Robust Lane Detection via Expanded Self Attention

Minhyeok Lee Junhyeop Lee Dogyoon Lee Woojin Kim Sangwon Hwang
Sangyoun Lee*
Yonsei University School of Electrical and Electronic Engineering
{hydragon516,jun.lee,nemotio,woojinkim0207,sangwon1042,syleee}@yonsei.ac.kr

Abstract

The image-based lane detection algorithm is one of the key technologies in autonomous vehicles. Modern deep learning methods achieve high performance in lane detection, but it is still difficult to accurately detect lanes in challenging situations such as congested roads and extreme lighting conditions. To be robust on these challenging situations, it is important to extract global contextual information even from limited visual cues. In this paper, we propose a simple but powerful self-attention mechanism optimized for lane detection called the Expanded Self Attention (ESA) module. Inspired by the simple geometric structure of lanes, the proposed method predicts the confidence of a lane along the vertical and horizontal directions in an image. The prediction of the confidence enables estimating occluded locations by extracting global contextual information. ESA module can be easily implemented and applied to any encoder-decoder-based model without increasing the inference time. The performance of our method is evaluated on three popular lane detection benchmarks (TuSimple, CULane and BDD100K). We achieve state-of-the-art performance in CULane and BDD100K and distinct improvement on TuSimple dataset. The experimental results show that our approach is robust to occlusion and extreme lighting conditions.

1. Introduction

Advanced Driver Assistance Systems (ADAS), which are a key technology for autonomous driving, assists drivers in a variety of driving scenarios owing to deep learning. For ADAS, lane detection is an essential technology for vehicles to stably follow lanes. However, lane detection tasks, which rely on visual cues such as cameras, remain challenging owing to severe occlusions, extreme changes in the lighting conditions, and poor pavement conditions. Even in such difficult driving scenarios, humans can sensibly determine the positions of lanes by recognizing the positional relationship between the vehicles and surrounding environment. This remains a difficult task in image-based deep learning.

Baseline Ours

Figure 1: Compare our method with the baseline model. Our approach shows robustness in a variety of occlusion and low-light conditions.

The most widely used lane detection approach in image-based deep learning is segmentation-based lane detection [17, 18, 10, 7, 14, 15, 2, 16, 3]. These works learn in an end-to-end manner whether each pixel of the image represents the lane. However, it is very difficult to segment lane areas that are not visible by occlusion. To solve this problem, the network must capture the scene context with sparse supervision. Therefore, some works [18, 10] also introduce message passing or attention distillation. In [7], adversarial learning was applied to generate lanes similar to the real ones. These approaches can capture sparse supervision or sharpen blurry lanes. However, segmenting every pixel to detect lanes can be computationally inefficient.

To simplify the lane detection process and increase efficiency, some works [20, 27, 4] consider the problem of lane detection a relatively simple task and adopt the classification method. In [20], a very fast speed was achieved by dividing the image into a grid of a certain size and determining the position of the lane with row-wise classification. However, these methods do not represent lanes accurately, nor do they detect relatively large numbers of lanes.

To address the shortcomings of the semantic segmenta-

tion and classification methods described earlier, we propose a novel self-attention module called the Expanded Self Attention (ESA) module. Our modules are designed for segmentation-based lane detection and can be attached to any encoder-decoder-based model. Moreover, our method does not increase the inference time because the ESA module is removed in the testing phase. To make the model robust to occlusion and difficult lighting conditions, ESA module aims to extract important global contextual information by predicting the occluded location in the image. Inspired by the simple geometry of lanes, ESA modules are divided into HESA (Horizontal Expanded Self Attention) and VESA (Vertical Expanded Self Attention). HESA and VESA extract the location of the occlusion by predicting the confidence of the lane along the vertical and horizontal directions, respectively. Since we do not provide additional supervisory signals for occlusion, predicting occlusion location by the ESA module is a powerful help for the model to extract global contextual information. Details of the ESA module will be presented in Section 3.2.

Our method is tested on three popular datasets (TuSimple, CULane and BDD100K) containing a variety of challenging driving scenarios. Our approach achieves state-of-the-art performance in the CULane and BDD100K datasets, especially in CULane, surpassing the previous methods with a F1 score of 74.2. We confirm the effectiveness of the ESA module in various comparative experiments and demonstrate that our method is robust under occlusion and extreme lighting conditions. In particular, the results in Figure 1 show that our module shows impressive lane detection performance in various challenging driving scenarios.

Our main contributions can be summarized as follows:

- We propose a new Expanded Self Attention (ESA) module. The ESA module remarkably improves the segmentation-based lane detection performance by extracting global contextual information. Our module can be attached to any encoder-decoder-based model and does not increase inference time.

- Inspired by the simple lane geometry, we divide the ESA module into HESA and VESA. Each module extracts the occlusion position by predicting the lane confidence along the vertical and horizontal directions. This makes the model robust in challenging driving scenarios.

- The proposed network achieves state-of-the-art performance for the CULane [18] and BDD100K [28] datasets and outstanding performance gains under low-light conditions.

