

BEVANET: BILATERAL EFFICIENT VISUAL ATTENTION NETWORK FOR REAL-TIME SEMANTIC SEGMENTATION

Ping-Mao Huang^{*} I-Tien Chao^{*} Ping-Chia Huang[†] Jia-Wei Liao[†] Yung-Yu Chuang[†]

^{*}Graduate Institute of Networking and Multimedia, National Taiwan University

[†]Department of Computer Science and Information Engineering, National Taiwan University

ABSTRACT

Real-time semantic segmentation presents the dual challenge of designing efficient architectures that capture large receptive fields for semantic understanding while also refining detailed contours. Vision transformers model long-range dependencies effectively but incur high computational cost. To address these challenges, we introduce the Large Kernel Attention (LKA) mechanism. Our proposed Bilateral Efficient Visual Attention Network (BEVANet) expands the receptive field to capture contextual information and extracts visual and structural features using Sparse Decomposed Large Separable Kernel Attentions (SDLSKA). The Comprehensive Kernel Selection (CKS) mechanism dynamically adapts the receptive field to further enhance performance. Furthermore, the Deep Large Kernel Pyramid Pooling Module (DLKPPM) enriches contextual features by synergistically combining dilated convolutions and large kernel attention. The bilateral architecture facilitates frequent branch communication, and the Boundary Guided Attention Fusion (BGAF) module enhances boundary delineation by integrating spatial and semantic features under boundary guidance. BEVANet achieves real-time segmentation at 33 FPS, yielding 79.3% mIoU without pretraining and 81.0% mIoU on Cityscapes after ImageNet pretraining, demonstrating state-of-the-art performance.

Index Terms— Real-time Semantic Segmentation, Large Kernel Attention, Adaptive Feature Fusion

1. INTRODUCTION

Semantic segmentation, assigning class labels to every pixel, is vital in computer vision. Initially reliant on hand-crafted features, it evolved with Fully Convolutional Networks [1] and models like UNet [2], improving performance with encoder-decoder structures and skip connections. PSPNet [3] enhanced results using pyramid pooling. However, computational demands of many models render them impractical for real-time applications like autonomous driving and robotics.

Real-time semantic segmentation seeks to balance speed and accuracy through lightweight architectures, often lever-

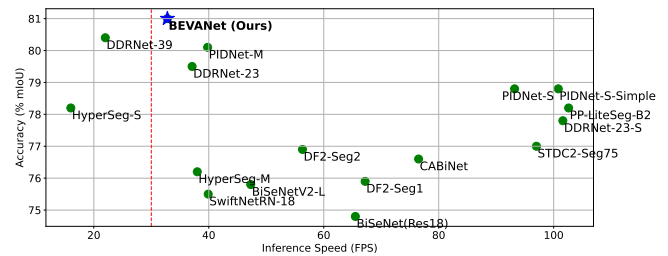


Fig. 1. Performance of real-time models on the Cityscapes [4] validation set, with our model in blue and others in green.

aging depth-wise separable convolutions. Recent models like BiSeNet [5], STDC [6], and DDRNet [7] blend spatial and semantic features to improve efficiency. PIDNet [8] achieves state-of-the-art results by applying Proportional-Integral-Derivative controller principles, incorporating three branches to process detailed, contextual, and boundary information. However, there are insufficient receptive fields for contours.

The Vision Transformer [9] and its variants introduced superior long-range dependency modeling but are computationally heavy. Efficient attention mechanisms such as SeaFormer [10] attempt to optimize this, but challenges remain.

Recent approaches like RepLKNet [11] highlight the advantages of Large Kernel Attention (LKA), integrating convolution and attention mechanisms to capture global context effectively. SLaK [12] expanded kernel sizes to 51 by replacing a large kernel with two long parallel kernels and a small kernel. VAN [13] and LSKA [14] further optimized this with dilated and strip convolution, reducing computational demands. LSKNet [15] introduced the selective kernel concept from SKNet [16]. However, these approaches still lack multi-scale feature integration and receptive field adjustment.

