Reverse and Boundary Attention Network for Road Segmentation

Jee-Young Sun, Seung-Wook Kim, Sang-Won Lee, Ye-Won Kim, and Sung-Jea Ko*
Korea University
Seoul, Korea
{jysun, swkim, swlee, ywkim}@dali.korea.ac.kr, sjko@korea.ac.kr

Abstract

Road segmentation is an essential task to perceive the driving environment in autonomous driving and advanced driver assistance systems. With the development of deep learning, road segmentation has achieved great progress in recent years. However, there still remain some problems including the inaccurate road boundary and the illumination variations such as shadows and over-exposure regions. To solve these problems, we propose a residual learning-based network architecture with residual refinement module composed of the reverse attention and boundary attention units for road segmentation. The network first predicts a coarse road region from deeper-level feature maps and gradually refines the prediction by learning the residual in a top-down approach. The reverse and boundary attention units in residual refinement module guide the network to focus on the features in the previously missing region and the region near the road boundary. In addition, we introduce the boundary-aware weighted loss to reduce the false prediction. Experimental results demonstrate that the proposed approach outperforms the state-of-the-art methods in terms of the segmentation accuracy in various benchmark datasets for traffic scene understanding.

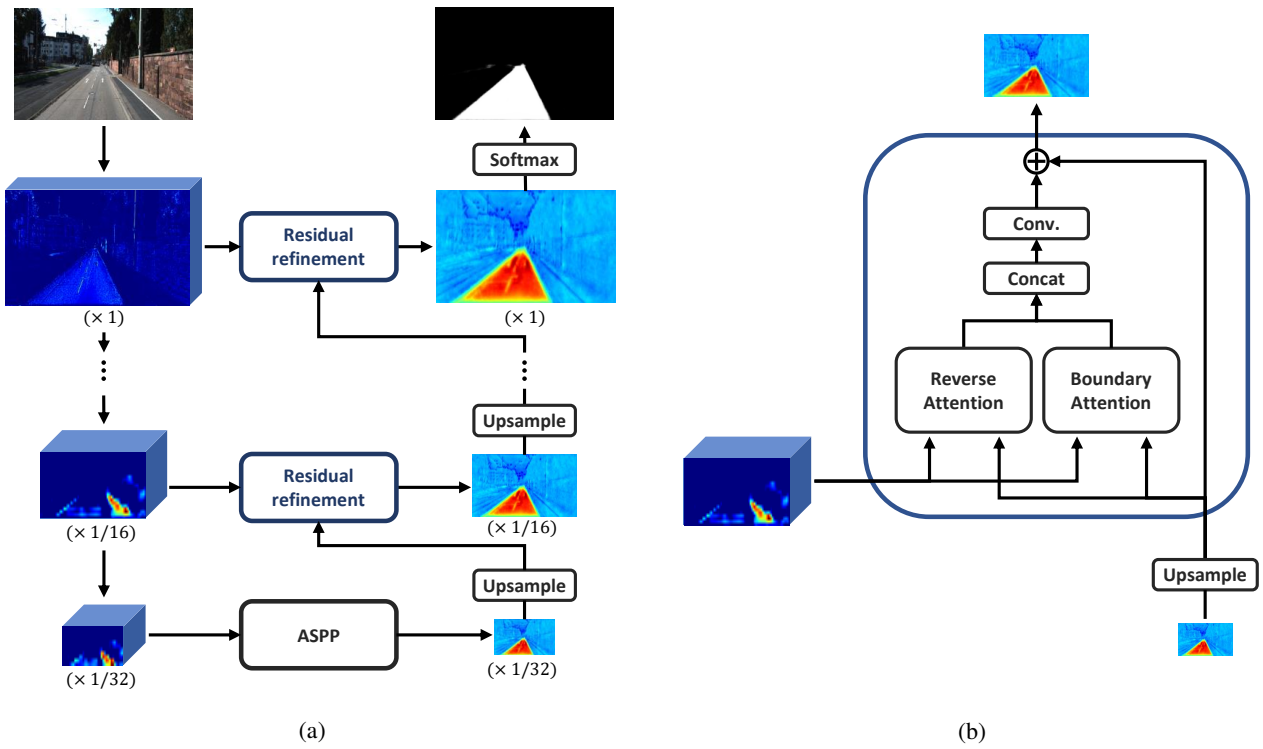
1. Introduction

In recent years, there has been considerable research interest in autonomous driving and advanced driver assistance systems (ADAS). As an essential component for autonomous driving, road segmentation aims to perceive the drivable area and provide crucial information for path planning of autonomous vehicles [21, 32]. Besides, road segmentation results can offer valuable prior knowledge for various cognitive tasks such as vehicle detection [4] and pedestrian detection [12].