2. Related Work

Lane Detection. The use of deep learning for lane detection has been increasingly popular. Owing to the success of deep learning in the computer vision field, many studies have been proposed by adopting deep learning technique on lane detection for advanced driving assistant system, particularly for autonomous driving [17, 18, 10, 20, 27]. This approach performs better than hand-crafted methods [5, 22, 25, 13]. There are two main deep-learning-based approaches: 1) classification-based and 2) segmentation-based approaches.

The first approach considers lane detection a classification task [20, 27, 4]. Some works [20, 27] applied row-wise classification for the detection of lanes, thereby excluding unnecessary post-processing. In particular, [20] achieved high-speed performance by lightening the model. However, in the classification method, the performance depends on how many times the position of the lane is subdivided. In addition, it is difficult to determine the shape of the lane accurately.

Another approach to lane detection is to consider it a semantic segmentation task [17, 18, 10, 9, 14]. Neven *et al*. [17] performs instance segmentation by applying a clustering method to line mark segmentation. Moreover, Lee *et al*. [14] proposes multi-task learning that simultaneously performs grid regression, object detection, and multi-label classification guided by the vanishing point. Multi-task learning provide additional supervisory signals. However, the additional annotations required for multi-task learning are expensive. Pan *et al*. [18] applies a message passing mechanism between adjacent pixels. This method overcomes lane occlusion caused by vehicles and obstacles on the road and recognizes lanes in low-light environments. However, this message passing method requires considerable computational cost. To solve the slow speed of the method in [18], Hou *et al*. [10] proposes the Self Attention Distillation (SAD) module and achieve a significant improvement without additional supervision or labeling while maintaining the number of parameters in the model. However, in the SAD module, knowledge distillation is conducted from deep to shallow layers, which only enhances the inter-layer information flow for the lane area and does not provide an additional supervisory signal for occlusion. Our work is similar to [10], in that it uses the self-attention module. However, it adopts a new self-attention approach in a completely different way. To overcome occlusion problems, the proposed ESA module calculates the confidence of the lane that is deeply related to the occlusion. By using lane confidence, the model can reinforce the learning performance for these areas by providing a new supervisory signal for occlusion.

Self-attention. Self-attention has provided significant improvements in machine translation and natural language processing. Recently, self-attention mechanisms are used

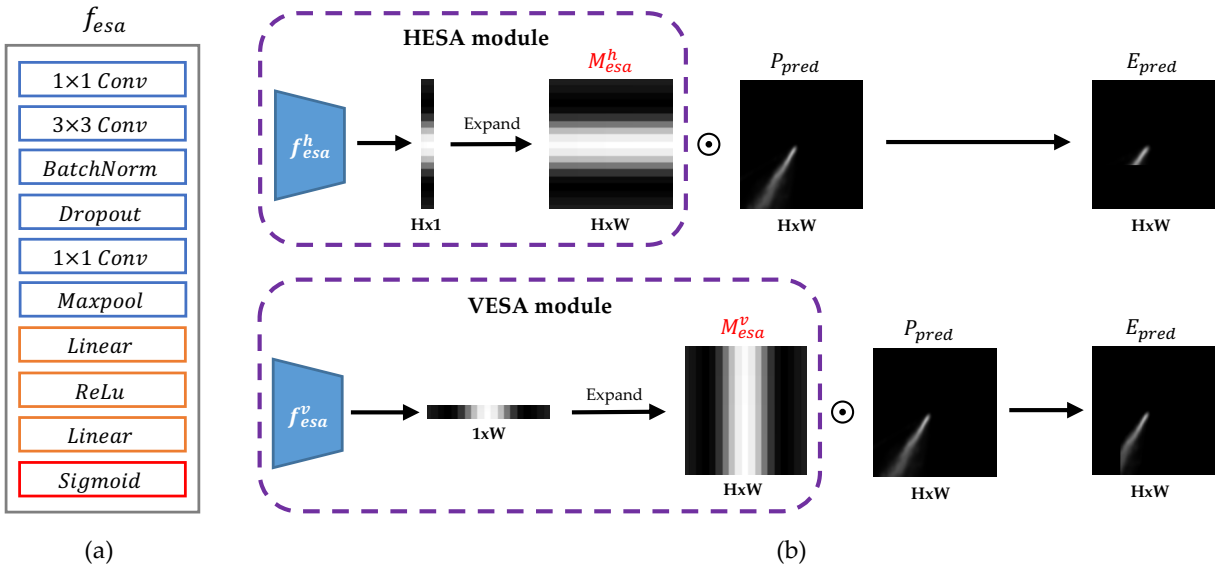


Figure 2: (a) Structure of ESA encoder f_{esa}. (b) Details of the Horizontal Expanded Self Attention (HESA) module (top) and Vertical Expanded Self Attention (VESA) module (bottom). The only difference between the two modules is the expansion direction of the ESA encoder output. Operator $\odot$ is defined as an element-wise product.

in various computer vision fields. The non-local block [24] learns the relationship between pixels at different locations. For instance, Zhang *et al*. [29] introduces a better image generator with non-local operations, and Fu *et al*. [6] improves the semantic segmentation performance using two types of non-local blocks. In addition, self-attention can emphasize important spatial information of feature maps. [19, 26] showed meaningful performance improvement in classification by adding channel attention and spatial attention mechanisms to the model.

