PyraBiNet: A Hybrid Semantic Segmentation Network Combining PVT and BiSeNet for Deformable Objects in Indoor Environments

Zehan Tan[1], Weidong Yang[1,2(✉)], and Zhiwei Zhang[3]

[1] School of Computer Science, Fudan University, Shanghai, China
{18110240062,wdyang}@fudan.edu.cn
[2] Zhuhai Fudan Innovation Institute, Hengqin New Area, Zhuhai, Guangdong, China
[3] Gree Electric Appliances, INC. of Zhuhai, Zhuhai, China
zzwyyds0606@gmail.com

Abstract. In this study, we introduce PyraBiNet, an innovative hybrid model optimized for lightweight semantic segmentation tasks. This model ingeniously merges the merits of Convolutional Neural Networks (CNNs) and Transformers. We propose a dual-branch structure that strategically employs the global feature extraction capabilities of the Pyramidal Vision Transformer (PVT) and the local feature extraction proficiency of BiSeNet. Specifically, the global feature branch employs a transformer from PVT to harness high-level patterns from input images, while the local feature branch utilizes a CNN, inspired by BiSeNet, to extract fine-grained details. Comprehensive evaluations conducted on the ADE20K and DOS datasets underscore PyraBiNet's superior performance compared to the existing state-of-the-art methods. With its effective and efficient performance, PyraBiNet proves to be an invaluable asset in the domain of mobile robotics, particularly beneficial for applications such as sweeping robots. The code source and dataset are open at https://github.com/zehantan6970/PyraBiNet.

Keywords: Image processing · Semantic Segmentation · Real-time processing

1 Introduction

Semantic segmentation is a task within the field of computer vision, the goal of which is to classify each pixel in an image, dividing it into distinct semantic categories, thereby enabling a deeper understanding of the image. The challenge of semantic segmentation lies in the precise delineation of object boundaries and assigning them the correct category labels. This necessitates the model to possess substantial perceptual capability, allowing it to comprehend the various objects, colors, textures, and shapes within an image, as well as the relationships among them. Concurrently, the model must be capable of classifying each pixel within

B. Luo et al. (Eds.): ICONIP 2023, CCIS 1968, pp. 552–564, 2024.
https://doi.org/10.1007/978-981-99-8181-6_42

the image since the same object may appear in different locations, sizes, and orientations. With the increasing demand of intelligence, semantic segmentation has become the basic perception component for applications such as autonomous driving [6], medical imaging diagnosis [1] and indoor robot [3,15]. To meet real-time or mobile requirements, researchers have come up with many efficient and effective models in the past for semantic segmentation. The field of lightweight semantic segmentation models has experienced significant evolution, characterized by shifts in underlying network architectures. These transitions can be seen from the initial utilization of Convolutional Neural Networks (CNNs) as typified by Fully Convolutional Networks (FCNs) [21] and extended in BiSeNet series [40,41] and PIDNet [37]. The focus later moved to transformer-based methods, exemplified by LeViT [8] and Pyramid Vision Transformer (PVT) [34]. The latest developments showcase hybrid architectures that combine CNNs with Vision Transformers (ViTs). These include models like the MobileViT series [23,24,33] and Convolutional Vision Transformer (CVT) [36]. Thus, the development of lightweight semantic segmentation models has seen a significant transformation, marked by diverse architectural designs to optimize performance.

By rethinking previous successful lightweight semantic segmentation works with reference to SegNeXt's research [9], we found that these works all face the challenge of how to balance accuracy, parameter scale and inference speed, and improve the fusion of different features. We argue a successful lightweight semantic segmentation model should have the following characteristics:
(i) Feature Extraction: Robust feature extractors not only capture a global features but also discern local detail features. These can acquire features of varying scales.
(ii) Feature Fusion: A rational approach is needed for the integration of local detail features and global features.
(iii) Feature Enhancement: Enhancing the diversity and detailed spatial information of features is essential. Lightweight models have limited capabilities in modeling global relationships, leading to insufficient attention to details in segmentation tasks and often unclear edges.
(iv) Network Architecture Design: The optimization of network structure is necessary, ensuring not only the reasonable utilization of global and local detail features but also control over the number of parameters, while maintaining network inference speed. The key to this network structure is to balance accuracy, parameter scale, and inference speed, while improving the fusion of different features. Given the yearly increase in memory with the widespread use of embedded systems, the size of the parameter scale should be a limiting factor. However, keeping the model size small at the cost of relatively high computation, which also means high latency, is not a sound practice. The parameter volume should not be blindly reduced. Similarly, the network structure should not simply trade accuracy for speed, or vice versa.

