Lightweight Prediction and Boundary Attention-Based Semantic Segmentation for Road Scene Understanding

JEE-YOUNG SUN, (Graduate Student Member, IEEE), SEUNG-WON JUNG, (Senior Member, IEEE), AND SUNG-JEA KO, (Fellow, IEEE)
Department of Electrical Engineering, Korea University, Seoul 02841, South Korea
Corresponding author: Seung-Won Jung (swjang83@korea.ac.kr)

This work was supported by the Ministry of Trade, Industry Energy (MOTIE), South Korea, through the Technology Innovation Program (or Industrial Strategic Technology Development Program) (Development of Deep Learning-Based Open EV Platform Technology Capable of Autonomous Driving) under Grant 10082585.

ABSTRACT
Semantic segmentation is one of the most commonly used techniques for road scene understanding. Recently developed deep learning-based semantic segmentation networks are typically based on the encoder-decoder structure and have made great progress in road scene understanding. However, these conventional networks still encounter difficulties in recovering spatial details. To overcome this problem, we introduce a lightweight prediction and boundary-aware refinement module that can hierarchically refine the segmentation results with spatial details. The proposed refinement module has two attention units called the upper-level prediction attention unit and the upper-level boundary attention unit. The upper-level prediction attention unit emphasizes the features in the regions that need to be refined by using predicted class probability from the upper-level, whereas the upper-level boundary attention unit focuses on the features near the semantic boundary of the upper-level segmentation result. By using the proposed prediction and boundary-aware refinement module in the decoder network, the segmentation result can gradually be improved in a top-down manner to a finer and more complete one. Experimental results on the Cityscapes and CamVid datasets demonstrate that the proposed prediction and boundary attention-based refinement module can achieve considerable performance improvement in segmentation accuracy with a marginal increase in computational complexity.

INDEX TERMS
Attention mechanism, boundary attention, deep learning, feature refinement, real-time semantic segmentation, residual learning, road scene understanding.

I. INTRODUCTION
In recent years, a large number of studies have addressed in the field of autonomous driving [1], [2]. Road scene understanding, an essential component of autonomous driving, involves multiple tasks, such as detecting drivable areas [3]–[8], pedestrians, and vehicles [9]–[11]; it provides crucial information for the path planning of autonomous vehicles. One of the most commonly used techniques for road scene understanding is semantic segmentation, which aims to classify each pixel of an image into one of the predefined classes.

In this paper, we focus on the semantic segmentation of road scenes, which means segmenting on-road objects and background materials such as road, sidewalk, sky, and vegetation. Deep learning has made remarkable progress in computer vision, and deep convolutional neural network (CNN)-based methods [9]–[22] have achieved high performance in semantic segmentation [23], [24]. Most of these deep CNN architectures for semantic segmentation [9]–[19] are based on the encoder–decoder structure. The encoder network extracts meaningful features using convolutional (Conv) and pooling layers, and the decoder network reconstructs the original resolution by using deconvolution or interpolation layers. However, these decoding operations cannot adequately recover
the spatial details that are lost by the pooling layers in the encoder network, thereby yielding unsatisfactory results with corrupted segmentation boundaries. To overcome this problem, various recent approaches have attempted to extract more informative features from the encoder network. Pyramid scene parsing network (PSPNet) [22] and Deeplab-series [19]–[21] proposed a pyramid pooling module (PPM) and an atrous spatial pyramid pooling (ASPP) module, respectively, to harvest multi-scale context information. Several attention-based networks, such as dual attention network (DANet) [25] and criss-cross attention network (CCNet) [26], have been introduced to integrate local features with global dependencies. Although these approaches have shown high segmentation accuracy, the aforementioned networks have enormous numbers of parameters, and are computationally inefficient to be implemented in practical applications including autonomous driving.

In general, low-level feature maps extracted from shallow layers of the encoder network have rich spatial details. Thus, the lack of details in high-level feature maps can be complemented by fusing the fine details in low-level feature maps. Based on this, we propose a lightweight feature refinement module, called prediction and boundary-aware refinement module (PBRM), that can effectively recover the spatial details of segmentation result in a top-down manner. The proposed PBRM contains two attention units, i.e., the upper-level prediction-based attention (UPA) and upper-level boundary attention (UBA) units. Most conventional attention-based segmentation networks [25]–[27] are based on the self-attention mechanism [28] that obtains attention masks from the input feature map. However, the proposed UPA and UBA units generate the attention masks by directly injecting the upper-level coarse prediction, rather than using the same level of feature maps. Specifically, the UPA unit, whose attention mask is obtained from the upper-level predicted probability, enables the network to discover the features in the regions that need to be refined. To address the corrupted semantic boundary, the UBA unit emphasizes the features near the boundary of the upper-level segmentation result. In the proposed PBRM, the UPA and UBA units transfer the information from the upper-level prediction to lower-level feature maps, which can guide the network to capture informative lower-level features necessary to supplement the upper-level prediction. The proposed PBRM is a lightweight module containing only 0.01M parameters and 0.3G floating point operations per second (FLOPS) on average, and it can be integrated into any encoder–decoder structure-based CNNs.