The proposed ESA module operates in a different way than the previously presented module. The ESA module extracts the global context of congested roads to predict areas with high lane uncertainty and to emphasize those lanes.

3. Proposed Approach

3.1. Overview

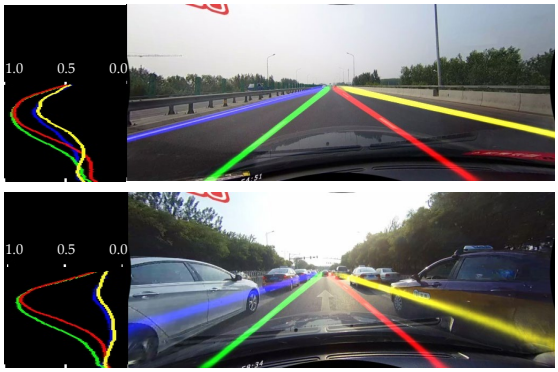
Unlike general semantic segmentation, lane segmentation conducts segmentation by predicting the area in which the lane is covered by objects. Therefore, lane segmentation tasks must extract global contextual information and consider the relationship between distant pixels. In fact, self-attention modules with non-local operation [24] can be an appropriate solution. Several works [30, 6, 11] prove that non-local operations are effective in semantic partitioning where global contextual information is important. However, in contrast to the complex shape in general semantic segmentation, the lane has a relatively simple geometric shape in lane segmentation. This makes non-local operations inefficient.

If the network can extract occluded locations, lanes that are invisible owing to occlusions are easier to segment. The location information of occlusions becomes more important than their shape owing to the simple lane shape. Therefore, rather than extracting the high-level occlusion shape, it is more effective to extract the low-level occlusion position. By using this positional information, the ESA module can extract the column or row-wise confidence of lanes by itself. The confidence indicates that the model knows the location of the occlusion based on the global contextual information of the scene.

3.2. Expanded Self Attention

The ESA module aims to extract global contextual information by recognizing the occluded area. The structure of the ESA module is inspired by the fact that the lane is a line that spreads from the vanishing point. Due to the simple shape of the lane, it is efficient to predict the confidence along the vertical or horizontal direction of the lane in order to estimate the location of the occlusion. Therefore, we divide the ESA module into HESA and VESA according to the direction to extract the lane confidence. Furthermore, all ESA modules are only used in the training phase and are removed in the testing phase. Therefore, our method has the same inference time and number of parameters as the baseline model.

Figure 2 shows two types of ESA modules, HESA and VESA. Both modules have an ESA encoder f_{esa} consisting of convolution layers and fully connected layers. The ESA



(a) (b)

Figure 3: (a) Graph of ESA encoder f_{esa}^{h} output and (b) predicted lane probability map. The output of the ESA encoder represents the lane confidence of each row. The graph and lane are matched with the same color, and the graph shows only the area in which the lane exists.

encoders of the HESA and VESA modules are defined as f_{esa}^{h} and f_{esa}^{v}, respectively. The only difference between the two encoders is the length of the output vector. For the HESA modules, the output shape of f_{esa}^{h} is $\mathbb{R}^{C\times H\times 1}$, where C is the maximum number of lanes, and H is the height of the image. This output will be expanded horizontally and will be equal to the original image size. More specifically, as shown in Figure 2 (b) of the paper, this output is duplicated with size W in the horizontal direction, where W is the width of the input image. The expanded matrix is ESA matrix, $M_{esa}^{h} \in \mathbb{R}^{C\times H\times W}$. It should be noted that each row of M_{esa}^{h} has the same element value, as shown in Figure 2 (b). Similarly, regarding the VESA module, the output of f_{esa}^{v} of size $\mathbb{R}^{C\times 1\times W}$ is vertically expanded to ensure that the ESA matrix is $M_{esa}^{v} \in \mathbb{R}^{C\times H\times W}$, where W is the width of the image. Therefore, as illustrated in Figure 2 (b), each column of M_{esa}^{v} has the same value. The ESA matrix has a value between 0 and 1 owing to the sigmoid layer of f_{esa} and highlights a part of the predicted probability map via the element-wise product between the predicted probability map and ESA matrix. If the predicted probability map is $P_{pred} \in \mathbb{R}^{(C+1)\times H\times W}$, the weighted probability map E_{pred} is formulated as $E_{pred} = P_{pred} \odot M_{esa}^{h}$ for the HESA module and $E_{pred} = P_{pred} \odot M_{esa}^{v}$ for the VESA moduel, where the operator $\odot$ describes an a element-wise product. The reason that the number of channels in P_{pred} is $C+1$ is that C lane classes and one background class are included in the dataset. Therefore, element-wise product is performed only on lane channels except for a background channel, and the size of E_{pred} is $C \times H \times W$.

