Importance-Aware Semantic Segmentation with Efficient Pyramidal Context Network for Navigational Assistant Systems

Kaite Xiang, Kaiwei Wang* and Kailun Yang

***Abstract*— Semantic Segmentation (SS) is a task to assign semantic label to each pixel of the images, which is of immense significance for autonomous vehicles, robotics and assisted navigation of vulnerable road users. It is obvious that in different application scenarios, different objects possess hierarchical importance and safety-relevance, but conventional loss functions like cross entropy have not taken the different levels of importance of diverse traffic elements into consideration. To address this dilemma, we leverage and re-design an importance-aware loss function, throwing insightful hints on how importance of semantics are assigned for real-world applications. To customize semantic segmentation networks for different navigational tasks, we extend ERF-PSPNet, a real-time segmenter designed for wearable device aiding visually impaired pedestrians, and propose BiERF-PSPNet, which can yield high-quality segmentation maps with finer spatial details exceptionally suitable for autonomous vehicles. A comprehensive variety of experiments with these efficient pyramidal context networks on CamVid and Cityscapes datasets demonstrates the effectiveness of our proposal to support diverse navigational assistant systems.**

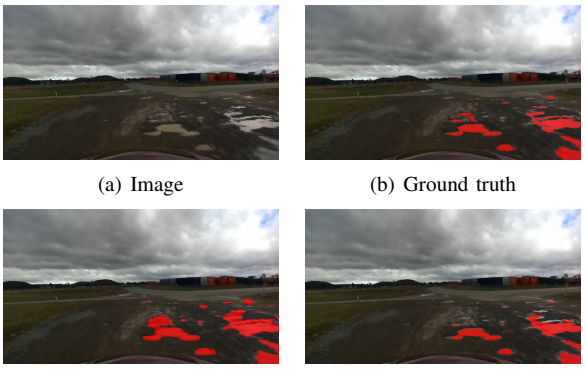
I. INTRODUCTION

Semantic Segmentation (SS) is a task to assign semantic labels to each pixel of the images, which is of crucial significance for autonomous vehicles, robotics and navigation assistance systems for vulnerable road users like visually impaired pedestrians, where safety is critical [1].

In recent years, with the development of deep learning, SS has come into the stage based on deep convolutional neural networks (CNNs) since the milestone created by Fully Convolutional Networks (FCN) [2]. The performance of FCN is surpassed by subsequent PSPNet [3] and DeepLab [4], which can perform semantic segmentation with high accuracies and huge numbers of parameters. Inevitably, the complex calculation keeps SS from being put into practice for real-time applications in devices with limited computation resources. In previous works, we propose ERF-PSPNet [5][6], a real-time SS network especially designed for navigation assistance systems supporting the visually impaired, which largely sacrifices the resolution and accuracy of edges extraction, resulting in coarse segmentation maps. However, applications like autonomous vehicles and driving assistance require high-resolution semantic maps and highly accurate road boundary segmentation. To address this problem, we extend ERF-PSPNet with novel efficient pyramidal context network design by proposing BiERF-PSPNet, which can yield high-quality semantic segmentation maps with finer spatial details, while maintaining real-time inference.

This work has been partially funded through the project "Research on Vision Sensor Technology Fusing Multidimensional Parameters" (111303-I21805) by Hangzhou SurImage Technology Co., Ltd and supported by Hangzhou KrVision Technology Co., Ltd (krvision.cn).

Kaite Xiang, Kaiwei Wang and Kailun Yang are with State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou, China {katexiang, wangkaiwei, elnino}@zju.edu.cn



(a) Image (b) Ground truth (c) Cross-entropy Loss (d) Focal Loss

Fig. 1. Effect of different loss functions: (c) Output of ERF-PSPNet trained with cross-entropy loss, (d) Output of ERF-PSPNet trained with focal loss. It could observed that the model trained with cross-entropy loss has a higher recall rate and lower precision than cross focal loss.