In the experiments, we compared the proposed PBRM with other attention modules proposed in state-of-the-art semantic segmentation networks, i.e., DANet [25] and CCNet [26], in terms of the number of parameters, FLOPS, and segmentation accuracy. Besides, we applied the proposed refinement module to several state-of-the-art real-time semantic segmentation networks including context guided network (CGNet) [29], LiteSeg-mobile [30], and fully-convolutional harmonic densely connected network (FC-HarDNet) [31] to verify the effectiveness of PBRM. Experimental results on the Cityscapes [32] and CamVid [33] datasets demonstrated that the proposed PBRM can further improve the segmentation accuracy of conventional networks with a slight increase in computational cost.

II. RELATED WORKS

A. SEMANTIC SEGMENTATION

Fully convolutional network (FCN) [12] has been a milestone in the field of semantic segmentation. By replacing fully connected layers with Conv layers, FCN can learn the pixel-level dense prediction, and has achieved considerable performance improvement in semantic segmentation. Including FCN, most deep CNN architectures developed for semantic segmentation [9]–[19] are based on the encoder–decoder structure. However, spatial details can be lost during the successive pooling in the encoder network, which leads to unsatisfactory segmentation results with corrupted boundaries. To overcome this problem, several methods [13]–[15] attempted to design better decoder structures. One of the most popular networks, called U-Net [15], has a decoder network with skip connections that deliver the high-resolution features from the encoder to the decoder, so as to compensate for the loss of spatial details. Other structures such as PPM [22] and ASPP module [19], [21] have been introduced to better extract multi-scale context information by enlarging the receptive field. Recently, attention mechanism-based semantic segmentation networks such as DANet [25] and CCNet [26] have been proposed to capture contextual information by considering full-image dependencies for each pixel. Although these networks have achieved high performance, they require large numbers of parameters and expensive computational costs, which can be a critical issue in real-time application including autonomous driving.

Real-time semantic segmentation networks [29]–[31], [34]–[38] have attempted to find a good trade-off between speed and accuracy; the main purpose of these networks is to reduce the number of network parameters and FLOPS while minimizing the loss of accuracy. Zhao et al. proposed a compressed-PSPNet-based image cascade network (ICNet) [37] that can perform real-time segmentation by extracting the semantic information in low-resolution images and the details in high-resolution images. Bilateral segmentation network (BiSeNet) [38] consists of the spatial path and context path to improve the inference speed by separately learning the spatial details and the contextual information. Emara et al. [30] attempted to reduce the model size by using lightweight backbone networks, such as MobileNetV2 [39] and DarkNet19 [40], and proposed a deeper version of ASPP module to enhance the segmentation accuracy. Wu et al. introduced CGNet [29] composed of context guided blocks that can effectively combine local features with the contextual information, and CGNet has shown high performance without using a backbone network. Chao et al. suggested memory
traffic as a dominating factor of the inference latency, and proposed FC-HarDNet [31] that shows high efficiency in terms of FLOPS and memory traffic.

**B. ATTENTION MECHANISM**

Attention mechanism is widely adopted and has shown significant performance improvement in various deep learning-based computer vision applications [25], [26], [41]–[47]. In vision tasks, attention mechanism first computes the attention weights that represent the degree of importance of features, and then extracts more informative features from the input feature maps by using the weight values. Additive attention and multiplicative attention are the two most well-known attention mechanisms and most of the recently developed attention-based deep CNNs [41]–[46] employ the latter one, i.e., performing element-wise multiplication of the attention mask and feature maps. Pang et al. [41] proposed a mask-guided attention network for occluded pedestrian detection. When a pedestrian detection network predicts the candidate region of pedestrians, the attention network emphasizes the features in the region containing pedestrians and suppresses those in the occluded areas to capture more important features related to the visible pedestrian. In the research field of salient object detection, reverse attention [42] was introduced to capture the features of the missing part of salient objects from the background region.

In recent years, several attention-based semantic segmentation networks have been introduced. DANet [25] used a dual attention module (DAM) that consists of spatial and channel attention modules to aggregate local features with global dependencies in spatial and channel dimensions. CCNet [26] proposed a recurrent criss-cross attention module (RCCAM) that can capture full-image dependencies from all pixels in a more efficient way. While these attention-based segmentation networks are based on the self-attention mechanism [28] that computes the attention weight from the input feature map itself, the proposed UPA and UBA units generate the attention masks from the upper-level coarse prediction to emphasize informative lower-level features to reconstruct the original-sized segmentation result with more spatial details.

**C. BOUNDARY-RELATED NETWORK**

Several recent methods have focused on utilizing boundary information in the field of semantic segmentation [48]–[52]. Bertrasius et al. [48] and Kokkinos [49] first trained their boundary detection networks to find the object boundary, and then integrated the detected boundary information into the conditional random fields [21] to refine the segmentation results. Bertrasius et al. [48] utilized deep object features to train the boundary detection network, whereas Kokkinos [49] proposed a multi-scale boundary detection network with deep supervision. Cheng et al. [50] presented a multitask network that performs boundary detection and semantic segmentation together and uses the detected boundary as a regularization term for semantic segmentation. Recently, Zou et al. [51] developed a boundary-aware CNN that predicts the semantic boundary by using RGB and depth images and employs the boundary information to guide the Conv layers to extract effective features for segmentation. Ding et al. [52] attempted to improve segmentation accuracy by adding high-resolution score maps obtained from very low-level feature maps. Since the high-resolution score maps may contain a large amount of noise, they introduced a boundary delineation module that selectively supplements the score maps only near the boundaries.

