BiAttnNet: Bilateral Attention for Improving Real-Time Semantic Segmentation

Genling Li, Liang Li, and Jiawan Zhang, *Senior Member, IEEE*

***Abstract*—Semantic segmentation requires both speed and accuracy. This paper presents a two-branch network BiAttnNet with a unique Bilateral Attention structure that separates all attention modules into the Detail Branch to contribute semantic detail selections for specialized detail exploring. Specifically, the Detail Branch comprises AttnTrans entirely, which provides a better alternate for regular convolution. AttnTrans is a computationally efficient filtration entirely composed of concurrent spatial and channel attention. Meanwhile, a Context Branch is implemented with FCN-ResNet for rough segmentation. By combining two branches' outputs, BiAttnNet achieves a good balance between speed and accuracy. Evaluations on the Cityscapes testing set conclude that BiAttnNet achieves 74.7% mIoU at 89.2 FPS at a quarter (512 $\times$ 1024) resolution with only 2.2 million parameters, running on a single GTX 2080 Ti card.**

***Index Terms*—Image segmentation, machine learning, real-time semantic segmentation.**

I. Introduction

SEMANTIC segmentation is to parse the input image into areas associated with fixed labels. Recently, attention mechanisms and bilateral structure growth widespread in effectively boosting real-time segmentation accuracy.

ContextNet [1] initially proposes separated Detail Branch and Context Branch design, whose main idea is to explore emphasized details in a parallel efficient quick-downsampling path. BiSeNet [2], [3] and its variants take advantage of this bilateral structure but use attention refinement modules to combine hierarchical features and outputs of two branches. ICNet [4] focuses on multi-resolution branches for reducing inference costs. DFANet [5] and HRNet [6] reuse multilevel features to enrich context. These approaches address the efficiency of enhancing context with multiple branches but mostly rely on regular convolutions for exploring details, which isn't fully benefited from the advantages of attention mechanisms.

Attention mechanisms act as essential parts in boosting segmentation accuracy. DANet [7] implements channel attention and spatial attention modules after backbone to provide correlations. TaNet [8] uses STN [9] as an attention mechanism placed before three parallel convolutional branches. Di-CNN [10] calibrates convolution outputs by dilations. Using attention can provide more effective fusions inside U-Net for medical segmentation [11], [12]. SENet [13] and DFN [14] enhance special channels by explicit feature selection. The scSE [15] block provides a concurrent spatial and channel attention solution. The attention modules inside them mainly act as auxiliary recalibrations of regular convolutions, which isn't fully benefited from multiple branches structures.

Based on existing approaches' analyses, we develop BiAttnNet to take full advantage of channel attention, spatial attention, and bilateral structure while maintaining a high inference speed for real-time segmentation. We plot overviews of comparison of segmentation network structures in Fig. 1. Viewing overall, BiAttnNet separates all attention modules into a parallel branch during downsampling. Specifically, BiAttnNet's Detail Branch entirely relies on AttnTrans, which provides a

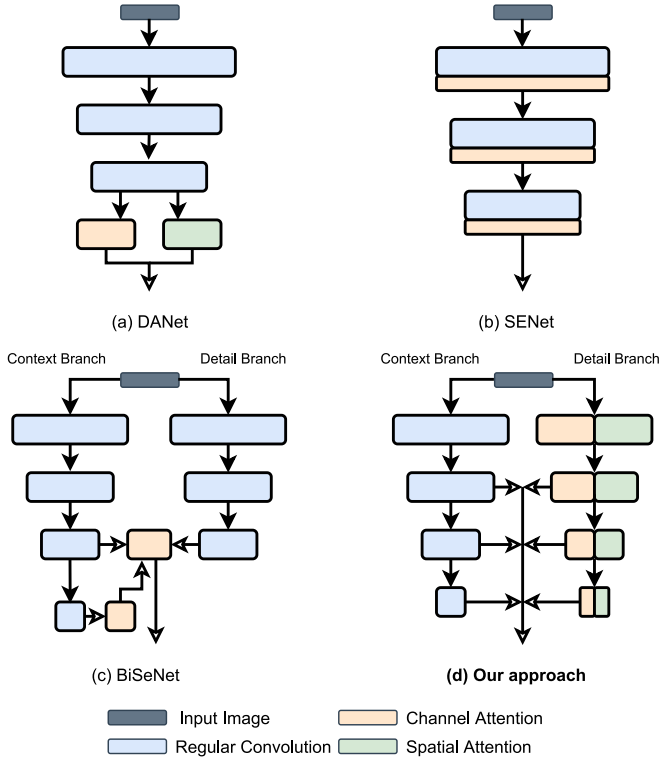

Fig. 1. Visual comparison of different segmentation network structures. Our approach (BiAttnNet) separates all attention modules into a parallel branch for specialized semantic detail selections.