We introduce the LKA mechanism into our Bilateral Efficient Visual Attention Network (BEVANet) to address challenges in real-time semantic segmentation, such as capturing contour details, semantic context, and fusing features at different levels. Our design features the Efficient Visual Attention (EVA) block with Sparse Decomposed Large Separable Kernel Attention (SDLSKA) to expand the receptive field, capture semantic context, and refine details. We design the

Thanks to NSTC, ASUSTek, and NTU for their support through grants 113-2634-F-002-007, 113-2221-E-002-112-MY3, 113L9009 and 114L892203.

Comprehensive Kernel Selection (CKS) mechanism, which integrates features from both small and large kernels using dynamic channel and spatial attention. Additionally, we propose the Deep Large Kernel Pyramid Pooling Module (DLKPPM) to enrich contextual features and prevent information loss typically caused by pooling and striding in traditional methods. We also develop a bilateral architecture that facilitates continuous communication between two branches and the Boundary Guided Attention Fusion (BGAF) module that adapts semantic and detail fusion with boundary information. The interactions of these branches improve the representation of features by integrating various features. As shown in Fig. 1, BEVNet offers a robust and efficient framework that achieves state-of-the-art real-time segmentation by effectively balancing accuracy and efficiency. Main contributions are listed as follows.

- **Efficient Attention Mechanisms.** We leverage large kernel attention to design the EVA block, SDLSKA, CKS, and DLKPPM modules. These components enlarge and dynamically adjust receptive fields, enhance feature representation, capture contextual information, and refine object details, improving spatial modeling.
- **Branch Interaction.** Frequent communication between high- and low-level branches through the bilateral architecture and the BGAF module enhances semantic concepts and detail contour by sharing information and shortcut, enabling adaptive feature fusion.
- **Performance.** BEVNet offers a superior balance of speed and accuracy to existing models. It achieves real-time segmentation over 30 FPS with 81.0% mIoU on Cityscapes after ImageNet pre-training and 79.3% without, indicating reduced dependency on large pre-training datasets. Its variant BEVNet-S further achieves 83% mIoU on CamVid, demonstrating its scalability.

2. METHODOLOGY

We propose the Efficient Visual Attention (EVA) module, utilizing Sparse Decomposed Large Separable Kernel Attentions (SDLSKA) and Comprehensive Kernel Selection (CKS) to adaptively enlarge the receptive field. The Deep Large Kernel Pyramid Pooling Module (DLKPPM) leverages large kernels for contextual enrichment. Additionally, the Bilateral Architecture (BA) and Boundary Guided Attention Fusion (BGAF) facilitate feature interaction across two branches.

2.1. Bilateral Architecture

Inspired by PIDNet [8], we propose BA, which frequently integrates semantic context into detailed features. As depicted in Fig. 3, it features a high-level branch to capture semantic concepts by reducing feature maps and a low-level branch to extract contours and detect boundaries while maintaining resolution. This enables continuous branch interaction to improve semantic understanding and object boundary.

