View Decomposition and Adversarial for Semantic Segmentation

He Guan[1,2,4] and Zhaoxiang Zhang[1,2,3,4]

[1] University of Chinese Academy of Sciences
[2] Research Center for Brain-inspired Intelligence, CASIA
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
[4] National Laboratory of Pattern Recognition, CASIA
{guanhe2015,zhaoxiang.zhang}@ia.ac.cn, tnt@nlpr.ia.ac.cn

Abstract. The adversarial training strategy has been effectively validated because it maintains high-level contextual consistency. However, limited to the weak capability of a simple discriminator, it is irresponsible and unreasonable to identify one from the sample source at a time. We introduce a novel discriminator module called Multi-View Decomposition which transforms the discriminator role from general teacher to specific adversary. The proposed module separates single sample into a series of class inter-independent streams and extracts corresponding features from current mask. The key insight in the MVD module is that the final source decision can be aggregated from all available views rather than a harsh critic. Our experimental results demonstrate that the proposed module can improve performance on PASCAL VOC 2012 and PASCAL Context dataset further.

Keywords: View decomposition · Adversarial · Semantic segmentation.

1 First Section

1.1 Introduction

Semantic segmentation is a fundamental computer vision problem where the goal is to assigns all pixels into different semantic classes. It enjoys a wide range of applications such as self-driving systems and scene parsing. There have been some recent efforts on adapting fully convolutional network for semantic segmentation and achieved state-of-the-art performance. Some of these work proposed various exclusive network architectures to embed higher-level convolutional features from multiple layers in CNN. Others are based on a variety of post-processing methods by integrating higher-order potentials as much as possible to enforce spatial contiguity in score maps. In both cases, the final semantic predictions attempt to overcome the limitation of the barrel-shaped network architecture and enhance the correlation among neighboring nodes.

Generative Adversarial Network is a powerful approach to identify the authenticity of sample. We expect the desirable state if a generated sample can

be confused with a genuine sample. In the process of alternating iterations, discriminator tends to converge rapidly while generator is hard to update. On the contrary, it reverses the dilemma when the generator is initialized with a well pre-trained segmentation model [12]. With the enhancement of model robustness and generalization, the appearance gap between real sample and synthetic ones is gradually disappearing and the discriminator will quickly obtain the local optimal solution. Compared with the standard GAN, the gradients of feedback are less of deliberation and more like random allocation in this state.

Based on above observation, we hypothesis that multi-discriminator variants can collect more specific representations from different subspace and approximate $max_D V(D, G)$ better by ensemble learning and propose a novel discriminator to deal with the above shortcoming. Different from previous work, we make a series of class inter-independent streams for each category and aggregate them into a more harsh critic via global optimization. The final paradigm is a pattern of one generator against one discriminator with multi-view. Even if the semantic segmentation model produce relatively realistic labeling decision, it can still exploit the potential differences in Contextual details and avoid misleading guidance. Extensive experiments on two benchmark datasets demonstrate the superiority of our Multi-View Decomposition module.

2 Related Work

Following Long [11] replaced fully-connected layers in classification into convolutional layers, various FCN architectures have made breakthroughs constantly in semantic segmentation task. To make use of global prior from complex objects and scenes, some methods [9, 18] extract Context information using global pooling branches while others embed or combined the feature map from multi-scale [15, 17] and multi-resolution [2, 8].

Recently, Various GAN variants are not satisfied with the structural limitations of one generator against one discriminator. Nguyen et al. propose D2GAN model [14] with different rewards using the complementary statistical properties of Kullback-Leibler divergence and reverse-KL divergence for single discriminator against double discriminators. Both the Multi-View GAN [4] and GMAN [5] extend to the universal paradigm of single generator against multiple discriminators. As for apply adversarial training to semantic segmentation task, Luc et al. [12] take the lead to combine both and prove its effectiveness by optimize a general multi class cross-entropy loss with an additional adversarial term. Natalia et al. [13] expand his work on predicting semantic segmentation maps in future video frames. Souly et al. [16] replace the discriminator with baseline segmentation model rather than generator to leverage from generated output or unlabeled data. These work shows that GAN framework can flexibly extend to semantic segmentation task.