Manuscript received September 9, 2021; revised October 26, 2021; accepted October 26, 2021. Date of publication November 2, 2021; date of current version January 20, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yue Deng. *(Corresponding author: Liang Li.)*

The authors are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China (e-mail: meetchuling@tju.edu.cn; liangli@tju.edu.cn; jwzhang@tju.edu.cn).

Digital Object Identifier 10.1109/LSP.2021.3124186

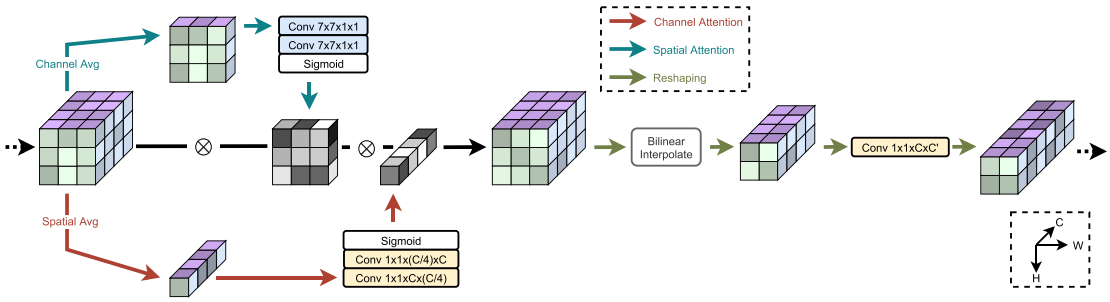

Fig. 2. An illustration of the proposed AttnTrans module. AttnTrans uses convolutions internally but is superior to regular convolution for detail selection. Our module is composed of three parts for different propose: spatial attention, channel attention tensor reshaping. The first two attention mechanisms compute independently. Then the selected details are multiplied with the input using broadcasting. Finally, the tensor is reshaped by bilinear interpolation and 1x1 convolution.

drop-in convolution replacement solution for specialized and fuller detail selection.

Our Bilateral Attention structure stacks AttnTrans modules as a separated Detail Branch for concentrating all attention mechanisms while reusing a reduced FCN-ResNet as a Context Branch. By combining the hierarchical outputs of two branches, the proposed real-time semantic segmentation network BiAttnNet archives 74.7% mIoU, running at 89.2 FPS at a quarter (512×1024) resolution with only 2.2 million parameters, evaluated by the Cityscapes testing set and a single GTX 2080 Ti card.

II. Proposed Method

A. Attntrans

As illustrated in Fig. 2, we propose AttnTrans module as a drop-in replacement solution for regular convolution for better detail selection. AttnTrans module can be split into three parts with different propose: spatial attention, channel attention, and tensor reshaping.

For spatial attention, first, the input tensor ($C \times H \times W$) is averaged on the channel dimension. Now each pixel of the averaged result ($1 \times H \times W$) holds the average value of all its channels. Then, it will be transformed into a spatial attention map ($1 \times H \times W$) by applying two 7×7 convolution filters and `Sigmoid` in sequence. `ReLU` activation is used between convolutions.

For channel attention, first, the input tensor ($C \times H \times W$) is averaged on the spatial dimensions. Now each channel of the averaged result ($C \times 1 \times 1$) holds the average value across all spatial positions. Then, it will be transformed into a channel attention map ($C \times 1 \times 1$) by applying two 1x1 convolutions and `Sigmoid` in sequence. The squeeze radio between convolutions is 4, and `ReLU` activation is used.

Two explored attention maps are applied to the input tensor by multiplying all three of them together. Because of their different shapes, broadcastings happen during multiplication.

