2018 21st International Conference on Intelligent Transportation Systems (ITSC)
Maui, Hawaii, USA, November 4-7, 2018

Attribute-aware Semantic Segmentation of Road Scenes for Understanding Pedestrian Orientations

M. D. Sulistiyo[1,2], Y. Kawanishi[1], D. Deguchi[3], T. Hirayama[4], I. Ide[1], J. Y. Zheng[1,5], H. Murase[1]

Abstract— **Semantic segmentation is an interesting task for many deep learning researchers for scene understanding. However, recognizing details about objects' attributes can be more informative and also helpful for a better scene understanding in intelligent vehicle use cases. This paper introduces a method for simultaneous semantic segmentation and pedestrian attributes recognition. A modified dataset built on top of the Cityscapes dataset is created by adding attribute classes corresponding to pedestrian orientation attributes. The proposed method extends the SegNet model and is trained by using both the original and the attribute-enriched datasets. Based on an experiment, the proposed attribute-aware semantic segmentation approach shows the ability to slightly improve the performance on the Cityscapes dataset, which is capable of expanding its classes in this case through additional data training.**

I. Introduction

In recent years, computer vision technologies by deep learning have become popular for various complex tasks. One of the most interesting topics is the semantic segmentation, which is a more complex task compared to *simple* image segmentation. Semantic segmentation is a pixel-wise labeling task to simultaneously classify pixels in the input image into predefined classes. Semantic segmentation with deep-learning approaches have been widely utilized for the purposes such as indoor and outdoor scene understanding as well as applications like autonomous vehicle sensing [1] and robotics navigation [2]. For instance, from a road scene, a model trained for semantic segmentation will inform an autonomous vehicle of surrounding objects before it acts.

Many studies on this task have been conducted so far. A Fully Convolutional Network was introduced for the semantic segmentation task as an end-to-end trained model [3]. Almost in the same period, SegNet with a different network architecture was also published [1]. It was then modified into Bayesian SegNet and obtained better performance by adding a probabilistic element [4]. Later on, the Pyramid Scene Parsing network (PSPNet) [5] outperformed other existing methods and achieved a state-of-the-art performance for PASCAL VOC 2012 [6] and Cityscapes datasets [7]. The Mask R-CNN presented a conceptual framework for instance-level segmentation, which combined object detection and semantic segmentation tasks. The approach outperformed some previous methods for several challenges [8]. Reference [9] provides a summary of popular deep neural networks that were adopted to solve the semantic segmentation.

However, the existing semantic segmentation methods only recognize an object's name as its class, e.g., road, building, car, person, and so on, while for autonomous vehicles, additional information describing a particular object in detail such as its attributes could help the better understanding of scenes. Therefore, this paper proposes an attribute-aware semantic segmentation method, which is a more difficult challenge. Yet we can obtain better evaluation results if the process is successfully performed. With simultaneous recognition and segmentation tasks, it can enhance environment perception tasks in Intelligent Transportation Systems that need precision such as parking, road following, pedestrian detection, action identification, and localization.

Here we chose pedestrian as the target class, out of other classes like vehicle and traffic signs, as it has many attributes to be explored. Since a pedestrian is one of the most important moving objects in a street scene, it should be useful for assisting an autonomous vehicle. In this paper, we particularly consider the body orientation of a pedestrian as the attribute and divide it into a predefined number of classes. In the autonomous vehicle system, being informed of pedestrian orientations can be helpful in collision avoidance. Although *only* the pedestrian body orientation was added as a new semantic information, it can give hint and insight on other object attributes when the method is extended.

The contributions of this paper are as follows:

1) We introduce a new concept of attribute-aware semantic segmentation that is general and possible to be applied to various tasks.
2) We design an attributes-dependent loss function that is modified from the original one for increasing the number of classes regarding the attribute-aware concept.
3) We modify the annotation of Cityscapes dataset for the attribute-aware semantic segmentation task.
4) By extending the Cityscapes to include additional attribute classes, we achieved better performance of semantic segmentation.

