Deep Convolutional Encoder-Decoder Network with Model Uncertainty for Semantic Segmentation

Shuya Isobe
Tokyo City University
Graduate School of Engineering
Setagaya, Tokyo, Japan
Email: isobe15@ipl.cs.tcu.ac.jp

Shuichi Arai
Tokyo City University
Graduate School of Engineering
Setagaya, Tokyo, Japan
Email: arai.s@cs.tcu.ac.jp

***Abstract*—We propose a new semantic segmentation method and the necessity of certainty for practical use of semantic segmentation in scene understanding. We implement a deep fully convolutional encoder-decoder neural network for semantic segmentation. This network architecture makes the segmentation accuracy improve by retaining boundary details in the extracted image representation. This accuracy means how much the segmentation results match to ground truth labels. However, the conventional evaluation method ignores unlabeled regions in ground truth labels. In other words, the segmentation results has not been evaluated in the regions of unknown objects. Toward practical use of the semantic segmentation, the evaluation should consider such regions. So it is necessary to recognize accurately whether the object is known or not. We call this factor certainty. Bayesian SegNet makes it possible to produce an uncertainty of the segmentation results with a measure of model uncertainty from the sampling of the posterior distribution of the model using Dropout. However, the uncertainty is not used for segmentation itself, and all pixels are classified into one of the predefined classes in this segmentation result. It means that the pixels within the regions of unknown objects are definitely misclassified as one of the predefined classes. Our study aims the improvement of certainty for semantic segmentation in road scene understanding with model uncertainty. Our method rejects the uncertain region and classifies it as an unknown object using the model uncertainty. We achieved improvement of certainty by our method as shown in the evaluation results. Furthermore, we indicated the possibility of the performance improvement on the deep convolutional encoder-decoder network architecture from the comparison of our network architecture with Bayesian SegNet architecture.**

***Keywords*—Semantic segmentation, Convolutional neural networks, Deep learning, Auto-encoder, Upsampling, Certainty**

I. Introduction

Semantic segmentation is a pixel-level labelling for image classification. This is an important technique for scene understanding because this makes it possible to recognize each object in per-class and divide the object region as the shape of the object contour. Recently, convolutional neural networks (CNNs) [1] are very popular in training for segmentation models. In particular, some architectures achieved good performance [2]–[8]. However, these architectures have tried to directly adopt deep convolutional architectures to pixel-wise labelling, so the segmentation results appear coarse. This is because that the results are obtained from the low-resolution feature maps by max-pooling. To solve this problem, SegNet [9] proposed a deep convolutional encoder-decoder network. This architecture improves boundary delineation and reduces the number of parameters enabling end-to-end training. As a result, the accuracy of prediction has improved considerably and exceeded 90% in the datasets for scene understanding. However, this accuracy ignores the regions of unknown objects. In other words, the unlabeled regions in the ground truth labels are ignored. This is because that models can not classify the regions into except for the predefined classes and it is hard to define all objects in the scene as classes. But it becomes a problem on a practical level so that the pixels within the regions of unknown objects are definitely misclassified as one of the predefined classes. It is necessary to recognize accurately whether the object is known or not. We call this factor $certainty$. Bayesian SegNet [10] which is an architecture that extends SegNet makes it possible to produce an uncertainty of segmentation results with a measure of model uncertainty from sampling of the posterior distribution of the model using Dropout [11]. However, the uncertainty is not used for segmentation itself.

The purpose of our study is the improvement of certainty for semantic segmentation in road scene understanding with model uncertainty. Our method rejects the uncertain region using the model uncertainty and makes the certainty improve. In a certain pixel, if the uncertainty is large, the model considers it as an unknown object and the pixel is not classified into any predefined classes. If the uncertainty is small, the model considers it as the object belonging to a class which has the highest probability among predefined classes. We evaluate our proposed segmentation method using 5 common measures of evaluation to show the improvement of certainty by our method. Furthermore, we indicate the possibility of the performance improvement on the deep convolutional encoder-decoder network architecture from the comparison of our network architecture with Bayesian SegNet architecture.

