GCCNet: Global Context Constraint Network for Semantic Segmentation

Hyunwoo Kim[1], Huaiyu Li[2] and Seok-Cheol Kee[3,*]
[1]*Beijing Advanced Innovation Center for Intelligent Robotics and Systems, Beijing Institute of Technology, Beijing, China*
[2]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[3]*Smart Car Research Center, Chungbuk National University, Cheongju, South Korea*

Keywords: Convolutional Network, Joint Training, Global Context, Semantic Scene Segmentation.

Abstract: The state-of-the-art semantic segmentation tasks can be achieved by the variants of the fully convolutional neural networks (FCNs), which consist of the feature encoding and the deconvolution. However, they struggle with missing or inconsistent labels. To alleviate these problems, we utilize the image-level multi-class encoding as the global contextual information. By incorporating object classification into the objective function, we can reduce incorrect pixel-level segmentation. Experimental results show that our algorithm can achieve better performance than other methods on the same level training data volume.

1 INTRODUCTION

Semantic segmentation is one of the key computer vision tasks with various applications including scene understanding, autonomous driving, and 3D reconstruction. It aims at parsing images into several regions and labeling them with their corresponding semantic categories, which can also be viewed as a pixel-wise classification problem.

Early segmentation methods mainly relied on low-level hand-crafted vision features combined with machine learning algorithms to merge image regions or classify pixels. Typically, CRF (Conditional Random Field) models are exploited and have plenty of effective extensions (Krähenbühl and Koltun, 2011).

However, both expensive human labors and expert knowledge are required in these methods and satisfactory results are still not obtained. In recent years, due to the powerful hierarchical feature learning ability of deep convolutional neural networks (CNN), traditional semantic segmentation methods are almost superseded by deep learning approaches, especially after fully convolutional neural networks (Long et al., 2015) (FCNs) were proposed. The FCNs structure formulated image semantic segmentation task as a pixel-wise labeling problem and many state-of-the-art algorithms are extended from it.

Although the state-of-the-art semantic segmentation tasks can be achieved by the variants of FCNs, which consist of the feature encoding and the deconvolution, they struggle with missing or inconsistent labels.

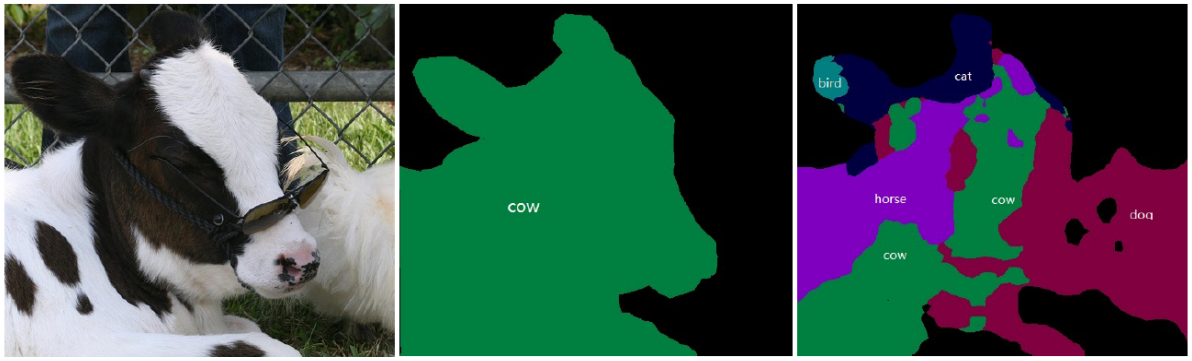
First, very large scale objects with complex texture and illumination conditions can be easily segmented into different categories. This problem may be caused by the fixed-size receptive field of CNN and it cannot sense the whole object. We have padded the large-scale object with zeros and resized the image to original size to reduce the object size. Then, we can get appropriate segmentation results using these preprocessed images. Second, tiny objects were often ignored and classified as background. We call these problems "confusion segmentation", and some typical visual segmentation examples of FCNs are shown in Figure 1.