TABLE I
FOCAL LOSS V.S CROSS-ENTROPY LOSS

Loss	Recall rate	Precision	IoU
Focal loss	0.80	**0.85**	**0.70**
Cross-entropy loss	**0.94**	0.63	0.60

In addition to the architecture design, we made key observation that loss function is the key element in the training procedure and greatly influences the SS output. Traditional cross-entropy loss function is broadly used for SS, but the loss function only pays attention to the frequencies of objects by applying weights for different classes with different frequencies. For another thing, the focal-loss [7] that is designed for difficulty-aware object detection task has also been utilized in existing researches to train SS networks. Evidently, the model trained with different loss functions will have a vast difference between recall rate and precision. Before our project, we conduct an experiment on a water puddle segmentation dataset [8]. We find that the model trained with focal loss possesses a higher precision while the model trained with cross-entropy loss has a higher recall rate, as shown in Fig. 1 and Table I. We argue that in some safety-critical application scenarios of autonomous vehicles, recall rate plays a more important role than precision for traffic objects like cars, buses, and pedestrians, as we need to focus on the detection of them, which need to be detected with high recall rate. In other words, it is preferred to detect it

wrongly rather than miss it, because these traffic objects will be dangerous if the algorithm misses them and predict them as safe roadways. In addition, hierarchical importance should be emphasized for different objects for autonomous vehicles. Taking it for granted, roadways and sidewalks are more important than sky and buildings, while cars and pedestrians are even more important and safety-critical than those flat classes.

Therefore, existing methodologies like focal loss and cross-entropy loss are not ideally suitable for SS associated with autonomous driving system. As explained above, an autonomous driving system needs to focus on some important objects for driving rather than segment all classes with the same level of importance. In this paper, inspired by [9], we adapt and re-design an importance-aware loss function (IAL), and perform a comprehensive set of experiments to prove its effectiveness and wide applicability.

The contributions of the paper lie in four key aspects:

- We adapt a real-time SS architecture named ERF-PSPNet, and extend the model into a bilateral architecture BiERF-PSPNet to recover better spatial details.
- We adapt and re-design an importance-aware loss function, and improve its stability and reliability.
- A series of experiments are conducted on two autonomous driving benchmark, *i.e.*, CamVid [10] and Cityscapes [11], which demonstrate the structure of the refined model can recover better spatial information and the effectiveness of the adapted IAL.
- We perform a systematic analysis of the experiment results, throwing insightful hints on how importance is assigned for real-world SS frameworks. Our implementations and codes are available at: `https://github.com/Katexiang/ERF-PSPNET`

II. Related work

A. Semantic Segmentation Neural Networks

Since the milestone created by FCN, SS has gain tremendous advances based on CNNs. The ConvNets first transfer known classification networks into SS by making them fully convolutional. Immediately following the success, UNet [12], DeepLab [4] and many other SS Networks were proposed. Many of them have achieved state-of-the-art performance on different benchmarks of SS task. Their normal procedure involves encoding more spatial information or enlarging the receptive filed at the expenses of huge operations and multiple parameters. Therefore, they normally perform inference at a low speed so that they can not be applied for real-time application like autonomous vehicles.

In order to put the SS networks into practice, many light-weighted real-time SS networks were proposed. ENet [13] is one of the first networks in pursuit of real-time inference, which is modified from ResNet structure [14] to perform SS with much fewer parameters. ERFNet [15] [16] and our previous ERF-PSPNet [5][6] utilize residual factorized module to reduce parameters and keep fine performance.

At the same time, some SS networks with multi-path structure were put forward to refine the spatial details of the output. ICNet [17] is one of the pioneer with multi-path structure, which uses multiple-size input image at shallow layers to get spatial information, while inputting small image to deep layers to extract semantic information. BiSeNet [18] works in a different way, which divides the network structure into two paths, one for spatial information to refine output, and another for excavating context information. ContextNet [19] combines a deep network branch at low resolution capturing global context efficiently with a shallow branch focusing on high-resolution segmentation details to reach competitive performance.

B. Somewhat-Aware Method for training CNNS

With the development of deep learning, many training methods are advanced to solve somewhat-aware problems like difficult-aware [20] and attribute-aware [21] SS. Li et al. [20] considered that different pixels own different ranks of difficulty and propose a difficulty-aware network to pay attention to more difficult pixels. At the same time, focal loss [7] acts as a loss function to cope with the detection of difficult objects. Inspired by attention mechanism, Chen et al. [22] designed a SS network to emphasize objects with different scales. Bulo et al. [23] introduced a novel loss max-pooling concept for handing imbalanced training data distributions. Following it, Importance-Aware-Loss (IAL) created by Chen et al. [9] was leveraged to distinguish important pixels from normal pixels. But IAL is unstable, sometimes the effect is remarkable, and sometimes it is unserviceable. To alleviate the shortcomings, in this paper we re-design and adapt to a more stable IAL, and prove the effectiveness.