**III. PROPOSED METHOD**

**A. OVERALL NETWORK ARCHITECTURE**

Fig. 1 presents the overall architecture of a common encoder-decoder network including the proposed PBRM; in this example, U-Net [15] is utilized as a baseline network. Any type of backbone network such as VGG16 [53], MobileNet [39], and ResNet variants [54] can be utilized as the encoder network. The proposed PBRM is inserted into the decoder part of the network. Unless otherwise mentioned, a “Conv” block contains 3 × 3 Conv, batch normalization, and ReLU activation layers. The encoded feature maps are down-sampled by pooling or strided Conv layers and the total number of feature levels is denoted as $N$. The $i$-th level feature map concatenated with the skip connection of U-Net is denoted as $F_i$ in Fig. 1. The top-most feature map, $F_N$, has the lowest resolution but high-level semantics. At the end of the encoder network, a coarse prediction, $P_N$, is obtained by using $F_N$. Then, $P_N$ is up-sampled, and fed into the proposed PBRM, as shown in Fig. 2. In the PBRM, UPA and UBA units emphasize the lower-level features in consideration of the coarse upper-level prediction. Specifically, the UPA unit utilizes the upper-level predicted class probability to learn the regions that need to be corrected, and refines the prediction by emphasizing or suppressing the lower-level features. The UBA unit highlights the features near the semantic boundary.

**FIGURE 1.** Overall network architecture of a common encoder-decoder network with the proposed PBRM. “∥” represents concatenation.
of the upper-level prediction to capture more spatial details near the boundaries. The structure of UPA and UBA units are illustrated in Figs. 3 and 4, respectively, and detailed explanation on the two attention units is provided in the following subsections. The attentive feature maps obtained by the UPA and UBA units are concatenated, and followed by Conv layers to obtain the residual prediction. By adding the residual, the coarse prediction map is gradually revised to a finer and more complete one in a top-down manner.

### B. UPPER-LEVEL PREDICTION ATTENTION

Reverse attention [42], [55] that uses the probability of belonging to the background as an attention mask was proposed to extract informative features for distinguishing the foreground and background regions in the field of binary image segmentation. Motivated by the reverse attention, we introduce the UPA unit that can be used for semantic segmentation of multiple object classes. The attention mask of the proposed UPA unit is generated from the predicted per-class probability, similar to the reverse attention. However, unlike the reverse attention that directly uses the probability map as the attention mask, the UPA unit trains the mask from the probability through a Conv layer. In other words, the UPA unit is designed to learn where to pay more attention to better refine the upper-level prediction, and to extract more informative features by using the leaned attention mask. As shown in Fig. 3, the per-class softmax probability of the up-sampled upper-level prediction is passed through the $1 \times 1$ Conv, batch normalization, and sigmoid activation layers to generate the UPA mask. Then, the mask is multiplied by the input feature maps to obtain the attentive feature maps. Given the $n$-th level feature maps, $F_n$, the $n$-th level output of the UPA unit, $h_{n}^{\text{UPA}}$, is formulated as

$$h_{n}^{\text{UPA}} = A_{n}^{\text{UPA}} \otimes F_n,$$

where $\otimes$ denotes element-wise multiplication, and the UPA mask, $A_{n}^{\text{UPA}}$, is defined as

$$A_{n}^{\text{UPA}} = \sigma(\text{Conv}_{1 \times 1}^m(\text{softmax}(P_{n+1}^\top))).$$

where $P_{n+1}^\top$ is the bilinearly up-sampled $(n + 1)$-th level prediction and $\sigma$ represents a sigmoid activation function. $\text{Conv}_{1 \times 1}^m$ denotes a Conv layer with $m$ $1 \times 1$ Conv kernels. Let $C_n$ denote the number of channels of $F_n$. The number of Conv kernels, $m$, can then be chosen among the divisors of $C_n$. When $m = 1$, a single $1 \times 1$ kernel generates one UPA mask, and then every channel of $F_n$ is multiplied by the same mask. When $m$ is set as $C_n$, the UPA unit learns $C_n$ different attention masks, each of which is appropriate for each channel of $F_n$.

Figs. 5(a), 5(c), and 5(e) illustrate the attention masks of UPA units when $m = 1, 8$, and $C_n$, respectively, where the input feature maps correspond to $F_3$ of the MobileNetV2-based encoder network [39]. As can be seen, the UPA masks have similar attention weights within the same semantic regions, such as road, sky, cars, and pedestrians. This is because these masks are obtained from the per-class probability maps. The feature maps presented in Fig. 5(b) are relatively similar to each other, whereas those in Figs. 5(d) and 5(f) have more diverse appearances. This implies that multiple UPA masks can extract more diverse features in consideration of semantic classes. However, using a large value of $m$ can also increase the number of parameters and FLOPS, thus it is important to find a balance between the accuracy and model size. The dependency of the segmentation accuracy and model size on $m$ is discussed in the ablation study of Section IV-C.

### C. UPPER-LEVEL BOUNDARY ATTENTION

In our previous work [55], BA unit was introduced for binary road segmentation. In this paper, we present the UBA unit, an extended version of BA unit [55], that can be applied not only to binary segmentation, but also to semantic segmentation with multiple classes. As shown in Fig. 4, from
the up-sampled upper-level prediction, $P_{n+1}^i$, UBA unit first generates a label map, $L_n(p)$, which is defined as

$$L_n(p) = \arg\max_i (P_{n+1}^i(p, i)), i = \{0, 1, \ldots, C-1\},$$

where $p$ and $C$ denote the pixel position and the number of semantic classes, respectively, and thus $L_n(p)$ represents the predicted class label at pixel position $p$ in the up-sampled $(n+1)$-th level prediction. Using the label map, a semantic boundary map, $B_n$, is obtained as

$$B_n(p) = \begin{cases} 1, & \text{if } \exists q \in \mathbb{N}_4(p), L_n(p) \neq L_n(q), \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbb{N}_4(p)$ represents the 4-neighborhood of $p$.