The rest of this paper will discuss related work in Sec. II, our proposed idea and its implementation in Sec. III, and our modified dataset annotation in Sec. IV. Experimental results and discussions are presented in Sec. V followed with a conclusion.

[1] Mahmud Dwi Sulistiyo, Yasutomo Kawanishi, Ichiro Ide, Hiroshi Murase are with Graduate School of Informatics, Nagoya University, Japan {kawanishi, ide, murase} @i.nagoya-u.ac.jp

[2] Mahmud Dwi Sulistiyo is also with School of Computing, Telkom University, Indonesia mahmuddwis@telkomuniversity.ac.id

[3] Daisuke Deguchi is with Information Strategy Office, Nagoya University, Japan ddeguchi@nagoya-u.jp

[4] Takatsugu Hirayama is with Institutes of Innovation for Future Society, Nagoya University, Japan takatsugu.hirayama@nagoya-u.jp

[5] Jiang Yu Zheng is with Dept. of Computer Science, Indiana University-Purdue University Indianapolis, USA jzheng@iupui.edu

978-1-7281-0322-8/18/$31.00 ©2018 IEEE

II. Related Works

Many methods in deep learning are trying to solve the semantic segmentation task [1], [3]–[5], [8]. Among them, SegNet was motivated mainly by road-scene understanding. It is a deep neural network architecture consisting of convolutional encoder-decoder layers, as shown in Figure 1.

Meanwhile, studies related to the pedestrian attribute recognition task have also been widely conducted. References [10] and [11] introduced a new dataset and presented a benchmark performance by an SVM-based method for pedestrian attribute recognition in far-view surveillance scenes and proposed an alternative approach for an improved attribute inference. To improve real-time applications, the Richly Annotated Pedestrian (RAP) dataset was proposed in reference [12]. It reported that attributes such as viewpoints, occlusions, and body parts information could help recognize pedestrian attributes in real applications.

Some related datasets with annotated images are publicly available for the challenges of semantic segmentation with various environments and lighting conditions. A number of datasets are used in reference [9] for a review on deep-learning techniques. Among them, Cityscapes is a benchmark suite and a large-scale dataset to train and test methods for pixel-level and instance-level semantic labeling. The dataset was highly motivated by the need for semantic urban-scene understanding applications. Images were collected to cover various street conditions in a number of different cities. It was initially published as an ongoing project in reference [13] and finally completed in reference [7]. Many studies are addressed and challenged using this dataset for comparison to other existing datasets. It is the largest and the most diverse dataset of street scenes with high-quality and coarse annotations [7].

For stand-alone objects such as bicycles, riders, and traffic signs, the network can be trained to pick up their unique texture and local shape as the features for their identification. These features through filtering at different layers are still very stable to the final layer. For extending regions such as road, tree, cars, and sidewalk, the network propagates more position and shape information through a linear combination of coefficients and max-pooling mechanism in the network according to annotated samples in the learning process. This "propagation" of segment id stops at strong edges but may fail to line out a uniformed region if there is no clear edge cue; the segmented region is thus noisy.

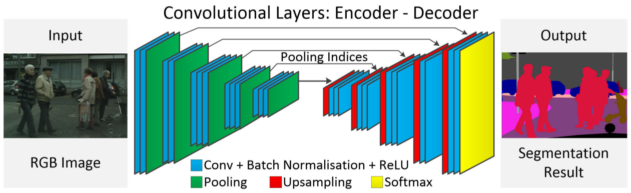

Fig. 1. SegNet's network architecture [1].

III. Attribute-Aware SegNet

A. Introducing Attributes-grouping Loss Function

We propose an attribute-aware semantic segmentation concept to improve the semantic segmentation task by simultaneously obtaining per-pixel class label as well as the attribute information of a particular object. Basically, it extends the classification from an object class to its sub-classes. For example, in addition to classifying vegetation, car, and pedestrian, the classifier should also distinguish between types of vegetation, types of car, and even between male and female pedestrians. We assume that some particular classes have disjoint sub-classes divided according to some attributes. We introduce this primary concept by modifying the neural network's structure and loss function so as to enrich the object understanding by their attributes.

