Fusion of Different Features by Cross Cooperative Learning for Semantic Segmentation

Ryota Ikedo and Kazuhiro Hotta
Meijo University, Japan

Keywords: Semantic Segmentation, Cooperative Learning, Multiple Backbones, Fusion of Different Features.

Abstract: Deep neural networks have achieved high accuracy in the field of image recognition. Its technology is expected to use the medical, autonomous driving and so on. Therefore, various deep learning methods have been studied for many years. Recently, many studies used a backbone network as an encoder for feature extraction. Of course, the extracted features are changed when we change backbone networks. This paper focused on the differences in features extracted from two backbone networks. It will be possible to obtain the information that cannot be obtained by a single backbone network, and we can get rich information to solve a task. In addition, we use cross cooperative learning for fusing the features of different backbone networks effectively. In experiments on two kinds of datasets for image segmentation, our proposed method achieved better segmentation accuracy than conventional method using a single backbone network and the ensemble of networks.

1 INTRODUCTION

Convolutional Neural Network (Krizhevsky, A., 2012) achieved high accuracy in various kinds of image recognition problems such as image classification (Szegedy, C.,2015)(Wang, F., 2017), object detection (Redmon, J. 2016)(Liu, W.,2016), pose estimation (Cao, Z., 2018) etc. In addition, semantic segmentation assigns class labels to all pixels in an input image. This task recognizes various classes at pixel level. Semantic segmentation using CNN is also applied to cartography (Isola, P., 2017)(Ronneberger, O., 2015), automatic driving (Chen, L.C., 2018) (Yang, M., 2018), medicine and cell biology (Ji, X., 2015)(Havaei, M., 2017). Especially in autonomous driving, it is necessary to instantly predict various classes such as people, cars and signs from in-vehicle images. Therefore, semantic segmentation technology is important to realize autonomous driving. In this paper, we work on semantic segmentation task for autonomous driving.
We proposed cooperative learning method (Ryota, I. 2021) as conventional study. Neural network is derived from the human brain structure. Cooperative learning was based on the group learning of humans. We used the learning method in a neural network. Basic cooperative structure is showed Figure 1.

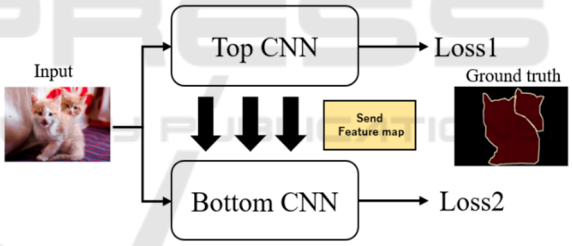

Figure 1: The structure of one-way cooperative network.

In Figure 1, we prepare two CNNs with the same structure. Then, we introduce paths between two networks for sending feature maps. Due to this structure, bottom CNN can obtain new feature maps from top network. Previous study used the exactly same CNN structure. In other words, previous cooperative learning is consulted with the same person. There is a problem that bottom CNN cannot get completely new information from top CNN. Therefore, we propose to give completely different feature maps in cooperative learning. We use two kinds of backbone networks and extract different features. Then, we use cross cooperative learning to effectively fuse those features obtained from different backbone networks. By sending the features mutually, each network has rich features and improves the accuracy.

Ikedo, R. and Hotta, K.
Fusion of Different Features by Cross Cooperative Learning for Semantic Segmentation.
DOI: 10.5220/0010889800003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 582-589
ISBN: 978-989-758-555-5; ISSN: 2184-4321

