Depth-Aware Feature Pyramid Network for Semantic Segmentation

Taehyeon Kim†, Seho Park, and Kyung-Taek Lee
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
{taehyeon.kim†, sehopark, ktechlee}@keti.re.kr

***Abstract*—Object segmentation based on multi-sensor fusion is a critical technique in autonomous vehicles, providing several benefits, including increased accuracy, robustness to adverse conditions, heightened situational awareness, and efficient processing. In this paper, we introduce a novel feature fusion-based object segmentation model named Depth-Aware Feature Pyramid Network that integrates RGB and depth information using a multi-scale feature fusion mechanism. As a result, the proposed algorithm can dynamically fuse features from multiple modalities, namely RGB and depth, to perform object segmentation with depth awareness. To validate the performance of the proposed algorithm, we conducted experiments on the Cityscapes benchmark and achieved a 72.4% mean Intersection over Union (mIOU), outperforming related object segmentation methods for autonomous vehicles.**

***Index Terms*—Autonomous Vehicle, Convolutional Neural Network, Object Segmentation, Feature Fusion, Multi Sensor Fusion**

I. Introduction

Object segmentation is an important task in autonomous vehicles as it enables the vehicle to accurately perceive and understand its environment. The ability to identify and segment objects in the scene, such as other vehicles, pedestrians, road signs, and lane markings, is crucial for the vehicle to make safe and informed decisions while driving. Recently, object segmentation based on deep learning has become a highly active area of research in recent years. The use of deep learning algorithms has led to significant advances in the accuracy and robustness of object segmentation, making it a promising solution for an autonomous driving system.

Specifically, RGB and depth fusion based object segmentation is a powerful approach for autonomous driving, offering improved accuracy, robustness, and efficiency compared to using a single sensor. By combining information from multiple sources, the vehicle can obtain a more complete and accurate representation of its environment, allowing it to make more informed decisions and navigate more safely and efficiently.

There are several studies on object segmentation based on the fusion of RGB and depth information. Choy, et al. [1] presents a real-time RGB-D object instance segmentation method based on a novel neural network architecture called Receptive Field BlockNet. The method fuses RGB and depth information at multiple levels of the network to achieve improved accuracy and robustness. Lin, et al. [2] proposed a depth-aware object segmentation method that uses a combination of RGB and depth information to improve the accuracy and robustness of object segmentation. The method uses a deep neural network to learn to make use of both RGB and depth information in an effective way. Weng, et al. [3] suggested a method for RGB-D object detection and segmentation using convolutional neural networks. The method combines RGB and depth information to achieve improved accuracy and robustness in object detection and segmentation.

In this paper, we propose a novel technique that combines RGB and depth information and is named "Depth-Aware Feature Pyramid Network (DAFPN)". The proposed method employs a multi-modal feature fusion mechanism with multi-scale features to effectively utilize RGB and depth information. The feature fusion layer is a specific type of layer in a neural network that combines information from multiple sources or modalities. The purpose of a feature fusion layer is to integrate information from different sources in a way that is useful for the object segmentation. Feature pyramid network is designed to effectively process multi-scale information by creating a pyramid-like structure of feature maps, where each level of the pyramid represents a different scale of features.

In summary, the main contributions of this study are as follows:

- We propose an effective feature fusion algorithm, called Depth-Aware Feature Pyramid Network (DAFPN), which appropriately combines the features from RGB and depth information.
- The use of an multi-scale feature fusion mechanism enables DAFPN to effectively filter out redundant features when combining information and retain features that enhance object segmentation performance.
- To demonstrate the effectiveness of DAFPN, we evaluate its performance using benchmark dataset. The results of extensive experiments suggest that DAFPN outperforms previous algorithms.

The remainder of this paper is organized as follows: Section II provides a brief introduction to feature fusion and attention

This research was financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute of Advancement of Technology (KIAT) through the International Cooperative R&D Program [P0019782, Embedded AI Based Fully Autonomous Driving Software and MaaS Technology Development].

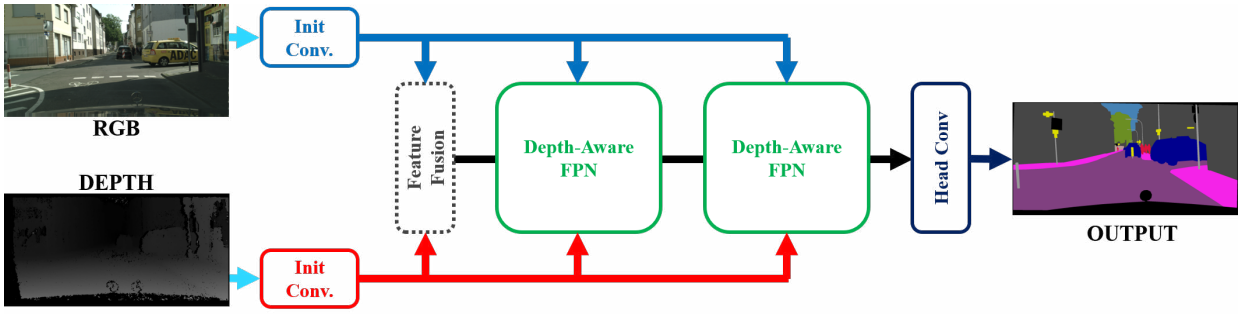

Fig. 1: Neural architecture for object segmentation with depth-aware feature pyramid network in autonomous vehicles.