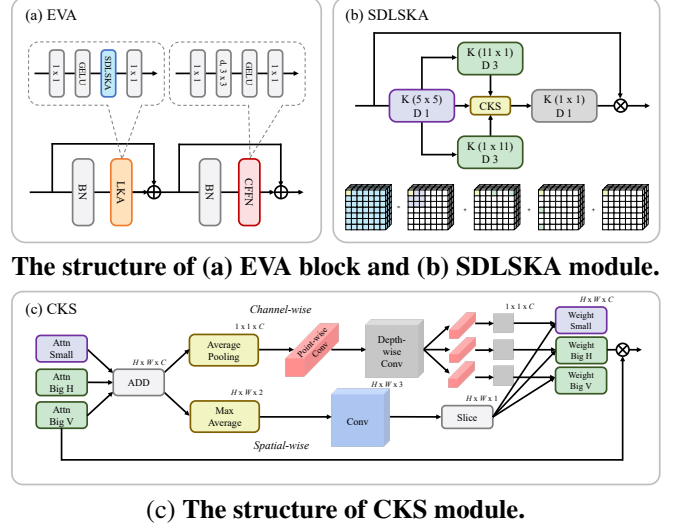


Fig. 2. Our EVA block (a) utilizes the SDLSKA (b) and CKS (c) modules to adaptively expand the vision. The SDLSKA module extracts large features using a 5×5 convolution, followed by 1×11 and 11×1 strip convolutions with a dilation rate of 3, achieving a receptive field of 35. These features are fused with the CKS mechanism and refined via a pointwise convolution. The CKS module computes weights across channel and spatial dimensions through pooling and aggregation. In the spatial branch, a convolution operation generates three feature channels. The channel branch refines features using pointwise and depthwise convolution. The fusion of spatial and channel weights, facilitated by the multiplication.

2.2. Efficient Visual Attention Block

Our EVA Block is inspired by the robust block design in VAN [13] and LSKA [14], demonstrated in Fig. 2. It consists of two main components: the Large Kernel Attention (LKA) and the Convolution Feed-Forward Network (CFFN).

The LKA block captures long-range dependencies. By utilizing SDLSKA, it expands the receptive field to a broader context, resulting in more effective semantic capture. It also uses the CKS mechanism to combine features from kernels of different sizes and adjusts the vision accordingly. This synergistic combination renders EVA Block an efficient module for robust feature extraction and precise feature representation.

The CFFN block refines and integrates features to ensure a well-balanced and informative output. In addition, it refines features channel-wise using pointwise convolution.

2.2.1. Sparse Decompose Large Separable Kernel Attentions

The SDLSKA module is designed to expand the receptive field to effectively capture semantic information and refine details. The structure is depicted in Fig. 2(b). Drawing from SLak [12], we simplify large-kernel computation through sparse grouping by decomposing them into a smaller convo-

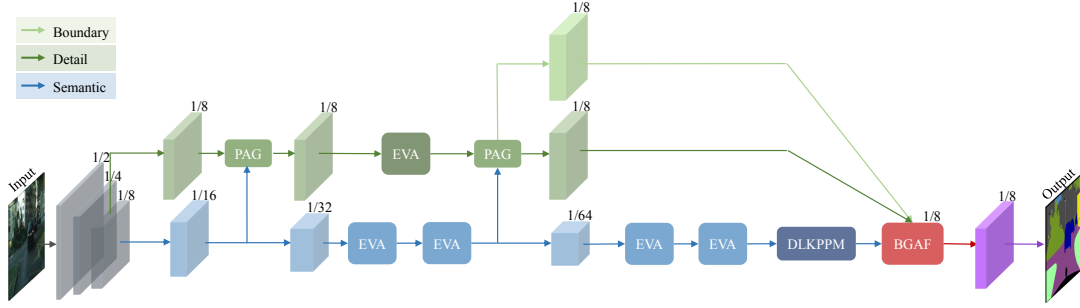


Fig. 3. The overall structure of the BEVNet. The high-level branch captures contextual information and long-range dependencies, enriching the spatial branches with semantics. Meanwhile, the low-level branch preserves high-resolution details and high-frequency features for accurate boundaries, maintaining a resolution of 1/8 of the original size to retain fine contours.

lution and two strip dilation kernels, then adaptively fusing them using CKS module. The smaller convolution helps focus on specific areas, while the two strip dilation kernels refine the focus, with low computation. Additionally, inspired by LSKA [14], we combine strip convolutions with depth-wise, pointwise, and dilated convolutions to capture large-kernel features efficiently. This approach reduces parameters while leveraging 2D structural information, resulting in better computational efficiency. It also adapts effectively to spatial and channel dimensions to capture long-range dependencies.

2.2.2. Comprehensive Kernel Selection

Our CKS module innovatively performs joint channel-wise and spatial-wise adjustments, crucial for the receptive field adaptation of SDLSKA and for dynamic multi-scale feature fusion from kernels of varying shapes. This integration surpasses SKNet [16] and LSKNet [15], which address these dimensions in a decoupled manner. This approach captures the interdependence between channels and spatial dimensions, which is crucial for effective feature integration. As shown in Fig. 2(c), the module efficiently manages complex fusion across these dimensions. By merging features across kernel scales and considering interdependencies, it enables a holistic, adaptable fusion that leads to richer representations.

