DCSN: DEFORMABLE CONVOLUTIONAL SEMANTIC SEGMENTATION NEURAL NETWORK FOR NON-RIGID SCENES

Bor-Sheng Huang[2], *Chih-Chung Hsu*[1], *Wo-Ting Liao*[3], *Han-Yi Kao*[1], *Xian-Yun Wang*[1]

[1]Institute of Data Science and Department of Statistics
[2]Department of Electrical Engineering
National Cheng Kung University
[3]Department of Management Information Systems
National Pingtung University of Science and Technology

ABSTRACT

This paper presents a novel semantic segmentation network for outdoor and unstructured scenarios for autonomous driving based on deformable convolution and geometric distortion pipelines. The semantic segmentation tasks for autonomous driving are generally designed for the urban scene, city-view, and highly structured scenarios, such as the CityScapes dataset, KITTI, and BDD, while rare study focuses on outskirts scenarios. Therefore, the performance of existing semantic segmentation networks on such datasets might be unreliable. To conquer this issue, a novel densely connected residual block (DCRB) with the deformable convolution is proposed to form our backbone for capturing the non-rigid feature representation. In this way, the gradient flow of our DCRB could be better back-propagated from the segmentation head, resulting in a stable training process. Second, geometric distortion augmentation is introduced in the data augmentation pipeline, simulating the possible deformation situations in real-world outdoor scenarios. The experiments are conducted that the proposed semantic segmentation network significantly outperforms the state-of-the-art methods for both Cityscapes and Outdoor scenarios.

Index Terms— Semantic segmentation, high-resolution network, deformable convolution, deep learning, autonomous driving.

1. INTRODUCTION

With the rapid growth of computer vision applications on autonomous driving, the higher performance of semantic segmentation scheme is highly desired for understanding the semantic information for the autonomous vehicle. Several public datasets were collected such as Berkeley Deep Drive (BDD) [1], nuScenes [2], Cityscapes [3], KITTI[4], and Drive360 [5]. However, most of them were designed for understanding the city-like and urban scenarios, in which the object and scene of the dataset are relatively highly structured. Several state-of-the-art semantic segmentation schemes were proposed to achieve promising results on those datasets recently. However, the performance of those schemes might be suppressed and unreliable for the outdoor and outskirts scenarios since they were designed for learning in the well-known datasets aforementioned. Since the outdoor scenarios are usually non-rigid and unstructured, the deformable scenes and objects might lead to a restricted performance in practice [6].

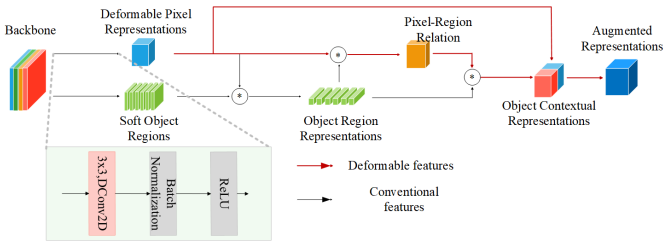

Fig. 1. Framework of the proposed deformable high-resolution network for outdoor scenarios semantic segmentation.

Toward the high-performance semantic segmentation networks, a High-resolution network (HR-Net) [7] was proposed to learn the high-resolution (HR) feature representation, in which the semantic segmentation neck and head will aggregate the multi-scale feature representation to achieve the promising result on various datasets such as Cityscapes [3], BDD [1], and KITTI [4]. In HR-Net, the convolution network will be separated into four branches with different spatial resolutions, where the first branch keeps the high-resolution feature representation, while the rest of the branches will learn the multi-scale feature representations. Therefore, all the spatial information in HR-NET could be kept during the

This study was supported in part by the Ministry of Science and Technology, Taiwan, under Grants MOST 109-2218-E-006-032, 107-2218-E-020-002-MY3, and 109-2634-F-007-013. We thank to National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

training and inference phases, leading to a promising performance on pixel-level tasks (i.e., semantic segmentation). Furthermore, object contextual representation (OCR) was proposed in [8] better to capture the context features between scenes and objects, leading to a significant improvement of the performance in semantic segmentation tasks. Moreover, Multi-scale attention (MSA) network [9] was proposed to boost the context feature learning, which the MSA is a variant of the multi-head attention mechanism in conventional Transformers. Although these advanced semantic segmentation methods show excellent performance on the existing dataset such as KITTI [4], Cityscapes [3], and BDD [1], their style of the scene are highly similar to each other and highly structured. In other words, the performance of other types of the dataset is unsure and unknown.