mechanisms. Section III details our proposed object segmentation algorithm, focusing on the Depth-Aware Feature Pyramid Network. The experimental results are presented and analyzed in Section IV, and the study is summarized in Section V.

II. Preliminaries

A. Feature Fusion

Feature fusion is a technique in deep learning that combines multiple representations of the same data to produce a more robust and accurate output. In the context of convolutional neural networks (CNNs), feature fusion can be achieved through various methods, such as concatenation, summation, and element-wise multiplication.

There are several mechanisms for feature fusion, including: **Concatenation**: This involves concatenating the feature maps from different layers or branches of the network along a specific axis to form a combined feature map. The concatenated feature map is then fed into the next layer of the network. **Summation**: In this mechanism, the feature maps from different layers or branches of the network are summed element-wise to form a combined feature map [5]. **Element-wise multiplication**: This involves element-wise multiplication of the feature maps from different layers or branches of the network to form a combined feature map [6]. **Attention Mechanism**: This mechanism involves learning a weighting factor for each feature map that determines its contribution to the final combined feature map. The attention mechanism can be implemented using various methods, such as the dot product, softmax activation, or neural network-based approaches. Each of these mechanisms has its own strengths and weaknesses, and the choice of which to use will depend on the specific problem being solved and the nature of the data [7].

B. Feature Pyramid Network

The feature pyramid network (FPN) architecture is designed to effectively process multi-scale information by creating a pyramid-like structure of feature maps, where each level of the pyramid represents a different scale of features.

In the FPN, the input image is first passed through a series of convolutional and pooling layers to extract features at multiple scales. These features are then up-sampled and combined to create a hierarchical pyramid of features, where each level of the pyramid contains information at a different scale. This allows the network to effectively process objects at multiple scales and handle objects of varying sizes in the image. The FPN architecture has several benefits compared to traditional CNNs. First, the FPN allows the network to process features at multiple scales, which can improve its ability to detect objects at different scales. Second, the FPN reduces the amount of computation required for processing large input images, as the pyramid structure allows for reusing the computed features at different scales. Finally, the FPN can help alleviate the issue of vanishing gradients that is commonly encountered in deep neural networks.

III. Proposed Method

The proposed model for object segmentation is designed specifically for autonomous vehicles, utilizing a multi-sensor fusion approach that combines RGB and depth data. Figure 1 illustrates the neural architecture of the model. As shown in the figure, it can be observed that the proposed technology utilizes a depth-aware feature pyramid network to extract and represent features suitable for object segmentation from various resolution image and depth information. By inputting image and depth information sequentially into the DAFPN in the proposed neural network, each DAFPN module can secure various gradient flows, ultimately enhancing the quality of convergence during the learning process. Therefore, we have not only designed a new feature fusion module for multi-sensor fusion but also proposed an object segmentation technique that incorporates depth information by utilizing multi-resolution features. This is one of the crucial technologies in the development of autonomous driving. The detailed structure of the proposed FPN can be found in Figure 2.

The proposed DAFPN consists of two parts: one that processes RGB information and another that processes depth information. For the part that processes the RGB input, a feature aggregation module based on concatenation is utilized to incorporate features of various resolutions. Although this method may have a negative impact on the computational complexity of the neural network, the performance of object segmentation techniques primarily relies on image information input from the camera. Based on various empirical studies, setting fewer trainable parameters in the image information input section can actually lead to a critical decrease in object segmentation performance. Therefore, the part that processes

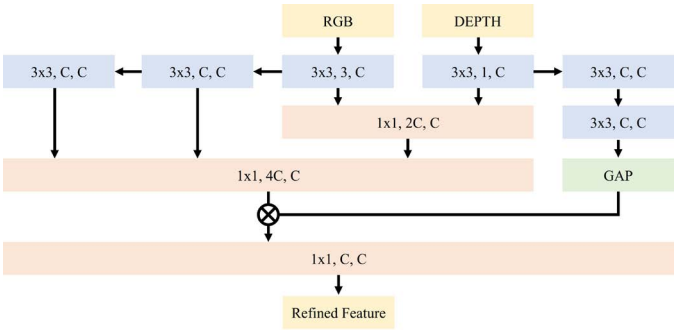

Fig. 2: Framework of the proposed depth-aware feature pyramid network (DAFPN) module. Blue- and red-colored boxes, which represent 3×3 convolutional layers with an input depth size of C and a number of filters of $2C$, denoted as "3×3, C, $2C$". The GAP means global average pooling layer. The $\otimes$ symbol is used to indicate channel-wise multiplication.

image features utilizes the multi-scale layer aggregation technique, the rationality of which can be confirmed in the [12].