Considering the analyses above, we reassess the design of lightweight network architectures for semantic segmentation in this paper. Instead of applying PVT or BiSeNet independently, we propose a novel hybrid architecture, PyraBiNet,

which integrates the strengths of both PVT and BiSeNet. The global feature branch of PyraBiNet, powered by a transformer from PVT, extracts the global features from the input images. Concurrently, the local feature branch, inspired by BiSeNet, utilizes a convolutional neural network (CNN) to capture the local detailed features. Subsequently, these two sets of features are fused to generate a final feature map that is utilized for semantic segmentation.

Our primary contributions are:

- We present a novel lightweight network architecture, termed PyraBiNet, which combines the strengths of convolution (inductive bias, translation invariance, exceptional local detail capture ability, and low computational complexity) and Transformers (ability to capture long-range dependencies) in a dual-branch structure optimized for embedded devices, bolstered by an efficiently parametrized Detail Feature Block that adjusts resolution to align with the global feature branch while effectively capturing local spatial information.
- We introduce the Parallel Dual-Feature CBAM (PDF-CBAM) that concurrently applies a Channel Attention Module to the transformer-derived global features and a Spatial Attention Module to the CNN-derived local features, resulting in an enhanced final feature map that effectively integrates detailed spatial information and diversity of features.
- Our experimental results demonstrate that our proposed architecture achieves state-of-the-art (SOTA) on different benchmarks of ADK20K [46] and our proprietary DOS dataset[1].

2 Related Work

The arena of lightweight semantic segmentation [25,32] has witnessed numerous advances over recent years. We primarily focus on three major neural network types in this context: 1) Convolutional Neural Networks (CNNs), 2) Vision transformers (ViTs), and 3) Hybrids of CNNs and ViTs.

2.1 Convolutional Neural Networks (CNNs)

CNN-based models, such as FCNs [21] and MobileNets [12,13,28], have greatly improved performance by encoding local features, replacing handcrafted [17,18,29,39] systems. Techniques like channel shuffle, micro-factorized convolution, and dynamic operators help enhance information flow and efficiency. Furthermore, novel methods like DDRNet [26] and BiSeNet [31,40,41] utilize bilateral connections and multi-path frameworks to blend low-level details and high-level semantics. PIDNet [37], one of the latest architectures, is composed of three branches to parse the detailed, context, and boundary information. However, despite these advancements, CNNs still have limitations like high computational time and disregard for global information.

[1] https://github.com/zehantan6970/DOS_Dataset.

2.2 Vision Transformers (ViTs)

Drawing from NLP success, transformers have been employed in computer vision tasks [11,14,19], yielding impressive results. Models like ViT [5], DeiT [30], T2T-ViT [44], and Swin Transformer [20] have significantly pushed the boundaries of image classification performance. To create lightweight ViTs, architectures like LeViT [8] and PVT [34] fuse standard convolution layers with improved ViT. PoolFormer [42] replaces the attention module in Transformers with an embarrassingly simple spatial pooling operator to conduct only basic token mixing. Despite these advances, ViTs still face challenges in dealing with visual features of different scales and are often inefficient in terms of memory usage.

2.3 Hybrids of CNNs and ViTs

To capitalize on the strengths of both CNNs and ViTs, hybrid models like MobileViTv3 [33], TopFormer [45], LVT [38], and others have been proposed. These models aim to combine the efficiency of convolution with the global receptive field of Transformers. Other architectures, such as the CeiT [43] and CVT [36], integrate convolutional and self-attention modules in the same architecture. Twins [2] builds upon PVT by substituting its absolute position embedding with relative conditional position embedding and incorporating separable depth-wise convolutions for capturing both local and global image contexts. DFvT [7] opens up the transformer block and enhance it with convolution, both before and after self-attention that tightly integrates transformer and convolution. Despite the efficiency of these hybrid models, they usually come at the cost of performance accuracy.