In the current work, we use SegNet as the baseline model since it is a low-cost design in terms of memory usage and computational time during inference [1], while maintaining better performance than several comparable methods. Regarding the loss function, there are several forms for existing deep neural network models; and particularly, SegNet uses a cross-entropy loss function in its Softmax layer to produce the output [1], which is defined as follows:

$$L(y, y') = -\sum_i y_i \log y'_i, \tag{1}$$

where y is the actual output of the network, and y' is the target output given in the dataset. N is the number of training samples, so y_i and y'_i are the actual and the target outputs of the i-th sample.

Considering that the attributed classes are actually derived from an object class, we extend the class loss function by adding a loss calculation unit for its attribute. The modified loss function can be written as follows:

$$L'(y, y') = L_c + L_a, \tag{2}$$

$$L_c = -\beta_c \sum_i y_{i_c} \log y'_{i_c}, \tag{3}$$

$$L_a = -\beta_a \sum_i y_{i_a} \log y'_{i_a}, \tag{4}$$

$$\beta_c + \beta_a = 1. \tag{5}$$

The class loss and attribute loss functions are denoted as L_c and L_a, respectively. We furthermore introduce parameters β_c and β_a as the *weights* for both class and attribute losses, respectively. They are constant numbers ranging from 0 to 1 to determine the influence level of each loss score for the combined loss function score. For classes with no attribute, β_c and β_a are set to be 1 and 0, respectively. However, determining those values between 0 and 1 is another issue to optimized in the future.

B. Implementation of the Proposed Method

Weighting the class loss as well as the attribute loss with appropriate values will efficiently solve this task. In addition, setting different values for β_c and β_a will also affect the network implementation, i.e. the network output units might be constructed in some different levels. On the other hand, assigning the parameters β_c with 0 and β_a with 1 implies that in the network implementation, there will be additional sub-classes in the same level of the output layer. This means the attribute-aware semantic segmentation task will be represented in a so-called 'flat classifier' instead of a hierarchical one [19].

For example, let's consider a set of classes $C = \{X, Y, Z\}$ and let class X have a set of subclasses $A_X = \{x, y, z\}$ divided by attribute A; while classes Y and Z do not have any attribute. Note that $|C|$ is the number of classes, while $|A_X|$ is the number of sub-classes determined by attributes. Thus, the number of output units in the final connected layer will be $(|C|+|A_X|-1)$. In this sample case, there will be five output units. By the proposed method, the network will be trained using target classes of $C' = C \cup A_X$. In the inference phase, the output of the trained model can be either $|C'|$ number of classes, consisting of object and attribute classes, in a flat model, or just $|C|$ by integrating the attribute classes into one object class in a hierarchical model. In the sampling case, the first model output will be $\{x, y, z, Y, Z\}$ and the second alternative output will be $\{X, Y, Z\}$ after combining $\{x, y, z\}$ into class X. Figure 2 depicts this proposed concept for the sample case compared to the basic concept.

In addition to the structure of the network, typically for the semantic segmentation, it is necessary to weight the losses differently between classes, since the sample numbers of classes (road, sidewalk, car, pedestrian, etc.) usually have a large variance [1] in a training set. Therefore, class balancing by calculating the loss weight for each class is needed, including the attribute classes since they are treated as object classes as well. The idea is to simply put more *attention* or weight on to the classes with pixels appearing infrequently, and give a smaller weight to a class that has a high pixel frequency. We use a median frequency balancing strategy by computing the pixel frequency of each class in the training set. The balancing factor on each class, denoted by α_c, is obtained from the median of all class frequencies divided by its own class frequency [14]. Here, α_c is defined as follows:

$$\alpha_c = \frac{\text{median\_freq}}{\text{freq}_c} \tag{6}$$

where freq_c is the number of pixels in class c divided by the total number of pixels in images where class c is present, while median_freq is the median value of all freq_c. This formula implies that the more frequent a pixel class occurs in the training set, the less loss weight it gains (the value is less than 1).