The most important role of the ESA module is extracting lane confidence. Figure 3 presents the predicted probability map of the model and output of the ESA encoder f_{esa}^{h}. The colors in the graph match the colors of the lane. The output of f_{esa}^{h} is identical to the height of the image. However, in Figure 3, only the location in which the lane exists is presented as a graph. If there is no occlusion on the road as shown in the first figure in Figure 3, the output of f_{esa}^{h} is overall high. If occlusion occurs, such as the blue and yellow lanes in the second figure, the measured f_{esa}^{h} value of the occluded area is small. This is how the ESA module measures the confidence of the lane. If the visual cues for the lane are abundant, the lane confidence at the location increases, and a great weight is output. Conversely, if there are few visual cues, the lane confidence decreases and a small weight is output.

3.3. Network Architecture

Our network architecture is illustrated in Figure 4. Our neural network starts with the baseline model, which consists of encoder and decoder. In this paper, since inference time is an important factor in lane detection, lightweight baseline models such as ResNet-18 [8], ResNet-34 [8], and ERFNet [21] are used. Inspired by the works [10, 15], we add the existence branch to the baseline model. Existence branch is designed for datasets in which lanes are classified according to their relative position, such as TuSimple and CULane. In the case of BDD100K, existence branch is not used because we consider all lanes as one class. We extract a total of four feature maps from the baseline model encoder. These feature maps are resized and concatenated to become input to the ESA module. We will discuss in detail how the ESA module output, baseline model output, and ground truth labels interact with each other in Section 3.4.

3.4. Objective Functions

Segmentation and existence loss. First we reduce the difference between the predicted lane segmentation map S_{pred} and the ground truth segmentation map S_{gt}. The segmentation loss $\mathcal{L}_{seg}$ is used as follows:

$$\mathcal{L}_{seg} = \mathcal{L}_{CE}\left(S_{pred}, S_{gt}\right), \tag{1}$$

where $\mathcal{L}_{CE}$ is the standard cross entropy loss. We apply cross entropy loss to C lane classes and one background class. In addition, the existence loss is proposed for the TuSimple and CULane datasets because lanes are classified by their relative positions. The existence loss $\mathcal{L}_{exist}$ is formulated as follows:

$$\mathcal{L}_{exist} = \mathcal{L}_{BCE}\left(l_{pred}, l_{gt}\right), \tag{2}$$

where $\mathcal{L}_{BCE}$ is the binary cross entropy loss, l_{gt} is a lane existence label, and l_{pred} is an output of the lane existence branch.

ESA loss. The ESA module aims to predict the confidence of the lane by recognizing occlusion with global contextual

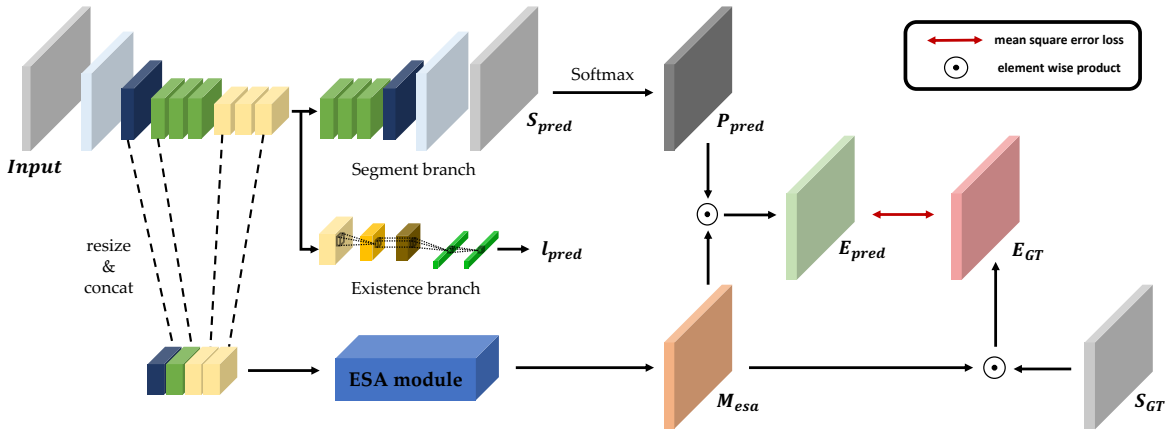


Figure 4: The neural network architecture. The model is a combination of the existence branch and ESA module in the baseline model. The existence branch outputs the probability of existence of each lane and the ESA module generates an ESA matrix.

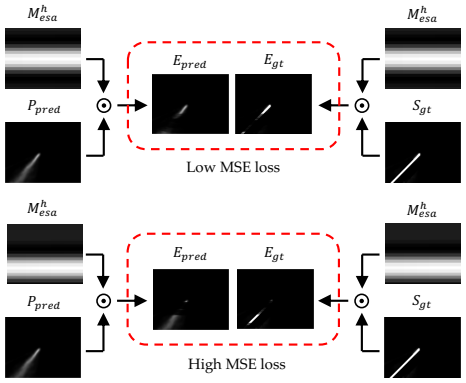


Figure 5: Comparison of low (top) and high (bottom) loss. The mean square error is determined according to the location in which the ESA matrix is active.

information. However, creating an annotation for the location information of the occlusion is time-consuming and expensive, and the consistency of the annotation cannot be guaranteed. Therefore, our module learns the occlusion location without additional annotations by reducing the mean square error between the weighted probability map E_{pred} and the weighted ground truth segmentation map E_{gt}. Figure 5 presents this process.