In this paper, we propose a global context constraint network for semantic segmentation in order to solve the confusion segmentation problem. We utilized the global contextual information and defined an objective function to learn it explicitly in order to eliminate the segmentation confusion in the encoded feature.

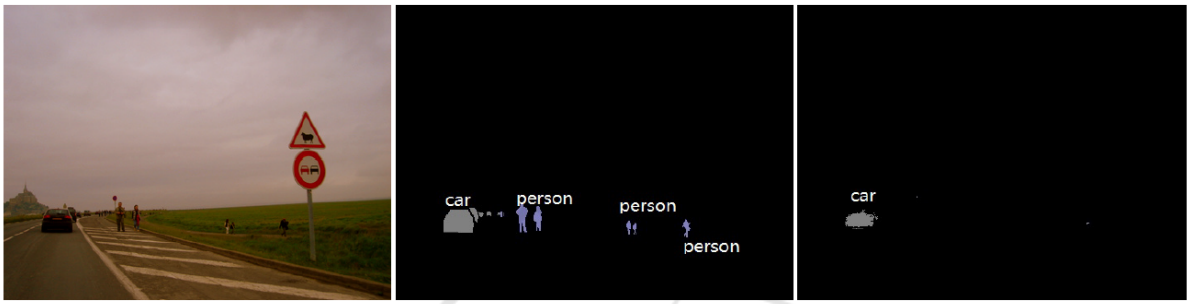
The intuition of the proposed network is as follows. We hypothesize that the joint learning of image-level class-specific features with baseline semantic segmentation can improve the semantic segmentation results while solving the confusion segmenta-

*Corresponding author

Kim, H., Li, H. and Kee, S.
GCCNet: Global Context Constraint Network for Semantic Segmentation.
DOI: 10.5220/0007705703800387
In *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2019)*, pages 380-387
ISBN: 978-989-758-374-2



(a) Inconsistent labels due to complex texture and large object



(b) Inconsistent labels due to complex texture and large object

Figure 1: Confusion segmentation problems of fully convolutional semantic segmentation networks. (Left) original image. (Center) ground-truth annotation. (Right) segmentation results.

tion problem. We will find that the addition of the image-level cross-entropy loss layer before deconvolution can give the better segmentation results even when the pixel-level segmentation information is not enough. Other recent results (Wang et al., 2016; Hong et al., 2015) can be interpreted in this perspective.

2 RELATED WORK

The DeepLab models (Chen et al., 2014) enlarged the receptive field to incorporate larger contextual information by using dilated convolution and utilize fully connected CRF (Krähenbühl and Koltun, 2011) as post-processing to refine the segmentation results. After FCNs (Long et al., 2015) was proved to be successful in semantic segmentation, their variants have been improved the accuracy. The FCNs structure formulated image semantic segmentation task as a pixel-wise labeling problem and many state-of-the-art algorithms are extended from it. The CRFas-RNN (Zheng et al., 2015) model integrated fully convolutional network with CRF algorithm into an end-to-end deep network that can be trained by the back-propagation algorithm. It can possess both the properties of CRF and FCNs and reach impressive segmentation results. The DeconvNet (Noh et al., 2015) approach got coarse feature map through convolution and pooling layers then recovered the dense prediction through symmetric up-convolution and unpooling operations. This network utilized the pooling mast with unpooling and the object proposals in order to solve inconsistent and missing labels problems. However, DeconvNet contains too many layers, and therefore, training and inference consumes too much time and memory. SegNet (Badrinarayanan et al., 2015) regarded fully convolutional network as the encoder network and the corresponding up-convolutional network as the decoder network. In essence, its network architecture was the same as DeconvNet, but disposed of several top layers to reduce the number of parameters. These approaches all required fixed size input images which will lose details of object in the images. ParseNet (Liu et al., 2015) captures the global context feature through global pooling and normalizing different features before fusion.