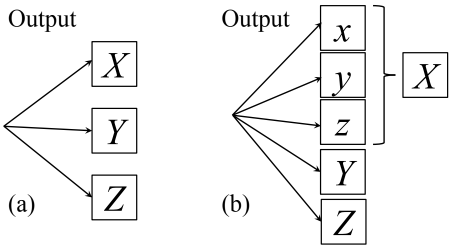

Fig. 2. Network structure comparison of approaches. (a) Basic output units for object level semantic segmentation; (b) Proposed concept for attribute-aware semantic segmentation.

IV. Attribute-enriched Cityscapes Dataset

A. Cityscapes Dataset and CityPersons

The Cityscapes dataset contains a total of 5,000 images, recorded in streets from 50 different cities. It is divided into 2,975 images for training, 500 for validation, and 1,525 for testing. It also provides fine annotations for training and validation sets, while the test annotation is not publicly available for benchmarking purpose. There are also 20,000 extra images with coarse annotations but not used in our current work. The Cityscapes dataset is annotated with 34 different class labels encoded by `label_ID`. However, since some labels are objects such as border, ground, and so on, which are unnecessary here and can be ignored, we use a simpler encoding version, namely `train_ID`, that reduces the task complexity into 19 class labels. The Cityscapes recommends users to use `train_ID` instead of `label_ID` during the training phase.

Our study also makes use of the CityPersons annotation [15], which is a new set of person annotations with high-quality bounding boxes. It was built upon the Cityscapes dataset and provides better data for improved pedestrian detection. The CityPersons bounding box annotation is utilized in our work to easily extract each pedestrian instance before conducting the pedestrian orientation annotation on the dataset.

The CityPersons dataset explains various conditions of a pedestrian class based on the person types and occlusion levels. Person types in this dataset consist of pedestrian, rider, sitting person, and unusual posture; while the occlusion levels range from 0.0 to 0.9. The occlusion level 0.0 indicates the person has no occlusion, while 0.9 indicates he/she is almost completely occluded by some objects. The bounding box annotation is also well-aligned with the dimension ratio of 0.41. The dataset provides annotations for 19,654 persons in the training set and 3,938 in the validation set [15]. In our work, we extract only pedestrian-type person images with occlusion level below 0.1.

B. Pedestrian Orientation Annotation

To add new annotations to each extracted Cityscapes pedestrian, we used the Pedestrian Direction Classification (PDC) dataset from reference [6] as an image reference. We also referred to reference [16] for the orientation rules. The annotation process to modify Cityscapes dataset is as follows.

1) Extract single pedestrian images that are bounding boxes from Cityscapes images; CityPersons annotation is used to locate such single person pixels.
2) Filter the extracted persons considering the type and the occlusion level; Here, the person type should be pedestrian with a maximum occlusion level of 0.1.

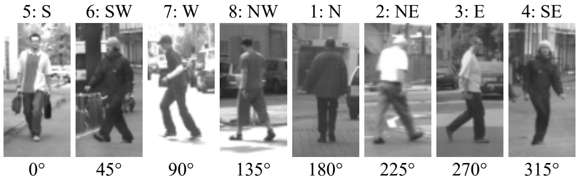

Fig. 3. References for annotating pedestrian directions [16].

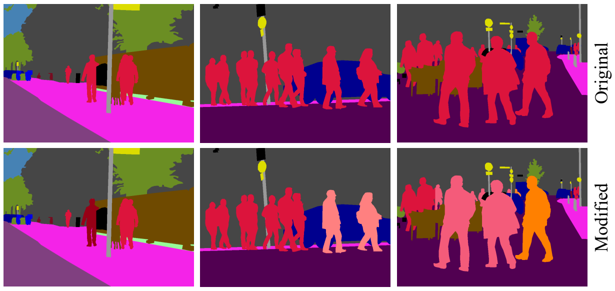

Fig. 4. New colored ground-truth examples with pedestrian orientations compared to the original ones; The new colors indicate variations of pedestrian orientations.

