PIDDNet: RGB-Depth Fusion Network for Real-time Semantic Segmentation

Yunsik Shin
dept. Electrical Engineering
Hanyang University
Seoul, South Korea
ysshin@spa.hanyang.ac.kr

Yongho Son
dept. Artificial Intelligence
Hanyang University
Seoul, South Korea
yhson@spa.hanyang.ac.kr

Junghee Park
dept. Ground Control System S/W
LIG Nex1 Co. Ltd.
Seongnam, South Korea
junghee.park@lignex1.com

Chaehyun Lee
dept. Ground Control System S/W
LIG Nex1 Co. Ltd.
Seongnam, South Korea
chaehyun.lee@lignex1.com

YangGon Kim
dept. Ground Control System S/W
LIG Nex1 Co. Ltd.
Seongnam, South Korea
yanggon.kim@lignex1.com

Jun Won Choi
dept. Electrical Engineering
Hanyang University
Seoul, South Korea
junwchoi@hanyang.ac.kr

***Abstract*—For RGB semantic segmentation, a two-branch network was proposed to effectively utilize both local detail information and global contextual information within an RGB image. This architecture combines a shallow spatial path with a deeper context path, resulting in high performance and FPS. Research on RGB-Depth segmentation has shown the performance gain that the depth map could provide complementary information to the RGB model. However, the advantage of fusing RGB and depth map within a two-branch network framework is unclear due to the distinct characteristics of these modalities. To address this, we present a novel fusion RGB-Depth architecture that takes into account the attributes of local context, global context, RGB, and depth map. Through the bidirectional image depth fusion technique, we effectively leverage each of the modalities, achieving a performance of 81.23 mIoU. This marks a gain of 1.27% when compared to the RGB-only model and 0.45% when contrasted with the element-wise feature addition fusion baseline.**

***Index Terms*—Semantic Segmentation, Deep Learning, RGB-Depth Fusion**

I. Introduction

Computer vision has gained a wide range of applicability in industries due to the advancement of deep learning technology. In particular, within the domain of autonomous driving, active research has been conducted on tasks such as object detection [13]–[15], image segmentation [10]–[12], and depth estimation [16]–[18] using cameras.

Image segmentation is a task that assigns classes for every pixel in an image, making it more challenging than object detection. Moreover, it needs to predict the results not only for thing classes but also for stuff classes, requiring global context information as well as fine-detailed information. To date, most image segmentation models are based on FCN(Fully Convolutional Network) [9] or Transformer backbones [5], [8]. FCN offers the advantages of relatively low computational complexity and high FPS (Frames per Second). Transformer-based models are more flexible framework which has low inductive bias, but they are hard to implement in real-time applications because of the computational cost and low FPS. Thus, it is important to develop a semantic segmentation model that can achieve the performance of Transformer-based models but the computation time of FCN.

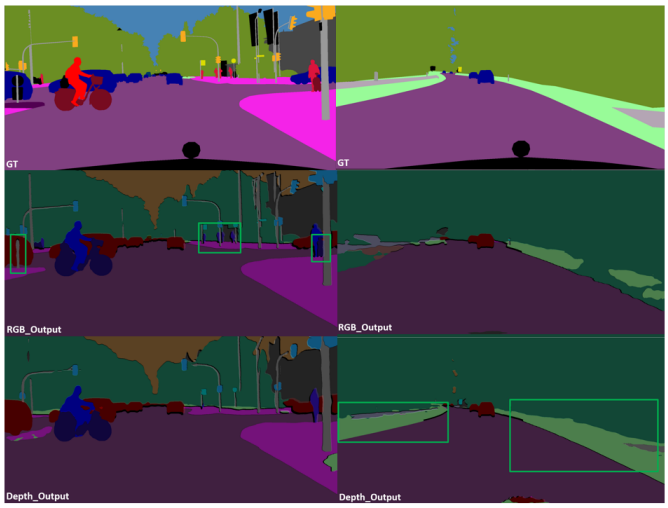

Fig. 1. From the first to the last row, the images refer to the GT segments, outputs from the RGB, and depth models.

Recently, RGB-Depth fusion networks that leverage depth map information to enhance performance have gained attention. The fusion of rich color information from a camera and depth information from a stereo disparity map could complement each other. While the RGB segmentation models outperform the depth segmentation models(Fig 1. left column), they occasionally suffer from inaccuracies in some parts of the image. On the other hand, the depth map model tends to produce segmentation results with coarse resolution, yet it outputs a more consistent performance for broader regions compared to the RGB model(Fig 1. right Column).