The final step is tensor reshaping. We implement spatial reshaping using bilinear interpolation and channel reshaping using 1x1 convolution. Our approach is more flexible than regular convolution because not being constrained by strides, paddings, or dilations. For downsampling, the bilinear interpolation can be moved to the front of two attentions for accelerating these two attentions by a smaller input size.

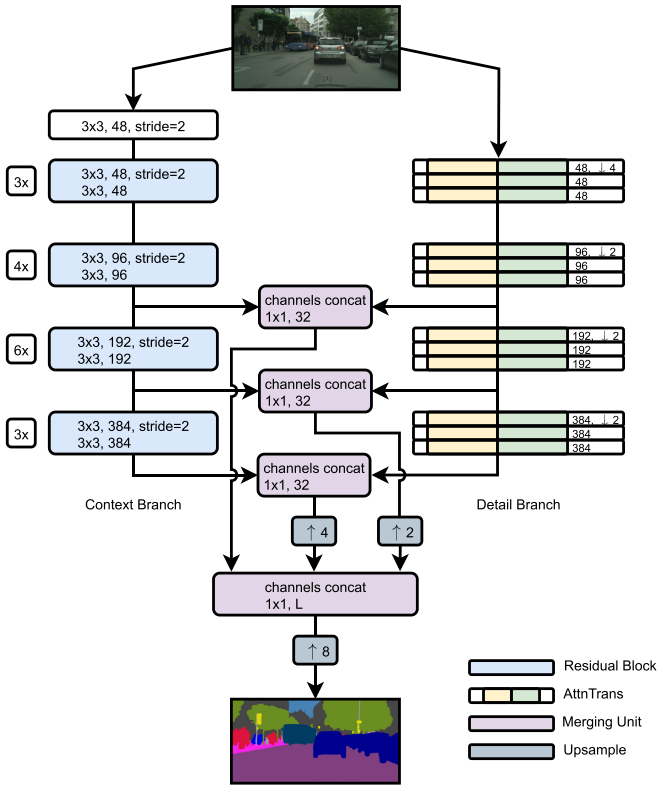

Fig. 3. Our BiAttnNet is benefited from the Bilateral Attention structure. With both channel and spatial attention packed into AttnTrans, the Detail Branch stacks AttnTrans modules for semantic detail selection specially and parallelly. The Context Branch is implemented with a reduced ResNet-34.

B. Bilateral Attention

As illustrated in Fig. 3, we propose the Bilateral Attention structure, in which all attention mechanisms (AttnTrans) are separated and stacked into a parallel branch during inference.

This parallel branch is named as Detail Branch in BiAttnNet, for its specialized detail selection purpose. On the other side, the Context Branch for capturing semantic context is implemented by a reduced ResNet-34 [16], [17]. It has smaller output channel sizes in convolutions and replaces all its regular convolutions with depth-wise separable ones [18].

The Detail Branch is much shallower than the Context Branch because rich details require the lowest possible features. We

balance detail maintaining and selections by three AttnTrans modules at each stage with Batch Normalization [19] and ReLU activation used between them.

Both branches' 8x, 16x, and 32x downsampled outputs are kept and merged with their corresponding ones at the same stages. Hierarchical merged feature maps are all interpolated to 8x downsampled resolution and merged into the semantic segmentation map, later 8x interpolated to input image's resolution. All merging units are implemented with a channels concatenation and 1x1 convolution. Batch Normalization [19] is applied after concatenation for stabilizing concatenated channels' distribution.

III. Experiments

Here we present ablation studies of BiAttnNet on the Cityscapes [20] validation set with only fine annotations used. We further report BiAttnNet's ultimate performance on the Cityscapes testing set with coarse annotations included during training. All experiments are performed on a Linux-based platform with a single GTX 2080 Ti card and CUDA 10.2.

A. Implementation Details

Dataset The Cityscapes dataset is a large-scale database focused on semantic understanding of street scenes. It has a training set with 2975 images, a validation set with 500 images, and a testing set with 1525 images. The first two sets are fine annotated. We only use the first two sets for ablation studies but adopt 20000 additional coarse annotated images while comparing with other methods. The whole database is resized to the quarter resolution 512×1024.