As is shown above, among the various existing notions of SS networks, they will be coming into use in the near future. Besides, the somewhat-aware method can be exploited to cope with certain application problems in somewhat-bias tasks like importance, scale, difficulty and so on. Based on these observations, this paper aims to cope with the problem that different objects own different levels of importance in autonomous driving systems.

III. Methodology

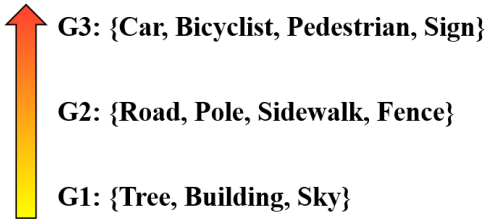
In this section, we firstly detail the modified version of Importance-Aware-Loss (IAL), and then illustrate our real-time SS networks, ERF-PSPNet and its extended version, BiERF-PSPNet.

A. Importance-aware Loss Function

IAL proposed by Bi et al. [9] is a modified version of entropy-cross loss function in practice. Making a brief introduction to traditional entropy-cross loss function $\mathbf{I}$ which is defined by:

$$\mathbf{I} = -\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{q}_{i,j} \cdot log(\mathbf{p}_{i,j}) \tag{1}$$

where $\mathbf{q}_{i,j}$ and $\mathbf{p}_{i,j}$ are the one-hot encoding label and output at i-th row and j-th column, both of which have the shape of (1,C) (C: the number of classes). When training the model,



(a)

0	1	1	1
0	1	1	1
0	1	1	1
0	1	1	1

(b) $\mathbf{M}_1$

X	0	0	1
X	0	0	1
X	0	0	1
X	0	0	1

(c) $\mathbf{M}_2$

Fig. 2. (a) shows the rankings of importance of CamVid classes, G3 is the most important group. (b) and (c) are the importance matrices, (b) illustrates the $\mathbf{M}_1$, (c) illustrates the $\mathbf{M}_2$, where the blue area belongs to G1, the yellow area belongs to G2, and the green area belongs to G3.

$\mathbf{I}$ needs to be divided by $\mathbf{H}$ and $\mathbf{W}$, which represents height and width of the image respectively. However, the model trained with the loss function may segment certain pixels into classes that occupies most pixels of the image. Therefore, weight $\omega_{i,j}$ can be employed in the loss function to enable the model to pay attention to the rare classes. It is defined by:

$$\omega_{i,j} = \frac{1}{ln(\mathbf{a} + \mathbf{f}_{i,j})} \tag{2}$$

where $\mathbf{a}$ is a hyper-parameter avoiding divided by zero, in this paper, we set it to 1.02. And $\mathbf{f}_{i,j}$ is the pixel sum of the class at i-th row and j-th column divided by the number of input image's pixels. Therefore, $\mathbf{I}$ can be modified into

$$\mathbf{I} = -\sum_{i=1}^{H}\sum_{j=1}^{W} \omega_{i,j} \cdot \mathbf{q}_{i,j} \cdot log(\mathbf{p}_{i,j}) \tag{3}$$

In order to enable models to focus on certain important objects, we need to exert dynamic weight for loss function. Taking CamVid as an example, the dataset has 11 classes, *i.e.*, sky, building, pole, road, sidewalk, tree, sign, fence, car, pedestrian and bicyclist, which will be detailed in Section IV. First, we categorize the classes into three importance groups as a hierarchical structure like Fig. 2. For autonomous driving systems, the traffic objects like cars, pedestrians are the most important, while the road or sidewalks are less important, the sky and buildings away from the passable area are the least important. Then we divide the $\mathbf{I}$ into three parts, *i.e.*, $\mathbf{I}_1$, $\mathbf{I}_2$ and $\mathbf{I}_3$, each of which stands for the certain part of the cross-entropy loss belonging to certain group. Following the rationale, here comes to importance matrix, $\mathbf{M}_i$. If we have three ranks of importance, we should construct two importance matrices, *i.e.*, $\mathbf{M}_1$ and $\mathbf{M}_2$ as shown in Fig. 2. Taking $\mathbf{M}_2$ as an example, for three ranks of importance category, the most important classes are assigned to 1 at both matrices, while the classes of middle rank are signed to 1 at $\mathbf{M}_1$ and 0 at $\mathbf{M}_2$, and the least important classes are assigned to 0 at $\mathbf{M}_1$ and $\mathbf{X}$ at $\mathbf{M}_2$ ($\mathbf{X}$ is a number either 0 or 1). They are the key elements to dynamically assign importance weights for the loss function.

