Towards Cross-View Consistency in Semantic Segmentation While Varying View Direction

Xin Tong, **Xianghua Ying***, **Yongjie Shi**, **He Zhao** and **Ruibin Wang**
Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
{xin_tong, xhying, shiyongjie, zhaohe97, robin_wang}@pku.edu.cn

Abstract

Several images are taken for the same scene with many view directions. Given a pixel in any one image of them, its correspondences may appear in the other images. However, by using existing semantic segmentation methods, we find that the pixel and its correspondences do not always have the same inferred label as expected. Fortunately, from the knowledge of multiple view geometry, if we keep the position of a camera unchanged, and only vary its orientation, there is a homography transformation to describe the relationship of corresponding pixels in such images. Based on this fact, we propose to generate images which are the same as real images of the scene taken in certain novel view directions for training and evaluation. We also introduce gradient guided deformable convolution to alleviate the inconsistency, by learning dynamic proper receptive field from feature gradients. Furthermore, a novel consistency loss is presented to enforce feature consistency. Compared with previous approaches, the proposed method gets significant improvement in both cross-view consistency and semantic segmentation performance on images with abundant view directions, while keeping comparable or better performance on the existing datasets.

1 Introduction

Semantic segmentation, which aims to label each pixel of a given image with a certain semantic class, is one of the fundamental missions in computer vision. Benefiting from the strong feature learning ability of the Convolutional Neural Networks (CNNs) and rich datasets, previous semantic segmentation methods have obtained promising performances. However, the generalization ability related to varying view direction of semantic segmentation algorithms has not been paid enough attention yet. When it comes to semantic segmentation, an intuitive motivation is that if we observe the same object via different view directions at the same viewpoint, the semantic labels of the object should be the same.

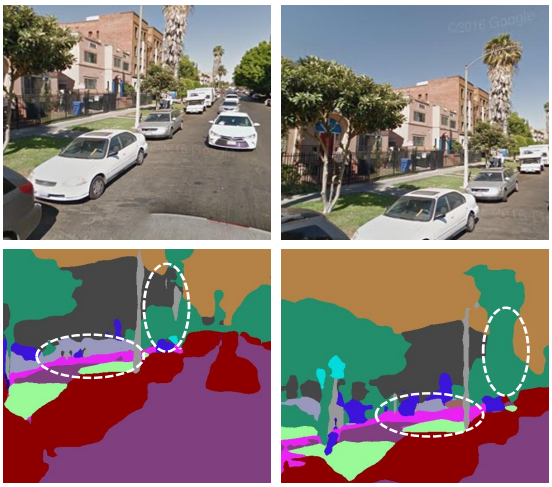

Figure 1: Two images of the same scene grabbed from "Google Street View", by keeping the viewpoint unchanged but only varying the view direction in the browser. Their semantic segmentation results are obtained by the same PSPNet method [Zhao *et al.*, 2017]. One may easily find that inferred labels in many parts of the two images are ambiguous for the same things.

We call it cross-view consistency while varying view direction. The consistency is crucial to keep stable and robust results in higher-level computer vision tasks such as 3D reconstructions, autonomous driving, AR/VR, etc. Unfortunately, as shown in Fig. 1, we find the existing semantic segmentation algorithms are often not able to keep cross-view consistency while varying view direction. Moreover, images in popular urban scene datasets such as Cityscapes [Cordts *et al.*, 2016] and CamVid [Brostow *et al.*, 2008] are acquired by mounted cameras in moving vehicles. View directions of these images are often near urban street directions, i.e., along driving directions, meanwhile images with other view directions are a minority. The CNN-based models trained on these datasets will be constrained and overfit in specific view directions due to the lack of data with diverse view directions. Specifically, the segmentation performance drops a lot on images with richer view directions. Meanwhile, the inconsistency problem is more severe in urban scenes.