Recently, researchers have been tried to actively use class information to semantic segmentation tasks. Objectness-aware Semantic Segmentation (Wang et al., 2016) combined faster R-CNN to generate object proposals. Surprisingly, the simple combination of the object detector and semantic segmentation achieve a top performance in PASCAL VOC2012 challenge. It can be evidence that the

consideration of each object class separately is very helpful. Also, Hong et. al. (Hong et al., 2015) decoupled classification and segmentation to reduce the search space for segmentation effectively by exploiting class-specific activation maps, contrary to existing approaches posing semantic segmentation as region-based classification.

More recently, PSPNet (Zhao et al., 2017) aimed to enforce global priors using global pooling for scene parsing. Additionally, in the context of object detection tasks, similar object or instance based methods have been proposed (Hu et al., 2017)(He et al., 2017). In contrast to those methods, we proposed a complementary module to be easily incorporated into the FCN-based methods.

3 PROPOSED METHOD

In this section, we discuss the architecture of our Global Semantic Context Constraint network and describe the overall semantic segmentation algorithm. In our network, we consider the segmentation as an encoding-decoding process and the two components are discussed in detail as follows.

In this section, we propose a global context constraint network (GCCNet) for semantic segmentation in order to solve the confusion segmentation problem. We consider the semantic segmentation network as a pairwise encoding-decoding process. We use a modified fully convolutional network utilizing global context information to encode segmentation features and take the up-convolution operation to decode the probability of each category for each pixel. We define an objective function to explicitly guide the training of the global contextual information. And then we incorporate the global contextual feature into the convolutional feature map in order to make the final encoding feature. The decoding process is equivalent to a parameter learnable upsampling procedure. The whole network is jointly trained with segmentation loss end to end. It is worth noting that our network neither needs to resize the input images to the same scale nor does it need to utilize objects proposals. So, our method is robust to reach better performance than previous methods on the same level training data volume. Furthermore, it consumes less time and memory during training and inference and contains much fewer parameters in the network than previous methods.

3.1 Network Analysis

The overall architecture is shown in Figure 2, we consider the semantic segmentation network as an encoding-decoding process. In our network, we regard forward convolutional operations as the process of encoding features from the original images and regard up-convolutional operations as the decoding process to predict the probability of each category for each pixel. In the encoding component, we make use of the modified version (Liu et al., 2016) of the VGG16 network (Simonyan and Zisserman, 2014) as the initialization convolutional network. The network adjusts the fully connected layers into fully convolutional layers in the same ways as FCNs (Long et al., 2015) and adds dilation (Yu and Koltun, 2015) operations for the top three convolutional layers in order to enlarge the receptive field. Due to the existence of dropout in the VGG16 network, we eliminated a portion of parameters in the adjusted convolutional layers so as to reduce the number of parameters significantly. This basic network is pre-trained on ImageNet dataset (Russakovsky et al., 2015). We used the up-convolution operations as the decoding process. Compared with DeconvNet (Noh et al., 2015) and SegNet (Badrinarayanan et al., 2015) which use tens of up-convolution layers, we used only one up-convolution layer in order to reduce the number of parameters and to speed up the training. We regard this model as our baseline network.

Our baseline network is the FCN. By comparing results of state-of-the-art FCNs segmentation networks qualitatively, we concluded that almost all of them have inconsistent-labels and missing-labels problems. We assumed that the limitation comes from the lack of the global context information. Incorporating the global context into semantic segmentation facilitates better feature encoding, and it will lead to better feature encoding will lead to better and easier encoding. The global context is known to be very useful for detection and segmentation tasks in deep learning and has been explored in several works (Liu et al., 2015) (Mottaghi et al., 2014) (Szegedy et al., 2014). In order to merge global context information, we applied global average pooling to get the contextual embedding from the last convolutional layer and use element-wise sum operation to combine it with the final encoded features. Since different level features have different level numerical scales, we add a normalizing layer (Liu et al., 2015) before feature combination, which is a learnable scaling transformation in essence. However, because of the limitation of the data or ability of back-propagation (BP) algorithm, the global context branch may be not able to