In the rest of this paper is organized as follows. In Section II, the progress of the performance improvement for semantic segmentation and a problem of the conventional evaluation are introduced. In Section III, we describe our network architecture and the comparison with Bayesian SegNet architecture. We propose the method of inference with model uncertainty in Section IV. In Section V, we describe the training condition of our network and Bayesian SegNet architecture for comparison of each network. We evaluate our method and our network

978-1-5090-5795-5/17/$31.00

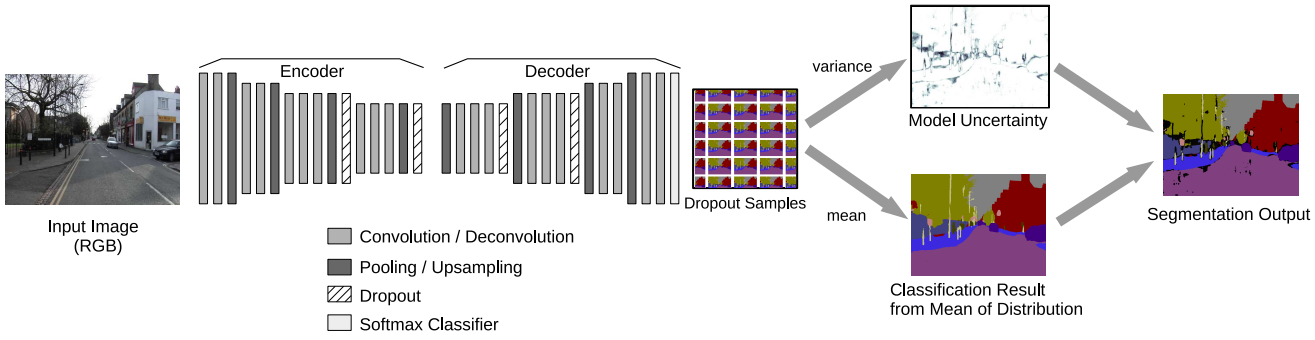

Fig. 1. A schematic of the our network architecture.

using 5 common measures of evaluation in Section VI. Finally, the paper is concluded in Section VII.

II. Related Works

Semantic segmentation originating from TextonForest [12] and TextonBoost [13] is a pixel-level labelling for image classification. This technique makes it possible to recognize each object in per-class and divide the object region as not a rectangle but the shape of the object contour and so it is widely used for scene understanding. In recent years, many approaches adopted convolutional neural networks (CNNs) to learning of the shape of each class object. The deeper a network was, the more accuracy improved. As a result, the network became complicated and the number of trainable parameters was increased. But pixel-level classification like semantic segmentation had a problem that training and inference needed lots of time for recalculation of a wide region during convolution. Fully Convolutional Network (FCN) [2] is a network that all layers consist of convolutional layers as the name suggests. The contribution of this architecture is to reduce the number of trainable parameters by replacing the fully connected layers with several convolutional layers which play the same role. In addition, the key to success is revealed that deepening a network while restraining the number of trainable parameters. For the purpose of further performance improvement, various methods had been proposed, and furthermore, the network became deeper than previously. However, it is difficult to train end-to-end and so it has led to multi-stage training [14], [15], fine-tuning a pre-trained network, appending other networks, preprocessing like region proposal [4], [14], [16], and postprocessing using graphical models like conditional random fields (CRFs) [17]. SegNet [9] inspired by the unsupervised feature learning architecture proposed by Ranzato et al. [18] is a deep convolutional encoder-decoder network by combining CNN with Auto-encoder. The decoder produces dense high-resolution feature maps through upsampling and deconvolution of low-resolution feature maps obtained from convolution and pooling in the encoder. This architecture enables to retain boundary information and deep end-to-end training by reducing the number of parameters. Consequently, the performance is further improved and inference time is shortened so as to inference in real time. It has led to expectations for autonomous driving.