The key feature of the proposed feature network in this paper is processing depth information. Depth information in autonomous driving vehicles can be obtained through LiDAR. However, utilizing depth information in the form of point clouds requires the use of complex convolutional layers such as sparse convolutional layers or 3D convolutional layers. Furthermore, the features extracted using these layers are difficult to fuse with the feature-maps of general 2D convolutional layers. On the other hand, replacing point cloud information with depth image information has a higher computational efficiency compared to using complex neural networks. Therefore, the proposed object segmentation technique easily utilizes widely used convolutional neural network techniques by using image-formatted depth information.

In contrast to the structure that processes image information, the structure that processes depth information does not utilize cross-stage connections. The depth information is processed through a single convolutional layer, and the resulting feature information is fused with the features extracted from the image information. Furthermore, the depth information is transformed into a vector that can perform channel-wise multiplication through a different branch structure. This process gives channel-wise attention to the fused features that are extracted from both image and depth information. The reason for this approach is that depth information represents information about objects that are relatively close more clearly than objects that are farther away, as the depth information may lose information or have sparse features for objects that are far away. Therefore, instead of using various spatial features extraction and representation methods as used for image information processing in the proposed module, a global average pooling layer is utilized to fuse channel-wise features.

Ultimately, using the methods described above, the image and depth information are combined to form a single feature map, which is then used to construct the depth-aware feature pyramid network. To summarize, for image information, the proposed feature fusion method that extracts and integrates multi-resolution features is used to enhance the gradient flow and capture detailed features. For depth information, the our method that forms channel features rather than spatial features is used to overcome the limitation of resolution. Using these methods, the proposed depth-aware feature pyramid network achieves multi-sensor fusion of depth and RGB information in the form of neural feature fusion, including multi-resolution feature fusion to improve object segmentation performance.

TABLE I: Quantitative Experimental Results on the Cityscapes Benchmark using an RTX 3090.

Network	Input Size	mIoU	# Param.	FPS
[4]	1024×512	58.3	0.36E9	83.7
[8]	640×360	56.1	29.5E9	17.3
[9]	1536×768	68.4	5.8E9	**107.8**
[10]	640×360	64.8	**0.53E9**	52
[11]	2048×1024	70.1	0.55E9	31.2
Ours	2048×1024	**72.4**	0.58E9	30.4

IV. Experimental Results

A. Implementation Details

The Cityscapes dataset focuses on semantic understanding of urban scenes and includes a total of 5,000 fine annotation and 20,000 coarse annotation images. The images were captured from 50 different cities under various seasons and weather conditions. The fine annotation set includes 2975 training, 500 validation, and 1525 testing images, with an original resolution of 1024 × 2048. The dataset encompasses 19 classes that are grouped into 7 categories, such as vehicles, which include cars, trucks, and buses. All experiments were conducted using code written in PyTorch and the results were obtained using an RTX 3090. The experimental settings, including the loss function, were consistent with those described in [11].

B. Quantitative Experimental Results

The quantitative evaluation results of the proposed image segmentation technique are presented in Table 1. The quantitative evaluation was conducted based on three evaluation criteria: mIOU (mean Intersection Over Union), which represents object segmentation performance, the number of parameters, which reflects the storage complexity of the neural network, and FPS, which represents the time complexity of the neural network. To ensure a fair comparison between object segmentation techniques, the results presented in Table 1 are the average values obtained from a total of 10 independent experiments.

The quantitative evaluation results showed that the proposed technique achieved higher mIOU performance than all the compared techniques. This demonstrates that the proposed