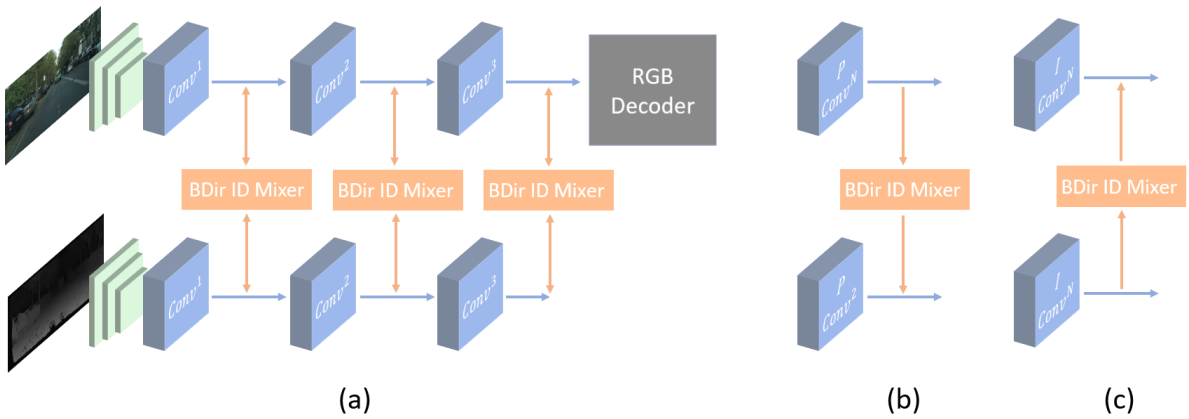

Fig. 2. (a) Overview of the PIDDNet. (b-c) BdirIDMixer Structure

In this paper, we propose an RGB-Depth sensor fusion network, referred to as PIDDNet, that incorporates depth information into the pipeline to get a robust segmentation performance for the broad areas. The RGB and depth map are fed into separate backbones and the fusion modules follow to combine the output of two backbones. We evaluate the performance of our fusion method in comparison with other methods on Cityscapes [6]. Our proposed method achieved an 81.23 mIoU (mean Intersection over Union), which is a mIoU gain of 1.27% over the RGB baseline model's 79.96 mIoU.

II. Related Work

A. RGB Semantic Segmentation

In the early field of image segmentation, methods using the FCN (Fully Convolutional Network) architecture predominated. The approach involved obtaining high-level features through an encoder and predicting pixel classes through a decoder. Additional techniques were proposed, such as dilated convolutions for contextual features and auxiliary losses for learning boundary information. PSPNet [10] proposed a pyramid pooling module based on a ResNet-based FCN. This module enabled the effective extraction of global context, structuring the decoder to utilize richer context features. Deeplabv3+ [11] employed an FCN structure with atrous convolutions, proposing a simple yet effective feature extraction technique for segmentation. Transformer-based models also have been widely leveraged to retrieve informative regions of input signals. Segformer [19] introduces a hierarchical Transformer encoder, which does not need additional positional encoding and both local and global attention representations are aggregated through a lightweight MLP decoder.

To accomplish the semantic segmentation task in real-time, PIDNet [12] proposed three PID branches to fuse detailed, context, and boundary information and achieved the state-of-the-art trade-off between inference time and accuracy.

B. RGB-Depth Semantic Segmentation

The development of depth sensors has led to a recent surge of interest in leveraging depth information for RGB-Depth semantic segmentation models. Typically, the methods for combining RGB and depth map are categorized into early, middle, and late fusion. As an early-stage fusion method, ShapeConv [1] concatenated the RGB and depth map and extracts semantic information through Shape-aware Convolution layers, which focuses on the inherent shape component in the depth information. However, the two distinct data modalities contain disparate features, which cannot be effectively processed via shared network feature extractors. For the late fusion, [2] proposed the geometry-aware propagation architecture to distill depth-aware embedding in the late stage instead. The middle-stage fusion outperforms the aforementioned methods through the interaction between intermediate information of the two modalities. [3] proposed a cross-modality guided encoder to fuse the different modality feature maps and propagate them to the next branch. This method efficiently reduces the domain gap in the middle stage and aggregates the two recalibrated representations. Moreover, FokenFusion [4] proposed a fusion method of multi-modal data using Transformers [5], which recombines the tokens in the feature tensor to strengthen the interaction of other informative multi-model tokens. EMSANet [7] proposed an efficient semantic segmentation network by adopting factorized non-bottleneck blocks in convolution neural networks and operating on a mobile platform in real time.

In this literature, we mainly focus on the middle fusion strategy to recalibrate RGB feature maps from a depth representation, which effectively enhances the representations for the outdoor semantic segmentation *e.g.,* CityScapes dataset [6].