In our previous work [55], the attention weight was simply defined using the distance from each pixel to its closest boundary, where the distance was obtained by distance transformation (DT) [56] of $B_n$. However, in the proposed UBA unit, we introduce a lightweight boundary attention mask generation (BAMG) block that consists of Conv layers. Consequently, the UBA mask becomes trainable using the upper-level boundary map. As shown in Fig. 6, the BAMG block has multiple parallel $3 \times 3$ Conv layers with different dilation rates to spread the attention weight from the boundary region, and the parallel feature maps are combined by a $1 \times 1$ Conv layer. The $n$-th level output of the proposed UBA unit is formulated as follows:

$$h_n^{\text{UBA}} = A_n^{\text{UBA}} \otimes F_n,$$

where the UBA mask, $A_n^{\text{UBA}}$, is defined as

$$A_n^{\text{UBA}} = \sigma \left( \text{Conv}_{1 \times 1} \left( \sum_{i=1}^{K} \text{Conv}_{3 \times 3, d_i}(B_n) \right) \right),$$

where $K$ is the number of $3 \times 3$ Conv layers in the BAMG block, $\text{Conv}_{3 \times 3, d_i}$ denotes $3 \times 3$ Conv layer with a dilation rate of $d_i$. We compare four different designs for the BAMG block by changing $K$ from 1 to 4. Fig. 7 shows the learned UBA masks obtained from different BAMG blocks and their corresponding output feature maps. As $K$ increases, the UBA mask has slightly wider middle gray regions. As can be seen from the output feature maps in Fig. 7, the proposed UBA unit effectively emphasizes the features near the segmentation boundary. Note that there is no significant differences in the output feature maps. This is because even if the dilation rate is set differently, the network is trained to emphasize only the region contains the features needed for supplementing the upper-level prediction.

### D. Auxiliary Supervision

The proposed PBRM can be included in any level of the conventional decoder network except for the top-most $N$-th level, and the successive decoding blocks with PBRMs can hierarchically improve the segmentation results. In the
FIGURE 7. Examples of attention masks (top) and output feature maps (bottom) of UBA units for (a) $K = 1$, (b) $K = 2$, (c) $K = 3$, and (d) $K = 4$.

deep CNN with the proposed PBRM, the N-th level feature maps generate a coarse prediction, $P_N$, by passing through a Conv layer. Then, $P_N$ is gradually refined by the hierarchical decoding blocks with PBRMs in a top-down manner, resulting in the lower-level fine-grained predictions, $P_{N-1}$ to $P_1$. Likewise with previous deep supervision techniques [22], [57], in addition to the main loss calculated by using $P_1$, we assign the auxiliary supervision to the predictions of the other levels with PBRMs to help optimize the overall learning process. We utilize the cross-entropy loss function, $l_{CE}$, for both the main and auxiliary losses. The total loss, $l_{Total}$, is defined as

$$l_{Total} = l_{CE}(P_1, y) + \alpha \sum_{i \in N_{PBRM}} l_{CE}(U(P_i), y),$$

where $y$ represents the ground-truth class label, $U$ is an up-sampling operation to make the auxiliary predictions to have the same resolution as $y$, and $N_{PBRM}$ is a set of levels where PBRMs are applied. The weight $\alpha$ is set to 0.5 in our method to balance the main and auxiliary losses.

IV. EXPERIMENTS
A. DATASET AND EXPERIMENTAL SETUP

We evaluated the proposed PBRM on Cityscapes [32] and CamVid [33] datasets. These two large-scale urban street scene datasets are widely used in road scene understanding and semantic segmentation. The Cityscapes and CamVid data have 19 and 11 class labels, respectively. Cityscapes dataset consists of 5,000 fine-annotated images and 20,000 coarse-annotated images. In our experiment, we trained all networks using only fine-annotated images. The Cityscapes fine dataset is split into training, validation, and test sets with 2,975, 500, and 1,525 images, respectively. The CamVid dataset consists of 367 training, 101 validation, and 233 test images.

We first investigated the impact of each component in the proposed PBRM on the Cityscapes validation dataset. An encoder-decoder-structured network that consists of a MobileNetV2-based encoder and a FCN-like segmentation head was employed as the baseline for the ablation study. To verify the effectiveness of the proposed PBRM, the baseline network with the proposed PBRM, called “PBR-Net”, was compared with two state-of-the-art attention-based semantic segmentation networks, i.e., DANet [25] and CCNet [26]. For fair comparison, ResNet101 [54] backbones of DANet and CCNet were replaced with MobileNetV2 [39] backbone.