3) The pedestrian image collection is then manually annotated into orientation classes; The CityPersons annotation is then modified by adding the attribute class labels to the bounding boxes. We refer to references [6] and [16] for this process.
4) Change the person label in Cityscapes into one of eight person classes based on the orientation; This process adds a new Cityscapes pixel annotation.

The pedestrian orientation annotation process yielded 2,216 and 458 well-aligned pedestrian images for training and validation images, respectively. Table I shows the number of pedestrians annotated in each orientation class. We annotated cropped pedestrian images one-by-one solidly. Our orientation annotation is slightly different to that in reference [16], where we used label '1' to denote direction N, '2' for NE, and so on until '8' for NW, respectively, as shown in Figure 3.

The number of classes was increased from 19 to 27 since we replaced class 'person' with nine attribute classes (eight orientation classes and one class for unknown orientation). Figure 4 depicts samples of new ground-truth images compared to the original ones after additional annotations were applied. In the original ground-truth, all pedestrians had the same color, while in the modified ones, one pedestrian might have a different color with another.

V. Experiments and Discussion

A. Experiment Design and Implementation

The experiments are conducted to 1) train the proposed model for attribute-aware semantic segmentation, 2) analyze whether the attribute-aware concept can help increase the semantic segmentation performance, and 3) compare the proposed method with the original SegNet on the same dataset.

TABLE I
Class Distributions of Annotated Pedestrian Orientations

Set	1	2	3	4	5	6	7	8
Training	466	131	327	126	545	115	334	172
Validation	98	28	66	49	79	30	81	27
Total	**564**	**159**	**393**	**175**	**624**	**145**	**415**	**199**

The original SegNet is built on Caffe [17] implementation[1]. However, some adjustments are needed because it was built for another dataset with less complexity. After some technical modification on the code to satisfy the requirements of the Cityscapes dataset, we trained models using the training set and its annotations.

In this experiment, we compare two models: the original SegNet trained using the `19_label_set` and the proposed attribute-aware SegNet trained using `27_label_set`. The `19_label_set` is a set of annotations based on the `train_ID` encoding scheme consisting of 19 labels to classify; while the `27_label_set` is a set of annotation based on the `train_ID` encoding scheme with additional eight new labels for attribute annotated-person classes.

Training: We re-trained *VGG 16 layers* as the initial model to build our models with 50,000 steps for each training phase. Since we trained SegNet with a large dataset and image dimension, it was necessary to decrease the computational load to deal with the space complexity issue. The size of all images was reduced into half of the original image during the training phase, while the batch number was set to be 2 for smaller memory usage.

Testing: To evaluate the trained models, we used `train_set` and `val_set` images including their available annotations to calculate the performance metrics and compare the results. We also tested the original and the modified models using `test_set` and then submitted the results to the Cityscapes Web site[2] for benchmarking purpose. Before submitting the results, all labels in output images should be converted into the original labels and all output images were resized into the original dimensions. Note that all pedestrian classes (a pedestrian with additional eight orientation labels) were combined into one pedestrian class label to comply with the Cityscapes labels and to compare the performances between the original SegNet and the proposed method.

B. Performance Metric on Segmented CityScapes

We measured the performance of the models using two common metrics, which are global accuracy and mean Intersection over Union (IoU) [3], defined as follows:

$$\text{glob\_acc} = \frac{\text{true\_predicted\_pixels}}{\text{all\_predicted\_pixels}}, \tag{7}$$

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c}, \tag{8}$$

$$\text{mean\_IoU} = \frac{1}{C}\sum_c \text{IoU}_c. \tag{9}$$

[1] Publicly available at https://github.com/alexgkendall/caffe-segnet/.

[2] https://www.cityscapes-dataset.com/submit/.