Afterwards, we need to utilize the matrices to construct the dynamic importance weights. The dynamic weight of a group f_t is defined as

$$\frac{\sum\sum[(\mathbf{M}_{t,i,j} + \lambda)^{0.5} \cdot (\mathbf{p}_{c,i,j} - \mathbf{M}_{t,i,j}) \cdot (\mathbf{M}_{t,i,j} \neq \mathbf{X})]^2}{N_t} \tag{4}$$

where f_t (t can be chosen as 2 or 3 in three-rank importance system as the G1's importance weight is 0) is the dynamic importance weight; λ is a tuning parameter set to 0.5 in order to take the lower-importance category into consideration and avoid ignoring them when calculating the dynamic importance weight. $\mathbf{M}_{t,i,j}$ is the value of the importance matrix, while $\mathbf{p}_{c,i,j}$ is the ground-truth channel value of the output at i-th row and j-th column; $\mathbf{p}_{c,i,j}$ is the key element of the weight pushing the loss function focusing on important category. And the value of ($\mathbf{M}_{t,i,j} \neq \mathbf{X}$) is 0 if the value of the matrix is $\mathbf{X}$ else the value is 1. N_t is a normalization factor, which is the pixel sum of the full image when t is 2 and the pixel sum of G2 and G3 when t is 3.

At present, the proposed loss function can be defined by

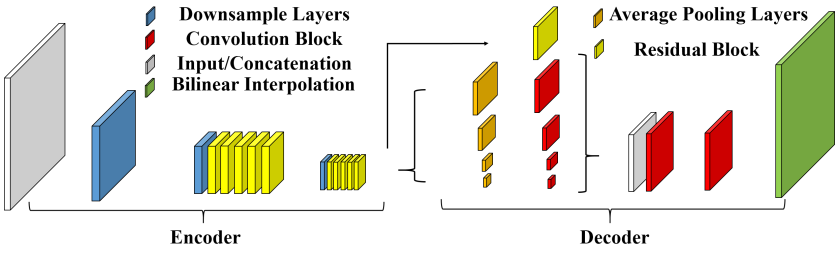
$$\mathbf{IAL} = \mathbf{I}_1 + (f_1 + \alpha) \cdot \mathbf{I}_2 + (f_2 + \alpha) \cdot (f_3 + \alpha) \cdot \mathbf{I}_3 \tag{5}$$

where $\mathbf{IAL}$ is the ultimate loss function, α is a tuning parameter being set to 1 in our experiment.