3 Approach

In this study, we propose PyraBiNet, a novel dual-branch architecture designed to address the challenges of Feature Extraction, Fusion, Enhancement, and Network Architecture Design in lightweight semantic segmentation. PyraBiNet integrates the broad-scale feature extraction capability of PVT with the detailed extraction prowess of BiSeNet. The architecture leverages a transformer from PVT in the global feature branch and a BiSeNet-inspired CNN in the local feature branch for comprehensive Feature Extraction. We utilize a Parallel Dual-Feature Convolutional Block Attention Module for efficient Feature Fusion and a Detail Feature Block for Feature Enhancement. The design of integrating the global feature branch and the local feature branch into a dual branch helps optimize the Network Architecture and provides a robust solution for lightweight semantic segmentation. In this work, the loss function used for training the model is cross-entropy.

Proposed Method: PyraBiNet is a hybrid model that combines the strengths of CNNs and transformers. Figure 1 illustrates the architecture of PyraBiNet, in which the input image is separately processed by the global feature branch and the local feature branch. In our global feature branch, designed following the

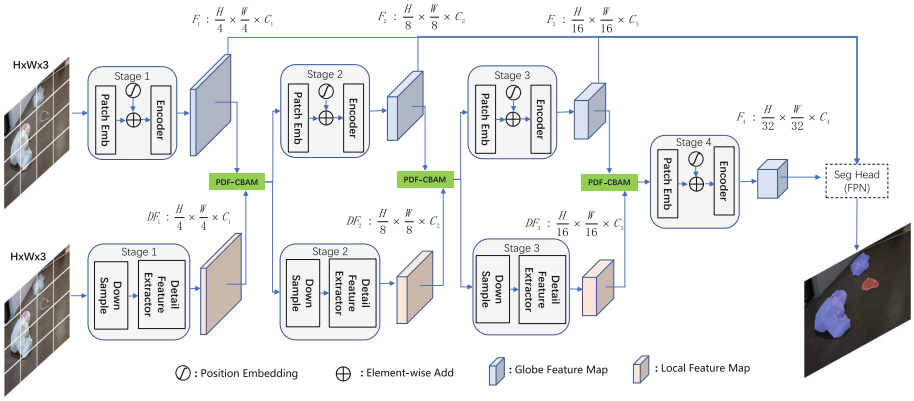

Fig. 1. The pipeline of the proposed PyraBiNet for semantic segmentation. The global feature branch (PVT) contains global branch (up); The local feature branch(reference BiSeNet) contains local branch (down); PyraBiNet contains fusion block and segmentation head.

PVT, the image is processed by a series of self-attention modules. Each of these self-attention modules focuses on different spatial regions of the input image, enabling the PVT to learn global features that are invariant to changes in pose and scale. Following this, the global and local features are fused through a Parallel Dual-Feature Convolutional Block Attention Module (PDF-CBAM), and then separately input into the global feature branch (PVT) and the local feature branch. Notably, in stage 4, only the global feature branch (PVT) is used to generate the final feature map for semantic segmentation. Ultimately, we employ a Semantic FPN [16] as the segmentation head to achieve the final segmentation outcome. Our local feature branch, referenced from BiSeNet, consists of a downsampling module and a Detail Feature Block.

$$\begin{cases} S_i\left(F_g, F_l\right) = PDF\text{-}CBAM_i\left(Attention_i\left(F_g\right), DF_i\left(F_l\right)\right) & , i = 1, 2, 3 \\ S_i\left(F_g\right) = Attention_i\left(F_g\right) & , i = 4 \end{cases} \tag{1}$$

where S_i represents the i-th stage in the architecture of PyraBiNet. PDF-CBAM$_i$ refers to the i-th Parallel Dual-Feature Convolutional Block Attention Module, which is designed for feature fusion. Attention$_i$ symbolizes the self-attention operation implemented at the i-th stage. DF$_i$ denotes the detail feature block deployed at the i-th stage. F_g refers to the global feature map obtained from the global feature branch, and F_l symbolizes the local feature map derived from the local feature branch in our architecture.