We described the progress of the performance improvement for semantic segmentation so far. This performance is evaluated based on accuracy which means how much the segmentation results match to ground truth labels. However, it is hard to define all objects in the scene as classes and so unknown objects definitely exist in the scene. Therefore, this accuracy ignores the regions of unknown objects. But it becomes a problem on a practical level so that the pixels within the regions of unknown objects are definitely misclassified as one of the predefined classes. It is necessary to recognize accurately whether the object is known or not. We call this factor $certainty$. However, none of these proposed architectures generate a measure of certainty. Bayesian SegNet [10] which is an architecture that extends SegNet makes it possible to produce an uncertainty of segmentation results with a measure of model uncertainty with sampling using Dropout from Gal and Ghahramani proposed methods [19], [20]. But the uncertainty is not used for segmentation itself. Then we aim the improvement of certainty for semantic segmentation in road scene understanding with model uncertainty.

III. Network Architecture

In this section, we present our network architecture. Our network is a deep convolutional encoder-decoder network like SegNet and Bayesian SegNet. Encoder network is composed of 10 convolutional layers, 4 pooling layers, and 2 dropout layers. Decoder network corresponding encoder network is 10 convolutional layers, 4 upsampling layers, and 2 dropout layers. The end of the network is softmax classifier. Fig. 1 illustrates our network architecture. Rectified Linear Unit (ReLU) and Batch normalization [21] are applied to each convolutional layer for solving the non-linear problems and preventing over-fitting.

A. Encoder

The encoder performs convolution with a 3×3 filter bank to extract the feature of object shapes in input RGB images. Following that max-pooling with a 2×2 window and stride 2 is applied to gain translation invariance over small spatial shifts in the input image for robust classification by reducing spatial resolution of the feature maps obtained from convolution. However, boundary details are lost in general max-pooling. Therefore, our method retains spatial information in a window to restore lost resolution of the feature maps by upsampling in the decoder. The encoder extracts features from input RGB images using convolution and max-pooling in this way.

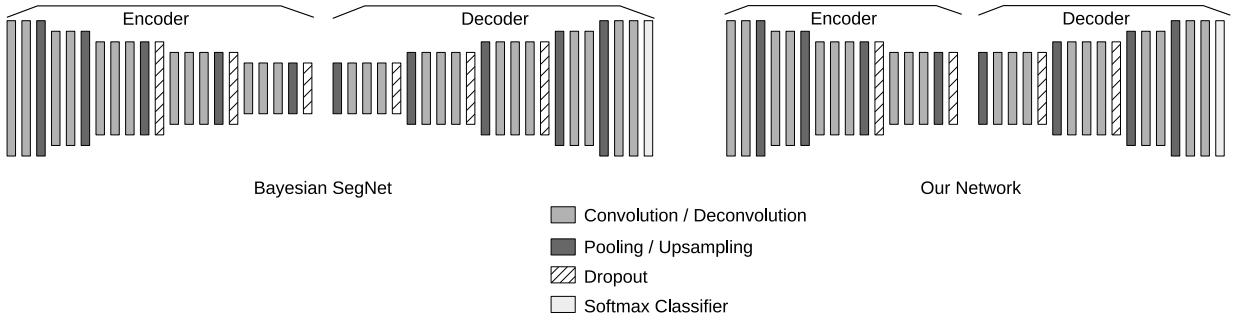

Fig. 2. Comparison of our network architecture with Bayesian SegNet architecture.

B. Decoder

The decoder performs upsampling to restore resolution of input feature maps using the memorized max-pooling indices from the corresponding encoder feature maps. The upsampling expands the resolution of input feature maps from subsampled values and spatial information in max-pooling. This step can obtain high-resolution feature maps but they are sparse. So the decoder performs convolution of the sparse feature maps to obtain the dense high-resolution feature maps. This is often called $deconvolution$. The decoder finally outputs the dense feature maps which are the same resolution as the input RGB images by upsampling and deconvolution.