Traditional road segmentation methods are mainly based on low-level features such as color, edge, and texture [2, 26, 23, 40]. Some of them employ the location prior, assuming that the road is located at the bottom of the image [1, 6]. However, the low-level hand-crafted features are not robust to illumination changes, and the assumption of the location prior does not hold when there is an on-road object in front of the autonomous vehicle, resulting in unsatisfactory segmentation results. Recently, deep learning has made remarkable progress in computer vision, and deep convolutional neural network (CNN)-based models [24, 3, 36, 45, 33, 8] have achieved high performance in semantic segmentation. Based on these models, a variety of deep-learning-based road segmentation methods [30, 42, 46, 25, 43, 44, 27, 10] have been introduced that attempt to reduce the processing time or improve the performance by injecting additional prior information. However, these methods require extra information such as the road boundary, contour map, and location prior. Besides, they still face the following significant challenges: illumination variations such as shadows and over-exposure regions, and the corrupted road boundary due to the loss of spatial details in the encoder-decoder structures of the segmentation network.

In this paper, we present a residual learning-based network architecture with the residual refinement module composed of the reverse attention (RA) and boundary attention (BA) units for road segmentation. First, we apply a residual learning scheme introduced in skeleton detection [18, 38] and super-resolution [19] to road segmentation. Specifically, the network first predicts a coarse prediction map from the top-level feature maps which have low resolution but high-level semantic information. This coarse prediction map is gradually refined by adding the residual predictions that are obtained from the lower-level feature maps in a top-down manner. To effectively recover the details lost in the encoder part, we employ the residual refinement module using the RA and BA units. The RA unit [9], whose attention mask is inversely related to the predicted probability belonging to the road class, enables the network to discover the missing part of the road in the previously estimated result. To handle the false predictions that frequently occur

*Corresponding author



(a) (b)

Figure 1: (a) Overall architecture of the proposed network for road segmentation. Based on the coarse prediction given by the backbone network at each level, (b) the residual refinement is conducted using the reverse and boundary attention units.

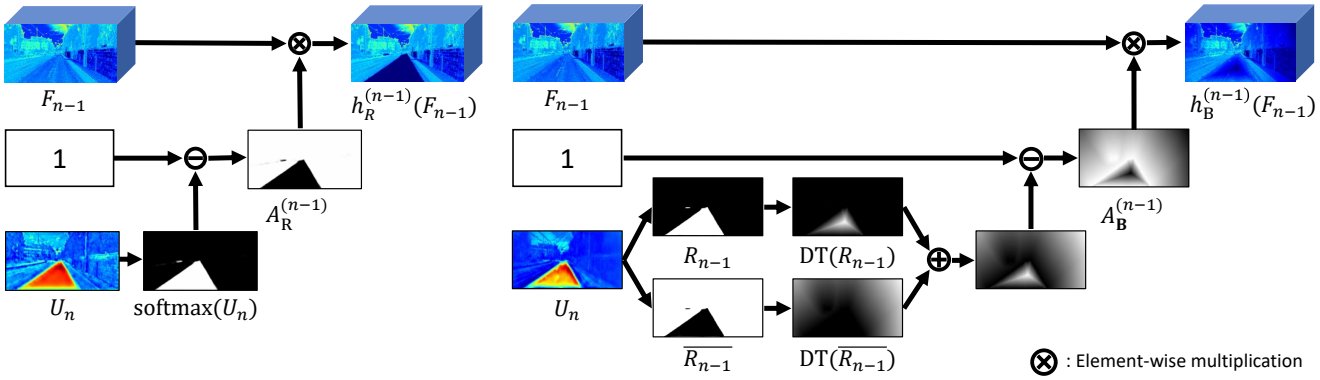
near the road boundary, the BA unit emphasizes the detailed features near the previously estimated road boundary. In addition, we present the boundary-aware weighted (BAW) loss function. The proposed BAW loss function is defined as the pixel-wise weighted cross entropy (CE) loss where the weight is the distance from the false pixel to the boundary of true region. The BAW loss function encourages the false negative (FN) and false positive (FP) regions to be included into the true positive (TP) and true negative (TN) regions, respectively. In summary, the contributions of this paper are the following:

- We present a residual learning approach for road segmentation, which improves the segmentation performance by gradually refining the prediction with the residual details.
- The residual refinement module with the RA and BA units is proposed to guide the network to effectively capture the residual features by emphasizing the missing road region and the road boundary. As compared to the baseline, the residual refinement module increases the F1-measure by 1.49%.
- We further introduce the boundary-aware weighted loss term to reduce the resultant FP and FN pixels. The proposed loss function achieves the performance improvement by 0.3% as compared to the baseline.

The overall network is evaluated on three common benchmarks, and the experimental results demonstrate that the proposed approach clearly outperforms the state-of-the-art methods in terms of F1-measure with 96.3% on KITTI [15], 98% on Cityscapes [11], and 96.72% on CamVid [5] datasets.