2.3. Deep Large Kernel Pyramid Pooling Module

DLKPPM preserves the hierarchical-residual structure and integrates kernels across various scales and depths, as illustrated in Fig. 4, inspired by DAPPM [7]. To mitigate the loss of spatial details typically induced by large strides and pooling operations in conventional pyramid modules, DLKPPM uniquely integrates dilation convolutions and LKA mechanism with LSKA [14], expanding the receptive field to 35. This efficient approach refines features and enhances semantic concept capture, while preserving crucial spatial information.

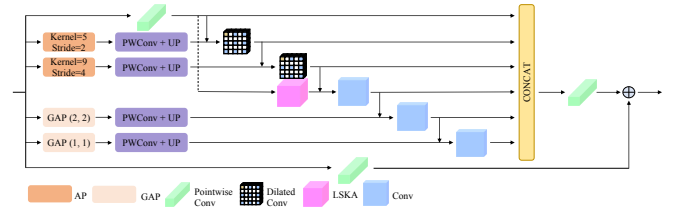


Fig. 4. The structure of DLKPPM. DLKPPM fuses multi-scale information using hierarchical kernels of various depths and sizes, minimizes the information loss from pooling with LKA, and uses dilated convolutions for small kernels.

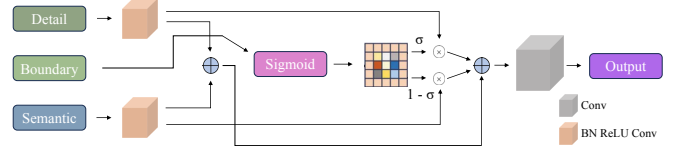


Fig. 5. The structure of BGAF module. Semantic and detail information are refined through BN, ReLU, and convolution, while the boundary feature employs Sigmoid activation to compute a balancing weight σ for adaptive merging. The balanced feature and shortcut are combined element-wise and processed through a final convolution to generate the output.

2.4. Boundary Guided Adaptive Fusion

BGAF provides efficient multi-branch aggregation, balancing contextual and spatial features with boundary information. As shown in Fig. 5, it employs a shortcut residual connection to preserve critical feature information, thereby avoiding degradation from simple weighted summation. The module dynamically fuses high-level and low-level features, integrating robust semantic understanding with precise contour details. It further mitigates the limitations of the context branch's spatial precision and the detail branch's shallow semantic representation by adaptively adjusting feature contributions based on boundary significance. This design ensures precise detection of complete object boundaries and fine-grained structures.

Model	Resolution	GPU	FPS \uparrow	#GFLOPs \downarrow	#Params (M) \downarrow	mIoU (%) \uparrow
BiSeNet(Res18) [5]	1536 \times 768	GTX 1080Ti	65.5	55.3	49	74.8
BiSeNetV2-L [5]	1024 \times 512	GTX 1080Ti	47.3	118.5	-	75.8
STDC1-Seg75 [6]	1536 \times 768	RTX 3090	74.8	-	-	74.5
STDC2-Seg75 [6]	1536 \times 768	RTX 3090	58.2	-	-	77.0
PP-LiteSeg-T2 [17]	1536 \times 768	RTX 3090	96.0	-	-	76.0
PP-LiteSeg-B2 [17]	1536 \times 768	RTX 3090	68.2	-	-	78.2
SFNet(DF2) [18]	2048 \times 1024	RTX 3090	87.6	-	10.53	77.8
SFNet(ResNet-18) [18]	2048 \times 1024	RTX 3090	30.4	247.0	12.87	78.9
DDRNet-23-S [7]	2048 \times 1024	RTX 3090	108.1	36.3	5.7	77.8
DDRNet-23 [7]	2048 \times 1024	RTX 3090	51.4	143.1	20.1	79.5
PIDNet-S-Simple [8]	2048 \times 1024	RTX 3090	100.8	46.3	7.6	78.8
PIDNet-S [8]	2048 \times 1024	RTX 3090	93.2	47.6	7.6	78.8
PIDNet-M [8]	2048 \times 1024	RTX 3090	39.8	197.4	34.4	80.1
BEVNet (Ours)	2048 \times 1024	RTX 3090	32.8	238.2	58.6	81.0

Table 1. Overall Quantitative Comparisons on Cityscapes [4] Validation Set. The bold numbers, indicating the best performance, emphasize the overall superiority of BEVNet. Most of the results are adopted from PIDNet [8].