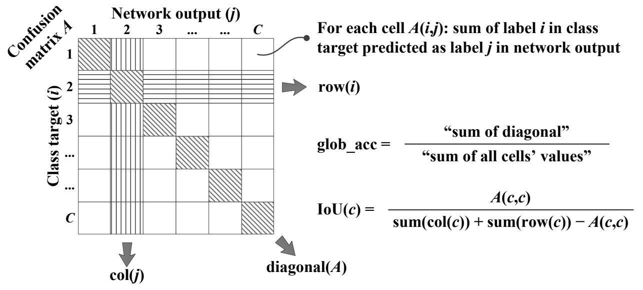
Confusion matrix A
Network output (j)
1 2 3 C
Class target (i)
1
2
3
...
...
C
For each cell A(i,j): sum of label i in class target predicted as label j in network output
row(i)
glob_acc = "sum of diagonal" / "sum of all cells' values"
IoU(c) = A(c,c) / sum(col(c)) + sum(row(c)) − A(c,c)
col(j)
diagonal(A)

Fig. 5. Calculating global accuracy and IoU using confusion matrix.

where IoU_c is the IoU for class c and C is the number of classes, while TP, FP, and FN respectively denote the number of 'true positive', 'false positive', and 'false negative' pixels.

Based on Formulae 7, 8, and 9, we use a confusion matrix in the implementation to calculate the global accuracy and the mean IoU. Figure 5 illustrates how we calculate the measuring metrics in this study.

C. Results of Semantic Segmentation

Table II shows the evaluation results for both the original SegNet and the proposed method. The original SegNet was trained using the Cityscapes dataset, while the proposed method was trained using our modified Cityscapes dataset. We use the global accuracy (glob_acc) and the mean of IoU (mean_IoU) to measure the performance of both methods for the validation set. We also add the accuracy of pedestrian orientation (mean_$\mathrm{IoU}_{\mathrm{or}}$) computed exclusively for the proposed method. Meanwhile, Table III shows the per-class accuracy of our method on the testing set, which is benchmarked for comparison between existing methods for the Cityscapes dataset. The benchmarking system uses the IoU metric for measuring the performance. We can see the qualitative results of some output samples in Figure 6 selected from the validation set.

It is observable that the performance of our proposed method outperforms the baseline method. Our proposed method improves 1.9 for the global accuracy and 8 for the mean of IoU. Meanwhile, the IoUs obtained for orientations 1 to 8 are 38.1, 17.1, 56.6, 0, 37.1, 0, 1.4, and 0.1, respectively. The mean_$\mathrm{IoU}_{\mathrm{or}}$ is calculated from those IoUs plus the IoU of the person with unknown orientation. Although the mean_$\mathrm{IoU}_{\mathrm{or}}$ is still very low, it generally shows an improvement on class accuracy by varying the person class into several sub-classes.

The time for segmentation after training is 1.7 frames per second, while the off-line network training certainly takes much more time lasting one day or so. The detection time is short to be considered in use on a slow-moving vehicle. Note that we used NVIDIA's Titan X GPU in our current study for training and segmentation processes.

D. Discussion on Accuracy and Data

The experimental results show that the proposed method outperforms the original SegNet slightly. For both global accuracy and mean IoU, the detailed sub-classes identification has gained higher values than the original SegNet on average.

TABLE II

PERFORMANCE EVALUATION ON VALIDATION SET

Original SegNet		Proposed Method		
glob_acc	mean_IoU	glob_acc	mean_IoU	mean_$\mathrm{IoU}_{\mathrm{or}}$
84.4	42.3	**86.3**	**50.3**	27.8

This is achieved by the proposed concept and implementation of the attribute-aware semantic segmentation.

Applying a conventional semantic segmentation to a number of object classes usually implies in high deviation between classes. It causes the difficulty in classifying each pixel into a given set of labels, especially for a class with a low ratio in the training set. Meanwhile, the proposed method divides a particular class into some sub-classes that correspond to its attribute classes. This proposed concept implies that between-classes deviations are decreased. Accordingly, the semantic segmentation becomes easier. Thus, the concept to provide object attributes with additional semantic information improves the semantic segmentation. The training process can also start from an existing model to refine the detailed aspects or attributes in a shorter time.