Detail Feature Block (DF): The overall framework of the proposed DF is presented in Fig. 2. For each stage, the process begins with a downsampling module to match the resolution of the PVT branch, followed by the use of a detail feature extractor to capture local spatial information. As this involves low-level information, the module requires a substantial channel capacity to encode rich

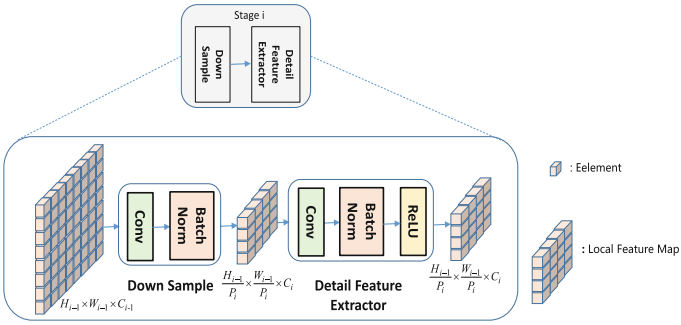

Fig. 2. Detail Feature Block (DF) is meticulously constructed to fulfill two main tasks at each stage: adjusting the resolution to align with the PVT branch, and capturing local spatial information. 'i' represents the corresponding stage. The numbers of Detail Feature Block is set to 1, 2, 2, corresponding to satege1, stage2, stage3.

spatial detail information. More specifically, wide channels and shallow layers are used to process spatial details, following the structural design of BiSeNetV2. Each extractor is composed of 'n' blocks, with each block comprising a convolution, Batch Normalization, and ReLU sequence. Our local feature branch, while inspired by the design philosophy of BiSeNet, diverges significantly from its prototype. BiSeNet consists of two branches, where its global branch employs CNNs with large receptive fields for the implementation. In contrast, our model utilizes PVT for global feature extraction. In terms of the local branch, our design also deviates from BiSeNet in terms of the parameters and quantity of CNN kernels, as well as the overall structure. Additionally, while BiSeNet does not have staged architecture and adopts a pyramid-like method without downsampling, our model is organized into stages, with fusion performed at each stage. Furthermore, the fusion strategy in BiSeNet is achieved by Aggregation Layer before the final feature map, while our approach incorporates a fusion process in each stage.

$$DF_i\left(F_l\right) = Local\_Extractor_i\left(DS_i\left(F_l\right)\right) \ , i = 1, 2, 3 \tag{2}$$

where Local_Extractor signifies the mechanism within our model that facilitates the extraction of local features from the input. DS is an acronym for downsampling in each stage. F_l represents the local features that are extracted and processed within our architecture.

Parallel Dual-Feature CBAM (PDF-CBAM): The feature fusion module, an integral part of semantic segmentation, enhances feature representations. However, in our ablation studies (Table 3), we discovered that straightforward strategies such as element-wise summation, multiplication, and concatenation did not yield satisfactory results when fusing local and global features. Considering that VIT has strong attention on space but weak attention on channels, and CNNs, with their local convolution operations, can naturally capture local

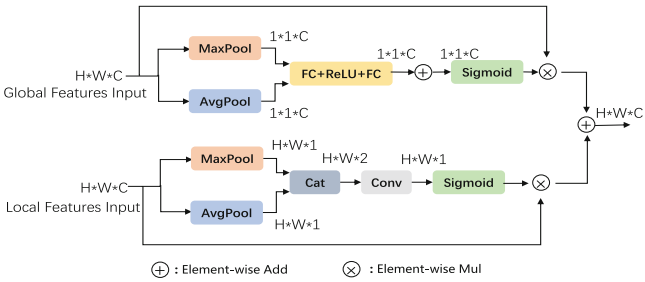

Fig. 3. The architecture of PDF-CBAM. Input the reshaped global feature map (H*W*C) and local feature map (H*W*C) respectively, and the output is a feature map (H*W*C) that combines global features and local features.

details and handle channel-wise information well, but handle space information weak. Notably, the Convolutional Block Attention Module (CBAM) [35] utilizes a sequential combination of the Channel Attention Module and Spatial Attention Module, each processing the input feature layer independently. Aiming for better integration of local detail features with global features and enhancement of the diversity and detailed spatial information of features, we opted for a parallel approach. Our Parallel Dual-Feature CBAM (PDF-CBAM) takes as input local detail features derived from a Convolutional Neural Network (CNN) and global features derived from a transformer. The global features are reshaped into a convolutional feature map (H*W*C), which is then subjected to a Channel Attention Module. Concurrently, a Spatial Attention Module is applied to the local feature map (H*W*C). Finally, the outputs from the Channel Attention Module and Spatial Attention Module are combined via element-wise summation to produce the final feature map, thereby resolving the issue of lack of information screening inherent to simple element-wise summation, multiplication, and concatenation. The architecture of PDF-CBAM is shown in Fig. 3.

$$PDF\text{-}CBAM = (CA(F_g) \otimes F_g) \oplus (SA(F_l) \otimes F_l) \tag{3}$$

where CA denotes Channel Attention, a CBAM that focuses on the channel-wise information of the input features. SA stands for Spatial Attention, another component of the CBAM, which pays attention to the spatial arrangement of the features. F_g refers to the global features derived from the transformer in the PyraBiNet architecture, whereas F_l represents the local features extracted by the CNN within the same architecture.