The predicted probability map of the lane is $P_{pred} = \Phi(S_{pred})$, where $\Phi(.)$ is the softmax operator. In addition, the ESA loss $\mathcal{L}_{esa}$ is formulated as follows:

$$\begin{aligned}\mathcal{L}_{esa} &= \mathcal{L}_{MSE}(E_{pred}, E_{gt}) \\ &+ \lambda \left| \Psi(E_{pred}) - \Upsilon \Psi(S_{gt}) \right|,\end{aligned} \tag{3}$$

where the ESA matrix is M_{esa}, the weighted probability map $E_{pred} = P_{pred} \odot M_{esa}$, the weighted ground truth map $E_{gt} = S_{gt} \odot M_{esa}$, and $\mathcal{L}_{MSE}$ is the mean square error loss. Moreover, the operator $\Psi(.)$ calculates the average of all values of the feature map, and λ is a regularization coefficient. The coefficient Υ has an important effect on the performance of the model, and it determines the proportion of the weighted lane area. The first term on the right-hand side of Equation (3) is visualized in Figure 5. In general, the lane probability map is blurred in areas with sparse supervisory signals. As shown in Figure 5, if a large weight is given to the accurately predicted region in the probability map, the mean square error is small. Conversely, when a large weight is given to an uncertainly predicted area, the mean square error is large. This is how to predict the confidence of the lane without additional annotations.

In fact, if only the mean square error loss is used as the ESA loss, the ESA module outputs are all zeros in the training. To solve this problem, a second term is added as a regularizer to the right-hand side of Equation (3). This regularization term keeps the average pixel value of the weighted probability map equal to a certain percentage of the average pixel value of the ground truth map. This ratio is determined by Υ, which has a value between 0 and 1.

It should be noted that although one ESA module is an HESA or a VESA module, both modules can be simultaneously attached to the model. In that case, the ESA loss is $\mathcal{L}_{esa} = \mathcal{L}^h_{esa} + \mathcal{L}^v_{esa}$, where $\mathcal{L}^h_{esa}$ is the ESA loss of the HESA module, and $\mathcal{L}^v_{esa}$ is the ESA loss of the VESA module. Finally, the above losses are combined to form the final objective function:

$$\mathcal{L} = \alpha \mathcal{L}_{seg} + \beta \mathcal{L}_{exist} + \gamma \mathcal{L}_{esa}. \tag{4}$$

The parameters α, β and γ balance the segmentation loss, existence loss, and ESA loss of the final objective function.

Category	**R-34-H&VESA**	**ERFNet-H&VESA**	SCNN [18]	ENet-SAD [10]	R-34-Ultra [20]	ERFNet [21]	ERFNet-E2E [27]
Normal	90.5	**92.0**	90.6	90.1	90.7	91.5	91.0
Crowded	68.3	**73.1**	69.7	68.8	70.2	71.6	**73.1**
Night	65.7	**69.5**	66.1	66.0	66.7	67.1	67.9
No line	42.2	45.8	43.4	41.6	44.4	45.1	**46.6**
Shadow	65.1	**75.1**	66.9	65.9	69.3	71.3	74.1
Arrow	85.4	**88.1**	84.1	84.0	85.7	87.2	85.8
Dazzle light	59.8	63.1	58.5	60.2	59.5	**66.0**	64.5
Curve	61.5	68.8	64.4	65.7	69.5	66.3	**71.9**
Crossroad	**1675**	2001	1990	1998	2037	2199	2022
Total	70.9	**74.2**	71.6	70.8	72.3	73.1	74.0
Runtime (ms)	**4.7**	8.1	133.5	13.4	5.7	8.1	-
Parameter (M)	2.44	2.68	20.72	**0.98**	-	2.68	-

Table 1: Comparison of F1-measures and runtimes for CULane test set. Only the FP is shown for crossroad. "R-" denotes "ResNet" and same abbreviation is used in the following sections.

4. Experiments

Datasets. We use three popular lane detection datasets TuSimple [23], CULane [18], and BDD100K [28] for our experiments. TuSimple datasets consist of images of highways with constant illumination and good weather, and are relatively simple datasets because the roads are not congested. Therefore, various algorithms [18, 17, 7, 10, 12] have been tested on TuSimple datasets since before 2018. CULane is a very challenging dataset that contains crowded environments with city roads and highways with varying lighting conditions. The CULane dataset and TuSimple dataset are officially labeled with up to four lanes and one background excluding lanes. These datasets focus on the detection of four lane markings, which are paid most attention to in real applications. The BDD100K dataset also consists of images captured under various lighting and weather conditions. In addition, the largest number of lanes among the three datasets is labeled. However, because the number of lanes is large and inconsistent, we detect lanes without distinguishing instances of lanes.

Evaluation metrics.

1) TuSimple. In accordance with [23], the accuracy is expressed as $Accuracy = \frac{N_{pred}}{N_{gt}}$, where N_{pred} is the number of predicted correct lane points and N_{gt} is the number of ground truth lane points. Furthermore, false positives (FP) and false negatives (FN) in the evaluation index.