2. Related Works

Traditional road segmentation methods detect the road region by the pixel/block-wise classification using a location prior and low-level features such as color, edge, and texture [2, 26, 23, 40, 1, 6]. Alvarez and Lopez [2] proposed illumination invariant features to deal with shadowed street scenes. Mendes *et al*. [26] presented a block-wise classification approach using low-level cues such as color intensity, entropy, and local binary pattern histograms. Sturgess *et al*. [40] classify the road based on the motion and appearance features and then refine the road boundary with the conditional random field. Alvarez *et al*. [1] and



(a) RA unit (b) BA unit

Figure 2: Illustration of (a) reverse attention (RA) and (b) boundary attention (BA) units.

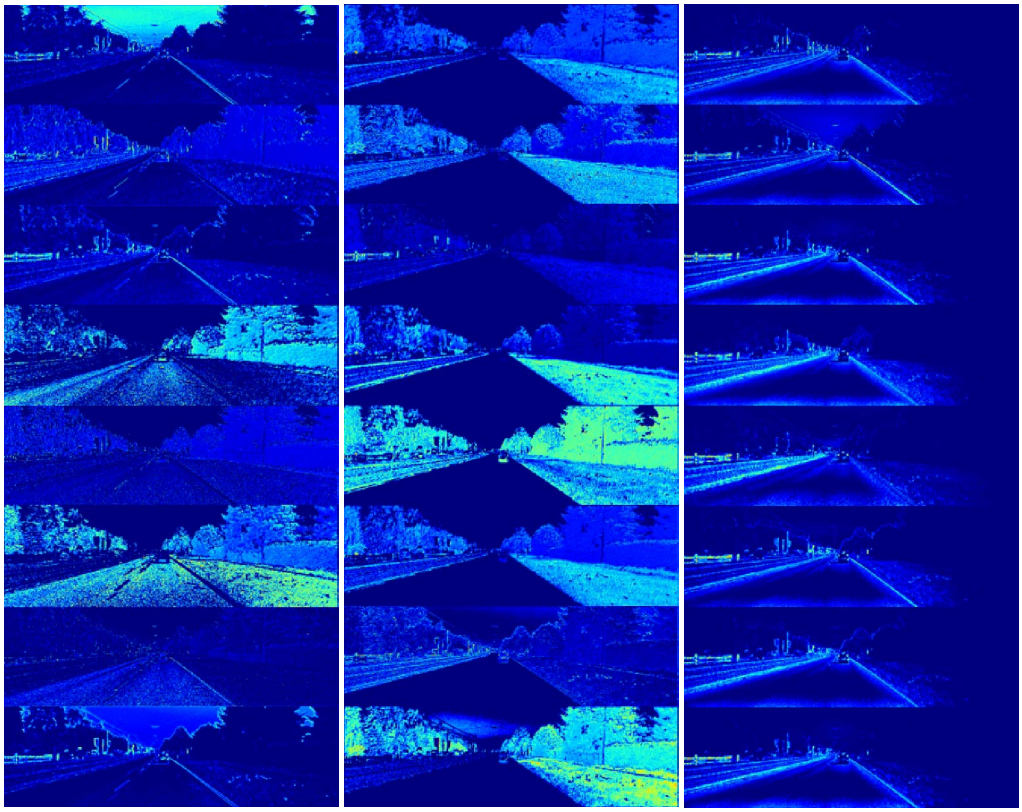
Chacra and Zelek [6] employ the location prior to detect the road area, assuming that the road exists at the lower part of the image. However, the low-level hand-crafted features are not robust to illumination changes. Moreover, the assumption of the location prior does not hold when the road region is occluded with other on-road objects such as vehicles and pedestrians. Therefore, the road regions predicted by the traditional road segmentation methods are somewhat inaccurate.

With the recent development of the deep CNNs, many deep CNN-based approaches including [24, 3, 36, 45, 33, 8] have achieved high performance in the field of semantic segmentation. Based on these networks, there have been many deep CNN-based road segmentation methods [30, 42, 46, 25, 43, 44, 27, 10] to improve the road segmentation performance and reduce the processing time. Oliveira *et al*. [30], Lyu and Huang [25], and Mendes *et al*. [26] proposed smaller CNNs based on an encoder-decoder network architecture to reduce the processing time of road detection. Wang *et al*. [42], Zohourian *et al*. [46], and Yadav *et al*. [43] inject extra knowledge such as contour priors and location priors into their network architectures to improve the road segmentation performance. Zhang *et al*. [44] and Chen and Chen [10] presented a network which simultaneously performs road segmentation and road boundary detection. This method improved the segmentation accuracy by utilizing the predicted results of the two tasks as the guidance for each other. However, these methods require extra information such as road contour map and location priors. Also, they may result in missing regions when there exist illumination variations including shadows and overexposure regions. Moreover, the encoder-decoder structure of the aforementioned networks often causes inaccurate road boundary due to the loss of details; the spatial information can easily be lost because of the down-sampling and interpolating operations contained in the encoder-decoder structure.