In [6], a fine-grained semantic segmentation dataset for outdoor scenarios, termed TAS500, was proposed to explore the suitable and effective semantic segmentation schemes to deal with the non-rigid and unstructured scenes in the wild. In TAS500, the types of the scene are significantly different from conventional datasets like Cityscapes [3], leading to the fact that the performance of existing semantic segmentation schemes is not promising. Specifically, the reported best mean intersection-over-union (mIoU) in [6] is less than 68%, while the performance of Cityscapes [3] is over 85%. It is also verified that existing peer methods might be ineffective for unstructured scenarios. As our observations of TAS500, the unstructured scene is non-rigid and deformable phenomenons. In [10], the restricted deformable convolution was introduced in the segmentation network for a 360-degree image captured by surround-view cameras. Since the 360-degree image is highly distorted by optical distortion, the restricted deformable convolution was proposed to learn the deformed scenarios. The phenomenons of the outdoor scenarios are similar to 360-degree images since both of them share unstructured and deformable scenarios. However, the nature of the unstructured scenarios of these two datasets are different; one is based on optical distortion, and the nature scene is naturally unstructured in another one. Therefore, the performance of [10] for the TAS500 dataset is still unclear.

In this paper, we introduce the deformable convolution operation [11][12] to the backbone network of HR-Net to adapt the non-rigid scene and unstructured objects in TAS500 to improve the performance further. Moreover, we also propose a novel densely-connected residual block (DCRB) to better capture the rich features based on multi-receptive fields. In this way, the gradient flow is easier to back-propagate from the segmentation head, resulting in the fact that the training process is relatively stable. A geometric distortion-based data augmentation pipeline is then proposed to simulate the non-rigid scenes during the training phase. In this fashion, our deformable convolution semantic segmentation network (DCSN) is effective for outdoor scenarios. In summary, the primary contribution is threefold:

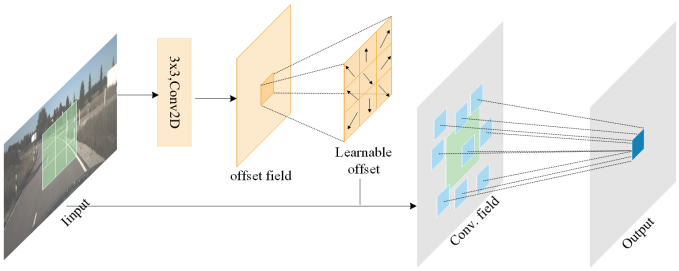

Fig. 2. The advantages of the deformable convolution for non-rigid scene adaptation.

- To the best of our knowledge, our DCSN is the first semantic segmentation scheme integrating deformable convolution-v2 [12] and high-resolution backbone network to deal with the non-rigid scenes segmentation issues.
- The geometric distortion is adopted in the data augmentation, which can be used to significantly improve the robustness of the model training in non-rigid and unstructured scenes.
- The proposed DCRB effectively captures the multi-scale feature representation based on feature reuse with multi-receptive fields property, which can improve the performance of semantic segmentation tasks.

The rest of this paper is organized as follows. Section II presents the DCSN for outdoor scene semantic segmentation tasks. In Section III, the superiority of the proposed method over peer methods is demonstrated. Finally, conclusions are drawn in Section IV.

2. THE PROPOSED DEFORMABLE CONVOLUTION SEMANTIC SEGMENTATION NETWORK

2.1. Overview

The overall network structure is illustrated in Fig. 1. In the first stage, the semantic features are extracted based on the proposed deformable backbone network with our DCRB. Then, the deformable pixel representation block is adopted to capture the non-rigid features extracted from our deformable backbone and followed by multiplying the weighting values to the soft object regions. Meanwhile, the deformable pixel features are re-weighted based on the object regional representation block to form the pixel-region relation information. Finally, the deformable pixel representation and the pixel-region relation feature are concatenated in channel-wise dimension to obtain the final contextual feature representation.

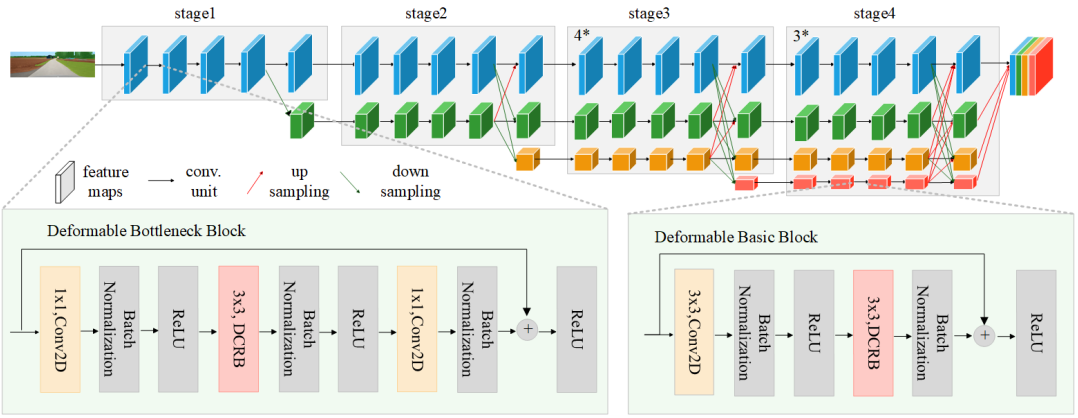

Fig. 3. The proposed backbone network architecture of the deformable high-resolution convolution for semantic segmentation.