2) CULane. In accordance with the evaluation metric in [18], each lane is considered 30 pixel thick, and the intersection-over-union (IoU) between the ground truth and prediction is calculated. Predictions with IoUs greater than 0.5 are considered true positives (TP). In addition, the F1-measure is used as an evaluation metric and is defined as follows:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{5}$$

where $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$.

Algorithm	Accuracy	FP	FN	Runtime (ms)
ResNet-18 [8]	92.69%	0.0948	0.0822	**3.4**
ResNet-34 [8]	92.84%	0.0918	0.0796	4.7
LaneNet [17]	96.38%	0.0780	0.0244	19.0
EL-GAN [7]	96.39%	0.0412	0.0336	-
ENet-SAD [10]	**96.64%**	0.0602	0.0205	13.4
SCNN [18]	96.53%	0.0617	**0.0180**	133.5
R-18-H&VESA	95.70%	0.0588	0.0622	**3.4**
R-34-H&VESA	95.83%	0.0587	0.0599	4.7
ERFNet-HESA	96.01%	**0.0329**	0.0458	8.1
ERFNet-VESA	95.94%	0.0340	0.0451	8.1
ERFNet-H&VESA	96.12%	0.0331	0.0450	8.1

Table 2: Comparison of performance results of different algorithms applied to TuSimple test set.

Algorithm	Accuracy	IoU	Runtime (ms)
ResNet-18 [8]	54.59%	44.63	**2.7**
ResNet-34 [8]	56.62%	46.00	4.1
ERFNet [21]	55.36%	47.04	7.3
ENet-SAD [10]	57.01%	47.72	12.1
SCNN [18]	56.83%	47.34	123.6
R-18-H&VESA	57.03%	46.50	**2.7**
R-34-H&VESA	59.93%	49.51	4.1
ERFNet-HESA	57.47%	48.97	7.3
ERFNet-VESA	57.51%	48.24	7.3
ERFNet-H&VESA	**60.24%**	**51.77**	7.3

Table 3: Comparison of results for BDD100K test set.

3) BDD100K. In general, since there are more than 8 lanes in an image, following [10], we determine the pixel accuracy and IoU of the lane as evaluation metrics.

We used different evaluation method for fair comparisons with previous studies. We evaluated with the same method as [7, 10, 17, 18] for TuSimple and [10, 18, 20, 27] for CULane.

Implementation details. Following [18, 10], we resize the images of TuSimple, CULane, and BDD100K to 368×640, 288×800, and 360×640, respectively. The original BDD100K images label one lane with two lines. Because this labeling method is difficult to learn, so we drew new 8 pixel thick ground truth labels that pass through the center of the lane. The new ground truth labels are applied equally

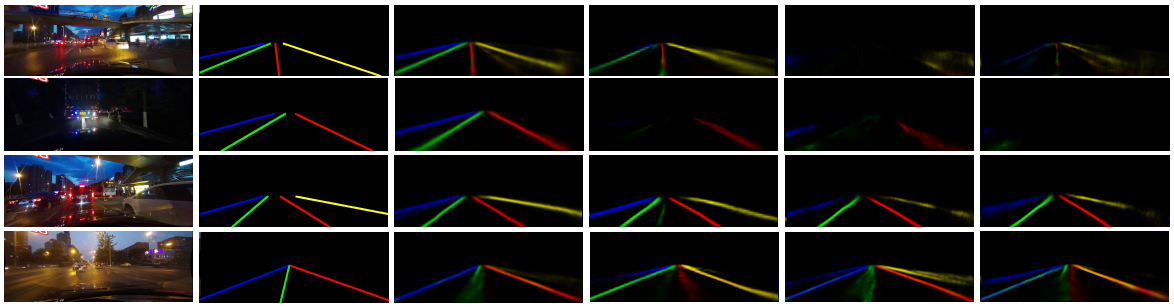


Figure 6: Comparison of the output probability maps of different algorithms applied to CULane test set. The third column is the result of the proposed ERFNet-HESA. The probability maps of the four lane classes are displayed in blue, green, red, and yellow, respectively.

to both train and test sets. Moreover, SGD [1] is used as the optimizer, and the initial learning rate and batch size are set to 0.1 and 12, respectively. The loss balance coefficients α, β, and γ in Equation (4) are set to 1, 0.1, and 50, respectively. The regularization coefficient λ in Equation (3) is 1. It is experimentally verified whether the value of the coefficient Υ in Equation (3) has a significant effect on the performance of the model. In CULane and BDD100K, the optimal Υ value is set to 0.8, and TuSimple is set to 0.9. The effect of Υ on the performance is discussed in detail in Section 4.2. Because the BDD100K experiment regards all lanes as one class, the output of the original segmentation branch is replaced with a binary segmentation map. In addition, the lane existence branch is removed for the evaluation. All models are trained and tested with PyTorch and the Nvidia RTX 2080Ti GPU.