Deformable Object Segmentation Dataset for Sweeping Robots (DOS Dataset): We present a novel dataset, designed specifically to serve as a benchmark for semantic segmentation of deformable objects within the context of obstacle avoidance in indoor robotic sweeping scenarios. DOS dataset comprises 3,056 images, We used the open-source LabelMe [27] annotation toolkit, to manually collect the polygon annotations of deformable objects of four types: faeces, socks, plastic bag, and rope. DOS dataset has 7687 annotated object instances.

4 Experiments

PyraBiNet was evaluated on the ADE20K dataset [46] and DOS dataset. ADE20K is a demanding scene parsing dataset designed to benchmark the performance of semantic segmentation. This dataset comprises 150 highly-detailed semantic categories and features 20,210 training images, 2,000 validation images, and 3,352 testing images. DOS dataset comprises 3,056 images, which have been randomly partitioned into training and validation sets at 8:1 ratio. The training set contains 6,800 semantic segmentation labels, while the validation set includes 887 annotated labels. The experiments were performed using an Intel Core i7-10700 CPU, Nvidia V100 16G GPU, and 16 GB memory.

For the quantitative evaluation, we report the performance of baseline methods and the proposed method by three metrics: mean intersection over union ($mIoU$).

Let TP, FP, and FN denote the total number of true positive, false positive, and false negative pixels, respectively. The Intersection over Union (IoU) is calculated as follows:

$$IoU_i = \frac{GT_i \cap \text{Pred}_i}{GT_i \cup \text{Pred}_i} \tag{4}$$

$$mIoU = \frac{1}{n} \sum_{i=1}^{n} IoU_i \tag{5}$$

where GT stands for ground truth, i denotes the semantic categories, and n symbolizes the total number of classes.

Params refers to the number of parameters in the model. FLOPs stands for "Floating Point Operations," and it is used as a measure of computational complexity or the number of calculations the model needs to perform during inference. GFLOPs represents for "Giga Floating Point Operations," equivalent to one billion FLOPs.

4.1 Semantic Segmentation on ADE20K and DOS

The experiments are carried out on semantic segmentation task. We employ the proposed PyraBiNet as our backbone architecture. To ensure a uniform evaluation metric, we strictly adhere to the training configurations set by PVT [34], utilizing a Semantic FPN [16] as our segmentation head. Our PyraBiNet is pre-trained on the ImageNet dataset [4]. The pre-training steps and parameters of our model are the same as the PVT. To entail a fair comparison, we keep the same data augmentation and training settings as the other vision transformers as far as possible. The competitors are all competitive vision transformers, including ResNet18 [10], PVT [34], BiSeNetv2 [40], PoolFormer [42]. PyraBiNet achieved state-of-the-art results on the ADE20K and DOS dataset, in Table 1 and Table 2.

As shown in Table 1, with the exception of BiSeNetV2, all models employ Semantic FPN as their segmentation head. In the nearly equivalent parameter range of 10M-20M, our model achieves the highest mIoU of 37.7. Remarkably,

Table 1. Performance comparisons on the test set of ADE20K. For each method, we report the mean intersection over union ($mIoU$), $Params(M)$, and $GFLOPs$.

Method	$Params(M)$	$GFLOPs$	$mIoU$
ResNet18 [10]	15.5	32.2	32.9
BiSeNetv2 [40]	**14.8**	**12.3**	19.5
PoolFormer [42]	15.7	30.7	37.2
PVT-Tiny [34]	17.0	33.2	35.7
Ours	19.4	37.3	**37.7**

our model outperforms PVT-Tiny by 2.0 points, highlighting the effectiveness of our proposed dual-branch architecture which fuses local and global features. The enhanced global feature extraction of ViT supplemented by the local feature extraction of the CNN increases segmentation precision. Furthermore, compared to the pure CNN-based dual-branch model, BiSeNetv2, our semantic branch possesses a global receptive field, resulting in superior segmentation accuracy in our model. Here, BiSeNetV2 is not pre-trained.