3. Proposed Method

3.1. Network Architecture

Figure 1(a) shows the overall architecture of the proposed network, which is built upon SegNet [3]. The encoder network of the SegNet consists of 13 convolutional (Conv) layers which correspond to the first 13 layers of the VGG16 network [39]. After passing through five pooling layers in the encoder network, the output feature maps with 1/32 scale resolution is obtained. Whereas the last feature maps are directly decoded to the original resolution in SegNet, we insert an atrous spatial pyramid pooling (ASPP) block similar to [7]. The ASPP block can effectively resample features at different scales and incorporate global context information. This block contains a 1×1 Conv layer, three 3×3 Conv layers with dilation rate of 2, 4, and 6, and the global context layer consisting of average pooling and interpolation layers. The resulting feature maps from the five layers are concatenated and passed through another 3×3 Conv layer to generate a top-level prediction. The top-level prediction is up-sampled and followed by the residual refinement module as shown in Figure 1(b). During the residual refinement, the RA and BA units emphasize specific regions of the lower-level feature maps by using the up-sampled upper-level prediction. Detailed descriptions on the RA and BA units are provided in the following subsection. The feature maps weighted by the two attention units are concatenated, and followed by Conv layers to predict the residual. Then, the residual prediction is summed up with the up-sampled upper-level prediction to obtain the finer lower-level predic-



(a) (b) (c)

Figure 3: Visualization of learned residual feature maps using the network (a) without the attention unit, (b) with the RA unit, and (c) with the BA unit. (Best viewed in color.)

tion. In this way, the road prediction map is gradually refined through the residual refinement module in a top-down manner.

3.2. Residual Refinement Module

The VGG16 backbone network pre-trained for image classification is highly responsive to sparse discriminative features to classify the object category. However, the proposed network aims to perform a pixel-level road segmentation which needs to extract dense features. Thus, in addition to the residual learning framework, we introduce the residual refinement module with RA and BA units to enable the network to explore more necessary search region.

3.2.1 Reverse Attention Unit

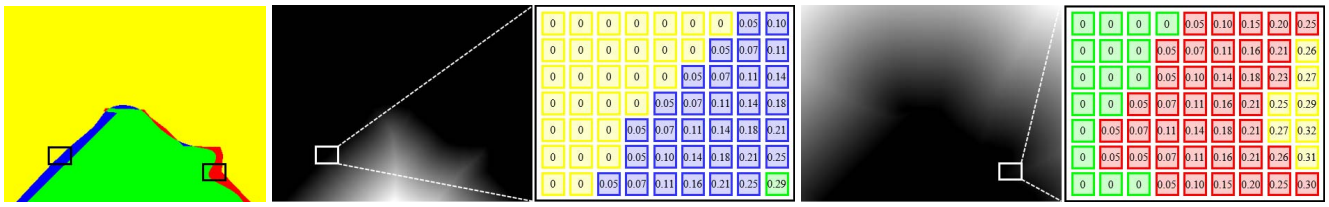
Inspired by [9] in the field of saliency detection, we utilize the RA unit to encourage the network to focus on the complementary road region. Figure 2(a) illustrates the RA unit. Given the $(n-1)$th level feature maps, F_{n-1}, the $(n-1)$th level output of the RA unit is defined as

$$h_{\mathrm{R}}^{(n-1)}(F_{n-1}) = A_{\mathrm{R}}^{(n-1)} \otimes (F_{n-1}), \tag{1}$$

where $\otimes$ denotes the element-wise multiplication and $A_{\mathrm{R}}^{(n-1)}$ is the RA mask. The RA mask is obtained using the up-sampled feature map for the nth level prediction, U_n, whose entry is given by

$$\begin{aligned} A_{\mathrm{R}}^{(n-1)}(p) &= 1 - \mathrm{softmax}(U_n(p))_i|_{i=1} \\ &= \mathrm{softmax}(U_n(p))_i|_{i=0}, \end{aligned} \tag{2}$$

where p denotes the index of the pixel position and the labels of i, one and zero, denote road and non-road regions, respectively. Thus, the RA unit aims to emphasize the feature maps corresponding to the previously estimated non-road region, and guides the network to discover the residual features related to the missing road region.



(a) (b) (c)

Figure 4: (a) A segmentation result generated by the baseline model. TP, TN, FP, and FN regions are denoted as green, yellow, blue, and red, respectively. (b) and (c) The DT results of the complementary TN and TP regions, respectively. (d) and (e) Visualization of the normalized distance values in the rectangle regions of (b) and (c), respectively, and the TP, TN, FP, and FN pixels are displayed in the corresponding colors.

(a) (b) (c)

Figure 5: Comparative segmentation results using different loss functions: (a) CE loss, (b) CE+IoU loss, and (c) CE+BAW loss. (Best viewed in color.)