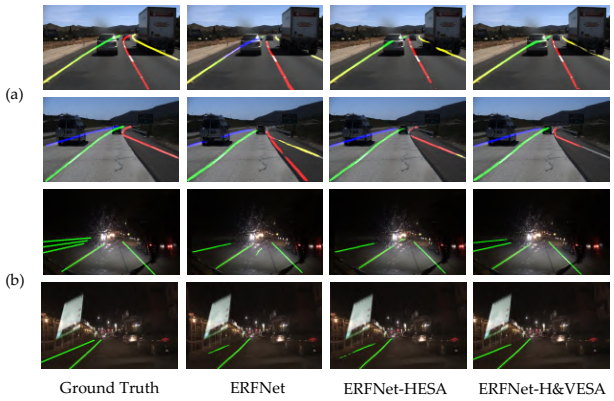


Figure 7: Performance of different algorithms for (a) TuSimple and (b) BDD100K test sets.

4.1. Results

Tables 1-3 compare the performance results of the proposed method and previously presented state-of-the-art algorithms for CULane, TuSimple, and BDD100K datasets. The proposed method is evaluated with the baseline models ResNet-18 [8], ResNet-34 [8], and ERFNet [21], and each model is combined with either an HESA or a VESA. Moreover, the use of both HESA and VESA modules is denoted as "H&VESA". The effects of using both modules simultaneously are presented in Section 4.2.

The combination of the baseline model ERFNet and ESA module outdoes the performance of the ERFNet and achieves state-of-the-art performance for CULane and BDD100K. In particular, ERFNet-H&VESA provides significant performance gains for almost all driving scenarios in the CULane dataset compared to ERFNet. However, the runtime and number of parameters remain unchanged. In addition, ERFNet-H&VESA surpasses the existing methods by achieving an F1-measure of 69.5 in the challenging low-light environment in the lane detection with the CULane dataset. It has a fast runtime similar to those of the previous state-of-the-art methods in Table 1. Thus, the proposed method is much more efficient than the previously proposed methods. As shown in Table 3, compared to ERFNet, ERFNet-HESA increases accuracy from 55.36% to 57.47% with the BDD100K dataset. In addition, ERFNet-H&VESA achieves the highest accuracy of 60.24%. These results show that the HESA and VESA modules work complementarily. The regarding details are covered in Section 4.2. The results of the TuSimple dataset in Table 2 show the effect of the ESA module, but it does not achieve the highest performance. The TuSimple dataset contains images of highways with bright light, and generally less occlusion. Because the ESA module extracts global contextual information by predicting the occluded location, our method is less effective for datasets with less occlusion. Unlike our method, ENet-SAD [10] provides additional supervision signals to lane areas and SCNN [18] applies a message passing mechanism between adjacent pixels in visible lanes. EL-GAN [7] uses adversarial learning to capture sparse supervision from visible lanes or sharpen blurry lanes. Therefore, these methods are effective in

Baseline	HESA	VESA	TuSimple			CULane	BDD100K	
			Accuracy	FP	FN	F1 (total)	Accuracy	IoU
ResNet-18 [8]			92.69%	0.0948	0.0822	67.8	54.59%	44.63
	✓		**95.73%**	0.0590	0.0643	70.4	56.68%	46.11
		✓	95.70%	0.0598	**0.0615**	70.3	56.70%	46.08
	✓	✓	95.70%	**0.0588**	0.0622	**70.7**	**57.03%**	**46.50**
ResNet-34 [8]			92.84%	0.0918	0.0796	68.4	56.62%	46.00
	✓		95.68%	0.0584	0.0634	70.7	58.36%	47.29
		✓	95.70%	**0.0533**	0.0681	70.7	58.11%	47.30
	✓	✓	**95.83%**	0.0587	**0.0599**	**70.9**	**59.93%**	**49.51**
ERFNet [21]			94.90%	0.0379	0.0563	73.1	55.36%	47.04
	✓		96.01%	**0.0329**	0.0458	**74.2**	57.47%	48.97
		✓	95.94%	0.0340	0.0451	74.1	57.51%	48.24
	✓	✓	**96.12%**	0.0331	**0.0450**	**74.2**	**60.24%**	**51.77**

Table 4: Performance comparison of various combinations of HESA and VESA modules with TuSimple, CULane, and BDD100K test sets

datasets with less occlusion.

We provide qualitative results of our algorithm for various driving scenarios in three benchmarks. In particular, the first and second rows of Figure 6 show that our method can detect sharp lanes even under extreme lighting conditions and in situations in which the lanes are barely visible owing to other vehicles. Figure 7 (a) shows that the ESA module can connect the lanes occluded by vehicles without interruption. According to Figure 7 (b), the approach achieves more accurate lane detection in low-light environments. Thus, compared to the baseline model, the ESA module can improve performance in challenging driving scenarios with extreme occlusion and lighting conditions.

4.2. Ablation Study

Combination of HESA and VESA. Table 4 summarizes the performance characteristics of different combinations of HESA and VESA. The following observations can be made. (1) The performance characteristics of the HESA and VESA modules are similar. (2) In general, the performance of H&VESA with HESA and VESA modules applied simultaneously is better. In addition, H&VESA results in a remarkable performance improvement for BDD100K. The reason why the HESA and VESA modules lead to similar performance characteristics is that the predicted direction of the lane confidence is not important for extracting the low-level occlusion location because the lane has a simple geometric shape. Because the HESA and VESA modules complement each other to extract more abundant global contextual information, it is not surprising that H&VESA generally achieves the highest performance. Therefore, global contextual information is more important for the BDD100K dataset, which includes many lanes.