Implementation Protocols We implement BiAttnNet using the PyTorch framework with only 2.2 million parameters used. SGD with a momentum rate of 0.9, a batch size of 8, and cross-entropy loss are used for optimization [21]. A polynomial policy [22] in which the initial learning rate (0.025) is multiplied by $(1 - \frac{epoch}{total\_epochs})^{0.9}$ after each epoch is adopted as an adaptive learning rate strategy. Images for training are loaded in [0, 1] then normalized using mean = [0.2841, 0.3227, 0.2817] and std = [0.1858, 0.1887, 0.1855]. Random scaling between $[448 \times 896, 640 \times 1280]$, horizontal flipping, random Gaussian noise disturbing with 0.5 probability, and random 448×448 cropping are applied for data augmentation [23]. BiAttnNet is trained for 500 epochs for ablation studies and 1000 epochs for achieving its ultimate performance as much as possible while comparing other real-time segmentation works.

B. Ablation Studies

Ablation for AttnTrans Modules We gradually reduce the number of participated stages from 3 to 0in Detail Branch to verify the effectiveness of AttnTrans modules. As summarized in Table I, the segmentation score rises to 71.4% as more AttenTrans modules participated. These results indicate AttenTrans modules' ability to boost semantic segmentation. We visualize heatmaps of the output from AttnTrans of Detail Branch's first stage in Fig. 4.

TABLE I
Studies of AttnTrans Modules. We Gradually Reduce Their Numbers in BiAttnNet by Removing Participated Stages in Detail Branch. Stage 1 is Not Removed Because It Doesn't Contribute Feature Maps to Output. Embedding All Three Stages Achieves the Highest Segmentation Score

Stage 4	Stage 3	Stage 2	mean IoU
			66.3%
✓			68.5%
✓	✓		69.1%
✓	✓	✓	**71.4%**

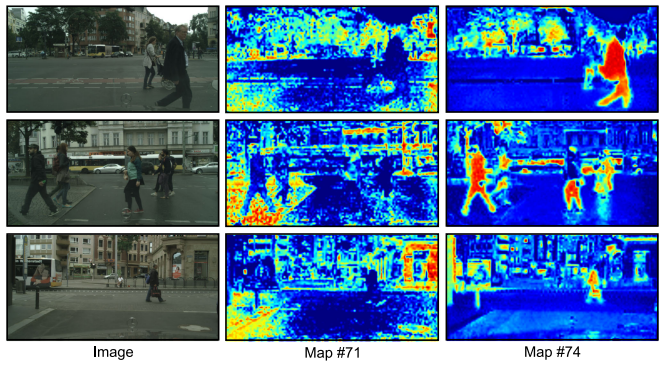

Fig. 4. Heatmaps of the outputs from AttnTrans: Map 71 channel's spatial heatmaps focus on capturing sidewalks and buildings, Map 74 channel's spatial heatmaps focus on capturing humans. Both types of heatmaps capture globally across spatial positions.

TABLE II
Studies of ReLU Activations Inside AttnTrans. as the ReLU Activation Ablated, the Semantic Segmentation Performance Also Drops, Indicating That Internal Non-Linear Activations are Necessary for Unleashing the AttnTrans Module's Full Ability. Since Internal Batch Normalization Has Negative Impacts, We Drop It From the AttnTrans Module

ReLU	Batch Normalization	mean IoU
		69.7%
✓	✓	68.2%
✓		**71.4%**

Ablation for ReLU activations inside AttnTrans The AttnTrans module uses ReLU activations internally between convolutions during its spatial and channel attention map generating. We verify the ReLU activations' necessity by ablating them from the AttnTrans module. As summarized in Table II, adopting ReLU activation without Batch Normalization achieves the highest segmentation performance with 71.4% mean IoU. Embedding Batch Normalization inside the AttnTrans module can even weaken its capacity.

Ablation for spatial filter size inside AttnTrans As summarized in Table III, we measure BiAttnNet's performance with different spatial filter sizes of 5, 7, and 9. Enlarging filter size [24] doesn't consistently boost segmentation performance in our case. Studies show that a spatial filter size of 7 achieves the highest score.