The purpose of GANs above are to enhance the robustness and generalization for the generator and obtain enough realistic samples to deceive the discriminator. Due to the situation of image blurring, deformation and distortion, it is

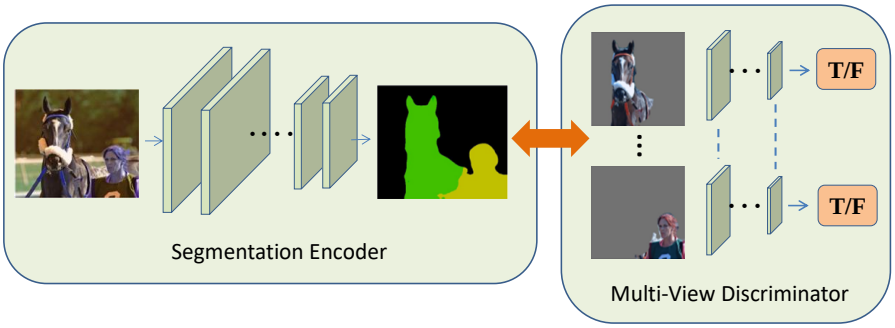

Fig. 1. Overview of the adversarial training using Multi-View Decomposition module. Given an input image, we first feed it to segmentation generator to product a dense prediction map, then the MVD module is applied to extract independent features from one-hot encoding mask on image, followed by concatenation layers and the rest of convolution layers to form our discriminator.

obviously easy for simple classifier to refuse these fake case. The difference is that the rough segments discard all high-frequency details except the structural relationship. It is not representative enough to pick out the fakes once the generation results has a rough prototype. Therefore, we exploit the capability of discrepancy by multiple discriminators collaborate via our Multi-View Decomposition module.

3 Approach

In this section, we propose a effective module as feature extractors called the Multi-View Decomposition as the discriminator to discern faint differences. An overview of the MVD module is shown in Fig 1. Our method has few differences from advances in several architectural choices for the discriminator such as convolutional Patch-GAN [7]. Unlike past work, the first few layers of MVD are replaced to penalizes large portions of image patches. Here we show that the multi-view learning method captures the characteristics of a particular category through a dedicated hierarchical structure. The timing of separation and fusion is explored in Section 4.3.

The MVD module is motivated by the following observations: The structure of semantic graph only contains spatial continuity information about the objects without enough precise texture details. It is no significant difference when one-hot coding of the ground truth ones or the probability maps as the input no matter whether the corresponding RGB images or not. Part of the explanations for this low gain is that the discrete or continuous maps have no enough driving force for discriminator to distinguish them. However, the essential aim of introducing adversarial training is to promote the segmentation network output

more realistic and reasonable. Thus, it should not be limited its potential by the trivial judgment.

The discriminator of our network has eight convolution layers and we split front-end network as the MVD before the k-th convolutional layer ($k = 2, 3, 4$). Excessive proportion of this module will leads to some signal dominates the others if propagate to subsequent layers. It is also necessary to extract features by hierarchical branches according to categories because the next following layer will mix the inter-class features immediately, which is equal to directly input the original RGB images or label maps. Weights sharing is one effective method for learning approximate features and greatly increases the versatility of the filters. We expect MVD to be more sensitive in local differences than stacking multiple convolution blocks by single stream, so the weights of different branches are not shared. The cost is we need to slightly increase the scanning frequency for one image. We believe it is more likely to train alternately from distinct views of the sample and jointly optimize the others to maximize the consensus in a series of subspace.