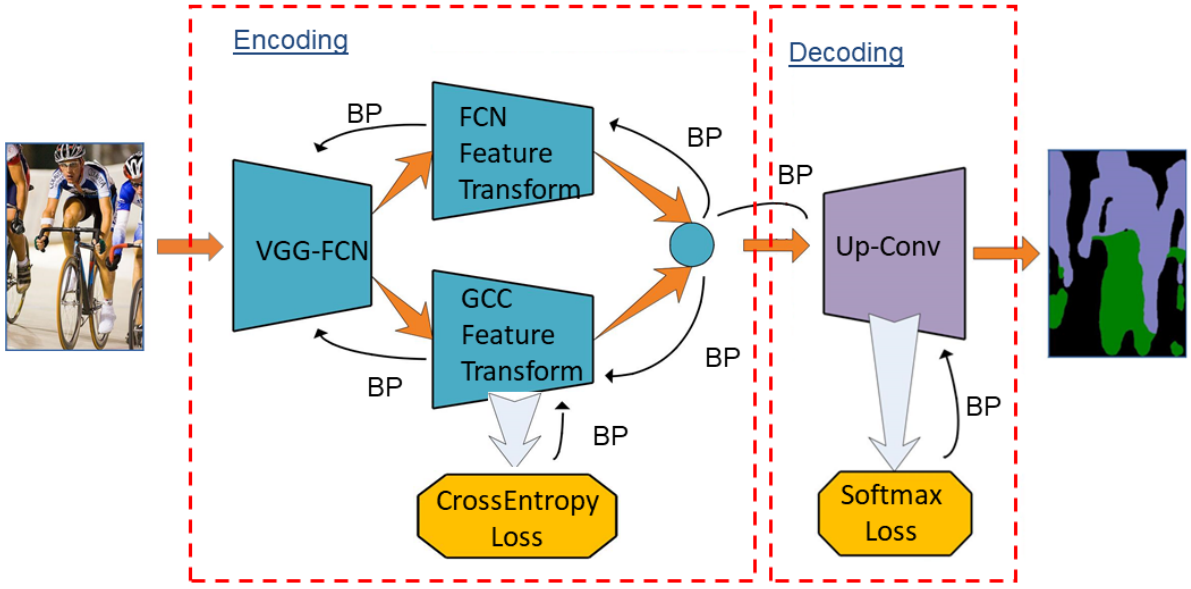


Figure 2: The architecture of Global Context Constraint Network. In the encoding process, the cross-entropy loss is employed to guide the global context features extraction and merge with fully convolutional features to obtain encoded features. And in the decoding process, up-convolution operation is utilized to decode the merged features to get segmentation results.

learn proper contextual information, so we added a constraint for the global context information to explicitly guide what the contextual information is to learn and this method leads to a huge improvement in the segmentation performance compared with the baseline network.

3.2 Global Semantic Context Constraint Encoding

Herein, we discuss how to constrain the global context and merge it into the encoded features. In order to solve the previous problems, we demand the network to encode the categories in a scene. So we define a multi-label classification loss which is a kind of cross-entropy loss to predict the possible categories in a scene and merge the predicted score of each category into the encoded features. The predicted scores denote the image contextual information. The objective loss function can be described as follows:

$$\begin{aligned} L_1 &= \frac{-1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} u_{i,k}\log\hat{p}_{i,k} \\ s.t. \quad \hat{p}_{i,k} &= \frac{1}{1+e^{-g_{i,k}}}, \quad g_i = t_1(W_1, I_i), \end{aligned} \tag{1}$$

where N and C are the batch size and the number of classes, respectively. $g_{i,k}$ denotes the global contextual score of image I_i for class $k \in \Omega$ (where Ω is a collection of all categories). And $\hat{p}_{i,k}$ denotes the probability of category k for image I_i in image level. We define $u_{i,k} = \mathbf{1}_{i,k}$ be the indicator function of Ω_i (all categories assigned to image I_i), then $u_{i,k} = \begin{cases} 1 & if \quad k \in \Omega_i \\ 0 & otherwise \end{cases}$. t_1 denotes the transformation of a neural network from the image I_i to global context constraint features and W_1 contains parameters in this transformation. We denote the fully convolutional features as follows:

$$h_i = t_2(W_2, I_i) \tag{2}$$

where t_2 denotes the transformation of a neural network from the image I_i to fully convolutional features and W_2 designates parameters in this transformation. We resized the global context scores to the same size as the fully convolutional features and use element-wise sum to get the final encoded features F_e.