IV. Inference with Model Uncertainty

In this section, we propose the method of segmentation by inference with the model uncertainty. The technique we use to perform probabilistic inference is a dropout. Dropout is used in many models in deep learning as a way to prevent over-fitting by dropping units randomly. The dropping units randomly means the same as learning with various models. We use this property as a way of obtaining samples from the posterior distribution of models from Gal and Ghahramani proposed methods [19], [20]. This method enables to perform probabilistic inference over our segmentation model. We find the posterior distribution over the convolutional weights W given input data X and its labels Y in training using dropout as shown in Eq. 1.

$$p(W \mid X, Y) \tag{1}$$

However this form of posterior distribution is hard to manage, so we prepare the approximating distribution $q(W)$. Then we use variational inference [22] to approximate $q(W)$ and the posterior distribution as shown in Eq. 1 by minimizing the Kullback-Leibler (KL) divergence as shown in Eq. 2.

$$D_{KL}(q(W) \,||\, p(W \mid X, Y)) \tag{2}$$

We also perform minimizing the cross entropy loss objective function to encourage the model to learn a distribution of weights in minimizing the KL divergence term. We train the model with dropout and sample the posterior distribution over the weights at test time using dropout to obtain the posterior distribution of softmax class probabilities. We take the mean of these samples for classification prediction and use the variance to model uncertainty for each class based on Bayesian SegNet. Fig. 1 shows a schematic of the segmentation prediction and model uncertainty estimate process. We perform segmentation using the model uncertainty. In a certain pixel, if the variance is large, the model considers it as an unknown object and the pixel is not classified into any predefined classes. If the variance is small, the model considers it as the object belonging to a class which has the highest probability among predefined classes. To our knowledge, there has been no methods to reject the uncertain region like our method for semantic segmentation. Our method is different from some methods which require a large training dataset to let the model learn the regions of unknown objects as a $background$ class.

V. Training of our network

We train our network with the CamVid [23] dataset for inference using our method at test time. We implement our network using the Caffe [24] which is common deep learning framework. CamVid is a scene understanding dataset with day and dark road scene images. We use 367 training and 233 testing RGB images in this dataset and resize these images to 360×480 pixels. Classes are 11 objects in the road scene such as sky, building, pole, road, pavement, tree, sign symbol, fence, car, pedestrian, bicyclist. We train our network end-to-end using stochastic gradient descent (SGD) with momentum 0.9, weight decay of 0.0005, and dropout ratio of 0.5. We performed training until the cross-entropy training loss converged on such conditions.

We also train the network of Bayesian SegNet on the same learning conditions to evaluate our network. Bayesian SegNet is composed of 26 convolutional layers, 10 pooling layers, and 6 dropout layers. On the other hand, we implement the smaller network than Bayesian SegNet because we consider that the more convolutional layers are deep, the more difficult it is to restore abstract features in the decoder. So our network composed of 20 convolutional layers, 8 pooling layers, and 4 dropout layers. Fig. 2 shows each network architecture.

VI. Evaluations

In this section, we evaluate improvement of certainty by our proposed segmentation method and performance of our network architecture. Fig. 3 shows the segmentation results obtained from classification by our model. In model uncertainty in Fig. 3, darker colors indicate a larger value of the variance and it means more uncertain predictions. We can see that our method prevents labelling in darker color pixels in