TABLE III
STUDIES OF SPATIAL FILTER SIZE INSIDE ATTNTRANS. A LARGER SPATIAL FILTER SIZE DOESN'T ALWAYS RESULT IN BOOSTING PERFORMANCE. FILTER SIZE OF 7 ACHIEVES THE HIGHEST MEAN IOU

Spatial Filter Size	Number of Filters' Parameters	mean IoU
5	$(5 \times 5 \times 2) \times 3 \times 4 = 600$	69.9%
7	$(7 \times 7 \times 2) \times 3 \times 4 = 1176$	**71.4%**
9	$(9 \times 9 \times 2) \times 3 \times 4 = 1944$	70.8%

TABLE IV
STUDIES OF BILATERAL ATTENTION STRUCTURE. BOTH STRUCTURES USE ATTNTRANS AS THE ATTENTION MECHANISM FOR EXPLORING RICHER DETAILS. PERFORMANCE RESULTS CONCLUDE THAT THE BILATERAL ATTENTION STRUCTURE IS SUPERIOR IN FEATURE EMPHASIZING

Structure	mean IoU
Single-branch Attention Intergrading	68.6%
Bilateral Attention	**71.4%**

TABLE V
STUDIES OF CONCURRENT SPATIAL AND CHANNEL ATTENTION. ATTNTRANS OUTPERFORMS SCSE FOR ITS FULL ABILITIES IN FILTERING BOTH DIRECTIONS

Concurrent Spatial and Channel Attention	mean IoU
scSE	67.7%
AttnTrans	**71.4%**

Ablation for Bilateral Attention Structure Our BiAttnNet's core concepts is separating all attention mechanisms into a parallel shallow branch for specialized detail selection. We verify the superiority of this design by comparing our Bilateral Attention structure with the regular single-branch attention intergrading approach. The latter's all attention mechanisms are attached between convolutional downsample blocks for recalibration. We implement the Bilateral Attention structure as the BiAttenNet. Single-branch attention intergrading is implemented by moving all groups of AttnTrans of BiAttenNet's Detail Branch to the back of their corresponding Context Branch's residual stages. As summarized in Table IV, the Bilateral Attention structure outperforms the single-branch approach by 2.8%, indicating BiAttnNet's design is more effective in boosting semantic segmentation.

C. Studies of Concurrent Spatial and Channel Attention

The scSE [15] mechanism implements concurrent spatial and channel attention as a recalibrating mechanism with a different approach. It is not designed as a replacement for regular convolution like AttnTrans. The scSE's spatial recalibrating is implemented by a scalar multiplication with bias. Its spatial and channel attention maps are added element-wise, and it doesn't have any built-in reshaping function. We compare scSE with AttnTrans by replacing BiAttnNet's concurrent attention with scSE's approach. As summarized in Table V, AttnTrans outperforms scSE by 3.7%, indicating AttnTrans's superiority.

TABLE VI
COMPARING WITH OTHER METHODS. OUR APPROACH ACHIEVES A COMPETITIVE TRADE-OFF BETWEEN ACCURACY AND SPEED WITH 512×1024 INPUTS FROM THE CITYSCAPES TESTING SET. ALL MODEL'S INFERENCE TIME IS MEASURED ON ONE GTX 2080 TI

Method	FPS	Para(M)	mean IoU
ENet [25]	51.9	**0.4**	58.3%
ContextNet [1]	**95.5**	0.9	67.3%
ERFNet [26]	42.7	2.1	68.0%
BiSeNet [2]	93.7	14.1	68.4%
ESNet [27]	41.7	1.6	69.1%
ICNet [4]	40.1	26.5	69.5%
LRR-4x [28]	-	-	69.7%
DFANet A [5]	43.5	7.8	70.3%
LEDNet [29]	77.1	0.94	70.6%
FBSNet [30]	-	0.6	70.9%
STDC1-50 [31]	87.1	8.4	71.9%
FPANet A [32]	-	14.1	72.0%
LRNNet [33]	-	0.7	72.2%
BiSeNetV2 [3]	84.9	27.7	72.6%
Faster BiSeNet [34]	-	3.2	72.8%
SegBlocks-RN18 [35]	-	> 11	73.8%
Ours	89.2	2.2	**74.7%**

D. Comparing With Other Methods

We compare BiAttnNet with other real-time semantic segmentation methods on the Cityscapes testing set. We measure all available models' inference speed with the same hardware on our platform and inputs at 512×1024 from the Cityscapes testing set. The number of parameters is counted in millions.