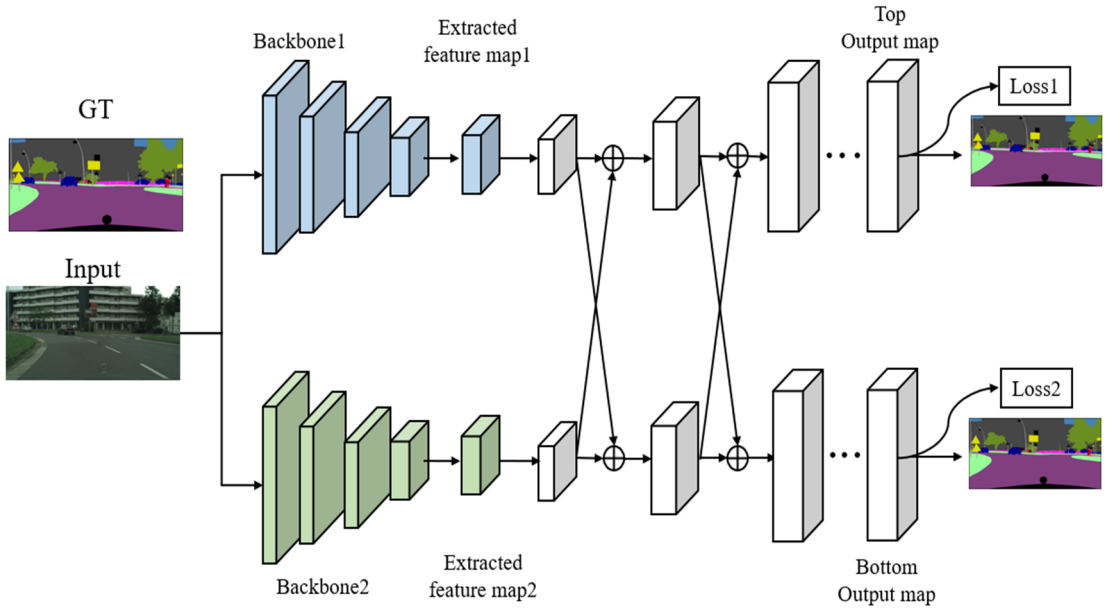

Figure 2: The overview of our cross cooperative network method.

We conducted the experiments on two kinds of famous datasets. The first dataset is the Pascal VOC 2012 (Everingham, M., 2010). The second one is the Cityscapes(Cordts, M., 2016) which is captured by in-vehicle camera. We see that the proposed method achieved higher accuracy than “single network”,” previous cooperative network” and ”the ensemble of networks”.

This paper is organized as follows. In section 2, we describe related works. The details of proposed method are explained in section 3. In section 4, we evaluate our proposed cross cooperative learning on segmentation tasks. Finally, we describe conclusions in section 5.

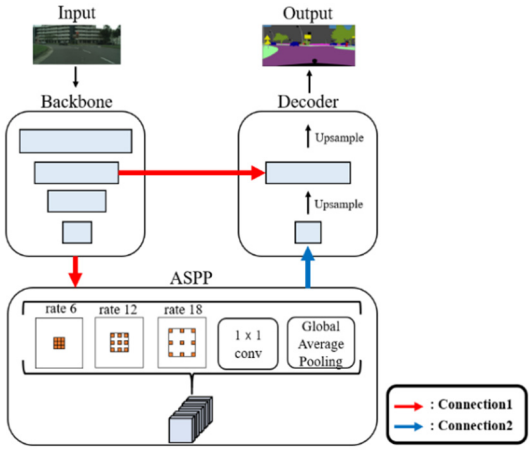

Figure 3: The structure of DeepLabV3+.

2 RELATED WORKS

The state-of-the-art approaches for semantic segmentation are based on CNNs. The famous approach is based on Fully Convolutional Network (FCN) such as SegNet (Badrinarayanan, V., 2017), U-net (Ronneberger, O., 2015) and so on. They had the simple structure of FCN but sharp accuracy improvements have been achieved by new architectures in recent years. One of the problems in semantic segmentation is that CNN lost spatial information by reducing the resolution in feature extraction process. Dilated convolution was proposed to solve this problem. It can extract the features while preserving spatial information by expanding receptive fields sparsely without reducing resolution. In the other works, PSPNet (Zhao, H., 2017) and DeepLab (Chen, L.C., 2018) proposed ASPP module. This module aggregates feature information at multiple scales. Thus, these works can get multi-scale contextual information and achieved high accuracy.

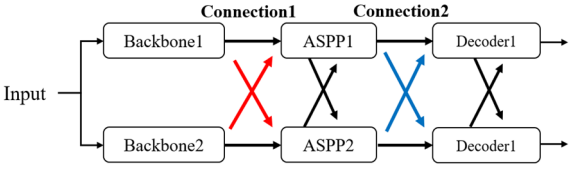

Figure 4: Cross cooperative connection in our method.