We assign a multi-class cross-entropy loss L_{mce} and a auxiliary adversarial regularization loss such as L_1 to approximate between the model output $\hat{Y}$ with ground truths Y. An additional weight is responsible for balancing the influence of auxiliary losses. We provides the optimal balance value using $\lambda = 0.5$.

$$\mathcal{L}_G(\hat{Y}, Y) = \mathcal{L}_{mce}(\hat{Y}, Y) + \lambda \cdot \mathcal{L}_{L1}(D(\hat{Y}), \alpha) \tag{1}$$

$$\mathcal{L}_D(\hat{Y}, Y) \ = \mathcal{L}_{L1}(D(Y), \alpha) + \mathcal{L}_{L1}(D(\hat{Y}), 0) \tag{2}$$

Given the number of annotations class C and input RGB image size $H \times W$, we minimize the $\mathcal{L}_G$ to produce more precise segmentation maps to fool the adversarial model by generative part, while the $\mathcal{L}_D$ is trained to ferret out the attacker from the pairs of feeding samples by the discriminator D. Using $|\cdot|$ denotes the absolute value function following with spatial position i, j and class index c. In addition, one-side label smoothing technique [1] is used to reduce the vulnerability of the neural network to counterattacks. We restrict positive cases with $\alpha = 0.9$ and negative cases with $\beta = 0$ because moderately positive sample smoothing is helpful to obtain stronger gradient feedback.

Since convolution with stride change can maintain more details than max pooling, we implement down-sampling operation by stride convolution layer rather than max-pooling layer. Convolution-ReLU form is also used to most layer except the first and final one. After eight convolution layers with 3×3 spatial filters, the fields-of-view has reached 34×34 pixels at the top of the discriminator which is similar to LargeFOV [12]. Note that current state-of-the-art segmentation network generally shrink eight times as output scale and zoom back to original size during test [3, 17, 18]. We do not expand the receptive field deliberately in the MVD further since it is sufficient to detect the sharpness of class boundaries and avoid tiling artifacts. Beyond this scale, to cover larger image patch, will bring considerably worse results. This may be because this front-end has many more parameters than before in the same depth of the architecture, which may be harder to converge.

4 Experiments

In this section, we first briefly describe the implementation details and the baseline model. Then we evaluate our method on two standard benchmarks: PASCAL VOC 2012 dataset and PASCAL Context dataset. In the end, we give a contrast verification on the impact of the module fusion depth and visual results analysis.

4.1 Implementation Details and Baselines

All experiments are done on the open source framework **Tensorflow**. The training optimizer selects Adam for generator and SGD for discriminator. Moreover, the learning rate is set to 0.0001 and the momentum remains 0.9. For the PASCAL VOC 2012 and PASCAL Context dataset, we randomly crop a 321×321 region from each image during training and substrate image pixel mean to normalize at every position. In contrast, we generate a broader coverage map of 512×512 then crop it to remove useless areas during test.

We have tried three alternative updating strategies to explore the effect of the switching speeds between G and D. The fast version implements a $1:1$ switching rate, while the constant version prefers 20 iterations intervals and the last one corresponds to slow frequency version such as $1:500$. Through verification, we found that switching from the slow to constant mode during training phase can help accelerate the network convergence.

We only focus on the extra gain by using our novel module in adversarial training so that the generator can be transferred into any semantic segmentation baseline model directly. Here we employ the state-of-the-art semantic segmentation network of [17] or [3]. One concern is that the former can compare with [12] and the latter can apply more pressure on the discriminator to test its strength. Both [3, 17] are based on the VGG-16 pre-trained model and apply several dilation convolution layers as substitutes for max-pooling. Additional Context module or multi-scale convolution branches can capture both short and long range contexts by expand its receptive fields.

The rest of the network parameters are initialized by COCO data pre-trained, which can suppress the interference caused by the cold start of the discriminator. Then we turn on branches alternately to model each particular class and follow the co-training strategy in multi-view learning to maximize the mutual agreement, which is similar to fine-tune current path only on current class information. More frequent alternating scheme is more applicable for solving the class imbalance problem.