In this work, we aim to modify existing CNN-based semantic segmentation algorithms by enhancing their ability re-

*Corresponding Author

lated to cross-view consistency while varying view direction. The modification can be easily applied in any CNN-based semantic segmentation methods. No extra annotated images are necessarily needed during training, since new manually annotated images for semantic segmentation are usually costly and time-consuming to acquire. In the modification, we propose three general add-on modules. We first present an online varying view direction (VVD) data generator to provide images with abundant view directions for training. The key insight is to utilize carefully generated homography to generate new image of the same scene in another view direction and get correspondent pixel pairs. Then we propose gradient guided deformable convolution module, which can learn dynamic proper receptive field from feature gradients. Furthermore, a novel consistency loss is proposed to enforce the consistency between features at corresponding positions in different feature maps. For evaluation, We derive VVD datasets which contain annotated images with diverse view directions from common-used urban datasets. The proposed method efficiently improves semantic segmentation methods, especially in terms of the cross-view consistency.

The contributions of this paper can be summarized below: (1) Our work is the first one that focuses on the problem that previous semantic segmentation methods may not keep cross-view consistency even if the viewpoint is unchanged and only the view direction is varied; (2) To alleviate the inconsistency, we propose a novel online varying view direction (VVD) data generator to dynamically generate images in different view directions using carefully generated homography. The generated images can be treated as real photos of the scene taken by the camera with a new direction exactly. We also construct VVD datasets that contain more challenging annotated images in plentiful view directions for evaluation; (3) We propose a gradient guided deformable convolution module and a novel consistency loss, which can be added to segmentation network easily and enforce network to predict consistent semantic predictions; (4) The proposed method achieves significant improvement on images with abundant directions in cross-view consistency and segmentation performance, while keeping comparable or better performance on common-used datasets.

2 Related Work

Semantic segmentation algorithms. Semantic segmentation methods based on convolutional neural networks have got promising performance. The seminal work [Long *et al.*, 2015] proposed fully convolutional network (FCN) to train the segmentation network end to end and fit for images of any size. [Chen *et al.*, 2017a] and [Yang *et al.*, 2018] applied Atrous/Dilated convolution to enlarge receptive field. Global information and multi-scale features were incorporated to boost the network in [Zhao *et al.*, 2017; Ronneberger *et al.*, 2015; Chen *et al.*, 2016]. Conditional random field (CRF) was used as post processing in [Chen *et al.*, 2017a] to refine the predictions or embedded into the network in [Zheng *et al.*, 2015] to enable end-to-end training. Recent approaches used attention [Huang *et al.*, 2019; Fu *et al.*, 2019; Li *et al.*, 2019; Zhao *et al.*, 2018; Yu *et al.*, 2018], dictionary learning [Zhang *et al.*, 2018], context relationship [Yuan and Wang, 2018; Zhang *et al.*, 2019] and get better results. We focus on improving the cross-view consistency in CNN-based image semantic segmentation, which has not been paid enough attention in the previous methods.

Cross-view consistency. Cross-view consistency is a basic and natural constrain in computer vision. It requests each projection of the same 3D object represents the same semantic meaning. Xiao *et al.* [Xiao and Quan, 2009] proposed a graph-based optimization approach to enforce consistency of the segmentation result across multiple views. Ma *et al.* [Ma *et al.*, 2017] warped feature maps into a common reference view and enforced multi-view consistency with various constraints based on RGB-D images. In this work, we use the cross-view consistency in the condition that the view direction is varied and the viewpoint is unchanged. The consistency can be defined by a homography transformation. We apply it in semantic segmentation of 2D RGB images by evaluating and enhancing the cross-view consistency while varying view direction of any given CNN-based algorithm.

Deformation of convolution. Deform the convolution for a better receptive field has been well studied. [Jaderberg *et al.*, 2015] proposed a Spatial Transformer Network (STN) which estimates a group of global parameters to warp the input feature maps. [Jeon and Kim, 2017] proposed a convolution unit with learned offsets to obtain better receptive field for object classification, by learning fixed offsets for feature sampling on each convolution. [Dai *et al.*, 2017] proposed a a more dynamically deformable convolution unit where the image offsets are learned through a set of parameters. The offsets in above methods are learned from the feature maps directly. For our task, when view direction varies, the local receptive fields are supposed to have similar semantic information with a local spatial deformation, e.g., homography transformation. The changes of feature map gradients can reflect the deformation directly. Thus, the offset is learned from the gradient of the feature maps in our approach.