In the latest semantic segmentation, they used deep and large backbone network such as ResNet (He, K., 2016), VGG (Simonyan, K., 2014), Xception(Chollet, F., 2017). By using those backbone networks, recent works achieved high accuracy. For example, PSPNet used the ResNet101, and it showed high accuracy in in-vehicle dataset. DeepLabV3+ used ResNet101 or Xception for feature extraction at encoder. On the other hand, when we want to reduce inference time, we should use light backbone networks such as MobileNet (Sandler, M., 2018) and EfficientNet (Tan, M., 2019). As described above, many works use backbone network that suits their purpose.

Basically, features are different when we changed backbone networks. There is information which are easy to extract and difficult to extract by the kind of backbone network architecture. Therefore, it is important for us to select backbone architecture. In this paper, we focused on various information obtained from different backbone networks. We improved the segmentation accuracy by effectively fusing the features of different backbones.

3 PROPOSED METHOD

3.1 Overview

Conventional cooperative learning (Ryota, I., 2021) was used completely same two networks like figure1. In other words, this structure is learning by two exactly same persons. Thus, there was a problem that networks have only similar information even if we share feature maps between two CNNs. To overcome the problem, we propose that not use the same person's information but use the information of other persons in new cooperative learning. The aim of new cooperative learning is making a cooperative learning that two networks are different like another person each other. Therefore, we considered use different backbones in cooperative network to obtain different information in each network. In addition, we integrate different features from each backbone network by cooperative learning to solve a segmentation task.

In our method, we introduce different backbones. Different backbone networks can extract different features. But, there are easy to extract information and hard to extract information for backbone network. If we can supplement each with information from two different backbones by using cooperative learning, we can overcome this weakness.

In addition, using various kinds of features from two backbones, our method can use the features that a single backbone network cannot extract. From this above, we thought our method improves the segmentation accuracy.

We explain the details of networks in section 3.2. We explain the connection methods between two backbone networks in section 3.3.

3.2 Details of Network

Our proposed method was created based on DeepLabV3+. We use the ResNet and the Xception as backbones because the ResNet is used as the backbone network in many tasks and the Xception is higher extraction ability than the Resnet by separating channel convolution and spatial convolution. We show the overview of the proposed method in Figure 2. In particular, backbone1 is the Xception-65 and backbone2 is the ResNet-101. We used two backbone networks pretrained by ImageNet.

Next, we explain the structure of our cooperative learning in Figure 2. Previous cooperative connection was only one-way path from top network to bottom network (Cordts, M., et. al.,2016). Our method uses cross connections which can send feature maps each other. We introduce the cross connection because we would like to fuse two different information in top and bottom networks. We used this connection at all layers in decoders and ASPP module of DeepLabV3+. By using the connection, top and bottom networks are expected to obtain information that single network cannot have. Therefore, cross connection is more effective than conventional one-way cooperative network. Finally, we obtain two outputs from both CNNs for calculating losses. We use these two losses to let the network learn simultaneously for cooperative learning. We use SoftMax Cross Entropy (CE) as a loss function.

$$Loss = Loss1 + Loss2 \quad \cdots (1)$$

where Loss1 is the CE loss for Top CNN and Loss2 is that for Bottom CNN. Both losses are optimized simultaneously.

3.3 Connection Method

The structure of DeepLabv3+ is shown in Figure 3. This model has backbone and ASPP as encoder, and uses a decoder for predicting segmentation result. We introduce cross cooperative connections to each layer in the ASPP module and Decoder to effectively use the features from different backbones.

Table 1: Accuracy on the PASCAL VOC2012 dataset.