On the other hand, the number of annotated pedestrians in this study is still limited. This can still be improved by collecting more attributed-object samples from the Cityscapes dataset. For example, adjusting the constraints of person type or occlusion level is necessary to increase the number of well-aligned pedestrian images for training and validation.

After the pedestrians are detected with orientation attributes, it is important to pay more attention on the persons facing the road, because they are perhaps walking towards the vehicle path and the vehicles may prepare a sudden breaking. In the future, we will further target the motion information [18] in the driving video to feed in the recognition of pedestrian's walking direction such that the segmentation of pedestrians can be more certain and robust.

VI. CONCLUSION

We proposed a new approach of the attribute-aware semantic segmentation to enrich recognition classes. A general concept was introduced to modify a deep neural network model to achieve this goal. We conducted an experiment using the Cityscapes dataset and the modified SegNet resulting in better performance as compared to the original one. We inserted the pedestrian body orientation as additional semantic information in the annotations to modify the Cityscapes dataset. This has proved that the proposed method not only performed an attribute recognition task but also enhanced the semantic segmentation. The proposed concept should be able to handle more object classes and attributes than pedestrians.

ACKNOWLEDGMENT

The first author would like to thank Indonesia's BUDI-LN scholarship for the generous supports to his doctoral study. Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research in Japan.

TABLE III

EVALUATION RESULTS FOR THE TESTING SET SUBMITTED TO CITYSCAPES BENCHMARKING

Method	**Average**	**road**	**side walk**	**building**	**wall**	**fence**	**pole**	**traffic light**	**traffic sign**	**vegetation**	**terrain**	**sky**	**person**	**rider**	**car**	**truck**	**bus**	**train**	**motorcycle**	**bicycle**
Original SegNet	44.25	85.78	51.78	**75.48**	**23.36**	**26.46**	26.22	33.48	**35.30**	83.85	48.66	**86.93**	57.37	29.35	76.61	5.30	**27.15**	0.03	19.68	**47.99**
Proposed Method	**45.92**	**89.94**	**54.64**	73.76	20.50	17.95	**29.70**	**33.53**	34.97	**84.34**	**51.14**	86.51	**59.58**	**37.96**	**81.90**	**11.98**	26.76	**1.75**	**31.59**	43.91