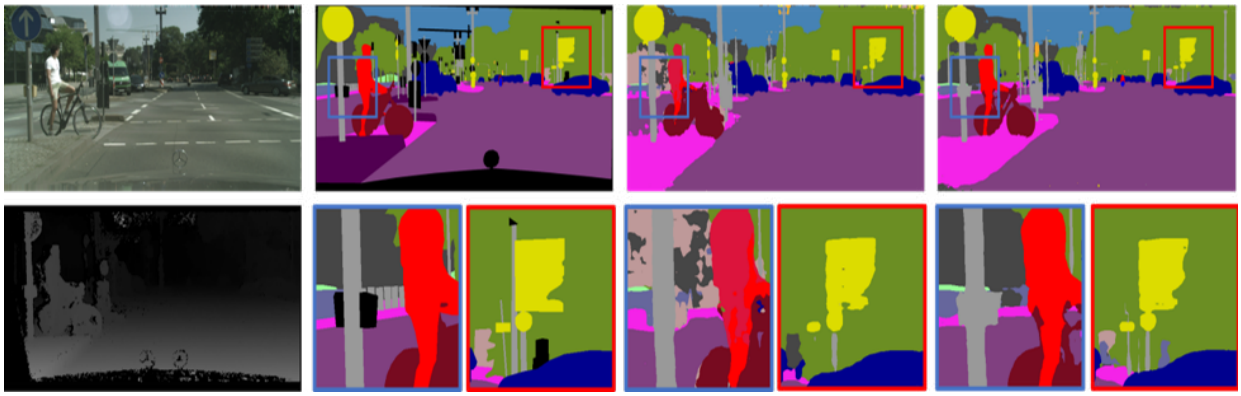

Fig. 3: Qualitative Experimental Results on the Cityscapes Benchmark: (First Column) RGB image (above) and depth image (below); (Second Column) Ground truth images; (Third Column) Images generated by [11]; (Fourth Column) Images generated by DAFPN.

technique successfully fused depth information in an appropriate form with the feature information extracted from the image, which led to the superior performance in mIOU. Furthermore, the fact that the proposed neural network architecture has a similar number of parameters to the existing techniques, despite the need for a separate neural network structure for receiving depth information to perform multi-sensor fusion, indicates that the proposed network structure has an efficient design. Although the FPS was measured low due to the computational process in the neural network structure that processes image information using DAFPN, considering both mIOU and the number of parameters, it can be seen that the proposed technique has an appropriate level of FPS.

C. Qualitative Experimental Results

The results of the qualitative evaluation are shown in Figure 3. In the qualitative evaluation, a comparative experiment was conducted between the proposed technique and [11], which showed the closest performance level to the proposed technique. Object segmentation, as it is also called pixel-wise classification, requires the ability to accurately extract the shape of each object. As can be seen in the enlarged images of the results presented in the second row, the proposed technique provides better quality object segmentation results than the existing techniques. Especially, by utilizing depth information, the proposed technique clearly represents the unknown structure installed on the signpost, and cleanly distinguishes human and bicycle objects by complementing various human shapes that are difficult to distinguish by image information. The above qualitative experimental results provide more clear evidence of the performance improvement effect of DAFPN, and demonstrate that the multi-sensor fusion technique through feature fusion of depth and image information has been implemented.

V. Conclusion

The proposed Depth-Aware Feature Pyramid Network (DAFPN) fuses features from RGB images and depth images through multi-scale feature fusion mechanisms to enhance object segmentation in autonomous vehicles. This novel feature pyramid network incorporates depth information to learn via supervision of 3D information, which is a key factor in performance improvement. The effectiveness of the proposed algorithm has been verified both quantitatively and qualitatively through extensive experiments on the Cityscapes benchmark. As future work, the DAFPN can be combined with an attention mechanism to adaptively fuse features.

References

[1] Bolya, Daniel, et al. “Yolact: Real-time instance segmentation.” Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[2] Wang, Weiyue, and Ulrich Neumann. “Depth-aware cnn for rgb-d segmentation.” Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[3] Qiao, Siyuan, et al. “Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[4] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “Enet: A deep neural network architecture for real-time semantic segmentation,” ar iv preprint ar iv:1606.02147, 2016.

[5] Calvi, Giuseppe G., Ilia Kisil, and Danilo P. Mandic. “Feature fusion via tensor network summation.” 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018.

[6] Fan, Jiahe, et al. “Multi-scale feature fusion: Learning better semantic segmentation for road pothole detection.” 2021 IEEE International Conference on Autonomous Systems (ICAS). IEEE, 2021.

[7] Dai, Yimian, et al. “Attentional feature fusion.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021.

[8] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation.” IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.

[9] Yu, Changqian, et al. “Bisenet: Bilateral segmentation network for real-time semantic segmentation.” Proceedings of the European conference on computer vision (ECCV). 2018.

[10] Wu, Tianyi, et al. “Cgnet: A light-weight context guided network for semantic segmentation.” IEEE Transactions on Image Processing 30 (2020): 1169-1179.

[11] Lou, Ange, and Murray Loew. “Cfpnet: channel-wise feature pyramid for real-time semantic segmentation.” 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021.

[12] Ma, Ningning, et al. “Shufflenet v2: Practical guidelines for efficient cnn architecture design.” Proceedings of the European conference on computer vision (ECCV). 2018.