3 Algorithm

In this section, we introduce three proposed modules for improving cross-view consistency in semantic segmentation. Training data are generated by online VVD data generator, then processed by the network modified with gradient guided deformable convolution, and finally supervised by additional consistency loss.

3.1 Online VVD Data Generator

Under the assumption of the pinhole model, we make the origin of the camera coordinate coincide with the origin of the world coordinate. The relation between a 3D point $\boldsymbol{P}$ and its 2D projection in an image $\boldsymbol{p}$ can be described as

$$\lambda \begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & s & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} [\, \boldsymbol{R} \mid \boldsymbol{0} \,] \begin{bmatrix} X_p \\ Y_p \\ Z_p \\ 1 \end{bmatrix}, \quad (1)$$

where $[x_p, y_p, 1]^T$ and $[X_p, Y_p, Z_p, 1]^T$ represent the homogeneous coordinate of $\boldsymbol{p}$ and $\boldsymbol{P}$, respectively. The matrix containing focal length f_x, f_y, principal point offset (u_0, v_0) and

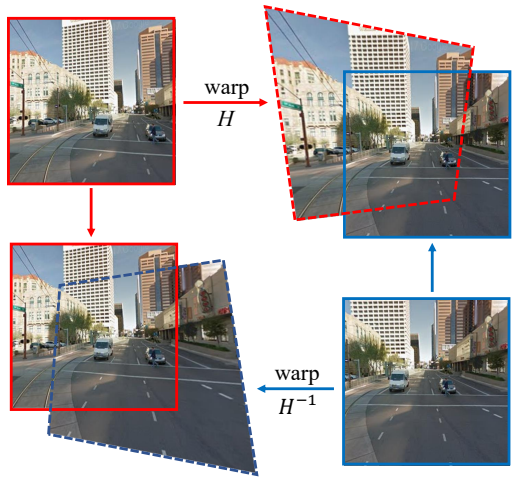

Figure 2: Relationship between two real images (in rectangle b of the same scene taken in different view directions, while ke the viewpoint unchanged. The two images can be warped to other using some specific homography H, exactly. There is n ference in the overlapping areas.

axis skew s is camera intrinsic matrix. $\boldsymbol{R}$ represents th tation matrix between the world coordinate system an camera coordinate system. λ is a normalization factor.

When 3D point $\boldsymbol{P}$ appears in two images with different views directions, the relation between the two projections (x'_p, y'_p) and (x_p, y_p) can be described as

$$\lambda \begin{bmatrix} x'_p \\ y'_p \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & u \\ 0 & f & v \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{R_z}(\alpha)\boldsymbol{R_y}(\beta)\boldsymbol{R_x}(\gamma) \begin{bmatrix} f & 0 & u \\ 0 & f & v \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix} = \boldsymbol{H} \begin{bmatrix} x_p \\ y_p \\ 1 \end{bmatrix}, \tag{2}$$

where $\boldsymbol{R_z}(\alpha)$, $\boldsymbol{R_y}(\beta)$, $\boldsymbol{R_x}(\gamma)$ represents the rotation matrix of the two view directions and can be decomposed into three rotation matrices based on three Euler Angles, respectively. $\boldsymbol{H}$ is the so-called homography with shape of 3×3. As shown in Fig. 2, two images of different views are connected with a homography $\boldsymbol{H}$. Equation (2) can be used to exhaustively and accurately find out corresponding pixel pairs that should have the same semantic label.

In order to get corresponding pixel pairs in images of different view directions, we propose an online VVD data generator. During training, the proposed online VVD data generator is used to produce a pair of image patches and the homography between them. We randomly crop a patch with the size of $2a \times 2a$ in the original image for the first one and record the crop center o. As to the second image, we firstly warp the original image with a generated homography $\boldsymbol{H}$. And we find the correspondent position of o in the warped image, which is denoted as o'. Then a patch of the same size is cropped in the warped image with center o'. Thus, the homography matrix