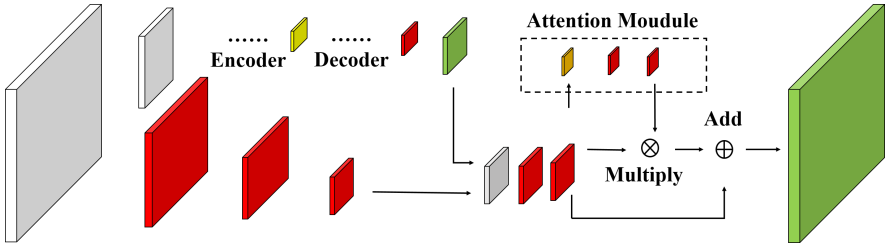
B. Architecture

In view of the trade-off between efficiency and accuracy, we select an efficient pyramidal context network, *i.e.*, ERF-PSPNet [5] as our base net. As is shown in Fig. 3(a), the model follows a typical encoder-decoder architecture. The model is a rational combination of efficient residual factorized network (ERFNet) and pyramid scene parsing network (PSPNet). The encoder originates from ERFNet, which utilizes a sequential architecture to produce down-sampled feature maps. The encoder first utilize the "down-sampler" block as detailed in [13] to down-sample the feature map quickly in order to reduce computation costs. The highlight of the encoder is "Non-bottleneck-1D" as detailed in [15] enabling an efficient utilization of minimized amount of residual layers to extract effective feature maps and achieve high efficiency. Following the pyramid pooling module modified from PSPNet, the decoder is designed to harvest contextual information among feature maps of varied sizes and attain larger receptive field. After that, the feature maps are bilinearly interpolated and cascaded to form the final feature representation. Following the concatenation layer, we append a convolution layer to re-weight the feature representation. In the end, we append a 1×1 kernel classification convolution layer, bilinear interpolation layer and softmax layer to output the final result.

The SS network is especially designed for assisted navigation of the visually impaired. Therefore, the segmentation



(a) ERF-PSPNet



(b) BiERF-PSPNet

Fig. 3. The architecture of our semantic segmentation networks: (a) ERF-PSPNet and (b) BiERF-PSPNet.

at the semantics boundaries (*e.g.*, road boundaries) remain coarse, because the decoder attains better contextual information at the sacrifice of texture and spatial information, which are less desired by visually impaired pedestrains, but these details are important for autonomous driving. Inspired by bilateral BiseNet and ContextNet, we advance ERF-PSPNet by proposing an important variant of the SS network with a different style. Making a brief introduction of BiSeNet, it is composed of spatial path and context path. The context path is a feature extractor to attain deep semantic information at low computation cost. In contrast, the spatial path is a concatenation of several convolution layers as shallow layers to extract spatial information fed with a high-resolution image. Therefore, BiSeNet's output is not only of high precision but also possesses fine textures. Therefore, another efficient pyramidal context network, *i.e.*, modified ERF-PSPNet is proposed concerning the textures of the output, named BiERF-PSPNet as shown in Fig. 3(b). Different from BiSeNet, our proposed model possesses two input images, the smaller one of which serves as the context information source going through deep neural network to attain semantic information, and the larger one of which is used for extracting spatial information to refine the results. In the end, feature maps of the two paths are fused by an attention module as shown in Fig. 3(b). In our experiments, $\mathbf{H}_1$ and $\mathbf{W}_1$ are 1024 and 2048 respectively. $\mathbf{H}$ and $\mathbf{W}$ are 360 and 720 respectively.

IV. Experiments

In order to verify our model and IAL, we conduct a series of experiments on dataset, *i.e.*, CamVid and Cityscapes.

A. Datasets

$CamVid$: The dataset is a street scene dataset from a driving vehicle's perspective containing 701 images of 720 $\times$ 960 and involving 11 semantic categories, 367 of which belong to its training set and 233 of which belong to the validation set.

$Cityscapes$: The dataset is a street scene dataset from the perspective of an intelligent car, which contains 2975 fine annotated images for training and another 500 images for validation. The resolution of the images is 1024 $\times$ 2048, we select 19 pre-defined classes for training and validation.

B. Implementation Details

We use Tensorflow and a NVDIA GeForce GTX 1080Ti GPU for training and validation. Due to the limited memory we set batch size to 8 when training ERF-PSPNet and 5 when training BiERF-PSPNet. ERF-PSPNet's mIoU can reach 59.7 at a highly efficient resolution of 360 $\times$ 720, but it can be enhanced to 64.5 when adopting data augmentations. In order to facilitate fair comparison, we abandon data augmentations for our experiments inspite of the effect for advancing performance and robustness of the models. The models are trained for 300 epochs (for both datasets) with Adam optimization algorithm. The initial learning rate is 0.001 divided by 10 every 100 epochs. For the sake of combatting overfitting, we use the L2 weight regularization with decay of 0.0002.

C. Quantitative Results

We conduct three groups of experiments, one of which is conducted on CamVid, and others are conducted on Cityscapes.