2.2. The Proposed Deformable Backbone

The network architecture of the backbone of the proposed DCSN is illustrated in Fig. 3. Inspired by HR-Net [7], the multi-scale convolutional sub-networks are constructed to capture the semantic features across different resolutions. In the first branch, the spatial resolution of the feature maps in each layer is identical to the original one to minimize the information loss as small as possible, while other branches focus on the coarser feature representation learning. Finally, the multi-scale feature is concatenated in channel-dimension to obtain the base feature representation. To better tickle the non-rigid scene issues, the deformable convolution operation is adopted in the basic and bottleneck blocks. In our experiments, the deformable block is used in the first and second stages to capture the non-rigid features for outdoor scenarios.

Specifically, deformable convolution has learnable offsets for feature maps, so in the same square-sized kernel, it contains more spatial information than CNN. Let $\mathbf{x}$ and $\mathbf{y}$ are the input data/feature map and its outcome after convolution operation, the deformable convolution can be described as

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in R} \mathbf{w}(\mathbf{p}_n) \dots \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n), \quad (1)$$

where the $\mathbf{p}_0$ is the center coordinate of the convolution kernel, $\mathbf{p}_n$ is the sampling point in convolutional region R, and the $\Delta\mathbf{p}_n$ is the offset of the sampling point. In practice, the offsets are usually in floating type, where the convolutional operation is hard performed in such non-integer position in practice since the values in those non-integer position do not exist. Inspired by Region-of-interests (ROI) align in Faster-RCNN [13], the bi-linear interpolation is introduced to solve this problem as follows:

$$\mathbf{x}(\mathbf{p}) = \sum_{\mathbf{q}} G(\mathbf{q}, \mathbf{p}) \dots \mathbf{x}(\mathbf{q}), \quad (2)$$

where $G(\mathbf{q}, \mathbf{p}) = g(q_x, p_x) \dots g(q_y, p_y)$ and $g(x, y) = \max(0, 1 - |x - y|)$. In this way, the non-rigid object and scene can be well-treated in convolutional operation, as an example shown in Fig. 2. Since the conventional convolution operations might be ineffective in the unstructured scene and object during the training phase, the proper offset of the sampling points in deformable convolution will benefit such an unstructured scene. It is clear that the deformable convolution can be automatically adapted to the non-rigid scenes with the learnable offset in the sampling position of the convolutional operation.

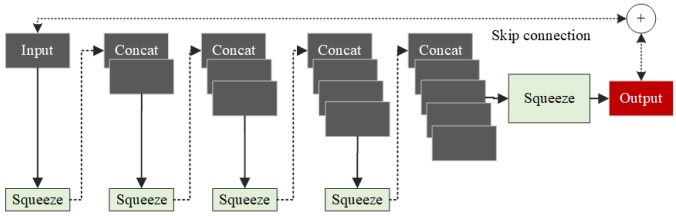

Fig. 4. The proposed densely-connected residual block in the backbone network.

2.3. Densely-connected Residual Block

To better aggregate the multi-scale feature information within each branch of the backbone, a novel DCRB is proposed based on residual learning and feature map reuse, as illustrated in Fig. 4. First, the input feature map $\mathbf{z}_0$ will be squeezed into a few-channeled feature map $\mathbf{z}_0^s$ and followed by concatenating the input feature map $\mathbf{z}_0$ and the squeezed feature map $\mathbf{z}_0^s$ in channel dimension so that $\mathbf{z}_1 = cat(\mathbf{z}_0, \mathbf{z}_0^s)$. In this way, the feature map can be reused in every layer with different receptive fields by $\mathbf{z}_{i+1} = cat(\mathbf{z}_i, \mathbf{z}_i^s)$, where $\mathbf{z}_{i+1}$ contains a 3×3 and 5×5 receptive fields, respectively. Finally, the shortcut connection between the input and the outcome of our DCRB is established to boost the performance, as well as improve the gradient flow back-propagation during the

Table 1. Performance comparison between the proposed method and other state-of-the-art semantic segmentation for the validation set in TAS500 dataset.