4.2 Datasets

PASCAL VOC 2012 is a generic segmentation benchmark, which contains 20 categories and the background. Following common practice [3, 10, 11], we use augmented data with extra SBD [6] resulting 10582 images for training. We also respectively validate and test on the original 1449 and 1456 images. PASCAL Context dataset contains 10103 images with 540-classes dense label, which is split

Table 1. Comparison results on the PASCAL Context using DeepLab-VGG. The preceding symbol † indicates fine-tuned on PASCAL Context dataset using original environment configuration.

Method	mIoU	Pixel Acc.
DeepLab-v2†	41.0	63.7
Ours(with BCE)	41.1	63.7
Ours(with L1)	**41.5**	**64.4**

Table 2. Comparison results on different dataset for fusion depth.

Fusion Depth	1st	2nd	3rd	4th	Avg
Pascal VOC mIoU	71.94	72.31	**72.86**	70.09	72.01
Pascal Context mIoU	41.0	41.2	**41.5**	40.4	40.9

to 4998 images for training and 5015 images for validation. We only consider the most frequent 60 classes (including background) for training and evaluation, regardless of those low-frequency object. As for evaluation, both mean intersection over union (mIoU) and pixel-wise accuracy (Pixel Acc.) are used. More results are shown in Table 1 and Table 3.

4.3 Fusion Strategy

We compare the adversarial training results by modifying the module depth in discriminator, so as to seek out the best-fit fusion depth in variations. The results are showed in Table 2. It is obvious that stacking more layers gives slight improvements over the shallower fusion mode in contrast with a negative effect on network once beyond the upper limit. We believe that excessive proportion of independent channels will leads to some signal dominates the others if propagate to subsequent layers. Besides, we further evaluate an additional framework that the final confidence score map is averaged by multiple discriminant outputs (Avg) at each position, which become a boosting algorithm in this extreme case. The module depth is eventually set to 3 to trade off the independent single-view feature extraction process and the multi-view feature integration process. Note that it doesn't matter about the receptive field scale, regardless of the merge position.

4.4 Visual Analysis

Here we analyze several visual examples shown in Fig 2, and demonstrate that multi-view decomposition and adversarial strategy can well cope with such problems as insufficient coherence of semantics and the misclassification of the pixels inside objects. As we observed that the completed human torso under the sleep cat (1st); or the correction of a wide range of sea background semantics (4th). However, there are still some samples that are corrected only in weak areas, especially when the fake samples were highly smooth such as shaded bus (2nd) or coupled girl bodies (5th).