3.3 Semantic Decoding

During the decoding process, we only use one up-convolution layer to decode the segmentation feature and use softmax loss to train the segmentation task. The objective loss function is defined as follows:

$$\begin{aligned} L_2 &= -\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{C}\mathbf{1}\{y_{ij}=k\}\,log(\hat{p}'_{ij,k}) \\ s.t. \quad \hat{p}'_{ij,k} &= \frac{exp(S'_{ij,k})}{\sum_{l=1}^{C} exp(S'_{ij,l})}, \quad S' = t_3(W_3, F_e), \end{aligned} \tag{3}$$

where M is the number of pixel in the image. $\hat{p}'_{ij,k}$ stands for the probability to assign category k for

Table 1: Evaluation results on PASCAL VOC 2012 test set.

Method	bkg	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	
FCN-8s (Long et al., 2015)	91.2	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	
ParseNet (Liu et al., 2015)	92.4	84.1	37.0	77.0	62.8	64.0	85.8	79.7	83.7	27.7	74.8	
DeepLab-CRF (Chen et al., 2014)	92.6	83.5	36.6	**82.5**	62.3	66.5	85.4	78.5	83.7	30.4	72.9	
DeconvNet-CRF (Noh et al., 2015)	92.9	**87.8**	**41.9**	80.6	63.9	67.3	**88.1**	78.4	81.3	25.9	73.7	
CRFasRNN (Zheng et al., 2015)	92.5	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	**78.2**	
GCCNet	**93.2**	85.5	38.6	81.4	**69.6**	**77.4**	84.8	**83.6**	**87.5**	**40.9**	78.0	

	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv		mean
	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1		62.2
	57.6	77.7	78.3	81.0	78.2	52.6	80.4	49.9	75.7	65.0		69.8
	60.4	78.5	75.5	82.1	79.7	58.2	82.0	48.8	73.7	63.3		70.3
	61.2	72.0	77.0	79.9	78.7	**59.5**	78.3	**55.0**	75.2	61.5		70.5
	60.4	**80.5**	77.8	**83.1**	80.6	**59.5**	**82.8**	47.8	78.3	67.1		72.0
	49.8	80.2	**78.7**	78.6	**83.5**	53.0	81.9	47.9	**80.5**	**70.9**		**72.6**

pixel j of image I_i in pixel level. S'_{ij} stands for the decoded feature vector of pixel j in image I_i, which is provided by the last up-convolutional layer. $S'_{ij,k}$ is the value in its kth channel. Let y_{ij} be the true label of pixel j in image I_i, we define $\mathbf{1}\{y_{ij} = k\}$ as the indicator function to judge whether the true label is k. t_3 denotes the decoding process from encoded feature F_e to final score map and W_3 contains parameters in this process.

Finally, we combine the two loss terms together, then the final loss function is given as $L = L_1 + L_2$. We can use the stochastic gradient descent algorithm to train our neural networks jointly.

Note that, from the perspective of the network structure, the proposed network is different from multi-task learning. We add the cross-entropy loss layer before the classification task. It can encode class-specific features before the pixel-level semantic classification.

4 EXPERIMENTS

In this section, we first describe our implementation details and experiments setup. Then, we analyze and evaluate the proposed network and make comparison with other methods.

4.1 Implementation Details

4.1.1 Dataset

We employed PASCAL VOC 2012 segmentation dataset (Everingham et al., 2010) for training and testing the proposed semantic segmentation network performance. Meanwhile, following (Long et al., 2015) (Chen et al., 2014), we employed the extra segmentation annotations from (Hariharan et al., 2011). Then, there are 10,582 training images and 1,449 testing images in our experiment. We employ images in original scale to maintain more details of objects, and images that are not compatible with (especially smaller than) the network's input sizes are resized, of which smallest dimension is less than 224 to 224 with a fixed ratio. The only data augmentation is to do randomly horizontal or vertical flip for images. Note that our experiment only uses PASCAL VOC 2012 augmented datasets for training and modified VGG16 as basic initialization network, whereas many state-of-the-art approaches also employ Microsoft COCO (Lin et al., 2014) which contains more than 80K images and deep residual networks (He et al., 2016) which more than hundreds of layers to improve performance.