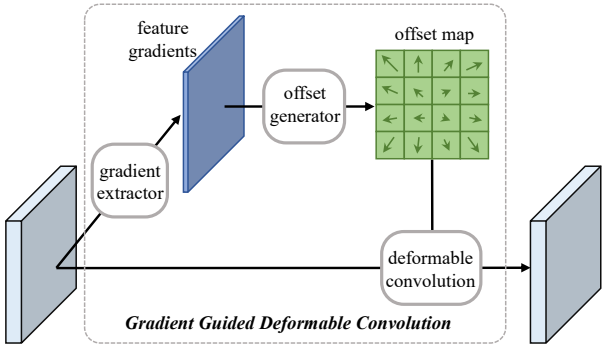

Figure 3: Illustration of the proposed gradient guided deformable convolution module. Feature gradients are obtained by gradient extractor from the input feature maps, and then are fed into offset generator to get offset map. Deformable convolution is used to get the final output feature maps. A difference between the proposed module and the standard deformable convolution is that the offset of the former is obtained by the gradient of the feature maps, while the latter is obtained by the feature maps directly.

between the two image patches can be described as

$$\boldsymbol{H_c} = \begin{bmatrix} 1 & 0 & a-x_{o'} \\ 0 & 1 & a-y_{o'} \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{H} \begin{bmatrix} 1 & 0 & x_o-a \\ 0 & 1 & y_o-a \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where x_o, y_o denote the x and y coordinate of point o. $x_{o'}$, $y_{o'}$ represent the x and y coordinate of point o', which can be calculated by

$$\lambda \begin{bmatrix} x_{o'} \\ y_{o'} \\ 1 \end{bmatrix} = \boldsymbol{H} \begin{bmatrix} x_o \\ y_o \\ 1 \end{bmatrix}. \tag{4}$$

3.2 Gradient Guided Deformable Convolution

Convolution can be seen as a feature sampling followed by a weighted sum operation. It gathers information from the receptive field of the convolution kernel in the input feature maps. After convolution operation, the value of location p on the output feature map y can be calculated as

$$y(p) = \sum_{\delta p \in \mathcal{R}} w(\delta p) \cdot x(p + \delta p), \tag{5}$$

where w is the weight of the convolution kernel and x is the input feature map. δp enumerates the pixel relative location in grid $\mathcal{R}$. For a standard 3×3 convolution, the regular grid $\mathcal{R}$ is represented as

$$\mathcal{R} = \{(-1,-1),(-1,0),...,(1,0),(1,1)\}. \tag{6}$$

For deformable convolution, $\mathcal{R}$ is not constrained in a rigid formulation and may vary when the position p changes. Thus, Eq. (5) becomes

$$y(p) = \sum_{\delta p \in \mathcal{R}} w(\delta p) \cdot x(p + \delta p + \mathcal{F}(x, p, \delta p)), \tag{7}$$

where $\mathcal{F}(x, p, \delta p)$ represents the learnable augmented offset. In standard deformable convolution, the grid is more likely to

Training Framework	VVD dataset			Cityscapes
	Segmentation	CVC/mIoU	CVC/pixAcc	Segmentation
PSPNet50 [Zhao *et al.*, 2017]	72.79	82.75	97.32	76.57
Online VVD	75.57	85.59	97.69	76.79
Online VVD + CLoss	76.35	87.46	97.84	77.14
Online VVD + GGDC + CLoss	**77.32**	**87.62**	**97.94**	**77.50**

Table 1: Results on Cityscapes dataset and the VVD dataset deriving from Cityscapes. From left to right, we show the segmentation results in mIoU and the cross-view consistency results (CVC) in mIoU and PixACC on VVD dataset, respectively, followed by the segmentation results in mIoU on Cityscapes. From top to bottom, We use PSPNet50 [Zhao *et al.*, 2017] as our baseline method. Online VVD, CLoss and GGDC represent the proposed online VVD data generator, consistency loss and gradient guided deformable convolution, respectively.

be driven by semantic information. For two images containing the same object via different view directions, the local receptive fields are supposed to have the similar semantic information with a local spatial deformation. The changes of feature map gradients reflect the deformation directly. Thus, we design a gradient guided deformable convolution module, in which the grid offset is driven by gradient information from the input feature maps. The output of the proposed gradient guided deformable convolution can be represented as

$$y(p) = \sum_{\delta p \in \mathcal{R}} w(\delta p) \cdot x(p + \delta p + \mathcal{F}(\mathcal{G}(x), p, \delta p)), \quad (8)$$

where $\mathcal{G}(x)$ represents the gradients of the input feature maps x. The gradient guided deformable convolution is illustrated in Fig. 3. In practise, the gradient extractor $\mathcal{G}$ and offset generator $\mathcal{F}$ are respectively implemented using Sobel operator and convolutions for end-to-end training. The novel gradient guided deformable convolution can readily replace standard convolution in existing segmentation networks.