3.2.2 Boundary Attention Unit

Figure 2(b) shows the BA unit. Similar to the RA unit, the $(n-1)$th level output of the BA unit is defined as

$$h_{\mathrm{B}}^{(n-1)}(F_{n-1}) = A_{\mathrm{B}}^{(n-1)} \otimes F_{n-1}, \tag{3}$$

where $A_{\mathrm{B}}^{(n-1)}$ is the BA mask. To obtain $A_{\mathrm{B}}^{(n-1)}$, the $(n-1)$th level binary road map R_{n-1} is formulated as

$$R_{n-1}(p) = \begin{cases} 1, & \text{if } \mathrm{softmax}(U_n(p))_i|_{i=1} > 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

By applying distance transformation (DT) [13], $\mathrm{DT}(\cdot)$, to R_{n-1}, each pixel position in the road region is filled with the distance to the road boundary. Similarly, the distance map of the non-road region can be simply obtained by DT of $\overline{R_{n-1}} = 1 - R_{n-1}$. After normalizing the two distance maps, they are summed to produce the overall distance map as follows:

$$D_{n-1} = \mathrm{DT}(R_{n-1}) / \max_p \mathrm{DT}(R_{n-1})(p) + \mathrm{DT}(\overline{R_{n-1}}) / \max_p \mathrm{DT}(\overline{R_{n-1}})(p). \tag{5}$$

In (5), D_{n-1} has the value of 0 at the road boundary and 1 at the farthest pixels from the boundary. Then, we define the BA mask as

$$A_{\mathrm{B}}^{(n-1)} = 1 - D_{n-1}, \tag{6}$$

which gives higher attention to the area near the previously predicted road boundary.

3.2.3 Discussion

Figure 3 shows the learned feature maps from the residual network with and without the two attention units. In Figure 3(a), the feature maps with no attention unit is activated around edge or texture details throughout the image, regardless of the road region. However, the RA unit erases the details on previously predicted road region and finds the residual features from the non-road region as shown in Figure 3(b). As shown in Figure 3(c), the BA unit well highlights the features near the road boundary. Note that, while the RA unit focuses on the complementary region of the road, the BA unit emphasizes both inner and outer regions near the previously predicted road boundary. In the proposed residual refinement module, the combination of RA

Loss	F1-measure	Precision	Recall	IoU
CE (baseline)	95.50±1.23	93.13±1.42	98.03±0.82	91.42±1.38
CE+IoU	95.65±1.16	93.80±1.03	97.60±1.13	91.69±0.99
CE+BAW	95.80±1.09	94.04±1.35	97.64±0.86	91.96±1.06

Table 1: Effectiveness of the proposed BAW loss compared with IoU loss.

ASPP	RA	BA	BAW	F1-measure	Precision	Recall	IoU
				95.50±1.23	93.13±1.42	98.03±0.82	91.42 ± 1.38
			✓	95.80±1.09	94.04±1.35	97.64±0.86	91.96 ± 1.06
✓				96.27±0.91	94.34±1.09	98.30±0.88	92.83 ± 1.15
✓	✓			96.79±0.52	95.18±0.69	98.66±0.53	93.97 ± 0.85
✓		✓		96.80±0.77	95.59±0.95	98.04±0.81	93.80 ± 1.04
✓	✓	✓		96.99±0.51	95.08±0.74	98.92±0.56	94.10 ± 0.69
✓	✓	✓	✓	97.06±0.50	95.38±0.58	98.80±0.86	94.29 ± 0.63

Table 2: Effectiveness of each component on KITTI cross-validation. (ASPP: atrous spatial pyramid pooing block; RA: reverse attention unit; BA: boundary attention unit; BAW: boundary-aware weighted loss)

and BA units encourages the network to effectively capture the residual details from the non-road region and the region near the road boundary. In this way, the coarse prediction map can be eventually refined into the complete one.

3.3. Boundary-Aware Weighted Loss

In the field of object detection, focal loss [22] has presented to address the class imbalance. The focal loss is defined as a weighted CE loss, and the weight is inversely related to the class probability of anchor boxes in order to focus on hard examples with low probability. Road segmentation networks trained by using only the CE loss often suffer from false predictions, *i.e.*, groups of FP and FN pixels, when the features are ambiguous to distinguish the road pixels. To deal with the false prediction, we introduce the BAW loss term in form of a weighted CE loss. Although both BAW loss and focal loss are expressed in a similar form, their scale factors are completely different. Whereas the weight of focal loss highlights the anchors with low class probability, the BAW loss gives higher weights on the hard examples for road segmentation, *i.e.*, false pixels near the road boundary. The BAW loss is defined as follows:

$$\ell_{\text{BAW}} = \frac{1}{|\mathcal{F}|} \sum_{k \in \mathcal{F}} \alpha_k \times \ell_{\text{CE}}(k), \tag{7}$$

where $\ell_{\text{CE}}(k)$ means the CE loss of kth pixel position, $\mathcal{F} = \text{FP} \cup \text{FN}$. The boundary-aware distance weight α_k is defined as follows:

$$\alpha_k = \begin{cases} 1 - \frac{\text{DT}(\overline{\text{TN}})}{\max_p \text{DT}(\overline{\text{TN}})(p)}, & k \in \text{FP}, \\ 1 - \frac{\text{DT}(\overline{\text{TP}})}{\max_p \text{DT}(\overline{\text{TP}})(p)}, & k \in \text{FN}. \end{cases} \tag{8}$$