Method			Backbound	Bicycle	Boat	Bus	Cat	Cow	Dog	Moterbike	Pottedplant	Sofa	Tvmonitor	Aeroplane	Bird	Bottle	Car	Chair	Diningtable	Horse	Person	Sheep	Train	Mean IoU
Single Network	Backbone : Resnet		94.6	42.7	75.9	95.5	93.6	92.7	89.1	88.5	67.0	50.2	76.3	91.1	89.0	82.3	90.7	43.2	57.8	89.5	87.1	90.3	87.7	79.7
	Backbone : Xception		94.8	43.8	70.6	95.4	94.5	92.0	89.7	86.6	66.4	51.0	76.1	92.7	90.6	81.9	90.4	46.7	58.4	90.3	87.7	90.9	90.7	80.1
Cooperative Network	conventional method	Top	94.3	41.7	70.3	95.0	92.6	87.4	87.4	86.3	58.8	48.5	72.6	91.9	87.8	81.8	87.6	37.5	61.9	85.8	86.5	58.3	88.4	77.6
		Bottom	94.8	40.7	74.4	95.5	94.8	91.2	90.9	86.3	69.5	52.7	76.4	92.3	89.4	83.4	89.0	42.9	59.9	78.4	88.2	89.5	91.0	80.1
	ours	Top	95.2	42.9	74.9	95.4	94.8	93.0	90.7	90.1	70.5	59.8	78.5	89.1	90.9	82.2	92.2	46.1	62.6	91.0	88.4	91.0	92.0	**81.5**
		Bottom	95.2	42.4	74.2	96.0	94.5	92.8	90.9	90.4	70.2	59.2	77.2	88.5	91.2	82.7	92.4	45.8	61.7	90.7	88.6	91.4	91.8	81.3
	w/o all path	Top	94.9	42.9	71.3	95.3	95.1	92.8	90.8	88.5	69.2	50.5	76.3	92.6	91.1	82.2	89.6	43.7	60.6	90.5	87.6	90.3	90.1	80.3
		Bottom	94.2	43.0	70.9	95.0	93.7	90.3	88.0	86.5	69.8	52.9	75.6	91.2	89.7	82.7	87.3	44.4	58.5	89.8	87.8	86.9	88.9	79.4
		ensemble	95.0	42.7	72.9	95.8	95.6	91.3	90.8	88.1	71.6	51.9	77.2	91.3	91.0	82.7	89.3	42.8	60.8	90.7	88.4	90.0	91.4	80.5

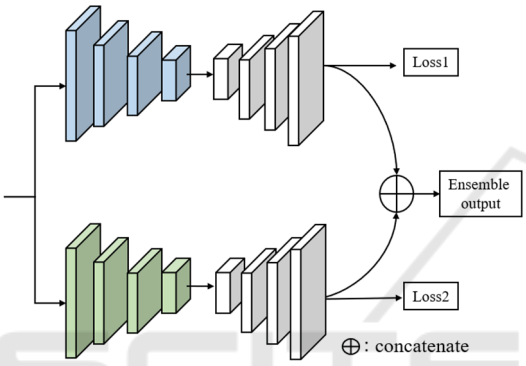

Figure 5: Ensemble output and ensemble method.

Connection1(Red Line) and Connection2(Blue Line) in Figure 3 are the outputs from backbone network and ASPP module. These connections have original information extracted from the encoder. Thus, if we also introduce cross cooperative connections after Connection 1 and 2 as shown in Figure 4, we can improve the accuracy further.

Table 2: Comparison with other models on the PASCALVOC2012 val set.

Method	Backbone	Mean IoU(%)
DeepLabV3	ResNet-101	76.5
WASPnet-CRF	ResNet-101	80.4
DFN	ResNet-101	80.6
HyperSeg-L	EfficientNet-B3	80.6
ResNet-GCN	ResNet-152	81.0
Ours	Res101+Xception	**81.5**

Therefore, we also add cross cooperative connections to the output of Connection 1 and 2. Cross cooperative connection after Connection 1 gives the information from two different backbones to each ASPP module. Cross cooperative connection after Connection 2 gives the information that enhanced by each ASPP to each decoder. This structure can be expected to provide more useful information for learning.

In experiments, we also evaluate the proposed method without Connection 1 and 2 to investigate the effectiveness of them.

4 EXPERIMENTS

In this section, we show experimental results. Section 4.1 describes the details of the dataset. Section 4.2 explains the implementation details. Section 4.3 shows the results on the PASCAL VOC dataset and section 4.4 presents the results on the Cityscapes dataset. Finally, in section 4.5, we show the comparison results about the connections.