3.3 Consistency Loss

To enforce the corresponding pixel pairs have the same inferred labels, we constrain output feature maps with a novel consistency loss in a deeply self-supervised way during training. When two images in different view directions are fed into the network, the feature columns at the corresponding positions in two feature maps are supposed to be the same. For the m-th selected layer, let $\boldsymbol{x}_i^m$ and $\boldsymbol{y}_i^m$ be the i-th positions of corresponding pixel pairs. We use cosine similarity to calculate the distance between two correspondence columns of the feature maps, followed by a linear operation to normalize and a log operation to magnify the penalty for the error. The consistency loss for the m-th selected layer can be represented as

$$\mathcal{L}_{cons}^{(m)} = -\frac{1}{N}\sum_{i=1}^{N}\log(0.5 + \frac{\boldsymbol{x}_i^{mT}\boldsymbol{y}_i^m}{2\left\|\boldsymbol{x}_i^m\right\|\left\|\boldsymbol{y}_i^m\right\|}), \quad (9)$$

where N is the number of corresponding pixel pairs which appear in both the images. Assuming that we consider total M layers in calculating the consistency loss, the final consistency loss can be represented as

$$\mathcal{L}_{cons} = \sum_{m=1}^{M}\mathcal{L}_{cons}^{(m)}. \quad (10)$$

Finally, the total loss $\mathcal{L}$ is the sum of the standard cross entropy loss in semantic segmentation $\mathcal{L}_{seg}$ and the proposed consistency loss $\mathcal{L}_{cons}$ with a weighting factor α, which can be written as

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha\mathcal{L}_{cons}. \quad (11)$$

α	0.1	0.5	1	2	5
Performance	76.82	77.03	**77.32**	77.18	76.73

Table 2: Parameter study of α for the proposed consistency loss. We train the network on Cityscapes dataset and evaluate it on VVD dataset. All semantic segmentation results are mIoU (in %).

4 Experiments

4.1 Implementation Details

Our training and evaluation is implemented in PyTorch. Resnet50 with the dilated network strategy is used as our backbone. We replace the last 3×3 convolution layer of each residual block with the proposed gradient guided deformable convolution module. Outputs of all residual blocks are selected to be supervised by the consistency loss and α is selected to 1 after a parameter study. For training, we use SGD optimizer and employ the polynomial learning rate policy [Chen *et al.*, 2017a; Liu *et al.*, 2015] where current learning rate equals to the initial one multiplying $(1 - \frac{iter}{max\_iter})^{power}$. The initial learning rate and the power are set to 0.01 and 0.9, while the momentum and weight decay are set to 0.9 and 0.0001 respectively. Due to GPU memory limitations, we use a batch size of 8 and crop size of 776 during training.

4.2 Ablation Study on Cityscapes

We conduct an ablation study on a widely adopted urban semantic segmentation dataset Cityscapes [Cordts *et al.*, 2016] to verify the effectiveness of our method. Cityscapes dataset contains 5000 high quality pixel-level finely annotated images including 2975 images for training and 500 for validation. For evaluating the cross-view consistency and the segmentation performance of images in diverse view directions, we derive varying view direction (VVD) dataset from Cityscapes. New images and annotations are generated by a warp followed by a crop operation. For each image in the datasets, we randomly generate a homography matrix according to Equation (2). The yaw, pitch and roll angles used in homography generation are set in range of $[-30, 30]$, $[-15, 15]$ and $[-3, 3]$ respectively. The focal length f is set to 2262 following [Godard *et al.*, 2017]. (u, v) are set to half of the width and length of the images. We use the matrix to warp the