Model	GPU	FPS \uparrow	#GFLOPs \downarrow	mIoU (%) \uparrow
PP-LiteSeg-T [17]	GTX 1080Ti	154.8	-	75.0
BiSeNetV2 [5]	GTX 1080Ti	124.0	-	76.7
BiSeNetV2-L [5]	GTX 1080Ti	33.0	-	78.5
DDRNet-23-S [7]	RTX 3090	182.4	-	78.6
DDRNet-23 [7]	RTX 3090	116.8	-	80.6
PIDNet-S [8]	RTX 3090	153.7	15.8	80.1
PIDNet-S-Wider [8]	RTX 3090	85.6	59.1	82.0
BEVNet-S (Ours)	RTX 3090	79.4	20.1	83.1

Table 2. Quantitative Comparisons on CamVid [19]. Most of the results are adopted from PIDNet [8].

Model	FPS \uparrow	#GFLOPs \downarrow	mIoU (%) \uparrow
PIDNet-S [8]	93.2	47.6	76.32
PIDNet-M [8]	39.8	197.4	78.22
PIDNet-L [8]	31.1	275.8	78.25
BEVNet (Ours)	32.9	238.2	79.27

Table 3. Quantitative Comparisons of Model Performance without Pretraining. Ours outperforms PIDNet [8].

3. EXPERIMENTS

3.1. The Datasets and Implementation Details

We mainly evaluated on Cityscapes [4], a widely recognized benchmark dataset for urban scene parsing, containing 2,975 training, 500 validation, and 1,525 testing images with a high resolution of 2048 \times 1024. It includes 19 classes for semantic segmentation evaluation. We also use the CamVid [19] dataset, consisting of 701 driving scene images at 960 \times 720 resolution. It selects 11 classes of 32 annotated categories.

Our model is pretrained on ImageNet [20] using random cropping to 224 \times 224 and horizontal flipping, consistent with prior works [5, 6]. Pretraining runs for 100 epochs with a batch size of 256, employing SGD optimizer with a learning rate of 0.1, weight decay of 0.0001, and momentum of 0.9.

The main training phase, which follows similar protocols as in previous studies [5, 6, 7, 8], runs for 484 epochs with a batch size of 12 and a learning rate of 0.008 for Cityscapes, and 200 epochs with a batch size of 24 and a learning rate of 0.003 for CamVid [19]. It employs a poly decay learning rate schedule (scaling from 0.5 to 2.0) and Online Hard Example Mining (OHEM) applied for challenging sample selection.

All inference benchmarks were conducted on an NVIDIA RTX 3090 GPU using PyTorch 2.4, CUDA 12.1, and Ubuntu 20.04, with a batch size of 1 for single-sample processing.

3.2. Quantitative Comparisons

Comparisons with pre-training. In this setting, all models are pre-trained on ImageNet [20] to ensure a fair comparison. As shown in Table 1, our BEVNet achieves state-of-the-art (SoTA) performance on Cityscapes [4] above the real-time threshold of 30 FPS. With 81% mIoU, our model outperforms similarly scaled benchmark models by leveraging optimized attention mechanisms and adaptive feature fusion, thereby delivering high accuracy and effectiveness without compromising speed. Furthermore, while PIDNet-M [8] gains 1.3% mIoU over PIDNet-S [8] at a cost of 53 FPS, BEVNet improves upon PIDNet-M [8] by 0.9% mIoU with only a 7 FPS reduction. Table 2 shows small-scale BEVNet-S also reaches SoTA on CamVid [19] with 20.1 GFLOPs, indicating scalability and suitability for edge applications. Notably, BEVNet-S reduces by nearly 40 GFLOPs compared to PIDNet-S-Wider [8], yet still achieves 1.1% higher mIoU, clearly demonstrating its superior efficiency and accuracy.

