	75.3	-	126.7
STDC2-Seg75 [4]	768×1536	GTX 1080Ti	76.8	-	97.0
RGPNet [50]	1024×2048	RTX 2080Ti	74.1	-	47.2
FDLNet-18	512×1024	TiTan X	76.3	104.2	-
		RTX 2080Ti		**150.4**	-
FDLNet-101	512×1024	TiTan X	**79.0**	27.9	-
		RTX2080Ti		41.7	-

FDLNet achieves slightly higher mIoU and faster inference FPS than the STDC2-Seg75 [4] even without TensorRT acceleration. Overall, our approach achieves state-of-the-art performance in real-time semantic segmentation.

V. CONCLUSION

In this paper, we revisit the real-time semantic segmentation task from two critical perspectives: global spatial structure description and multilevel information fusion. We first proposed the IS-Conv operator to capture global dependencies in a single shot via frequency domain learning. Then we constructed the GSRP based on IS-Conv for efficient spatial structure description with both edges and regions, and the FSA for light-weight but powerful multilevel feature fusion. Extensive experiments have demonstrated that FDLNet, consisting of the above modules, achieves state-of-the-art performance on the Cityscapes dataset. In the future, we will further explore vision tasks from a frequency domain learning perspective and take our study on IS-Conv further.

REFERENCES

[1] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[2] H. Li, P. Xiong, H. Fan, and J. Sun, "Dfanet: Deep feature aggregation for real-time semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9522–9531.

[3] Z. Zhang and K. Zhang, "Farsee-net: Real-time semantic segmentation by efficient multi-scale context aggregation and feature space super-resolution," *arXiv preprint arXiv:2003.03913*, 2020.

[4] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking bisenet for real-time semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9716–9725.

[5] S. Li, Q. Yan, C. Liu, M. Liu, and Q. Chen, "Holoseg: An efficient holographic segmentation network for real-time scene parsing," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2395–2402.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[7] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[8] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[10] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[11] R. P. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: fast semantic segmentation network," *arXiv preprint arXiv:1902.04502*, 2019.

[12] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, pp. 1–18, 2021.

[13] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

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

[19] T. Acharya and A. K. Ray, *Image processing: principles and applications*. John Wiley & Sons, 2005.

[20] S. Li, K. Xue, B. Zhu, C. Ding, X. Gao, D. Wei, and T. Wan, "Falcon: A fourier transform based approach for fast and secure convolutional neural network predictions," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8702–8711.

[21] C. Ding, S. Liao, Y. Wang, Z. Li, N. Liu, Y. Zhuo, C. Wang, X. Qian, Y. Bai, G. Yuan, X. Ma, Y. Zhang, J. Tang, Q. Qiu, X. Lin, and B. Yuan, "Circnn: Accelerating and compressing deep neural networks using block-circulant weight matrices," in *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2017, pp. 395–408.

[22] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on fourier domain analysis," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 330–339.

[23] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4084–4094.

[24] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1737–1746.

[25] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4479–4488, 2020.

[26] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 980–993, 2021.

[27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[29] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[30] M. Wang, B. Liu, and H. Foroosh, "Factorized convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 545–553.

[31] G. Li, I. Yun, J. Kim, and J. Kim, "Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation," *arXiv preprint arXiv:1907.11357*, 2019.

[32] Q. Yan, S. Li, C. Liu, M. Liu, and Q. Chen, "Roboseg: Real-time semantic segmentation on computationally constrained robots," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–11, 2020.

[33] W. Chen, X. Gong, X. Liu, Q. Zhang, Y. Li, and Z. Wang, "Fasterseg: Searching for faster real-time semantic segmentation," *arXiv preprint arXiv:1912.10917*, 2019.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[35] W. Wang and Z. Pan, "Dsnet for real-time driving scene semantic segmentation," *arXiv preprint arXiv:1812.07049*, 2018.

[36] R. P. Poudel, U. Bonde, S. Liwicki, and C. Zach, "Contextnet: Exploring context and detail for semantic segmentation in real-time," *arXiv preprint arXiv:1805.04554*, 2018.

[37] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[38] Y. Pei, B. Sun, and S. Li, "Multifeature selective fusion network for real-time driving scene parsing," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[39] T. Emara, H. E. A. El Munim, and H. M. Abbas, "Liteseg: A novel lightweight convnet for semantic segmentation," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2019, pp. 1–7.

[40] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[41] S. Hao, Y. Zhou, Y. Zhang, and Y. Guo, "Contextual attention refinement network for real-time semantic segmentation," *IEEE Access*, vol. 8, pp. 55 230–55 240, 2020.

[42] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[43] S. Kumaar, Y. Lyu, F. Nex, and M. Y. Yang, "Cabinet: efficient context aggregation network for low-latency semantic segmentation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 517–13 524.

[44] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 607–12 616.

[45] P. Hu, F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, and S. Sclaroff, "Real-time semantic segmentation with fast attention," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 263–270, 2020.

[46] J. Zhuang, J. Yang, L. Gu, and N. Dvornek, "Shelfnet for fast semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[47] P. Lin, P. Sun, G. Cheng, S. Xie, X. Li, and J. Shi, "Graph-guided architecture search for real-time semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4203–4212.

[48] P. Li, X. Dong, X. Yu, and Y. Yang, "When humans meet machines: Towards efficient segmentation networks." in *BMVC*, 2020.

[49] J. Lee, D. Kim, J. Ponce, and B. Ham, "Sfnet: Learning object-aware semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2278–2287.

[50] E. Arani, S. Marzban, A. Pata, and B. Zonooz, "Rgpnet: A real-time general purpose semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3009–3018.