4.1 Datasets

In this paper, we evaluate the proposed method using the PASCALVOC2012 and Cityscapes datasets. We describe the two datasets as follows.

4.1.1 Pascal Voc2012

This dataset includes various kinds of images. There are 10,582 images in training set, 1,449 images in validation set and 1,456 images in test set. These images involve 20 foreground object classes and one background class. In this study, we use validation set to get the best model. We evaluate the best model determined by validation set for test set. In addition, we randomly cropped the images of 513 × 513 pixels from training set, and we cropped a center region in validation and test phase.

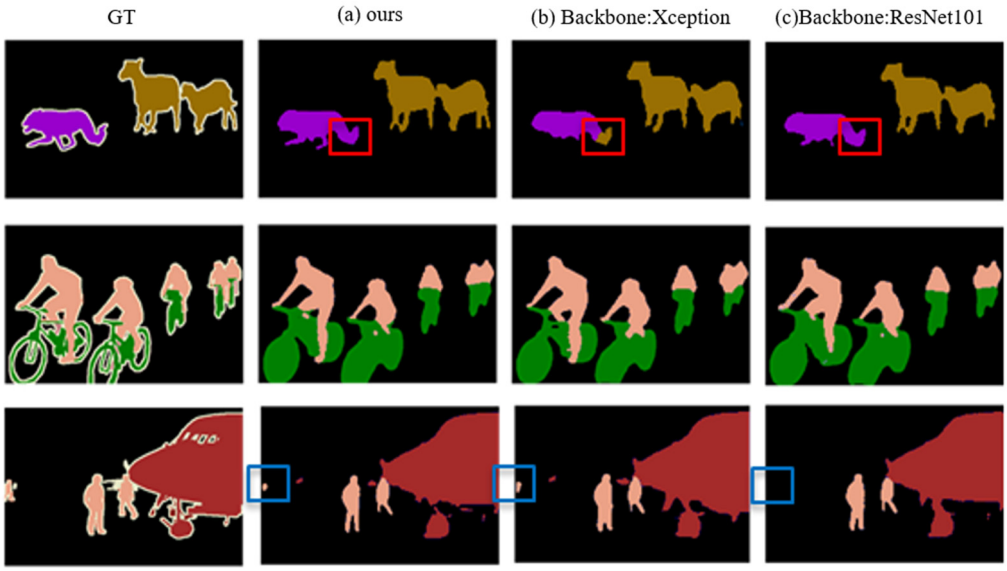

Figure 6: Segmentation result on the PASCALVOC 2012 dataset (val).

4.1.2 Cityscapes

This dataset includes the images captured by in-vehicle camera in Germany. All images are 2048 × 1024 pixels in which each pixel has high quality 19 class labels. There are 2,979 images in training set, 500 images in validation set. In this paper, we randomly cropped images of 768×768 pixels and used them for training.

4.2 Implementation Details

We implement our method by the Pytorch library and cross cooperative learning based on DeepLabV3+. To do fair comparison, we evaluated DeepLabV3+ and the proposed method under the same conditions on the same PC. We used single Deeplabv3+ (Resnet101 and Xception-65) and ensemble method model (figure 5) as a baseline for comparison. In our method, the batch size was fixed to 6 and SGD was used as the optimizer. The learning rate was set to 0.007 for PASCALVOC and 0.035 for Cityscapes. We used intersection over union (IoU) and mean IoU (mIoU) as evaluation measures.

4.3 Evaluation Result on Pascalvoc2012 Dataset

We evaluated the accuracy on validation set in the PASCAL VOC dataset. We compared four methods; a single network, the conventional cooperative learning, our proposed method, and our method without all cross cooperative paths between two networks. Our method has two outputs from top and bottom network. We have shown the results of each output in the Table 1.

The red number in Table 1 represents the maximum accuracy. We see that our proposed method achieved the highest accuracy in Table 1. The accuracy was improved more than 1.4% in comparison with a single network (Deeplabv3+). This result shows the effectiveness of the proposed method which fuses the features extracted from different backbones.