Comparisons without pre-training. In this setting, we compare our model primarily against the SoTA model PIDNet [8]. As shown in Table 3, without pre-training on ImageNet [20], BEVNet achieves 79.3% mIoU, outperforming PIDNet-M [8]. Furthermore, it surpasses PIDNet-L [8] by 1% mIoU while delivering higher FPS. These results highlight the model’s efficiency and reduced dependency on large-scale pre-training, suggesting its potential for real-time applications where computational resources and labeled data are limited.

Architecture	Block	Selection Kernel	Branch Fusion	FPS \uparrow	mIoU (%) \uparrow
PIDNet [8]	Convs	-	Bag [8]	42.64	78.22
BA (Ours)	Convs	-	Bag [8]	44.25	77.77
BA (Ours)	SLaK[12]	-	Bag [8]	37.76	77.84
BA (Ours)	LSKA[14]	-	Bag [8]	41.16	77.93
BA (Ours)	SDLSKA (Ours)	Addition	Bag [8]	37.79	78.60
BA (Ours)	SDLSKA (Ours)	LSKNet [15]	Bag [8]	37.46	78.72
BA (Ours)	SDLSKA (Ours)	CKS (Ours)	Bag [8]	37.29	78.86
BA (Ours)	SDLSKA (Ours)	CKS (Ours)	Light_Bag [8]	38.48	78.39
BA (Ours)	SDLSKA (Ours)	CKS (Ours)	BGAF (Ours)	32.85	79.27

Table 4. The ablation study comparisons of our modules without pretraining. Each color represents a different module in the ablation study, with our module located at the bottom of each colored area.

PPM	FPS \uparrow	mIoU (%) \uparrow
DAPPM [7]	32.85	80.43
PAPPM [8]	33.38	79.97
DLKAPPM (Ours)	32.47	80.96

Table 5. The ablation study comparisons of PPM.

3.3. Ablation Study

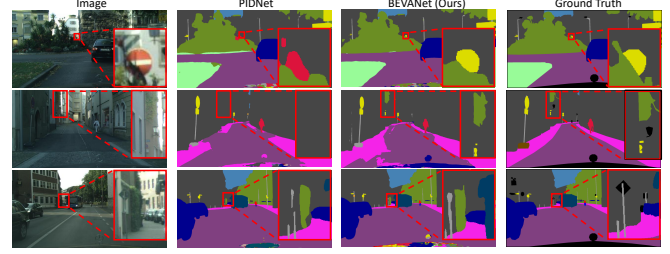
Architecture Efficiency. As shown in the first two rows of Table 4, our architecture demonstrates increased speed with only a marginal reduction in performance, showcasing the efficiency of its semantic concept and spatial detail branches.

Large Kernel Attention. Our EVA block integrates the SDLSKA and CKS modules. The SDLSKA module, which combines LSKA [14] and SLaK [12], significantly outperforms its components by improving the attention of the large kernel. As indicated by the orange section in Table 4, this improvement boosts mIoU by 0.8% to 78.6% while effectively capturing global context through an expanded receptive field. The pink section in Table 4 indicates CKS achieves a 0.26% mIoU gain over LSKNet [15] with a marginal 0.5 FPS reduction, highlighting its efficient fusion of multi-scale kernels across varying shapes and its integration of small and strip dilation kernels to balance spatial and channel information. Collectively, the EVA block improves mIoU by 1.1% with only a minor decrease in speed, thereby preserving real-time performance.

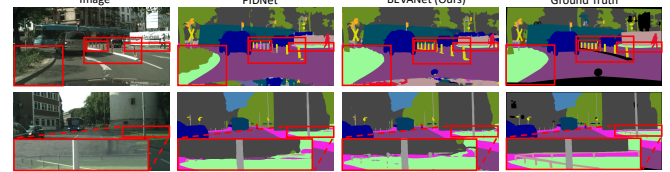
Branch Fusion. Our BGAF module outperforms BAG [8] by 0.4% mIoU, as highlighted in the purple area of Table 4. It enhances feature representation with shortcut paths and adaptive merging low- and high-level features with boundary information, emphasizing the value of balanced fusion and shortcuts.