In (8), the boundary-aware distance weight is inversely related to the distance from FP and FN pixels to TN and TP regions, respectively. In other words, false pixels near the road boundary have relatively high α, while those far from the road boundary have low α values. For example, the green, yellow, blue, and red regions in Figure 4(a) correspond to groups of TP, TN, FP, and FN pixels, respectively. The normalized distance maps for the complementary TN and TP regions are obtained by DT, as shown in Figures 4(b) and (c). The proposed BAW loss is calculated by (7) and (8), and then the summation with the conventional CE loss is used as the total loss of our network.

4. Experiments

4.1. Training Details

The VGG16 model pre-trained on ImageNet [37] was employed as a backbone network, and all the new Conv layers were initialized using the Xavier method [16]. Since the proposed network does not need extra knowledge and formed as an end-to-end model, we trained the overall parameters at once. The proposed network was implemented using Pytorch [31] and trained on a single NVIDIA Titan XP GPU. We used a mini-batch size of 2, the Adam optimizer [20] with the base learning rate of 1e-4, and the 'poly' learning rate policy with the power of 0.9. Data augmentation contains random mirroring and random scaling between 0.5 and 2.

4.2. Dataset and Evaluation

We evaluate the proposed method on KITTI [15], Cityscapes [11], and CamVid [5] datasets. The KITTI

(a) RBNet (b) SSLGAN (c) Proposed

Figure 6: Visual comparison with state-of-the-art methods on KITTI benchmark: (a) RBNet, (b) SSLGAN, and (c) the proposed method. The red, blue, and green areas correspond to FN, FP, and TP, respectively. (Best viewed in color.)

Methods	Metrics (%)					Runtime
	F1-measure	Precision	Recall	FPR	FNR	
ALO-AVG-MM [35]	92.03	90.65	93.45	5.31	6.55	0.03 s
s-FCN-loc [42]	93.26	94.16	92.39	3.16	7.61	0.4 s
DDN [28]	93.43	95.09	91.82	2.61	8.18	2 s
Up-Conv-Poly [30]	93.83	94.00	93.67	3.29	6.33	0.08 s
DEEP-DIG [29]	93.98	94.26	93.69	3.14	6.31	0.14 s
MultiNet [41]	94.88	94.84	94.91	2.85	5.09	0.17 s
StixelNet II [14]	94.88	92.97	96.87	4.04	3.13	1.2 s
RBNet [10]	94.97	94.94	95.01	2.79	4.99	0.18 s
SSLGAN [17]	95.53	95.84	95.24	2.28	4.76	0.7 s
Proposed	96.30	95.14	97.50	2.75	2.50	0.16 s

Table 3: Leaderboard of the Top-10 published monocular vision-based algorithms on the URBAN ROAD category of the KITTI vision benchmark suite server.

dataset consists of 289 training images and 290 test images. KITTI benchmark is one of the most popular datasets for road segmentation, but the evaluation server is restricted from repeatedly uploading models. Thus, we investigate the effect of each component in the proposed approach through 10-fold cross-validation on the KITTI training dataset, and only the final model is evaluated on the KITTI benchmark server. Cityscapes and CamVid are urban street scene datasets from the perspective of a vehicle, widely used in the field of semantic segmentation. Cityscapes and CamVid data have 19 and 11 class labels, respectively, including the road category. We change the class label for the road to 1 and the others to 0 because the proposed approach focuses on road segmentation. For the Cityscapes dataset, the proposed network is trained using 2,975 training images and evaluated on 500 images in the validation dataset. The CamVid dataset consists of 367 training, 101 validation, and 233 test images. We train our model on the training images, and the performance on the test set is compared with that of the state-of-the-art road segmentation methods. For quantitative evaluation, we follow the widely agreed pixel-wise segmentation metrics, precision, recall, F1-measure, FP rate (FPR), FN rate (FNR), and intersection over union (IoU). Processing time is evaluated with 360×720 input images.

4.3. Comparative Study with IoU Loss

Some recent CNN-based segmentation methods [44, 34] directly optimize the IoU metric to encourage the predicted region to be similar to the ground-truth region, which is referred to as IoU loss. Both the proposed BAW loss and IoU loss are designed for a similar training purpose: reducing false pixels by pushing them into true regions. Therefore, we compare the performance of the same networks with three different loss functions: CE loss (baseline), CE loss

Methods	F1-measure	Precision	Recall
FCN [24]	94.68	93.69	95.70
s-FCN-loc [42]	95.36	94.63	96.11
Zohourian et al. [46]	92.44	89.08	96.76
SegNet [3]	95.81	94.55	97.11
Proposed	98.00	97.87	98.13

Table 4: Road segmentation results on the Cityscapes validation dataset.