As summarized in Table VI, our BiAttnNet archives a competitive accuracy of 74.7% mean IoU. Meanwhile, BiAttnNet can run at 89.2 FPS at the quarter resolution of 512×1024. Although BiAttnNet's inference speed is about 6.3 FPS slower than ContextNet, which holds the highest inference speed, BiAttnNet's segmentation accuracy outperforms ContextNet by 7.4%, indicating that BiAttnNet achieves a better balance between speed and accuracy. Our BiAttnNet is not as lightweight as ENet, but parameters of 2.2 million are acceptable enough for equipment with limited resources. BiAttnNet's segmentation precision outperforms SegBlocks-RN18 by 0.9% while using much lesser parameters, proving that BiAttnNet is more parameter effective.

E. Conclusion

This work proposes BiAttnNet based on Bilateral Attention structure and AttnTrans to improve real-time semantic segmentation. Ablation experiments have verified the superiority of the Bilateral Attention structure over single-branch attention intergrading and the rationality of AttnTrans' concurrent spatial and channel attention design. BiAttnNet achieves outstanding performance on Cityscapes testing set with an accuracy of 74.7% mIoU and a speed of 89.2FPS at 512×1024 resolution, with only 2.2 million parameters and a single GTX 2080 Ti card for inference. Additionally, AttnTrans' generic recalibration ability and potential to work as a standalone sampling mechanism can be studied further, which will be carried out in future work.

REFERENCES

[1] R. P. K. Poudel, U. Bonde, S. Liwicki, and C. Zach, "ContextNet: Exploring context and detail for semantic segmentation in real-time," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–11.

[2] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 325–341.

[3] 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," *Int. J. Comput. Vision*, vol. 129, no. 11, pp. 3051–3068, 2021.

[4] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 418–434.

[5] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 9522–9531.

[6] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[7] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 3146–3154.

[8] G. Zamzmi, V. Sachdev, and S. K. Antani, "Trilateral attention network for real-time medical image segmentation," 2021,*arXiv:2106.09201*.

[9] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[10] R. Irfan, A. A. Almazroi, H. T. Rauf, R. Damaševičius, E. A. Nasr, and A. E. Abdelgawad, "Dilated semantic segmentation for breast ultrasonic lesion detection using parallel feature fusion," *Diagnostics*, vol. 11, no. 7, pp. 1–20, 2021.

[11] S. Kadry, F. Al-Turjman, and V. Rajinikanth, "Automated segmentation of COVID-19 lesion from lung CT images using U-Net architecture," in *Proc. 6th EAI Int. Conf. Sci. Technol. Smart Cities*, vol. 372, pp. 20–30, 2020.

[12] S. Kadry, R. Damasevicius, D. Taniar, V. Rajinikanth, and I. A. Lawal, "U-Net supported segmentation of Ischemic-Stroke-Lesion from brain MRI slices," in *Proc. 7th Int. Conf. Bio Signals, Images, Instrum.* 2021, pp. 1–5.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7132–7141.

[14] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1857–1866.

[15] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze excitation' in fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2018, pp. 421–429.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431–3440.

[18] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1800–1807.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[20] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 3213–3223.

[21] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.

[22] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[23] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[24] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters - improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1743–1751.

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

[26] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized convNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.

[27] Y. Wang, Q. Zhou, J. Xiong, X. Wu, and X. Jin, "EsNet: An efficient symmetric network for real-time semantic segmentation," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2019, pp. 41–52.

[28] G. Ghiasi and C. C. Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 519–534.

[29] Y. Wang *et al.*, "LEDNet: A lightweight encoder-decoder network for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1860–1864.

[30] G. Gao, G. Xu, J. Li, Y. Yu, H. Lu, and J. Yang, "FBSNet: A fast bilateral symmetrical network for real-time semantic segmentation." 2021, *arXiv:2109.00699*.

[31] M. Fan *et al.*, "Rethinking biSeNet for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 9716–9725.

[32] Y. Wu, J. Jiang, Z. Huang, and Y. Tian, "FPANet: Feature pyramid aggregation network for real-time semantic segmentation," *Appl. Intell.*, pp. 1–18, 2021, doi: 10.1007/s10489-021-02603-z.

[33] W. Jiang, Z. Xie, Y. Li, C. Liu, and H. Lu, "LRNNet: A light-weighted network with efficient reduced non-local operation for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–6.

[34] T. Verelst and T. Tuytelaars, "SegBlocks: Block-based dynamic resolution networks for real-time segmentation," 2020, *arXiv:2011.12025*.