In conventional one-way cooperative learning (Cordts, M., et. al.,2016), we send the feature maps in top network to only bottom network. However, the method induced the accuracy difference between top and bottom networks, because top network cannot receive additional feature maps. Here we introduce cross connection to overcome the problem. Top network can get information of bottom networks Therefore, we achieved high accuracy in both networks. We see that the proposed method overcomes the weakness of the conventional method and improved the accuracy. Next, we reveal the effect of cooperative connection by comparing with the ensemble of two networks as shown in Figure 5. The ensemble method is just adding the final output to the outputs of two networks. Table 1 showed that our proposed method is 1.0% higher than the ensemble. This result indicated more effective than standard ensemble of two networks. Thus, our proposed cross connection is useful for fusing the feature maps of different backbone networks.

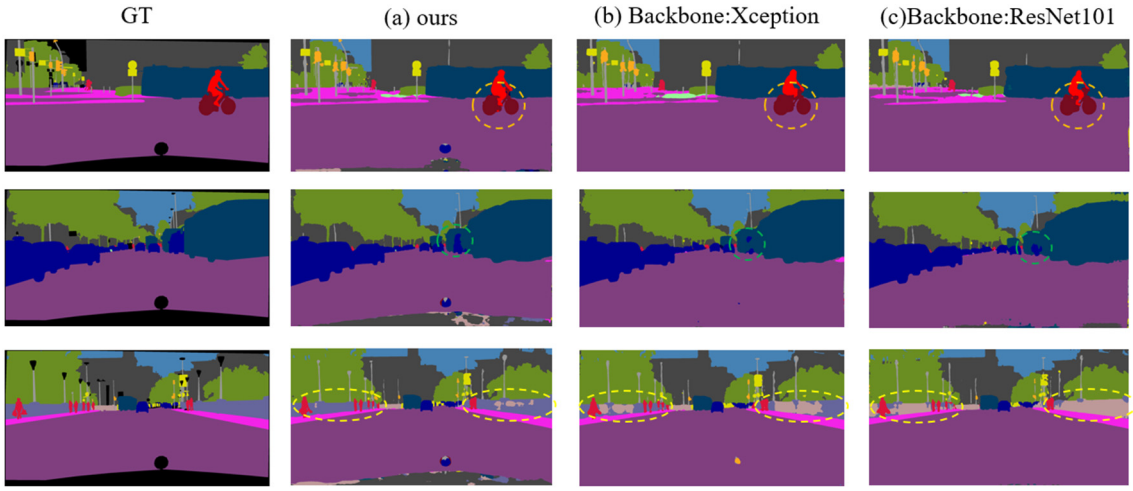

Figure 7: Segmentation results on the Cityscapes dataset (val).

Table 3: Accuracy on the Cityscapes dataset using DeepLabV3+.

Method	Backbone	Mean IoU%
DeepLabV3+(ours)	Res101	75.1
DeepLabV3+(ours)	Xception	76.2
Ours	Res101+Xception	**78.2**

Table 4: Comparison with other models on the Cityscapes val set.

Method	Backbone	Mean IoU(%)
WASPNet	ResNet-101	74.0
GSCNN	ResNet-101	74.7
U-Net++	ResNet-101	75.5
Dilated-ResNet	Dilated-ResNet-101	75.7
Multiscape DEQ	MEDQ-large	77.8
Ours	Res101+Xception	**78.2**

We show the comparison results with other segmentation models in Table 2. Conventional segmentation models have only one backbone network. Our method with two backbone networks achieved higher accuracy than those methods. These results demonstrated the effectiveness of the proposed cross cooperative learning method.

Figure 6 shows the segmentation results. In the first row, the red area is recognized correctly by single Deeplabv3+ using Xception, but not recognized by that with ResNet101. On the other hand, the blue area is recognized correctly by that with ResNet101 though Xception based Deeplabv3+ cannot recognize well. Many segmentation results showed the improvement of our method though we discovered a few results with bad influence. Thus, qualitative results also demonstrated that our proposed method could incorporate two feature maps effectively.

4.4 Evaluation Result on Cityscapes Dataset

We also evaluated the proposed method on the Cityscapes dataset. For fair comparison under the same condition, single Deeplabv3+ network was evaluated with own implementation.