Methods	F1-measure	Precision	Recall
SegNet [3]	93.95	93.07	94.86
Yadav *et al*. [43]	94.14	93.31	94.99
Proposed	96.72	97.14	96.30

Table 5: Road segmentation results on the CamVid test dataset.

with IoU loss, and CE loss with the proposed BAW loss. As shown in Table 1, the model trained by using both CE and BAW losses outperforms the others on KITTI dataset. Figure 5 illustrates comparative segmentation results with the three different loss functions. In Figure 5, the blue FP region and the red FN region are well-refined with the help of the BAW loss, achieving the most accurate road boundary.

4.4. Ablation Study

We further examine the effectiveness of each component in the proposed approach, *i.e*., the ASPP block, RA and BA units, and BAW loss. As shown in Table 2, the ASPP block can improve the F1-score from 95.50% to 96.27%. This indicates that the global context information extracted by the ASPP block is desirable for road segmentation. The RA and BA units with the ASPP block increase the F1-measure by 1.29% and 1.30%, respectively, as compared to the baseline. When these two attention units are utilized together, the performance is further improved to 96.99%, which demonstrates that the RA and BA units can achieve synergy in extracting residual features. Finally, when we train the network with the whole components, the proposed approach exhibits the best performance of 97.06%.

4.5. Performance comparison with state-of-the-arts

4.5.1 KITTI benchmark

KITTI benchmark is one of the most popular datasets for road segmentation. The evaluation server ranks all submitted methods according to maximum F1-measures on bird-eye-view transformed results. The performance of the top-10 published monocular vision-based road segmentation algorithms on URBAN ROAD category is reported in Table 3. As shown in the table, the proposed approach exhibits the best performance with 96.30% of F1-measure among the state-of-the-art methods. Figure 6 illustrates the road segmentation results of the top-3 algorithms in the leaderboard of KITTI benchmark. As can be seen, the proposed method yields finer road boundary with less false pixels than the others.

4.5.2 Cityscapes and CamVid datasets

We also examine the performance of the proposed road segmentation method on Cityscapes and CamVid datasets to verify the effectiveness of our method on various road scenes. Since these two datasets are relatively less utilized for road segmentation than KITTI benchmark, we compare the experimental results with a few methods that provide their performance on each of the two datasets. For a fair comparison, the training and evaluation are conducted under the same experimental settings with the competing methods. As shown in Tables 4 and 5, the proposed approach clearly outperforms other methods in terms of all metrics on both Cityscapes and CamVid datasets.

5. Conclusion

In this paper, we have proposed a residual learning-based network architecture for effective road segmentation. The main contributions of the proposed method is the residual refinement module with the RA and BA units. The RA unit gives weights to the previously predicted non-road region to help the network to discover the missing road region. The BA unit emphasizes the regions around the previously estimated road boundary to obtain finer and complete road prediction. We have also presented the boundary-aware weighted loss to handle the groups of the false predictions. By utilizing the proposed refinement module and the boundary-aware weighted loss function, our road segmentation network has shown state-of-the-art performance with 96.3% on KITTI, 98% on Cityscapes, and 96.72% on CamVid in terms of F1-measure. The proposed framework mainly focuses on binary road segmentation. In future work, we will attempt to extend the residual refinement module from binary to multi-class segmentation to apply it to lane detection or road instance segmentation.

Acknowledgments

This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program)(10082585, Development of deep learning-based open EV platform technology capable of autonomous driving) funded By the Ministry of Trade, Industry Energy (MOTIE, Korea).

References

[1] J. M. Alvarez, M. Salzmann, and N. Barnes. Learning appearance models for road detection. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 423–429. IEEE, 2013. 1, 2

[2] J. M. Á. Alvarez and A. M. Lopez. Road detection based on illuminant invariance. *IEEE Trans. Intell. Transp. Syst.*, 12(1):184–193, 2011. 1, 2

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017. 1, 3, 8

[4] M. Betke, E. Haritaoglu, and L. S. Davis. Real-time multiple vehicle detection and tracking from a moving vehicle. *Machine vision and applications*, 12(2):69–83, 2000. 1

[5] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.*, 30(2):88–97, 2009. 2, 6

[6] D. A. Chacra and J. Zelek. Road segmentation in street view images using texture information. In *2016 13th Conference on Computer and Robot Vision (CRV)*, pages 424–431. IEEE, 2016. 1, 2, 3

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 3

[9] S. Chen, X. Tan, B. Wang, and X. Hu. Reverse attention for salient object detection. In *Proc. European Conf. Comput. Vis.*, pages 234–250, 2018. 1, 4

[10] Z. Chen and Z. Chen. Rbnet: A deep neural network for unified road and road boundary detection. In *International Conference on Neural Information Processing*, pages 677–687. Springer, 2017. 1, 3, 7