Table 2. Performance comparisons on the test set of DOS. Our model is trained on a single v100 Gpu with 40k iterations, a batchsize of 4, a learning rate of 1e-4, and an input image size of 512*512.

Method	$Params(M)$	$GFLOPs$	$mIoU$
ResNet18 [10]	15.5	31.9	65.2
BiSeNetv2 [40]	**14.8**	**12.0**	67.3
PoolFormer [42]	15.7	30.4	71.0
PVT-Tiny [34]	17.0	32.9	71.3
Ours	19.4	37.0	**72.8**

As illustrated in Table 2, when employing Semantic FPN for semantic segmentation, our model exhibits superior performance on the DOS dataset, achieving a maximum mIoU of 72.8. This score exceeds that of ResNet18 by 7.6 points and PVT-Tiny by 1.5 points, thereby further corroborating the efficacy of our proposed method of combining transformers and CNNs.

4.2 Ablation Study

We carry out ablation studies to validate the effectiveness of the feature fusion module. We compare our PDF-CBAM with several widely used methods, such as 'SUM': element-wise addition, 'MUL': element-wise multiplication, and 'Cat+1 $\times$ 1conv': concatenation followed by a 1×1 convolution. In addition, '+Stage4' refers to the incorporation of our detail module in the fourth stage of PVT. The results of these experiments can be seen in Table 3. The findings indicate that our mixed-attention feature fusion strategy outperforms simple addition, multiplication, or fusion through 1×1 convolution. This superiority

Table 3. Different designs of the feature fusion module to fuse the information from global features and local detail features. Δ denotes mIoU Variation. Ablations were tested on ADE20K.

Method	$Params(M)$	$GFLOPs$	$mIoU$	Δ
SUM	**19.36**	**37.31**	37.2	−0.5
MUL	**19.36**	**37.31**	37.0	−0.7
Cat+1 × 1conv	19.61	37.79	37.5	−0.2
+Stage4	24.79	38.69	37.3	−0.4
PDF-CBAM	19.38	37.32	**37.7**	-

can be attributed to the differing levels of global features extracted by transformers and local features extracted by CNNs, where the application of mixed attention enhances the model's capability to screen features. Furthermore, we discovered that introducing the detail module into the fourth stage of PyraBiNet does not enhance model performance, but instead causes a 0.4 drop in mIoU. This decline is due to the requirement for the extraction of spatial detail information: network depth should be relatively shallow, feature map size large, and a sufficient number of network channels should be available. In the fourth stage of our model, the feature map resolution is excessively small, leading to weakened ability to extract detailed information. This reduction in extraction ability could even introduce noise, resulting in performance degradation.

As shown in Fig. 4, we provide qualitative segmentation results on ADE20K and DOS datasets. The image on the left is the original image, and the image on the right is the semantic segmentation result. As can be observed from the Fig. 4, PyraBiNet demonstrates accurate segmentation of the edges of deformable objects, primarily attributed to the role played by our Detail Feature Block (DF). The DF module enhances the fine details of localized regions, thereby being particularly suited for fine-grained image segmentation tasks. Consequently, this leads to a more precise segmentation of the edges of deformable objects by PyraBiNet.

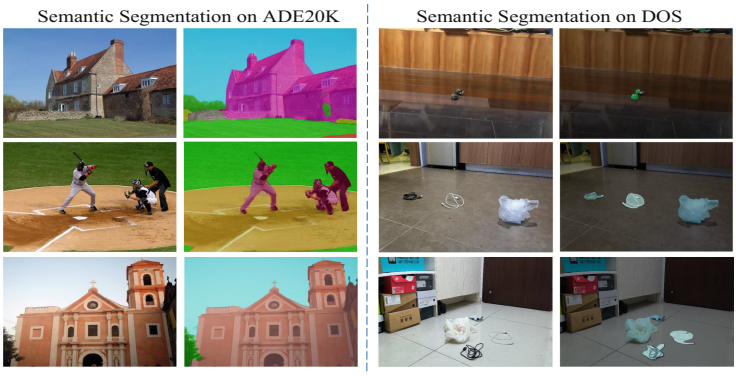

Fig. 4. Qualitative results of semantic segmentation on ADE20K and DOS datasets.