III. Proposed Method

A. Baseline Backbone Structure

Our PIDDNet builds on the baseline semantic segmentation model, PIDNet [12]. PIDNet [12] proposed a three-branch structure inspired by the PID controller. It redefined the existing two-branch network architectures into the P, and I branches, then introduced a new D branch. The P branch extracts fine-detailed information from the images, while the I branch takes charge of semantic information spanning across the entire image. By incorporating the D branch, PIDNet [12]

TABLE I
CITYSCAPES RESULTS ON VALIDATION DATASET.
LA: LINEAR ADDITION, SA: SPATIAL ATTENTION, BA: BIDIRECTIONAL ADDITION

Modalities	roa.	sid.	bui.	wal.	fen.	pol.	lig.	sig	veg.	ter.	sky	per.	rid.	car	tru.	bus	tra.	mot.	bic.	**mIoU**	**FPS**
RGB	98.2	85.9	93.2	56.9	66.6	67.4	73.9	81.6	92.7	64.8	94.8	83.8	68.0	95.6	80.7	90.0	78.5	66.8	79.4	79.96	-
Depth	96.8	74.2	83.7	30.3	43.4	48.6	41.3	47.1	80.4	42.0	83.8	65.0	46.4	88.5	49.1	58.9	49.8	27.6	50.5	58.29	-
RGB-D (EWA)	98.1	85.0	93.4	56.9	**70.7**	68.5	74.3	**81.9**	92.7	64.8	**95.2**	84.6	69.2	95.8	83.1	89.7	82.1	68.8	**79.9**	80.78	16.84
RGB-D (SA)	98.3	**86.4**	93.3	54.3	68.3	67.3	**74.7**	81.8	92.6	63.8	95.1	84.6	69.5	**96.0**	**88.9**	**92.3**	84.7	**71.4**	79.6	81.21	14.43
RGB-D (BA)	**98.4**	86.3	**93.5**	**61.1**	68.4	**67.5**	74.4	81.2	**92.8**	**64.9**	95.1	**84.7**	**69.6**	95.7	83.6	92.2	**85.7**	68.6	79.6	**81.23**	16.32

enhanced the model capacity for representing high-frequency features. This approach also facilitates real-time operation as well as strong segmentation performance.

B. PIDDNet

We propose PIDDNet, which extends PIDNet [12] to incorporate not only RGB images but also disparity depth map information. An overview of the model is shown in Fig 2 (a). The RGB image and the depth map are fed to the separate PIDNet-based backbones. Each modality input passes through the P, I, D three branches. Fig 2 (a) selectively visualizes one of the P, I, or D branch pipelines included in the two distinctive backbones. Then, the depth and RGB features are fused in the **BDirIDMixer** (Bidirectional Image Depth Mixer).

C. Bidirectional Image Depth Mixer

The three branches of PIDNet [12] are implemented for different purposes. The P branch is expected to learn fine-grained visual information. The I branch learns global and local contextual information, leading to richer information in the Depth I branch. To validate this, we compare three separately trained fusion methods in Table 1. Element Wise Addition **(EWA)** is a naïve approach that adds depth feature to RGB feature in an element-wise manner. Spatial Attention(**SA**) generates the attention weights for the depth map and updates the depth feature. Then, the updated depth features are fused with the RGB features in an element-wise manner. The bidirectional addition(**BA**) operates differently for the P and I branches as follows. For the P branches in the image backbone and depth backbone, the image feature map is added to the depth feature in an element-wise manner. In contrast, in the I branch pairs, depth features are added to the RGB features(Fig 2 (b,c)). This way of fusion method in the BDirIDMixer effectively transfers the fine-detailed and global contextual information to both backbones.

IV. EXPERIMENTS

A. Dataset and Training Details

Cityscapes dataset [6] is well-known urban scene parsing data, encompassing 5000 images from 27 cities. There are 2975 images for training, 500 for validation, and 1525 for testing. The image resolution is 2048×1024 and pixel-level labels of 19 semantic classes (11 "stuff" and 8 "thing") are provided.

Since we propose a fusion method into the baseline model architecture, we follow the proposed training details in PIDNet [12]. Cross-entropy loss and boundary-awareness CE loss are utilized at the decoder head. Additionally, auxiliary cross-entropy losses for the P and D backbones are adopted for better optimization.