For CamVid, we select the ranks of importance as depicted in Fig. 2(a). From the results of Table II and Table V, we observe that the recall rates of G3 have been advanced especially the classes like sign, car and pedestrian, and the mean recall rate has been improved by around 1 point. In view of the different categories between CamVid and Cityscapes,

TABLE II
CROSS-ENTROPY LOSS V.S IAL ON CAMVID BY ERF-PSPNET (%).
G3,G2 AND G1 ARE THE GROUPS OF THE CLASSES. G3 IS THE MOST IMPORTANT GROUP AND G1 IS THE LEAST IMPORTANT GROUP.

Group		G3				G2				G1		
Class		Sign	Car	Pedestrian	Bicyclist	Pole	Road	Sidewalk	Fence	Sky	Building	Tree
Cross-Entropy	Precision	33.0	86.2	47.8	69.4	35.9	93.8	83.3	39.9	94.5	85.6	76.7
	Recall rate	34.0	79.9	63.7	47.2	42.6	97.1	81.9	25.3	94.2	82.3	80.5
	IoU	20.1	70.9	37.6	39.1	24.2	91.2	70.3	18.3	89.3	72.3	64.7
IAL	Precision	31.0	83.1	39.3	65.6	31.3	**94.1**	81.1	34.1	**94.8**	**86.2**	76.6
	Recall rate	**47.0**	**83.4**	**69.1**	42.7	40.6	96.4	**83.3**	**27.3**	93.0	79.5	78.7
	IoU	**23.0**	**71.3**	33.5	34.9	21.5	90.9	69.7	17.8	88.5	70.6	63.4

TABLE III
CROSS-ENTROPY LOSS V.S IAL ON CITYSCAPES OF G3 BY BIERF-PSPNET (%)

Class		Traffic Light	Sign	Rider	Truck	Bus	Train	Motorcycle	Bicycle
Cross-Entropy	Precision	72.8	81.8	61.3	73.1	69.6	64.5	46.5	77.7
	Recall Rate	62.0	73.4	50.3	62.2	72.2	22.4	34.3	77.0
	IoU	50.4	63.1	38.2	50.6	54.9	19.9	24.6	63.0
IAL	Precision	63.5	73.6	**61.4**	72.6	**70.0**	**78.7**	**51.6**	76.2
	Recall rate	**68.9**	**78.4**	47.6	61.5	**76.7**	**46.0**	34.1	**79.7**
	IoU	49.4	61.2	36.6	49.9	**57.7**	**40.8**	**25.8**	**63.8**

TABLE IV
CROSS-ENTROPY LOSS V.S IAL ON CITYSCAPES OF G2 AND G1 BY BIERF-PSPNET (%)

Group		G2					G1					
Class		Car	Sidewalk	Fence	Pole	Pedestrian	Road	Building	Wall	Vegetation	Terrain	Sky
Cross-Entropy	Precision	94.1	84.1	65.4	70.0	71.4	98.7	92.9	63.6	94.3	73.9	94.2
	Recall rate	95.4	88.5	47.2	65.4	87.7	98.1	94.0	41.5	94.7	68.3	97.6
	IoU	90.0	75.8	37.7	51.0	64.9	96.8	87.7	33.6	89.6	55.0	92.1
IAL	Precision	93.0	81.1	63.7	65.1	69.4	**98.9**	**93.4**	**64.4**	**94.4**	70.0	**94.9**
	Recall rate	**96.1**	**89.6**	**47.5**	**67.0**	**88.6**	97.5	93.0	**45.8**	93.6	65.4	97.1
	IoU	89.5	74.2	37.4	49.3	63.7	96.5	87.3	**36.5**	88.7	51.0	**92.3**

TABLE V
CROSS-ENTROPY LOSS V.S IAL ON CAMVID AND CITYSCAPES OF MEAN GROUPS(%)

Dataset		Camvid by ERF-PSPNet				Cityscapes by BiERF-PSPNet			
Group		3	2	1	Mean	3	2	1	Mean
Cross-Entropy	Precision	59.1	63.2	85.6	67.8	68.4	77.0	86.3	76.3
	Recall rate	56.2	61.7	85.7	66.2	56.7	76.9	82.4	70.1
	IoU	41.9	51.0	75.4	54.3	45.6	63.9	75.8	60.0
IAL	Precision	54.8	60.2	**85.9**	65.2	68.4	74.5	86.0	75.6
	Recall rate	**60.6**	**61.9**	83.7	**67.4**	**61.6**	**77.8**	82.17	**72.3**
	IoU	40.7	50.0	74.2	53.2	**48.2**	62.8	75.4	**60.6**