Method	DCRB	DConv.	GDA	mIoU
HRNet [7]				51.27
OCR [8]				57.82
DeepLabv3+[14]				45.91
Fast-SCNN [6]				38.83
Our DCSN	✓			60.83
Our DCSN		✓		63.77
Our DCSN			✓	58.53
Our DCSN	✓	✓		66.91
Our DCSN	✓		✓	63.85
Our DCSN		✓	✓	67.33
Our DCSN	✓	✓	✓	**69.37**

Table 2. Performance comparison between the proposed method and other state-of-the-art semantic segmentation for Cityscapes dataset.

Method	DCRB	DConv.	mIoU
OCR [7]			81.6
Proposed DCSN	✓		81.8
Proposed DCSN		✓	80.9
Proposed DCSN	✓	✓	**82.4**

training phase.

2.4. Geometric Distortion-based Augmentation

Since the non-rigid scenes are hard to capture its semantic feature due to unstructured objects, the proper data augmentation pipeline is necessary to boost the robustness of the semantic segmentation models, termed GDA (geometric distortion augmentation). With the outdoor scenarios, we introduce the geometric distortion to the data augmentation pipeline to boost the diversity of the non-rigid scenes from the original data. Specifically, our data augmentation pipeline adopts the optical distortion with a slight distortion range, elastic transform, and grid-based distortion. Also, the randomized Affine transform is adopted to simulate the perspective changes in the wild. In this manner, the robustness of the proposed DCSN can be boosted further.

3. EXPERIMENTAL RESULTS

The datasets used for performance evaluation are TAS500 [6] and Cityscapes [3], respectively. In TAS500, the training, validation, and testing sets contain 440, 100, and 100 images, in which the testing set does not provide the label information so that the evaluation results are obtained from the validation set only. Similarly, the performance evaluation for the Cityscapes dataset is performed on the validation set. The main criterion of the performance evaluation is the mean Intersection-over-Union (mIoU) score, which is defined as follows:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}, \quad (3)$$

where p_{ij} indicates the number of pixels that the predicted label is i and its corresponding ground truth is j.

In our backbone network, HRNet-W48 with stride 4 [8][7] is adopted in our DCSN and other peer methods to have the similar number of parameters. In the training phase, the start learning rate is $1e-2$ with Cosine Annealing decay scheduling. The total number of epochs for TAS500 in the training phase is 2000 since the heavy data augmentation (i.e., geometric distortion) will lead to a higher diversity of the training data so that the number of the epochs could increase the performance and robustness. The weight decay is $1e-5$, and the optimizer is SGD with momentum 0.9. In the last 1000 epochs, the Online Hard Example Mining (OHEM) is adopted to increase the robustness of the trained model. Finally, a 10-step multi-scale inference is adopted to obtain the final output.

Table 1 presents the comparison between the proposed method with other state-of-the-art schemes. It is clear that the deformable convolution significantly improves the performance on TAS500, implying that our DCRB well captures the non-rigid shape and contours with deformable convolution. On the other hand, the proposed GDA is also beneficial to the model effectiveness due to the very limited number of training samples in TAS500 (say, 440 images).

Furthermore, we also conduct experiments for Cityscapes to evaluate the generalizability of the proposed DCSN. As a result shown in Table 2, the proposed DCSN could improve the performance of the Cityscapes dataset. The deformable convolution did not improve the performance well because the dataset is highly structured so that the advantages of deformable convolution are insignificant. However, the performance can be further improved while our DCRB is applied to capture the multi-receptive fields features, implying that the proposed DCSN is effective and robust.

4. CONCLUSION

In this paper, we have proposed a novel deformable convolution semantic segmentation network (DCSN) for outdoor and non-rigid scenarios. In the proposed DCSN, the deformable convolution is adopted in the proposed densely-connected residual block (DCRB) to form the backbone network to capture the non-rigid shapes and objects in outdoor scenarios. Furthermore, The proposed DCRB successfully discovered the feature map reuse and residual learning in one block, leading to a significant performance improvement. On the other hand, a geometric distortion augmentation has been proposed to boost the robustness of the proposed DCSN. Extensive experiments demonstrated that the proposed DCSN showed superior performance over existing state-of-the-art methods.

5. REFERENCES

[1] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell, "BDD100K: A diverse driving video database with scalable annotation tooling," *CoRR*, vol. abs/1805.04687, 2018.

[2] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *CVPR*, 2019.

[3] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[5] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 687–704.

[6] Kai A. Metzger, Peter Mortimer, and Hans-Joachim Wuensche, "A Fine-Grained Dataset and its Efficient Semantic Segmentation for Unstructured Driving Scenarios," in *International Conference on Pattern Recognition (ICPR2020)*, Milano, Italy (Virtual Conference), Jan. 2021.

[7] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[8] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2021, vol. 1.

[9] Andrew Tao, Karan Sapra, and Bryan Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821*, 2020.

[10] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang, "Restricted deformable convolution-based road scene semantic segmentation using surround view cameras," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4350–4362, 2020.

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[12] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.