Next, because the proposed PBRM can be integrated into any semantic segmentation network with an encoder–decoder structure, we compared the performance of the semantic segmentation networks with and without using PBRM. In particular, we chose three state-of-the-art real-time semantic segmentation networks, CGNet [29], LiteSeg-mobile [30], and FC-HarDNet70 [31], and measured the performance on the Cityscapes and CamVid test sets. The accuracy on the
Cityscapes test set was obtained by submitting our results to the official evaluation server.\footnote{https://cityscapes-dataset.com/benchmarks/}

For the quantitative evaluation of segmentation accuracy, we used the widely accepted metrics including the mean intersection over union (mIoU) and global accuracy (GA) [58]. To evaluate computational efficiency, we adopted three metrics, i.e., the number of model parameters (Params), FLOPS, and inference time (Time).

\section*{B. IMPLEMENTATION DETAILS}

The proposed and compared networks were implemented using Pytorch [59] and trained on a single NVIDIA Titan XP GPU. For the ablation study and comparative evaluation with state-of-the-art attention modules, we utilized a MobileNetV2 model pre-trained on ImageNet [60] as the backbone network and all the new Conv layers in the proposed PBRM were initialized using He’s initialization [61]. We set the initial learning rates of 0.001 and 0.01 for the MobileNetV2-based encoder network and the decoder network with PBRM, respectively. The stochastic gradient descent optimizer [62] with a momentum of 0.9 was utilized. We used the ‘poly’ learning rate policy with a power of 0.9 for a total of 200 epochs. Data augmentation including random horizontal flipping and random scaling between 0.5 and 2 was applied. We used different crop sizes for training the two datasets, i.e., 768 \(\times\) 768 for the Cityscapes dataset and 360 \(\times\) 480 for the CamVid dataset. To stay well within the hardware limitation, we used a mini-batch size of 8 for all networks. For comparison with the three state-of-the-art real-time segmentation networks, CGNet [29], LiteSeg-mobile [30], and FC-HarDNet70 [31], we followed the training strategy in the author-provided source codes.

\section*{C. ABLATION STUDY}

We first studied the performance dependence of UPA and UBA units on hyperparameter settings and then investigated the impact of each component of the proposed PBRM on the segmentation accuracy and computational cost. The baseline network for this study was defined as an encoder-decoder network consisting of a MobileNetV2-based encoder and a simple decoder that contains 3 \(\times\) 3 Conv layers. The output stride of original MobileNetV2 backbone is 32, which means the ratio of input image resolution to final output resolution is 32. However, we set the output stride as 16 to obtain denser feature maps. Also, the last Conv layer of the MobileNetV2 backbone was excluded because it is computationally inefficient due to its 1,280 Conv kernels. Thus, the modified MobileNetV2-based encoder contained four levels of feature maps, and the proposed PBRM was included into 1/8 and 1/4 scale feature maps as shown in Fig. 8.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{network_architecture.png}
\caption{Network architecture of baseline network and decoder with the proposed PBRM.}
\end{figure}

\begin{table}[h]
\centering
\caption{Hyperparameter analysis on UPA unit.}
\begin{tabular}{|c|c|c|c|c|}
\hline
Network & Overall FLOPS & Params & mIoU(\%) \\
\hline
Baseline & 12.71G & 2.13M & 68.55 \\
UPA (m=1) & 12.71G & 2.13M & 69.73 \\
UPA (m=8) & 12.71G & 2.13M & 70.37 \\
UPA (m=C_n) & 12.74G & 2.13M & 70.56 \\
\hline
\end{tabular}
\end{table}

1) \textbf{HYPERPARAMETER ANALYSIS ON UPA UNIT}

We examined the impact of the hyper parameter \(m\) of the UPA unit in terms of the accuracy and efficiency. The parameter \(m\) represents the number of 1 \(\times\) 1 Conv kernels that is utilized for generating the UPA masks. We compared three cases: \(m = 1, 8,\) and \(C_n\). When \(m = 1\), only one UPA mask was obtained; thus, each channel of the feature maps was emphasized by using the same attention mask. When \(m = C_n\), the UPA unit generated \(C_n\) different UPA masks and each of them was multiplied by each channel of the lower-level feature maps. In other words, channel-wise different UPA masks were utilized to assign different attention weights to each channel of the lower-level feature maps. When \(m = 8\), the UPA unit obtained 8 attention masks and each mask was repeated \(C_n/8\) times to have the same number of channels of the lower-level feature maps. As listed in Table 1, the three networks with the UPA units show higher mIoU values than the baseline, and the network with higher \(m\) shows better segmentation accuracy. This implies that channel-wise different UPA masks can effectively emphasize more informative features in consideration of semantic classes. The performance gap between the models with \(m = 8\) and \(m = C_n\) is not significant, as compared to the one with \(m = 1\). We also compared the three UPA units in terms of FLOPS and the number of parameters. As shown in Table 1, a larger \(m\) requires more parameters and FLOPS. Especially when
m = Cn, the refinement module contains 31.72M FLOPS, which is extremely larger than 0.45M FLOPS and 3.60M FLOPS of the other two cases. Based on this result, we finally chose m = 8 for the UPA unit considering both segmentation accuracy and computational efficiency.

2) HYPERPARAMETER ANALYSIS ON UBA UNIT

We compared the performance of the UBA unit with different K values of the BAMG block in (6), where K denotes the number of 3 × 3 Conv layers with different dilation rates. We evaluated four different cases by changing K from 1 to 4. The experimental results are provided in Table 2. The UBA unit with K = 3 achieves the best performance of 71.01% mIoU, which is 2.46% higher than the baseline. As K increases, the FLOPS and Params of the refinement module proportionally increase. However, the increments are negligible compared to the overall network because the 3 × 3 Conv layer in the BAMG block uses only a small number of Conv kernels. Thus, we selected the best performing condition, i.e., K = 3.