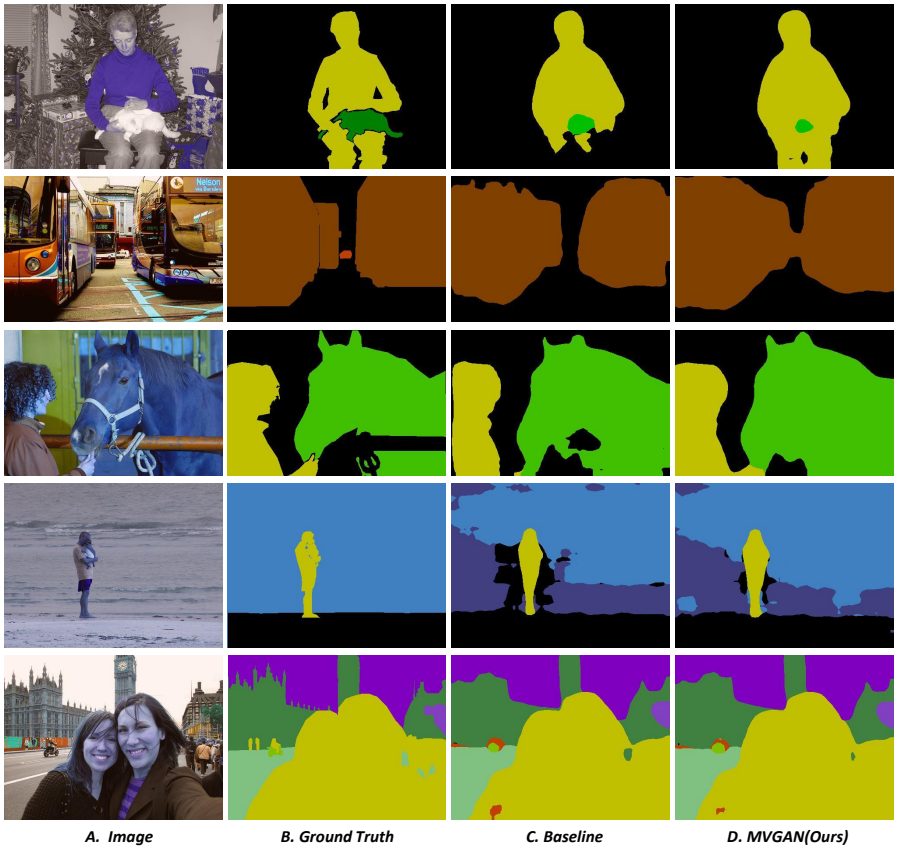

Fig. 2. Visual quality comparison results on different datasets.

Table 3. Per-class results on the PASCAL VOC 2012 test set

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
Dilation-8	87.2	38.2	84.5	62.3	69.7	88.0	82.3	86.4	34.6	80.5	60.5	81.0	86.3	83.0	82.7	53.6	83.9	54.5	79.3	63.7	73.1
SSGAN	87.1	38.5	84.9	63.2	69.7	88.0	82.5	86.8	34.5	80.3	61.5	80.9	85.8	83.3	82.6	55.0	83.5	54.7	79.7	62.9	73.3
Ours	87.2	38.7	84.8	63.5	69.8	88.4	82.5	87.1	34.7	80.4	61.7	81.3	86.2	83.4	83.3	54.8	83.8	54.6	79.9	63.8	73.7

5 Conclusion

Following the principle of multi-view learning, we propose a new module for adversarial training by update particular features alternately on split segmentation maps. It concentrates more on intra-class features and exploits the redundant views of the same input data to optimize the discriminant state without reducing the regularization property of higher-order statistics too much. Our results demonstrate that our proposed module enhance the discriminator in semantic segmentation task and still can improve the performance on several datasets even if the segment network enough powerful. For future work, we plan to ex-

plore multi-view adversarial training on different attributes of homologous data such as age, gender or expression of human face.

Acknowledgement

This work was supported in part by the National Key R&D Program of China(No. 2018YFB1004600), the National Natural Science Foundation of China (No. 61773375, No. 61375036, No. 61602481, No. 61702510), and in part by the Microsoft Collaborative Research Project.

References

1. M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.
2. A. Bansal, X. Chen, B. Russell, A. G. Ramanan, et al. Pixelnet: Representation of the pixels, by the pixels, and for the pixels. *arXiv preprint arXiv:1702.06506*, 2017.
3. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 2018.
4. M. Chen and L. Denoyer. Multi-view generative adversarial networks. In *ECML*, 2017.
5. I. Durugkar, I. Gemp, and S. Mahadevan. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016.
6. B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.
7. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
8. G. Lin, C. Shen, A. Van Den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.
9. W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.
10. Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *ICCV*, pages 1377–1385. 2015.
11. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
12. P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. In *NIPS Workshop on Adversarial Training*, 2016.
13. P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun. Predicting deeper into the future of semantic segmentation. In *ICCV*, 2017.
14. T. Nguyen, T. Le, H. Vu, and D. Phung. Dual discriminator generative adversarial nets. In *NIPS*, 2017.
15. B. Shuai, T. Liu, and G. Wang. Improving fully convolution network for semantic segmentation. *arXiv preprint arXiv:1611.08986*, 2016.
16. N. Souly, C. Spampinato, and M. Shah. Semi and weakly supervised semantic segmentation using generative adversarial network. *arXiv preprint arXiv:1703.09695*, 2017.
17. F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.
18. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017.