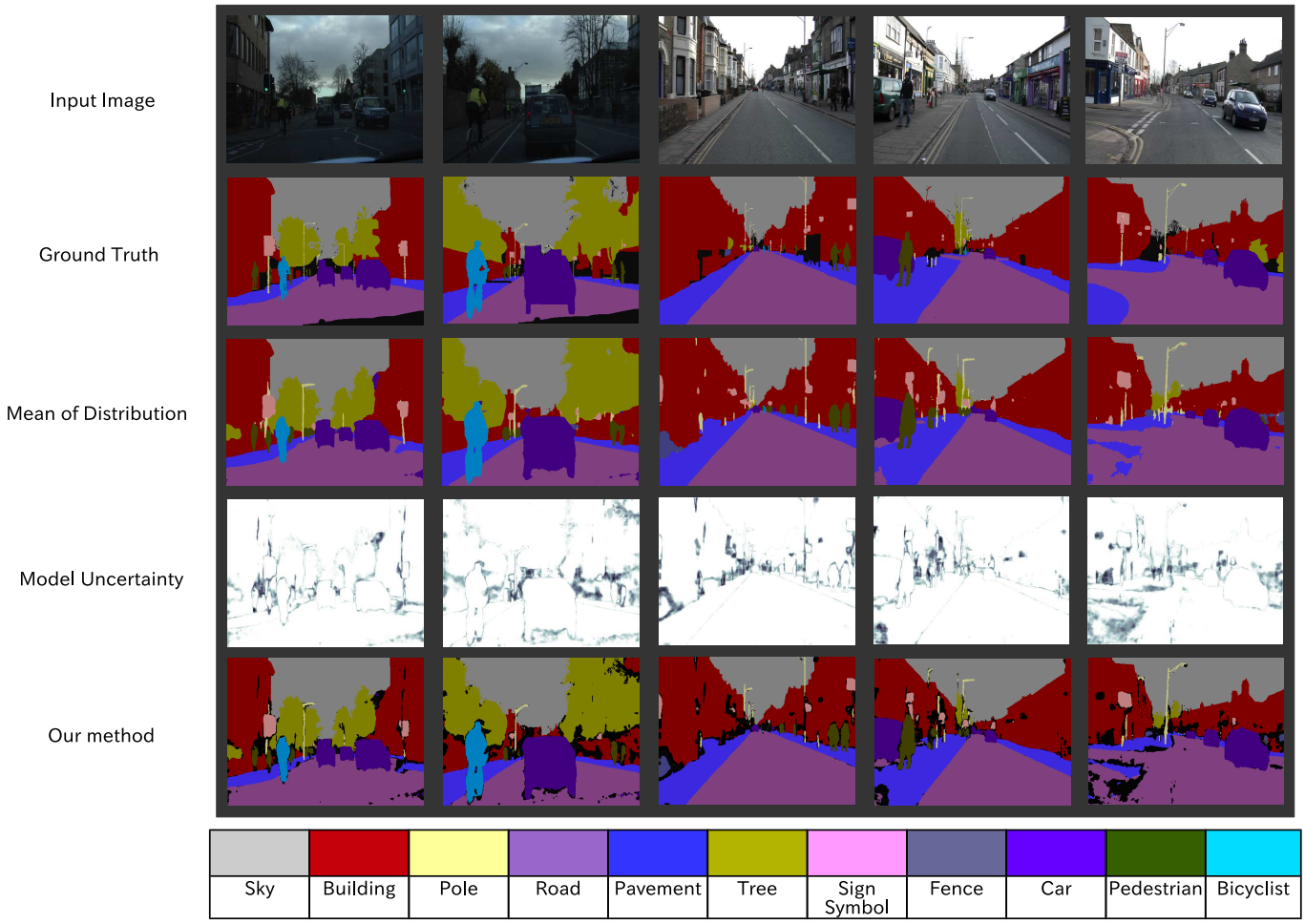

Fig. 3. Inference results by our model on CamVid road scene understanding dataset.

model uncertainty. It means success to the rejection of the uncertain region.

We evaluate our segmentation results to show the improvement of certainty by our proposed segmentation method. We use 5 common measures of evaluation: Global accuracy ($Global\ acc$) in Eq. 3 measures the percentage of pixels correctly classified by the division of the total number of pixels of true positive prediction (TP) and the total number of pixels of ground truth (GT).

$$Global\ acc \quad = \quad \frac{TP}{GT} \tag{3}$$

Class accuracy ($Class\ acc$) in Eq. 4 measures the percentage of pixels correctly classified in a class i.

$$Class\ acc_i \quad = \quad \frac{TP_i}{GT_i} \tag{4}$$

Class average accuracy ($Class\ avg$) in Eq. 5 is the mean of the predictive accuracy over all classes by the division of the sum total of class accuracy in all classes and the number of classes (C).

$$Class\ avg \quad = \quad \frac{\sum_i^C Class\ acc_i}{C} \tag{5}$$

Intersection over union (IoU) in Eq. 6 is a measure which imposes the penalty of false positive predictions (FP) on the class accuracy in a class i.

$$IoU_i \quad = \quad \frac{TP_i}{GT_i \ + \ FP_i} \tag{6}$$

Mean intersection over union ($Mean\ IoU$) in Eq. 7 is the mean of intersection over union in all classes.

$$Mean\ IoU \quad = \quad \frac{\sum_i^C IoU_i}{C} \tag{7}$$

we use these 5 measures to evaluate our method and our network.