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Comput. Vis. Pattern Recognit.*, pages 3213–3223, 2016. 2, 6

[12] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(12):2179–2195, 2009. 1

[13] R. Fabbri, L. D. F. Costa, J. C. Torelli, and O. M. Bruno. 2d euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys (CSUR)*, 40(1):2, 2008. 5

[14] N. Garnett, S. Silberstein, S. Oron, E. Fetaya, U. Verner, A. Ayash, V. Goldner, R. Cohen, K. Horn, and D. Levi. Real-time category-based and general obstacle detection for autonomous driving. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 198–205, 2017. 7

[15] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 6

[16] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 6

[17] X. Han, J. Lu, C. Zhao, S. You, and H. Li. Semisupervised and weakly supervised road detection based on generative adversarial networks. *Signal Process. Lett.*, 25(4):551–555, 2018. 7

[18] W. Ke, J. Chen, J. Jiao, G. Zhao, and Q. Ye. Srn: side-output residual network for object symmetry detection in the wild. In *Proc. IEEE Comput. Vis. Pattern Recognit.*, pages 1068–1076, 2017. 1

[19] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proc. IEEE Comput. Vis. Pattern Recognit.*, pages 1646–1654, 2016. 1

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 163–168. IEEE, 2011. 1

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[23] P. Lombardi, M. Zanin, and S. Messelodi. Switching models for vision-based on-board road detection. In *Proc. IEEE Intell. Transp. Syst.*, pages 67–72, 2005. 1, 2

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Comput. Vis. Pattern Recognit.*, pages 3431–3440, 2015. 1, 3, 8

[25] Y. Lyu and X. Huang. Roadnet-v2: A 10 ms road segmentation using spatial sequence layer. *arXiv preprint arXiv:1808.04450*, 2018. 1, 3

[26] C. C. T. Mendes, V. Frémont, and D. F. Wolf. Vision-based road detection using contextual blocks. *arXiv preprint arXiv:1509.01122*, 2015. 1, 2, 3

[27] C. C. T. Mendes, V. Frémont, and D. F. Wolf. Exploiting fully convolutional neural networks for fast road detection. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3174–3179. IEEE, 2016. 1, 3

[28] R. Mohan. Deep deconvolutional networks for scene parsing. *arXiv preprint arXiv:1411.4101*, 2014. 7

[29] J. Munoz-Bulnes, C. Fernandez, I. Parra, D. Fernández-Llorca, and M. A. Sotelo. Deep fully convolutional networks with random data augmentation for enhanced generalization in road detection. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 366–371. IEEE, 2017. 7

[30] G. L. Oliveira, W. Burgard, and T. Brox. Efficient deep models for monocular road segmentation. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4885–4891. IEEE, 2016. 1, 3, 7

[31] A. 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 *NIPS-W*, 2017. 6

[32] S. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y. Eng, D. Rus, and M. Ang. Perception, planning, control, and coordination for autonomous vehicles. *Machines*, 5(1):6, 2017. 1

[33] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017. 1, 3

[34] M. A. Rahman and Y. Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016. 7

[35] F. Reis, R. Almeida, E. Kijak, S. Malinowski, S. J. F. Guimaraes, and Z. Patrocinio, Jr. Combining convolutional side-outputs for road image segmentation. presented at *2019 International Joint Conference on Neural Networks*, 2019. 7

[36] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 1, 3

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 6

[38] W. Shen, K. Zhao, Y. Jiang, Y. Wang, X. Bai, and A. Yuille. Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images. *IEEE Trans. Image Process.*, 26(11):5298–5311, 2017. 1

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[40] P. Sturgess, K. Alahari, L. Ladicky, and P. H. Torr. Combining appearance and structure from motion features for road scene understanding. In *Proc. British Mach. Vis. Conf.* BMVC, 2009. 1, 2

[41] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1013–1020. IEEE, 2018. 7

[42] Q. Wang, J. Gao, and Y. Yuan. Embedding structured contour and location prior in siamesed fully convolutional networks for road detection. *IEEE Trans. Intell. Transp. Syst.*, 19(1):230–241, 2018. 1, 3, 7, 8

[43] S. Yadav, S. Patra, C. Arora, and S. Banerjee. Deep cnn with color lines model for unmarked road segmentation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 585–589. IEEE, 2017. 1, 3, 8

[44] J. Zhang, Y. Xu, B. Ni, and Z. Duan. Geometric constrained joint lane segmentation and lane boundary detection. In *Proc. European Conf. Comput. Vis.*, pages 486–502, 2018. 1, 3, 7

[45] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1, 3

[46] F. Zohourian, B. Antic, J. Siegemund, M. Meuter, and J. Pauli. Superpixel-based road segmentation for real-time systems using cnn. In *VISIGRAPP (5: VISAPP)*, pages 257–265, 2018. 1, 3, 8