B. Experiment Result

Table 1 compares the performance of three different types of models. The RGB-only and Depth-only models are trained in the same manners proposed in PIDNet [12]. The results for the two models are presented in the first and second rows, respectively in Table 1. From the third to the fifth row, the experimental results of the three fusion methods are in the table. It is evident that using the proposed bidirectional addition technique yielded the highest performance compared to the other fusion methods. With Bidirectional Addition (BA) fusion method, a performance gain of 1.27 mIoU (mean Intersection over Union) was achieved compared to the RGB-only model. When compared to the SA method, the BA method shows similar performance while exhibiting a 13% higher FPS, highlighting its computational efficiency.

C. Qualitative Result

We present qualitative results of our method on the CityScapes dataset [6] to demonstrate how our fusion method could improve semantic segmentation in various ways. From scene 1, we can observe that part of the bicycle is misidentified as a 'car' in the RGB model. Moreover, the pole occluded by cars was not identified from the RGB model but has been correctly identified with the RGB-D input. From scene 2, we can observe that semantic segmentation performance for vegetation intermingled with road area is more accurate with the RGB-D model. The bidirectional fusion method effectively delivers detailed information and also provides informative representations of expansive areas.

V. CONCLUSION

This paper presents the novel bidirectional fusion method. We extended the RGB segmentation model [12] to a sensor fusion model and introduced a new fusion method. Through

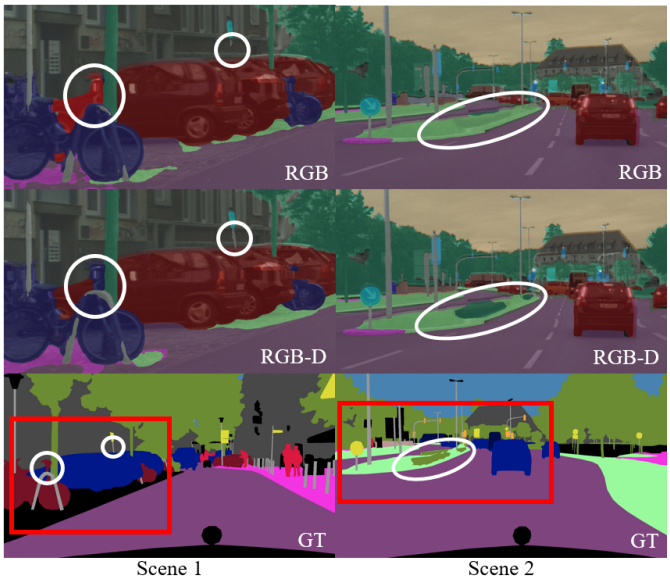

Fig. 3. From the first to the last row, the images refer to the outputs from baseline(RGB), outputs from ours, and ground truth.

the Bidirectional Addition method, which effectively conveys local and global contextual information compared to traditional methods, the proposed PIDDNet achieves better performance and computation time than the counterpart fusion methods.

Acknowledgment

This work was supported by Korea Research Institute for defense Technology planning and advancement(KRIT) grant funded by the Korea government(DAPA(Defence Acquisition Program Administration))(No. KRIT-CT-22-011-00, LiDAR-Camera Image Fusion, 2022)

References

[1] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7088–7097.

[2] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, "Geometry-Aware Distillation for Indoor Semantic Segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2869–2878.

[3] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-andaggregation gate for rgb-d semantic segmentation," in European Conference on Computer Vision. Springer, 2020, pp. 561–577.

[4] Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., Wang, Y. (2022). Multimodal token fusion for vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12186-12195).

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[6] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).

[7] Seichter, D., Fischedick, S. B., Köhler, M., & Groß, H. M. (2022, July). Efficient multi-task rgb-d scene analysis for indoor environments. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-10). IEEE.

[8] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[9] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[10] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[11] Yurtkulu, S. C., Şahin, Y. H., & Unal, G. (2019, April). Semantic segmentation with extended DeepLabv3 architecture. In 2019 27th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[12] Xu, J., Xiong, Z., & Bhattacharyya, S. P. (2023). 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).

[13] Xia, Z., Pan, X., Song, S., Li, L. E., & Huang, G. (2022). Vision transformer with deformable attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4794-4903).

[14] Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11976-11986).

[15] Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., ... & Qiao, Y. (2023). Internimage: Exploring large-scale vision foundation models with deformable convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14408-14419).

[16] Godard, C., Mac Aodha, O., Firman, M., Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3828-3838).

[17] Bae, G., Budvytis, I., & Cipolla, R. (2022). Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2842-2851).

[18] Li, R., Gong, D., Yin, W., Chen, H., Zhu, Y., Wang, K., ... & Zhang, Y. (2023). Learning to Fuse Monocular and Multi-view Cues for Multi-frame Depth Estimation in Dynamic Scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21539-21548).

[19] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems, 34, 12077-12090.