4.1.2 Optimization

We implemented the proposed network based on Caffe (Jia et al., 2014) framework and utilize the optimization strategy mentioned in (Liu et al., 2015). We employ the stochastic gradient descent with momentum strategy and "poly" learning rate policy for optimization, where initial learning rate, momentum, power and weight decay are set to $1e-8$, 0.9, 0.9 and 0.0005, respectively. We initialized the weights in the basic convolutional network by using a modified version of VGG16 network (Liu et al., 2016) pre-trained on the ILSVRC (Russakovsky et al., 2015) dataset. We employed *Xavier* initialization (Glorot and Bengio, 2010) method for other convolutional layers and bilinear initialization for the up-convolutional layer. We used gradient accumulation method to update the weights every 8 iterations.

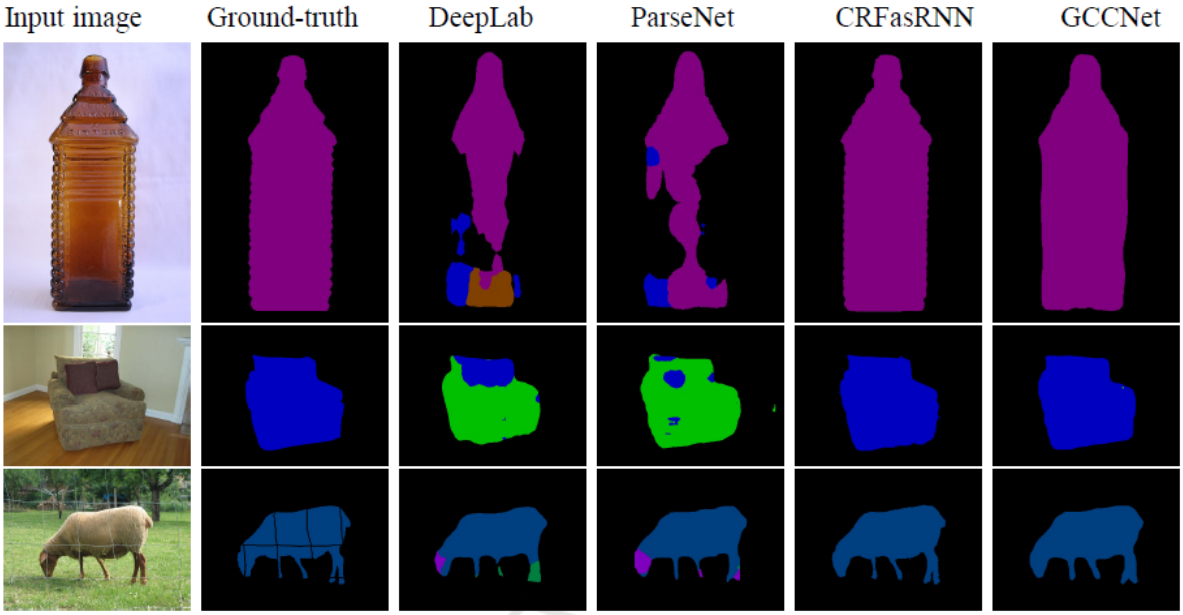


Figure 3: Example of semantic segmentation results on PASCAL VOC 2012 validation images. Note that the proposed method alleviates the confusion segmentation problem to some extent and have similar effect as CRFasRNN.

Table 2: Evaluation results on PASCAL VOC 2012 validation set for various network structure.