In Bayesian SegNet, the segmentation results are obtained from the mean of distribution like Fig. 3 showed, and this results show that all pixels are classified into one of the predefined classes in spite of the existence of the unlabeled regions in ground truth labels. So the conventional evaluation method ignores the unlabeled regions in ground truth labels and considers whether or not the segmentation results match to ground truth labels only in the labeled regions by true and false positives. However, it is hard to define all objects in the scene as classes, so the segmentation results should have the unlabeled regions having the same meaning as the regions of unknown objects. Toward practical use of the semantic

segmentation, it is necessary to evaluate the unlabeled regions in ground truth labels. Our method can infer the region which is not classified into any predefined classes and we newly consider the false positive in the case of misclassification in the unlabeled regions in ground truth labels. In evaluation of certainty, it is important that whether or not our method reduces 2 types of false positives: the case of the mismatch of the predicted class label and the ground truth class label in a pixel, and the case of the misclassification in a unlabeled pixel of the ground truth label.

A. Evaluation of certainty improvement by our method

We evaluate the segmentation results only in the labeled regions in ground truth labels to verify whether or not our method reduces false positives in the case of the mismatch of the predicted class label and the ground truth class label in a pixel. The results of the classification accuracy using 5 measures of evaluation are shown in Table. I, II, and III. We succeeded in reducing false positives in this case because our method exceeded conventional results obtained from the mean of distribution in all accuracy scores in Table I, II, and III.

We also evaluate the segmentation results only in the unlabeled regions in ground truth labels to verify whether or not our method reduces false positives in the other case. We treat the unlabeled pixels as the unknown class in this evaluation and if inference result in a pixel is an unknown class, this result is evaluated as a true positive. It is not necessary to evaluate using except global accuracy because the number of correct classes is only one unknown class. The result of the classification accuracy is shown in Table IV. We succeeded in reducing false positives in this case because the accuracy was improved from 0 % to 27.2 % by making it possible to reject the uncertain regions using our method. The accuracy of 27.2 % is far from the high score. But as we have shown in Section IV, our model has avoided learning of the regions of unknown objects so this score is appropriate.

However, if our method rejects large regions, it is not practical that our model classifies the most of the regions as unknown objects. So we show the percentage of the unlabeled regions and the labeled regions of each predefined class over the entire images in Table V. The each percentage in the mean of distribution exceeded the each percentage in ground truth in many classes. It means that the pixels within the each excess regions are definitely misclassified. In contrast, the exceeding class in our method is the only pedestrian. Our method has a lower probability of misclassification than the method based on the mean of distribution. Furthermore, our method restrains the percentage of the unlabeled regions to 10 % or less and this result is within the practical level.

From these evaluation results, we succeeded in the improvement of certainty by our method.

B. Evaluation of our network performance

We evaluate performance of our network architecture by comparing with Bayesian SegNet architecture. The results of the classification accuracy using 5 measures of evaluation are shown in Table VI, VII, and VIII.

Our network exceeded Bayesian SegNet architecture in most of accuracy scores in Table VI, VII, and VIII. On this condition, our network is better than Bayesian SegNet from these results. It means that there is room for improvement of performance on the deep convolutional encoder-decoder network architecture.

TABLE I. EVALUATION RESULTS OF OUR METHOD IN THE ENTIRE CLASS

	Global acc	Class avg	Mean IoU
Mean of distribution	89.6	75.2	62.2
Our method	92.5	79.5	68.7

TABLE II. EVALUATION RESULTS OF OUR METHOD USING CLASS ACCURACY

	Class acc	
Class	Mean of distribution	Our method
Sky	94.5	95.1
Building	87.1	91.4
Pole	50.0	54.2
Road	95.9	96.7
Pavement	90.5	93.5
Tree	86.3	89.6
SignSymbol	55.6	60.1
Fence	45.9	52.0
Car	87.4	90.3
Pedestrian	79.7	87.4
Bicyclist	53.9	64.1

TABLE III. EVALUATION RESULTS OF OUR METHOD USING INTERSECTION OVER UNION

	IoU	
Class	Mean of distribution	Our method
Sky	92.1	92.9
Building	78.3	84.1
Pole	32.2	38.0
Road	93.2	94.9
Pavement	79.6	84.0
Tree	72.9	78.0
SignSymbol	34.4	43.7
Fence	34.8	42.5
Car	76.1	83.1
Pedestrian	47.4	58.7
Bicyclist	43.1	55.5