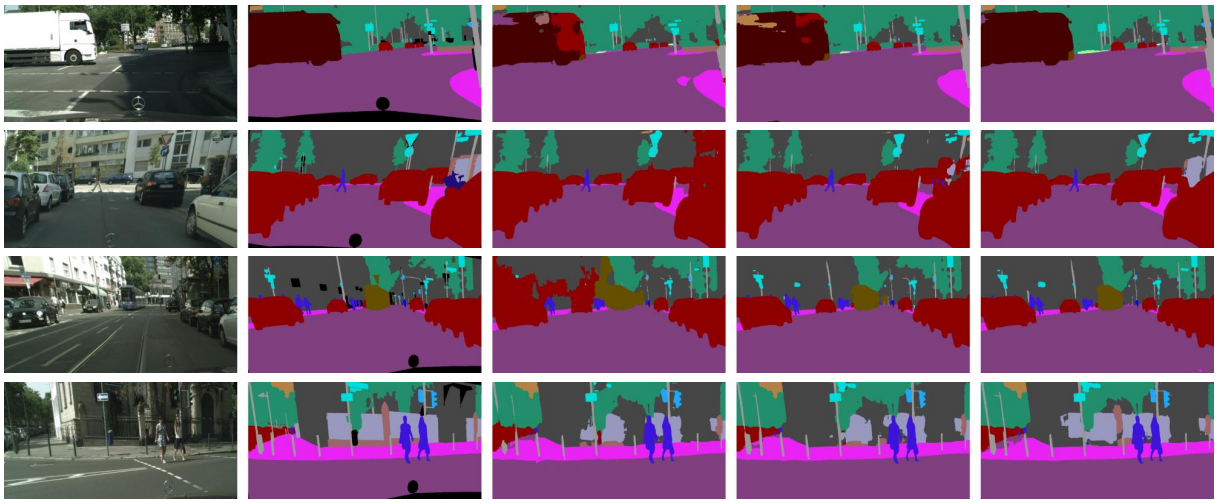

Figure 4: Visual comparisons in semantic segmentation on VVD dataset. (a) are input images and (b) are ground truth. From (c) to (e) are results of the baseline (PSPNet50) model, baseline model with online VVD data generator and the proposed approach with all three novel modules. We show the improvement in predicting vehicles and traffic signs.

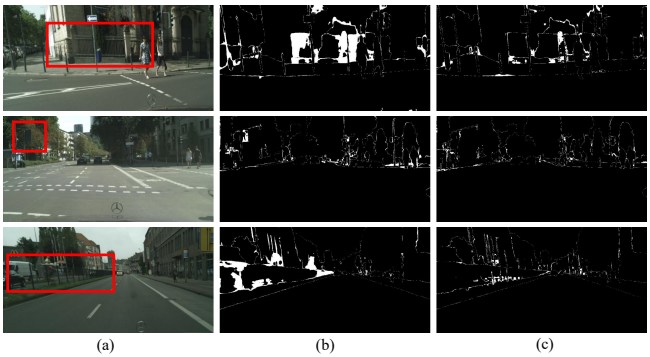

Figure 5: Visual comparisons in cross-view consistency on VVD dataset deriving from Cityscapes. Pixels not keeping the same semantic labels with their corresponding pixels in original images are shown in white color. The proposed method (c) keeps better cross-view consistency than the baseline method (b) in the shown images containing wall (top), sign (middle) and car (bottom).

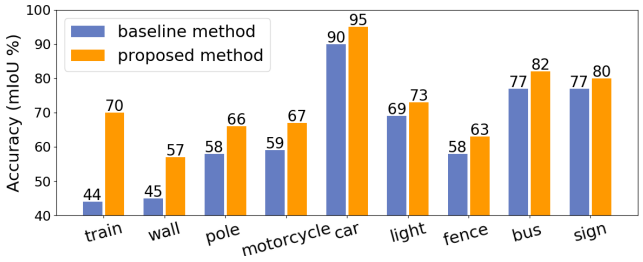

Figure 6: Per-class mIoU results on Cityscapes. We organize the classes in the order of results improvement from high to low. For each class, left(blue) shows the result of PSPNet50 [Zhao *et al.*, 2017] and the right(orange) shows that of the proposed method.

previous image and crop the maximum inscribed rectangle of it. The same operation is applied to the annotation. Correspondent pixel pairs of the previous image and the generated one can be found out using the homography. The constructed VVD dataset including annotated images containing the same scenes and in different view directions from Cityscapes and the corresponding pixel pairs between each image and its original image.