5 Conclusion

PyraBiNet is a groundbreaking dual-branch architecture adept at navigating the challenges inherent in lightweight semantic segmentation. By strategically integrating the global feature extraction capabilities of PVT with the meticulous local detail extraction of BiSeNet, we realized an efficient feature extraction process. Additionally, our innovative Parallel Dual-Feature Convolutional Block Attention Module facilitated optimal feature fusion while the Detail Feature Block enabled refined feature enhancement. PyraBiNet's superior performance compared to the existing state-of-the-art methods. With its effective and efficient performance, PyraBiNet proves to be an invaluable asset in the domain of mobile robotics, particularly beneficial for applications such as sweeping robots.

References

1. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep semantic segmentation of natural and medical images: a review. Artif. Intell. Rev. **54**, 137–178 (2021)
2. Chu, X., et al.: Twins: Revisiting the design of spatial attention in vision transformers. Adv. Neural. Inf. Process. Syst. **34**, 9355–9366 (2021)
3. Crespo, J., Castillo, J.C., Mozos, O.M., Barber, R.: Semantic information for robot navigation: A survey. Appl. Sci. **10**(2), 497 (2020)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. Ieee (2009)
5. Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Feng, D., et al.: Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. IEEE Trans. Intell. Transp. Syst. **22**(3), 1341–1360 (2020)
7. Gao, L., Nie, D., Li, B., Ren, X.: Doubly-fused vit: Fuse information from vision transformer doubly with local representation. In: Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII, pp. 744–761. Springer (2022). https://doi.org/10.1007/978-3-031-20050-2_43
8. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12259–12269 (2021)
9. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: rethinking convolutional attention design for semantic segmentation. arXiv preprint arXiv:2209.08575 (2022)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
12. Howard, A., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)

13. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
14. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. (CSUR) **54**(10s), 1–41 (2022)
15. Kim, W., Seok, J.: Indoor semantic segmentation for robot navigating on mobile. In: 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN), pp. 22–25. IEEE (2018)
16. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6399–6408 (2019)
17. Kohli, P., Ladickỳ, L., Torr, P.H.: Robust higher order potentials for enforcing label consistency. Int. J. Comput. Vision **82**, 302–324 (2009)
18. Ladickỳ, L., Russell, C., Kohli, P., Torr, P.H.: Associative hierarchical crfs for object class image segmentation. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 739–746. IEEE (2009)
19. Liu, Y., et al.: A survey of visual transformers. IEEE Trans. Neural Networks Learn. Syst. (2023)
20. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)
24. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 (2022)
25. Mo, Y., Wu, Y., Yang, X., Liu, F., Liao, Y.: Review the state-of-the-art technologies of semantic segmentation based on deep learning. Neurocomputing **493**, 626–646 (2022)
26. Pan, H., Hong, Y., Sun, W., Jia, Y.: Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. IEEE Trans. Intell. Transp. Syst. (2022)
27. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image. Int. J. of Comput. Vis. **77**(1) (2008). https://doi.org/10.1007/s11263-007-0090-8
28. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv 2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
29. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. Int. J. Comput. Vision **81**, 2–23 (2009)
30. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
31. Tsai, T.H., Tseng, Y.W.: Bisenet v3: bilateral segmentation network with coordinate attention for real-time semantic segmentation. Neurocomputing **532**, 33–42 (2023)

32. Ulku, I., Akagündüz, E.: A survey on deep learning-based architectures for semantic segmentation on 2d images. Appl. Artif. Intell. **36**(1), 2032924 (2022)
33. Wadekar, S.N., Chaurasia, A.: Mobilevitv3: mobile-friendly vision transformer with simple and effective fusion of local, global and input features. arXiv preprint arXiv:2209.15159 (2022)
34. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
35. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
36. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31 (2021)
37. Xu, J., Xiong, Z., Bhattacharyya, S.P.: Pidnet: a real-time semantic segmentation network inspired by pid controllers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19529–19539 (2023)
38. Yang, C., et al.: Lite vision transformer with enhanced self-attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11998–12008 (2022)
39. Yao, J., Fidler, S., Urtasun, R.: Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 702–709. IEEE (2012)
40. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: bilateral network with guided aggregation for real-time semantic segmentation. Int. J. Comput. Vision **129**, 3051–3068 (2021)
41. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 325–341 (2018)
42. Yu, W., et al.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10819–10829 (2022)
43. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 579–588 (2021)
44. Yuan, L., et al.: Tokens-to-token vit: training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 558–567 (2021)
45. Zhang, W., et al.: Topformer: token pyramid transformer for mobile semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12083–12093 (2022)
46. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 633–641 (2017)