Method	mean IOU
Baseline	0.55
Baseline+context	0.664
GCCnet	0.721
GCCnet+Compactness	0.729

4.1.3 Inference and Refinement

During inference, we also used the original scale images but resize images whose smallest dimension is less than 224 to 224 with a fixed ratio. We get the final score maps from our network and employ ArgMax for each pixel to get final segmentation results. According to many other kinds of literature, we tried the fully connected conditional random field algorithm (Krähenbühl and Koltun, 2011) as post-processing to refine and smooth the segmentation result. We used a grid search to adjust the hyper-parameters of CRF. Interestingly, we obtained similar accuracy segmentation results before and after post-processing. The result supports the fact that the global context constraints network possesses the ability of region smoothness. We further compare our segmentation result with some state-of-the-art methods qualitatively as illustrated in Figure 3, experimental results show that our algorithm can achieve better performance than other methods.

4.2 Evaluation

We evaluate our network on PASCAL VOC2012 segmentation benchmark (Everingham et al., 2010), which contains 1449 validation images and involves 20 object categories. We adopt the comp6 evaluation protocol to measure performance by using Intersection over Union (IoU) method between ground truth and predicted segmentation. The quantitative comparison of the result between the proposed algorithm and the competitors is shown in Table 1. The performance of GCCNet is competitive to the state-of-the-art methods using PASCAL VOC dataset. We also compare the performance of GCCNet with our baseline networks trained on the same condition. As demonstrated in the Table 2, we can see that using contextual information leads to 11% improvement on mean IoU compared with the baseline network. And when we constrain the global context with cross-entropy loss, the performance reaches another 5.7% improvement. By using Compactness post-processing, the best performance can reach 72.9% mean IoU. Both of the quantitative and qualitative comparisons demonstrate that constrained global context information are able to lead better results than the baseline.

Further analysis showed that the addition of cross-entropy loss can reduce the miss-classification because the global class information can estimate the class information without considering the segmentation boundaries. Therefore, the global class information is computed more robustly than the pixel-wise

location information. Moreover, we believe that by adding global context constraint to other FCN extension networks, better result can be achieved.

5 CONCLUSIONS

In this work, we propose the global context constraint network, which allows the direct inclusion of global semantic context constraint for the task of semantic segmentation. We have explicitly demonstrated that relying on constrained global context features can largely improve the segmentation result and eliminate semantic segmentation confusion because global context constraint loss explicitly predicts the global context information that merged into the final encoded feature. The result presented on PASCAL VOC 2012 dataset shows that our approach can also reach the state-of-the-art performance at the same training conditions and its simplicity and robustness of learning makes it more advantageous.

ACKNOWLEDGEMENTS

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No. R7117-16-0164, Development of wide area driving environment awareness and cooperative driving technology which are based on V2X wireless communication.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. volume abs/1511.00561.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. volume abs/1412.7062.

Everingham, M., Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S., and Malik, J. (2011). Semantic contours from inverse detectors. In *International Conference on Computer Vision (ICCV)*.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

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

Hong, S., Noh, H., and Han, B. (2015). Decoupled deep neural network for semi-supervised semantic segmentation. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 1495–1503, Cambridge, MA, USA. MIT Press.

Hu, H., Lan, S., Jiang, Y., Cao, Z., and Sha, F. (2017). Fastmask: Segment multi-scale object candidates in one shot. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2280–2288.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA. ACM.

Krähenbühl, P. and Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 109–117. Curran Associates, Inc.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755, Cham. Springer International Publishing.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., and Berg, A. C. (2016). SSD: single shot multibox detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 21–37.

Liu, W., Rabinovich, A., and Berg, A. C. (2015). Parsenet: Looking wider to see better. volume abs/1506.04579.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., and Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the 2015 IEEE International Conference*

on Computer Vision (ICCV), ICCV '15, pages 1520–1528, Washington, DC, USA. IEEE Computer Society.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. volume abs/1409.1556.

Szegedy, C., Reed, S. E., Erhan, D., and Anguelov, D. (2014). Scalable, high-quality object detection. volume abs/1412.1441.

Wang, Y., Liu, J., Li, Y., Yan, J., and Lu, H. (2016). Objectness-aware semantic segmentation. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 307–311, New York, NY, USA. ACM.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. volume abs/1511.07122.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. S. (2015). Conditional random fields as recurrent neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1529–1537.