We employ mean intersection-over-union metric (mIoU) and pixel accuracy for evaluating the consistency. mIoU measures them in a class-balance way while pixel accuracy does in a pixel-balance way. After inferring an image pair from the VVD datasets with a given algorithm, we can get two segmentation maps. As the positions of all correspondent pixel pairs are known according to the homography and the semantic labels of them should be the same, the cross-view consistency can be measured by considering one prediction as ground truth and evaluating another.

In the ablation study, we use PSPNet50 [Zhao *et al.*, 2017] consisted of Resnet50 and a pyramid pooling module as our baseline method. The baseline method is trained with standard cross entropy loss. We gradually add the proposed modules including online VVD data generator, consistency loss and gradient guided deformable convolution. We evaluate the cross-view consistency using VVD dataset and the segmentation performance using both Cityscapes and VVD dataset. For the method only using cross entropy loss, the training data is trained for 90K iterations. For the other methods, to ensure the network provides nontrivial solutions and make it easier to converge, the consistency losses are added after 45K iterations warm-up training. The total iterations are set to 90K to guarantee the fairness of comparison.

The semantic segmentation and cross-view consistency results are listed in Table 1. We find the segmentation performance of the baseline method drops severely when dealing with images in abundant view directions in VVD dataset. All

Method	VVD dataset		Cityscapes
	Seg	CVC	Seg
DeepLabV3	73.61	82.89	77.13
proposed	76.88	**87.24**	77.58
CCNet	73.94	81.47	77.62
proposed	**76.95**	86.82	**77.80**

Table 3: Comparison based on DeepLabV3 [Chen *et al.*, 2017b] and CCNet [Huang *et al.*, 2019]. We show segmentation results on Cityscapes dataset and both segmentation and cross-view consistency (CVC) results in VVD dataset. All values are mIoU (in %).

Method	VVD dataset		CamVid
	Seg	CVC	Seg
PSPNet50	74.08	83.89	77.32
proposed	**76.33**	**86.88**	**77.39**

Table 4: Comparison using PSPNet50 [Zhao *et al.*, 2017] on CamVid dataset. All values are mIoU (in %).

the metrics, especially in VVD dataset are improved with the use of different novel modules. The proposed method with all three modules gets the best performance in the comparison. Meanwhile, it obviously reduces the gap between the segmentation performance on VVD dataset and Cityscapes. For visual comparison, some segmentation examples are shown in Fig. 4. The proposed method improves the segmentation performance in vehicles and traffic signs compared with the baseline method. We also visualize the cross-view consistency results in Fig. 5. We gather the statistic of improvement for each class in Cityscapes. We get improvement in all 19 classes and present the highest 9 classes for clear visualization, as shown in Fig. 6.

A parameter study of α is performed for the consistency loss to investigate its impact on the performance in VVD dataset deriving from Cityscapes. Semantic segmentation performance is considered in this part. We use different α from 0.1 to 5 in the proposed consistency loss and keep other experiment details the same. The results are listed in Table 2. The best choice for α is 1 with respect to segmentation.

4.3 Performance on Different Architectures

We evaluate our method with different network architectures in this section. The three novel modules are applied on DeepLabV3 [Chen *et al.*, 2017b] and CCNet [Huang *et al.*, 2019], respectively. The results are listed in Table 3.

We also evaluate our method in CamVid dataset. CamVid dataset contains 701 images and their pixel-level segmentation annotation with size of 720×960. We use 468 images to train and 233 images to validate following [Badrinarayanan *et al.*, 2017]. We derive new VVD dataset from CamVid for evaluating the performance. The intrinsic matrix is obtained using [Li *et al.*, 2010]. Model pretrained on Cityscapes in used as a start point in the experiment. We train both the baseline method and our method for 10K iterations. The results are shown in Table 4.