Value of Υ. Figure 8 compares the total F1-score of the CULane validation set with respect to Υ in Equation (3). As shown in Figure 8, the model shows the best performance at $\Upsilon = 0.8$ in ERFNet-HESA. It is important to find a suitable

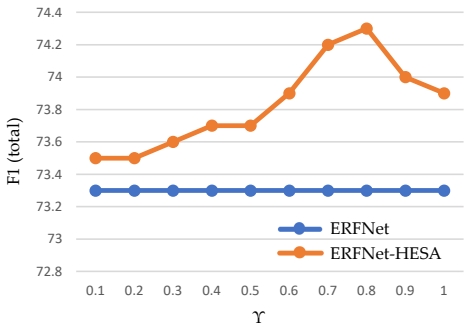


Figure 8: Comparison of performance characteristics with respect to Υ for the CULane validation set.

Υ value because it determines the ratio of occluded and normal areas. When the Υ is small (*i.e.*, when the predicted occlusion area is wide), the sensitivity to occlusion decreases, which makes it difficult to determine the occluded location accurately. Conversely, when Υ is large, the detected occlusion area becomes narrow, which makes it difficult for the network to reinforce learning for the entire occluded area.

5. Conclusion

This paper proposes ESA module, a novel self-attention module for robust lane detection in occluded and low-light environments. The ESA module extracts global contextual information by predicting the confidence of the lane. The performance of the model is evaluated on the datasets containing a variety of challenging driving scenarios. According to the results, our method outperforms previous methods. We confirm the effectiveness of the ESA module in various comparative experiments and demonstrate that our method is robust in challenging driving scenarios.

Acknowledgement. This research was supported by Multi-Ministry Collaborative R&D Program(R&D program for complex cognitive technology) through the National Research Foundation of Korea(NRF) funded by MSIT, MOTIE, KNPA(NRF-2018M3E3A1057289).

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[2] Donghoon Chang, Vinjohn Chirakkal, Shubham Goswami, Munawar Hasan, Taekwon Jung, Jinkeon Kang, Seok-Cheol Kee, Dongkyu Lee, and Ajit Pratap Singh. Multi-lane detection using instance segmentation and attentive voting. In *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, pages 1538–1542. IEEE, 2019.

[3] Ping-Rong Chen, Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, and Jing-Jhih Lin. Efficient road lane marking detection with deep learning. In *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2018.

[4] Shriyash Chougule, Nora Koznek, Asad Ismail, Ganesh Adam, Vikram Narayan, and Matthias Schulze. Reliable multilane detection and classification by utilizing cnn as a regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[5] Wang Y Teoh EK and D Sben. Lane detection and tracking using b-snake, image and vision computer. *Image and Vision computing*, 22:269–280, 2004.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[7] Mohsen Ghafoorian, Cedric Nugteren, Nóra Baka, Olaf Booij, and Michael Hofmann. El-gan: Embedding loss driven generative adversarial networks for lane detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12486–12495, 2020.

[10] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1013–1021, 2019.

[11] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 603–612, 2019.

[12] Seokwoo Jung, Sungha Choi, Mohammad Azam Khan, and Jaegul Choo. Towards lightweight lane detection by optimizing spatial embedding. *arXiv preprint arXiv:2008.08311*, 2020.

[13] ZuWhan Kim. Robust lane detection and tracking in challenging scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 9(1):16–26, 2008.

[14] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, and In So Kweon. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1947–1955, 2017.

[15] Tong Liu, Zhaowei Chen, Yi Yang, Zehao Wu, and Haowei Li. Lane detection in low-light conditions using an efficient data enhancement: Light conditions style transfer. *arXiv preprint arXiv:2002.01177*, 2020.

[16] Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, and Jing-Jhih Lin. Multi-class lane semantic segmentation using efficient convolutional networks. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019.

[17] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 286–291. IEEE, 2018.

[18] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. *arXiv preprint arXiv:1712.06080*, 2017.

[19] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[20] Zequn Qin, Huanyu Wang, and Xi Li. Ultra fast structure-aware deep lane detection. *arXiv preprint arXiv:2004.11757*, 2020.

[21] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017.

[22] Tsung-Ying Sun, Shang-Jeng Tsai, and Vincent Chan. Hsi color model based lane-marking detection. In *2006 IEEE Intelligent Transportation Systems Conference*, pages 1168–1172. IEEE, 2006.

[23] TuSimple. http://benchmark.tusimple.ai/#/t/1. Accessed: 2018-09-08.

[24] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[25] Yue Wang, Dinggang Shen, and Eam Khwang Teoh. Lane detection using spline model. *Pattern Recognition Letters*, 21(8):677–689, 2000.

[26] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[27] Seungwoo Yoo, Hee Seok Lee, Heesoo Myeong, Sungrack Yun, Hyoungwoo Park, Janghoon Cho, and Duck Hoon Kim. End-to-end lane marker detection via row-wise classification. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition Workshops, pages 1006–1007, 2020.

[28] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[29] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.

[30] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 593–602, 2019.