Comparison results with Deeplabv3+ are shown in Table 3. Our method which uses DeepLabv3+ as a baseline improved over 2% on mIoU than single Deeplabv3+ using each backbone. Table 4 shows comparison results with the other segmentation models. Table 4 show that our method is higher accuracy than the other models which use ResNet-101 or Dilated-ResNet as backbone. These results showed that our method using feature fusion is more effective.

Figure 7 shows the segmentation results on the Cityscapes dataset. Similarly with the PASCALVOC 2012 dataset, the proposed method can incorporate the advantages of each backbone network. We can show that our method was also useful for improving the accuracy on another dataset.

4.5 Ablation Study

As introduced in Section 3-3, the proposed cross cooperative learning contains two additional cross cooperative connection at Connection 1 and 2. Therefore, we study their contributions on

PASCALVOC2012 dataset. As shown in Table 5, when we did not introduce additional connections, the accuracy was 80.36%. The gain of cross cooperative connection at Connection1 is 0.16%. When we add cross connection at Connection2 to our method, it boosted 1.15% in comparison with the proposed method without additional connections. Especially, ASPP improved the feature extraction ability by performing some dilated convolutions and pooling, and it can obtain beneficial feature maps. Therefore, cross connection at Connection1 brings good effect to ASPP and cross connection at Connection2 brings good effect in decoding the extracted information. These results demonstrated the effectiveness of the additional cross cooperative connection.

5 CONCLUSION

In this paper, we proposed new cooperative learning method by fusing the features of different backbone networks for semantic segmentation. Especially, we used cross cooperative learning with two different backbones, and our method improved the conventional cooperative learning. We confirmed that our method improved the segmentation accuracy on the PASCAL VOC2012 dataset and the Cityscapes dataset.

The proposed cross cooperative network used much calculation resource because our method needs multiple backbone networks. Therefore, we would like to realize the cross cooperative learning with lower computational cost and high accuracy. This is a subject for future works.

ACKNOWLEDGEMENTS

This paper is partially supported by JSPS KAKENHI 18K11382.

REFERENCES

Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep con-volutional neural networks. In:Advances in Neural Information Processing Sys-tems. 1097–1105 (2012)

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2017)

Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once:unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788 (2016)

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision. pp. 21–37. Springer (2016)

Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)

Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7291–7299 (2017)

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017)

Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)

Chen, L.C., Collins, M., Zhu, Y., Papandreou, G., Zoph, B., Schroff, F., Adam, H., Shlens, J.: Searching for efficient multi-scale architectures for dense image prediction. In: Advances in Neural Information Processing Systems. pp. 8699–8710 (2018)

Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3684–3692 (2018)

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical image analysis 35, 18–31 (2017)

Ji, X., Li, Y., Cheng, J., Yu, Y., Wang, M.: Cell image segmentation based on an improved watershed algorithm. In: 2015 8th International Congress on Image and Signal Processing. pp. 433–437. (2015)

Ryota, I. and Kazuhiro, H.: Feature Sharing Cooperative Network for Semantic Segmentation. In: Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 577-584. (2021)

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision 88(2), 303–338 (2010)

Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12), 2481–2495 (2017)

Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017)

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4), 834–848 (2017)

Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 801–818 (2018)

He, K., Zhang, X., Ren, S., Sun, J: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)

Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scaleimage recognition. arXiv preprint arXiv:1409.1556 (2014)

Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258 (2017)

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L. C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 4510-4520 (2018)

Tan, M., Le, Q. V: Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International conference on machie learning. pp. 6105-6114 (2019)

Artacho, B., Savakis, A.: Waterfall atrous spatial pooling architecture for efficient semantic segmentation. Sensors, 19(24), 5361 (2019).

Takikawa, T., Acuna, D., Jampani, V., Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision pp. 5229-5238 (2019)

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In Deep learning in medical image analysis and multimodal learning for clinical decision support pp. 3-11(2018)

Bai, S., Koltun, V., Kolter, J. Z. Multiscale deep equilibrium models. In Neural Information Processing Systems, pp.33, (2020)

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1857-1866) (2018)

Nirkin, Y., Wolf, L., Hassner, T. Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4061-4070 (2021)

Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J. Large kernel matters--improve semantic segmentation by global convolutional network. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 4353-4361 (2017)