TABLE IV. EVALUATION RESULTS IN UNLABELED REGIONS USING GLOBAL ACCURACY

	Global acc
Mean of distribution	0.0
Our method	27.2

TABLE V. THE PERCENTAGE OF THE UNLABELED REGIONS AND REGIONS OF EACH CLASS OVER THE ENTIRE IMAGES

Class	Ground Truth	Mean of distribution	Our method
Sky	17.1	16.7	16.5
Building	24.6	25.2	22.1
Pole	1.2	1.3	0.9
Road	25.8	26.4	25.8
Pavement	9.3	10.0	9.1
Tree	11.3	12.4	11.2
SignSymbol	1.0	1.3	0.8
Fence	1.2	1.0	0.6
Car	4.0	4.4	3.8
Pedestrian	0.6	1.1	0.8
Bicyclist	0.2	0.2	0.1
Unlabeled	3.8	0.0	8.2

TABLE VI. EVALUATION RESULTS OF OUR NETWORK IN THE ENTIRE CLASS

	Global acc	Class avg	Mean IoU
Bayesian SegNet	87.9	72.3	58.1
Our Network	89.6	75.2	62.2

TABLE VII. EVALUATION RESULTS OF OUR NETWORK USING CLASS ACCURACY

	Class acc	
Class	Bayesian SegNet	Our Network
Sky	94.9	94.5
Building	83.1	87.1
Pole	45.1	50.0
Road	95.4	95.9
Pavement	89.8	90.5
Tree	83.2	86.3
SignSymbol	52.3	55.6
Fence	38.5	45.9
Car	87.7	87.4
Pedestrian	74.2	79.7
Bicyclist	50.6	53.9

TABLE VIII. EVALUATION RESULTS OF OUR NETWORK USING INTERSECTION OVER UNION

	IoU	
Class	Bayesian SegNet	Our Network
Sky	92.0	92.1
Building	74.2	78.3
Pole	26.9	32.2
Road	92.7	93.2
Pavement	76.7	79.6
Tree	69.5	72.9
SignSymbol	25.4	34.4
Fence	26.7	34.8
Car	76.6	76.1
Pedestrian	39.2	47.4
Bicyclist	38.9	43.1

VII. CONCLUSION

This paper proposed a new semantic segmentation method and the necessity of certainty for practical use of semantic segmentation in scene understanding. We aim the improvement of certainty for semantic segmentation by inference with model uncertainty. Our method can reject the uncertain region where the variance of the posterior distribution of the model is large. We achieved improvement of certainty by our method as shown in the evaluation results. Furthermore, evaluation results of network architecture showed that our network architecture was better than Bayesian Network architecture on the present conditions of learning. We indicated the possibility of the performance improvement on the deep convolutional encoder-decoder network architecture.

REFERENCES

[1] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," ECCV, 2014.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," CVPR 2015

[3] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," ICLR, 2016.

[4] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille. "Weakly- and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation," ICCV, 2015

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," ICLR, 2015.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," NIPS, 2012.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," CVPR, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CVPR, 2015.

[9] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," CVPR, 2015.

[10] V. Badrinarayanan, A. Handa, and R. Cipolla. "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding." arXiv preprint arXiv:1511.02680, 2016.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," JMLR, 2014.

[12] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," CVPR, 2008.

[13] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," IJCV, 2009.

[14] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," ICCV, 2015.

[15] S. Hong, H. Noh, and B. Han,"Decoupled deep neural network for semi-supervised semantic segmentation," NIPS, 2015.

[16] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," ECCV, 2014.

[17] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," ICCV, 2015.

[18] M. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," CVPR, 2007.

[19] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with bernoulli approximate variational inference," ICLR, 2016.

[20] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation:Representing model uncertainty in deep learning," ICML, 2016.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," ICML, 2015.

[22] A. Graves, "Practical variational inference for neural networks," NIPS, 2011.

[23] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," Pattern Recognition Letters, 30(2):8897, 2009.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," ACM MULTIMEDIA, 2014.