### Table 2. Hyperparameter analysis on UBA unit.

| Network | FLOPS | Params | mIoU (%) |
|---------|-------|--------|----------|
| Overall | UBA   | Overall | UBA      |         |
| Baseline | 12.71G | - | 2.13M | - | 68.55 |
| UBA (K=1) | 12.71G | 3.32M | 2.13M | 44.24K | 70.27 |
| UBA (K=2) | 12.71G | 6.43M | 2.13M | 44.40K | 70.94 |
| UBA (K=3) | 12.71G | 9.54M | 2.13M | 44.56K | 71.01 |
| UBA (K=4) | 12.71G | 12.66M | 2.13M | 44.72K | 70.78 |

3) COMBINATION OF UPA AND UBA UNITS

According to the experimental results from the two aforementioned studies, we determined m = 8 and K = 3 for the UPA and UBA units, respectively. There are two simple ways to combine the attentive feature maps obtained from the UPA and UBA units: summation and concatenation. In the case of summation, we separately obtained residual prediction from each attentive feature map and then added the residuals to the upper-level prediction. In the case of concatenation, we first concatenated the two attentive feature maps and obtained the residual by passing through a Conv layer together. We compared the two cases in terms of the accuracy and model size. As shown in Table 3, when the two attention units are utilized together, the mIoUs are further increased in both cases. The FLOPS and Params are slightly different for the two cases, but concatenation shows higher accuracy of 72.43% mIoU, which is 1.3% higher than summation. Thus, our proposed PBRM adopted to utilize concatenated attentive feature maps from the UPA (m = 8) and UBA (K = 3) units.

D. COMPARISON WITH STATE-OF-THE-ART ATTENTION MODULES

We compared the proposed PBRM with the state-of-the-art attention modules, i.e., DAM and RCCAM, proposed in DANet [25] and CCNet [26], respectively. The original DANet and CCNet are based on a ResNet101 [54] backbone network. However, ResNet101-based semantic segmentation networks have an enormous number of parameters and high computational costs, and thus they are not appropriate for real-time applications, such as autonomous driving.
Therefore, in this experiment, we replaced the ResNet101 backbone to the aforementioned modified MobileNetV2 backbone and evaluated the effectiveness of DAM and RCCAM on a lightweight backbone. The two attention-based segmentation networks were reproduced by using the author-provided source codes and the recurrence for RCCAM was set as 2 in our experiment. As listed in Table 4, all networks with attention modules improve the segmentation accuracy as compared to the baseline. The baseline network with the proposed PBRM outperforms those with DAM and RCCAM, showing the highest mIoU of 72.43%. Besides, PBRM requires much less parameters and FLOPS than the other two conventional modules. This is because DAM and RCCAM measure global dependency for each pixel to compute the attention weights, which requires a large number of FLOPS. The experimental result confirms that the proposed PBRM is a lightweight but effective module for improving the segmentation performance.

E. COMPARISON WITH STATE-OF-THE-ART SEMANTIC SEGMENTATION NETWORKS

To verify the applicability of the proposed PBRM, we compared the performance of the models before and after including PBRM on three state-of-the-art real-time semantic segmentation networks, i.e., CGNet [29], LiteSeg-mobile [30], and FC-HarDNet70 [31]. The encoder of CGNet [29] pools the feature maps to 1/8 scale, and the segmentation head predicts the final result using these feature maps. The proposed PBRM was applied at the 1/8 and 1/4 scale feature maps. LiteSeg-mobile [30] utilizes the MobileNetV2 backbone as the encoder whose top-most feature maps have 1/32 scale. In the decoder part, the top-most encoded feature maps are up-sampled as a factor of 8, and then concatenated with 1/4 scale feature maps extracted from the encoder. The final segmentation result is predicted by using these concatenated feature maps. Thus, we included PBRM to emphasize informative features in these concatenated feature maps. FC-HarDNet70 [31] employs a simple U-Net-like decoder and the final segmentation result is obtained at the feature maps of 1/4 scale. Therefore, the proposed PBRM was injected to the feature maps of 1/4 scale.

We compared the performance before and after including PBRM in terms of both accuracy and efficiency on the Cityscapes and CamVid test datasets. The performance comparison on the Cityscapes test dataset is provided in Table 5. The inference time was measured using input...
size of 2048 × 1024 with a single Titan XP GPU and FLOPS were estimated for an input size of 1024 × 512. As shown in Table 5, the proposed PBRM further increases the mIoU by 3.4%, 2.7%, and 1.5% from that of the original CGNet, LiteSeg-mobile, and FC-HarDNet70, respectively.

The FLOPS and Params are only slightly increased when the proposed PBRM is included into the conventional networks. PBRM requires 0.5G, 0.2G, and 0.2G additional FLOPS for CGNet, LiteSeg-mobile, and FC-HarDNet70, respectively. Although the increase amounts of FLOPS and Params are dependent on the original network architecture, all networks require only about 0.3G additional FLOPS and 0.01M additional Params on average to include PBRM. In terms of the inference time, as shown in Table 5, PBRM increases the inference time by about 2.03 ms on average, but all networks still perform in real-time for an input size of 2048 × 1024.