some categories' precision and recall rate are quite high for Cityscapes like pedestrians, cars and road. On the other hand, the extra categories which do not belong to CamVid are more important than them. Therefore, some categories need to be attributed into different importance rank or else it may lead to detrimental effect for training. For Cityscapes, we regard traffic light, sign, rider, truck, bus, train, motorcycle, bicycle as the most important classes, car, sidewalk, fence, pole, pedestrian as the second important classes, and road, building, wall, vegetation, terrain, sky as the least important classes. We conduct a series of experiments by using ERF-PSPNet and BiERF-PSPNet, both of them demonstrate the effectiveness of IAL. Taking BiERF-PSPNet's results as an example, the results are filled in Table III to Table V. From the results, what we can learn is that the IAL elevates the important classes' recall rates dramatically with few negative effect on the precision.

The speed of ERF-PSPNet is 74.1fps when inputting a 360 $\times$ 720 image on a GTX 1080Ti GPU and BiERF-PSPNet is 42.1fps. Their mIoU are 59.7 and 60.7 on Cityscapes validation set, respectively. In other words, the BiERF-PSPNet refine the spatial information by making a sacrifice for inference time, while still keeping above real-time inference.

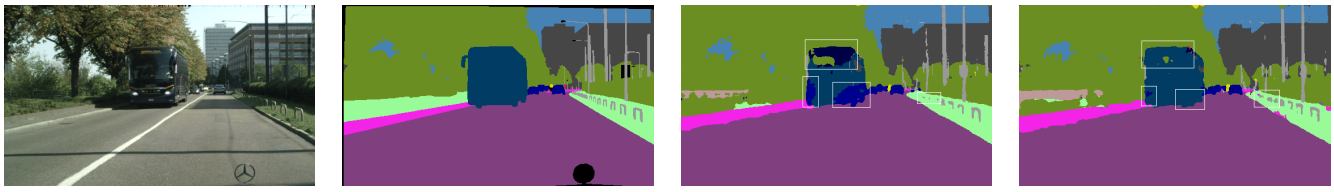
However, surprisingly, a by-product of IAL, attracts our interests, which is the promotion of the G1's precision on both datasets as displayed in Table II, Table IV and Table V. In other words, when categorizing the classes into three importance parts, although the original purpose is to advance the recall rate of G3, the precision of G1 has been advanced and even has a slight improvement on mIoU by accident, which is of practical significance for autonomous vehicles and other navigational assistant systems. We have emphasized the importance of recall rate for autonomous vehicles in Section I, but precision is another key point. In comparison, regarding navigation assistance for the visually impaired, we may underline the segmentation of sidewalks, which should be segmented with high precision, because the system must guarantee the visually impaired people navigate on safe sidewalks, in case the road is detected as sidewalks

(a) Image (b) Label (c) Result of ERF-PSPNet (coarse) (d) Result of BiERF-PSPNet (fine)

Fig. 4. The result comparison between ERF-PSPNet and BiERF-PSPNet.

(a) Image (b) Label (c) Cross-entropy Loss (d) IAL

Fig. 5. The result comparison between Cross-entropy loss and IAL in CamVid.



(a) Image (b) Label (c) Cross-entropy Loss (d) IAL

Fig. 6. The result comparison between Cross-entropy loss and IAL in Cityscapes.

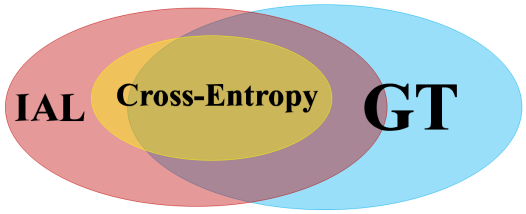


Fig. 7. Graphical illustration of the effect of IAL for G3. The blue area is ground truth, the yellow are is the output of model trained by cross-entropy loss, and the red area is the result of the model trained by IAL.

which will be dangerous for them.