Performance Enhancement. In the second and last rows in Table 4, under 30 FPS requirement, these innovations boost mIoU by 1.5% to 79.3%, demonstrating its effectiveness.

Multi-scale Fusion. For the first and last rows in Table 5, our DLKPPM improves contextual fusion, providing a richer context and boosting mIoU by 0.5% with a minimal 0.4 FPS reduction. Its larger receptive field reduces pooling information loss, making it well-suited for real-time applications.



(a) The visualization comparison for the small objects.



(b) The visualization comparison for the completeness.

Fig. 6. Visualization comparisons. Red denotes people, yellow indicates traffic signs, and green represents vegetation. From left to right, the columns display the image, PIDNet [8] (baseline), BEVNet (ours), and Ground Truth, with each row corresponding to a different instance.

3.4. Qualitative Analysis

Small Object. Our BEVNet outperforms PIDNet [8] in detecting small objects, which are inherently difficult to identify. As depicted in Fig. 6(a), it rectifies misclassifications such as confusing traffic signs for people, and detects plants and traffic signs that PIDNet [8] misses. BEVNet demonstrates the capability to identify unlabeled objects such as vegetation, indicating robust semantic understanding, accurate predictions, and a deeper comprehension of contextual information.

Completeness. In Fig. 6(b), our BEVNet leverages a large receptive field for thorough and accurate object detection, outperforming PIDNet in completely capturing traffic cones and grass. BEVNet also excels in handling larger objects and reliably detects sidewalks, where PIDNet [8] fails, demonstrating its superiority in capturing reliable spatial information.

4. CONCLUSION

Our BEVNet model achieves competitive performance compared to state-of-the-art methods while reaching real-time processing at 33 FPS. Its key features, including the SDLSKA block for expanding receptive fields and the CKS mechanism for dynamic adjustments, enable accurate small object detection and refined boundaries. The bilateral architecture efficiently communicates between feature levels, and the BGAF module further enhances feature fusion. Additionally, our DLKPPM enriches feature representations. Future research will aim to optimize fusion strategies and reduce computational overhead to develop a lightweight large kernel attention model to further improve BEVNet's efficiency.

5. REFERENCES

- [1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
- [2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer, 2015.
- [3] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
- [4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [5] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International journal of computer vision*, vol. 129, pp. 3051–3068, 2021.
- [6] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei, "Rethinking bisenet for real-time semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.
- [7] Yuanduo Hong, Huihui Pan, Weichao Sun, and Yisong Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," *arXiv preprint arXiv:2101.06085*, 2021.
- [8] Jiacong Xu, Zixiang Xiong, and Shankar P Bhat-tacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19529–19539.
- [9] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [10] Qiang Wan, Zilong Huang, Jiachen Lu, YU Gang, and Li Zhang, "Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation," in *The eleventh international conference on learning representations*, 2023.
- [11] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11963–11975.
- [12] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal Mocanu, and Zhangyang Wang, "More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity," *arXiv preprint arXiv:2207.03620*, 2022.
- [13] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu, "Visual attention network," *Computational Visual Media*, vol. 9, no. 4, 2023.
- [14] Kin Wai Lau, Lai-Man Po, and Yasar Abbas Ur Rehman, "Large separable kernel attention: Rethinking the large kernel attention design in cnn," *Expert Systems with Applications*, vol. 236, pp. 121352, 2024.
- [15] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li, "Large selective kernel network for remote sensing object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16794–16805.
- [16] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.
- [17] Juncai Peng, Yi Liu, Shiyu Tang, Yuying Hao, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Zhiliang Yu, Yuning Du, et al., "Pp-liteseg: A superior real-time semantic segmentation model," *arXiv preprint arXiv:2204.02681*, 2022.
- [18] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong, "Semantic flow for fast and accurate scene parsing," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* 16. Springer, 2020, pp. 775–793.
- [19] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern recognition letters*, vol. 30, no. 2, pp. 88–97, 2009.
- [20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, 2015.