As listed in Table 6, we compared the performance of state-of-the-art semantic segmentation networks on the Cityscapes test set in terms of the accuracy and model size. As a metric of the model size, we utilized Params because it is an objective metric that does not change depending on implementation environments. Here, the networks are separated into large and small models according to Params. Large models, such as PSPNet [22] and DeepLabV3plus [19], exhibit high mIoUs, but these networks are based on deep and heavy backbone networks such as ResNet101 and Xception71, which are not appropriate for real-time applications. Small models show relatively low accuracy but their Params are less than about 10% of those of large models. The proposed PBRNet, a network composed of the modified MobileNetV2 backbone and PBRM, achieves 72.4% mIoU with 2.14M Params. Compared to CARNet-mobile [67] that utilizes MobileNetV2 backbone, PBRNet shows better accuracy with less Params. CGNet with the proposed PBRM exhibits a higher mIoU than ENet [66] and ESPNetV2 [34] that have a similar model size. FC-HarDNet70 with the proposed PBRM outperforms other small models, showing the highest mIoU of 77.4%. Fig. 9 shows the semantic segmentation results of FC-HarDNet70 with and without PBRM. As can be seen, PBRM can effectively recover fine details, resulting in more accurate segmentation boundaries of road, pedestrians, traffic signs, and poles.

The effectiveness of PBRM can also be verified in the experimental results on the CamVid test set, as listed in Table 7. When PBRM is included in the conventional networks, the mIoU and GA values are improved by about 1.96% and 0.6% on average, respectively, with only a slight increase in Params. Based on these results, it is confirmed that the proposed lightweight PBRM can effectively improve the segmentation performance without a significant increase in computational cost.

V. CONCLUSION

In this paper, we proposed a prediction and boundary attention-based refinement module for recovering fine details in semantic segmentation networks for road scene understanding. The UPA unit focuses on extracting the informative features from the semantic regions by using the per-class probability of the upper-level prediction. The UBA unit highlights the features near the previously estimated upper-level segmentation boundary. By adding the proposed PBRM into the decoder network, the segmentation result can be gradually refined to a finer and more complete one in a top-down manner. We evaluated our proposed PBRM on public datasets for road scene understanding in terms of both accuracy and efficiency and confirmed that applying the proposed lightweight PBRM to the existing semantic segmentation networks can achieve significant performance improvements qualitatively and quantitatively with a small amount of increase in computational cost.

Since the PBRM was developed as a stand-alone module that can be applied to any encoder-decoder-structured segmentation networks, the performance depends on the adopted network. As a future work, we plan to devise a dedicated...
network that can take full advantage of PBRM. In addition, we aim to apply the PBRM to other tasks of road scene understanding including panoptic segmentation and semantic video segmentation.

**REFERENCES**

[1] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, M. Sokolsky, G. Stanek, D. Stavens, A. Teichman, M. Werling, and S. Thrun, “Towards fully autonomous driving: Systems and algorithms,” in *Proc. IEEE Intell. Vehi-
cles Sysmp. (IV)*, Jun. 2011, pp. 163–168.

[2] S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y. H. Eng, D. Rus, and M. H. Ang, “Perception, planning, control, and coordination for autonomous vehicles,” *Machines*, vol. 5, no. 1, p. 6, Feb. 2017.

[3] G. Máttyus, S. Wang, S. Fiidler, and R. Urtasun, “HD maps: Fine-grained road segmentation by parsing ground and aerial images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3611–3619.

[4] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual U-Net,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[5] O. Filin, A. Zapara, and S. Panchenko, “Road detection with DOSResUNet and post vectorizing algorithm,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 211–215.

[6] T. Sun, Z. Chen, W. Yang, and Y. Wang, “Stacked U-Nets with multi-
output for road extraction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 202–206.

[7] K. Kumari Eerapu, B. Ashwath, S. Lal, F. Dell’Acqua, and A. V. N. Dhan, “Dense refinement residual network for road extraction from aerial imagery data,” *IEEE Access*, vol. 7, pp. 151764–151782, 2019.

[8] J. Liang, N. Homayounfar, W.-C. Ma, S. Wang, and R. Urtasun, “Convolutional recurrent network for road boundary extraction,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9512–9521.

[9] W. Zhou, S. Lv, Q. Jiang, and L. Yu, “Deep road scene understanding,” *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 587–591, Apr. 2019.

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

[11] L. Deng, M. Yang, H. Li, T. Li, B. Hu, and C. Wang, “Restricted deformable convolution-based road scene semantic segmentation using surround view cameras,” *IEEE Trans. Intell. Transp. Syst.*, early access, Sep. 17, 2019. doi: 10.1109/TITS.2019.2939832.

[12] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[13] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[15] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Munich, Germany, 2015, pp. 234–241.

[16] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters–improve semantic segmentation by global convolutional network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4353–4361.

[17] A. Kendall, V. Badrinarayanan, and R. Cipolla, “Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding,” in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2017, p. 1–11.

[18] L. Fan, W.-C. Wang, F. Zha, and J. Yan, “Exploring new backbone and attention module for semantic segmentation in street scenes,” *IEEE Access*, vol. 6, pp. 71566–71580, 2018.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 801–816.

[20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” in *Proc. Int. Conf. Learn. Representations*, May 2015, pp. 1–14.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
W. Chen, K. Dasa, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, “Context encoding for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7151–7160.

S. Li, M. Dong, G. Du, and X. Mu, “Attention Dense-U-Net for automatic breast mass segmentation in digital mammograms,” IEEE Access, vol. 7, pp. 59037–59047, Dec. 2019.