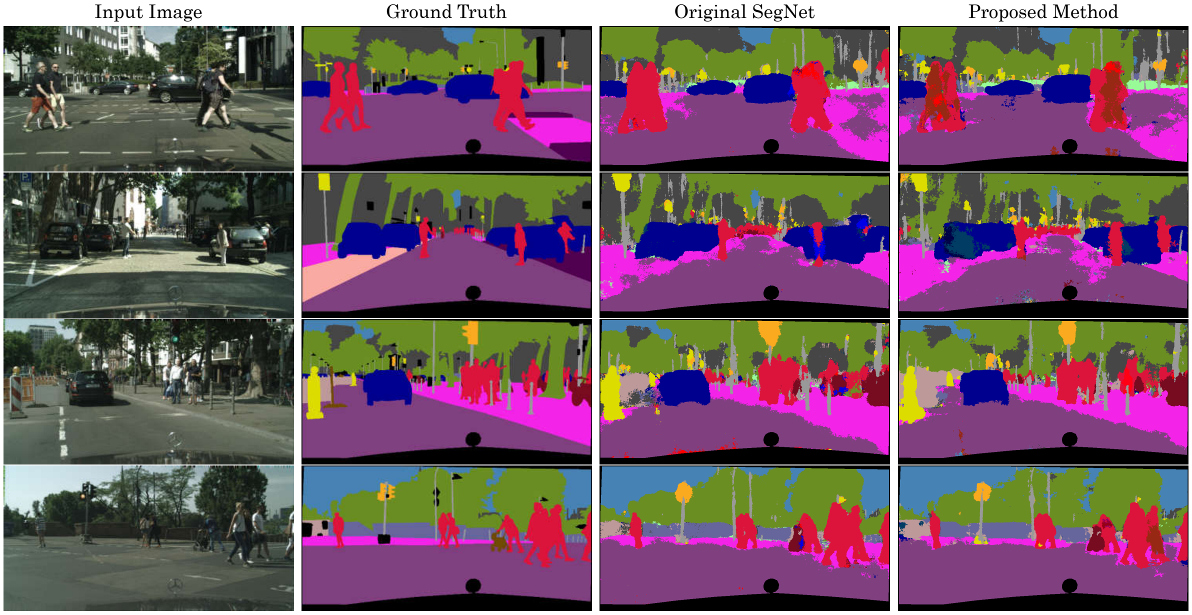

Fig. 6. Qualitative results for four input images selected from the validation set; The columns from left to right are the input image, the ground-truth image from the Cityscapes dataset, the result from the SegNet model trained using the original dataset (19 classes), and the result from the proposed method which is trained using a modified dataset (27 classes); The rightmost column contains information of pedestrian orientation.

REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” in IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.

[2] B. J. Meyer and T. Drummond, “Improved semantic segmentation for robotic applications with hierarchical conditional random fields,” in Proc. 2017 IEEE Int. Conf. on Robotics and Automation, 2017, pp. 5258–5265.

[3] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[4] A. Kendall, V. Badrinarayanan, and R. Cipolla, “Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding,” arXiv preprint arXiv:1511.02680, 2015.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[6] J. Tao and R. Klette, “Part-based RDF for direction classification of pedestrians, and a benchmark,” in Computer Vision ACCV 2014 Workshops, ser. Lecture Notes in Computer Science, vol. 9009. Springer, 2014, pp. 418–432.

[7] 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. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[8] K. He, G. Gkioxari, P. Dollr, and R. Girshick, “Mask R-CNN,” in Proc. 2017 IEEE Int. Conf. on Computer Vision, 2017, pp. 2980–2988.

[9] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, “A review on deep learning techniques applied to semantic segmentation,” arXiv preprint arXiv:1704.06857, 2017.

[10] Y. Deng, P. Luo, C. C. Loy, and X. Tang, “Pedestrian attribute recognition at far distance” in Proc. 22nd ACM Int. Conf. on Multimedia, 2014, pp. 789–792.

[11] Y. Deng, P. Luo, C. C. Loy, and X. Tang, “Learning to recognize pedestrian attribute,” arXiv preprint arXiv:1501.00901, 2015.

[12] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, “A richly annotated dataset for pedestrian attribute recognition,” arXiv preprint arXiv:1603.07054, 2016.

[13] M. Cordts, M. Omran, S. Ramos, T. Scharwchter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes dataset,” in Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition Workshops, vol. 1, no. 2, 2015, p. 3.

[14] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture” in Proc. 2015 IEEE Int. Conf. on Computer Vision, 2015, pp. 2650–2658.

[15] S. Zhang, R. Benenson, and B. Schiele, “Citypersons: A diverse dataset for pedestrian detection,” in Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, no. 2, 2017, p. 3.

[16] Y. Kawanishi, D. Deguchi, I. Ide, H. Murase, and H. Fujiyoshi, “Misclassification tolerable learning for robust pedestrian orientation classification,” in Proc. 23rd Int. Conf. on Pattern Recognition, 2016, pp. 486–491.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in Proc. 22nd ACM Int. Conf. on Multimedia, 2014, pp. 675–678.

[18] M. Kilicarslan, J. Y. Zheng, K. Raptis, “Pedestrain detection from motion,” in Proc. 23rd Int. Conf. on Pattern Recognition, 2016, pp. 1857–1863.

[19] P. Meletis and G. Dubbelman, “Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation” in Proc. 29th IEEE Intelligent Vehicles Symposium, 2018, pp. 1045–1050.