D. Qualitative Analysis

The effect of BiERF-PSPNet can be shown in Fig. 4. We find that the edge of objects is more accurate and refined because of the spatial path of the BiERF-PSPNet, especially at the edges of the pedestrians, riders, and some small and slender objects like telegraph poles, which are very important for safety-critical autonomous driving, as it is required to perceive pedestrians and poles at long distances, in order to take fast decisions in response to environmental events.

We find that IAL is of great effect in Fig. 5 and Fig. 6. In Fig. 5 from CamVid, we find the output of the cross-entropy loss ignores some pedestrians at further places and the segmentation of the truck is fragmented bounded by white boxes. But in the IAL's output, the fragmentation of the truck is refined, while the smaller pedestrians can be detected. Fig. 6 shows a representative example from Cityscapes, where it is obvious that the IAL is highly effective successfully to segment the bus and detect most of the poles while the cross-entropy loss's model segments part of the bus into car, part of the bus into truck. Moreover, it misses some poles as well. Therefore, we find our advanced IAL is not only effective in CamVid but also practical in challenging large-scale Cityscapes as well.

V. Conclusions and Future Work

SS is promising for many tasks especially navigational assistant systems like autonomous driving. As different objects possess different ranks of importance, the SS network for autonomous driving should be addressed by a different method, *i.e.*, our revised IAL can yet be regarded as a powerful technique. This paper re-design IAL so that the

loss function can make the training model focus on important classes and advance the classes' recall rate which will impose dynamic weights adaptively. In addition, we adapt ERF-PSPNet into BiERF-PSPNet for the sake of a finer spatial result, while maintaining above real-time inference.

As a saying goes, every rose has its thorn. The revised IAL can advance the recall rate of the important classes and precision of the least important classes without damaging the performance of the model, but it decreases the precision of the important classes which may category other objects into the important classes. Therefore, we summarize the performance of IAL as the Venn Diagram shown in Fig. 7. For G3, IAL has a higher recall rate and lower precision than cross-entropy loss, while keeping competitive mIoU. In other words, training with IAL yields models which would rather segment the important objects wrongly, than miss them for safety considerations. In the future, we aim to further optimize IAL and attempt to utilize new decision rules [24] to let model to improve certain categories recall rate and certain categories' precision.

REFERENCES

[1] K. Yang, X. Hu, L. M. Bergasa, E. Romera, X. Huang, D. Sun, and K. Wang, "Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 374–381.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3431–3440.

[3] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6230–6239.

[4] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[5] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen, and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1033–1038.

[6] K. Yang, L. M. Bergasa, E. Romera, D. Sun, K. Wang, and R. Barea, "Semantic perception of curbs beyond traversability for real-world navigation assistance systems," in *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*. IEEE, 2018, pp. 1–7.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2999–3007.

[8] X. Han, C. Nguyen, S. You, and J. Lu, "Single image water hazard detection using fcn with reflection attention units," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 105–120.

[9] B. Chen, C. Gong, and J. Yang, "Importance-aware semantic segmentation for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, no. 99, pp. 1–12, 2018.

[10] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *European conference on computer vision*. Springer, 2008, pp. 44–57.

[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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 3213–3223.

[12] 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*. Springer, 2015, pp. 234–241.

[13] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778.

[15] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2018.

[16] E. Romera, L. M. Bergasa, K. Yang, J. M. Alvarez, and R. Barea, "Bridging the day and night domain gap for semantic segmentation," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1184–1190.

[17] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 405–420.

[18] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.

[19] R. P. Poudel, U. Bonde, S. Liwicki, and C. Zach, "Contextnet: Exploring context and detail for semantic segmentation in real-time," *arXiv preprint arXiv:1805.04554*, 2018.

[20] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6459–6468.

[21] M. Sulistiyo, Y. Kawanishi, D. Deguchi, T. Hirayama, I. Ide, J. Zheng, and H. Mutase, "Attribute-aware semantic segmentation of road scenes for understanding pedestrian orientations," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 2698–2703.

[22] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 3640–3649.

[23] S. R. Bulo, G. Neuhold, and P. Kontschieder, "Loss max-pooling for semantic image segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 7082–7091.

[24] R. Chan, M. Rottmann, F. Hüger, P. Schlicht, and H. Gottschalk, "Application of decision rules for handling class imbalance in semantic segmentation," *arXiv preprint arXiv:1901.08394*, 2019.