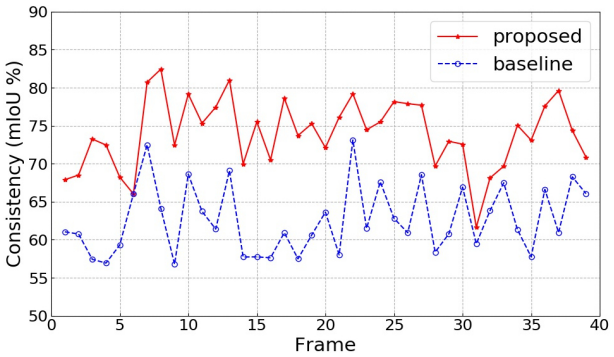

Figure 7: Cross-view consistency for a video sequence in Cityscapes video demo. The proposed method gains an obvious improvement over the baseline method [Zhao *et al.*, 2017].

4.4 Consistent Video Semantic Segmentation

To evaluate the effectiveness and applicability of the proposed method, we conduct an experiment on video semantic segmentation on Cityscapes dataset. We select 4 frame sequences with total 160 frames from the video demos. In this case the positions of the two cameras can be considered having a small baseline. Although a homography can approximately describe the relationship of the consecutive frames, we use dense optical flow to find the corresponding pixel pairs for more accurate measurement. We measure the consistency of the corresponding pixel pairs between consecutive video frames. Result of one sequence is shown in Fig. 7. The proposed method achieves an average of 70.4% on total sequences, which gains 8.2% over the baseline method. Results show our approach can improve the consistency in more general cases even the position of the camera is changed, though only view directions of images are varied in the training.

5 Conclusion

This paper focuses on improving cross-view consistency in semantic segmentation while varying view direction. We find that the view directions in previous urban datasets are often near directions of streets. In order to alleviate the bias, we propose an online VVD data generator that can modify annotated images to new plentiful view directions and calculate correspondent pixel pair indices, with carefully generated homography transformations. Furthermore, to improve the consistency, we present gradient guided deformable convolution that can readily replace standard convolution in existing segmentation networks. A novel consistency loss is also proposed for extra supervision during training. With the proposed consistency metrics, we evaluate the performance in the generated VVD datasets containing images with abundant view directions compared to the common-used datasets. Experimental results show that our method improves the segmentation performance of the convolutional neural network, making it more robust to view direction changes.

Acknowledgments

This work was supported in part by State Key Development Program Grand No. 2020YFB1708000, and NNSFC Grant No. 61971008.

References

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *PAMI*, 39(12):2481–2495, 2017.

[Brostow *et al.*, 2008] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, pages 44–57. Springer, 2008.

[Chen *et al.*, 2016] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, pages 3640–3649, 2016.

[Chen *et al.*, 2017a] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4):834–848, 2017.

[Chen *et al.*, 2017b] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017.

[Fu *et al.*, 2019] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019.

[Godard *et al.*, 2017] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017.

[Huang *et al.*, 2019] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, pages 603–612, 2019.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NIPS*, 28:2017–2025, 2015.

[Jeon and Kim, 2017] Yunho Jeon and Junmo Kim. Active convolution: Learning the shape of convolution for image classification. In *CVPR*, pages 4201–4209, 2017.

[Li *et al.*, 2010] Bo Li, Kun Peng, Xianghua Ying, and Hongbin Zha. Simultaneous vanishing point detection and camera calibration from single images. In *International Symposium on Visual Computing*, pages 151–160. Springer, 2010.

[Li *et al.*, 2019] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, pages 9167–9176, 2019.

[Liu *et al.*, 2015] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, June 2015.

[Ma *et al.*, 2017] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with rgb-d cameras. In *IROS*, pages 598–605. IEEE, 2017.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[Xiao and Quan, 2009] Jianxiong Xiao and Long Quan. Multiple view semantic segmentation for street view images. In *ICCV*, pages 686–693. IEEE, 2009.

[Yang *et al.*, 2018] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018.

[Yu *et al.*, 2018] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *CVPR*, pages 1857–1866, 2018.

[Yuan and Wang, 2018] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[Zhang *et al.*, 2018] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, pages 7151–7160, 2018.

[Zhang *et al.*, 2019] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *CVPR*, pages 548–557, 2019.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[Zhao *et al.*, 2018] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, pages 267–283, 2018.

[Zheng *et al.*, 2015] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pages 1529–1537, 2015.