G. Bertasius, J. Shi, and L. Torresani, “High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 504–512.

I. Kokkinos, “Pushing the boundaries of boundary detection using deep learning,” in Proc. Int. Conf. Learn. Representations, May 2016, pp. 1–12.

D. Cheng, G. Meng, S. Xiang, and C. Pan, “FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 12, pp. 5769–5783, Dec. 2017.

N. Zou, Z. Xiang, Y. Chen, S. Chen, and C. Qiao, “Boundary-aware CNN for semantic segmentation,” IEEE Access, vol. 7, pp. 114520–114528, 2019.

H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, “Semantic segmentation with context encoding and multi-path decoding,” IEEE Trans. Image Process., vol. 29, pp. 3520–3533, Jan. 2020.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

J.-Y. Sun, S.-W. Kim, S.-W. Lee, Y.-W. Kim, and S.-J. Ko, “Reverse and boundary attention network for road segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 876–885.

G. Bordefors, “Distance transformations in digital images,” Comput. Vis., Graph., Image Process., vol. 34, no. 3, pp. 344–371, Jun. 1986.

Z. Zhang, X. Zhang, C. Peng, D. Cheng, and J. Sun, “ExFuse: Enhancing feature fusion for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., Sep. 2018, pp. 269–284.

A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, “A review on deep learning techniques applied to semantic segmentation,” 2017, arXiv:1704.06857. [Online]. Available: http://arxiv.org/abs/1704.06857

L. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in Proc. Neural Inf. Process. Syst. Workshops, Dec. 2017, pp. 1–4.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 221–252, Dec. 2015.

K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1026–1034.

S. Ruder, “An overview of gradient descent optimization algorithms,” 2016, arXiv:1609.04747. [Online]. Available: http://arxiv.org/abs/1609.04747

F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in Proc. Int. Conf. Learn. Representations, May 2016, pp. 1–13.

G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 1925–1934.

X. Li, L. Zhang, A. You, M. Yang, K. Yang, and Y. Tong, “Global agglomeration then local distribution in fully convolutional networks,” in Proc. Brit. Mach. Vis. Conf., Sep. 2019, pp. 1–15.

A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “E-Net: A deep neural network architecture for real-time semantic segmentation,” 2016, arXiv:1606.02147. [Online]. Available: http://arxiv.org/abs/1606.02147

S. Hao, Y. Zhou, Y. Zhang, and Y. Guo, “Contextual attention refinement network for real-time semantic segmentation,” IEEE Access, vol. 8, pp. 55230–55240, 2020.

W. Chen, X. Gong, X. Liu, Q. Zhang, Y. Li, and Z. Wang, “FasterSeg: Searching for faster real-time semantic segmentation,” in Proc. Int. Conf. Learn. Representations, Apr. 2020, pp. 1–14.

S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, “The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jul. 2017, pp. 11–19.

JEE-YOUNG SUN (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2014, where she is currently pursuing the Ph.D. degree. She joined the Computer Vision and Image Processing Laboratory, Department of Electrical Engineering, Korea University, in March 2014. Her current research interests include image processing, computer vision, autonomous driving, and artificial intelligence.

SEUNG-WON JUNG (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2005 and 2011, respectively. He was a Research Professor with the Research Institute of Information and Communication Technology, Korea University, from 2011 to 2012. He was a Research Scientist with the Samsung Advanced Institute of Technology, Yongin, South Korea, from 2012 to 2014. He was an Assistant Professor with the Department of Multimedia Engineering, Dongguk University, Seoul, from 2014 to 2020. He is currently an Assistant Professor with the Department of Electrical Engineering, Korea University. He has published over 50 peer-reviewed articles in international journals. His current research interests include deep learning and computer vision applications. He received the Hae-Dong Young Scholar Award from the Institute of Electronics and Information Engineers, in 2019.

SUNG-JEA KO (Fellow, IEEE) received the B.S. degree in electronic engineering from Korea University, in 1980, and the M.S. and Ph.D. degrees in electrical and computer engineering from the State University of New York at Buffalo, in 1986 and 1988, respectively. In 1992, he joined the Department of Electronic Engineering, Korea University, where he is currently a Professor. From 1988 to 1992, he was an Assistant Professor with the Department of Electrical and Computer Engineering, University of Michigan-Dearborn. He has published over 210 international journal articles. He holds over 60 registered patents in fields, such as video signal processing and computer vision. He is a member of the National Academy of Engineering of Korea. He was a 1999 recipient of the LG Research Award. He received the Hae-Dong Best Paper Award from the Institute of Electronics and Information Engineers (IEIE), in 1997, the Best Paper Award from the IEEE Asia Pacific Conference on Circuits and Systems, in 1996, the Research Excellence Award from Korea University, in 2004, and the Technical Achievement Award from the IEEE Consumer Electronics (CE) Society, in 2012. He received the 15-Year Service Award from the TPC of ICCE, in 2014, and the Chester Sall Award (First Place Transaction Paper Award) from the IEEE CE Society, in 2017. He was honored with the Science and Technology Achievement Medal from the Korean Government, in 2020. He has served as the General Chairman of ITC-CSICC 2012 and the General Chairman of IEICE 2013. He was the President of the IEIE, in 2013, the Vice-President of the IEEE CE Society, from 2013 to 2016, and the Distinguished Lecturer of the IEEE, from 2015 to 2017. He is a member of the editorial board of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS.