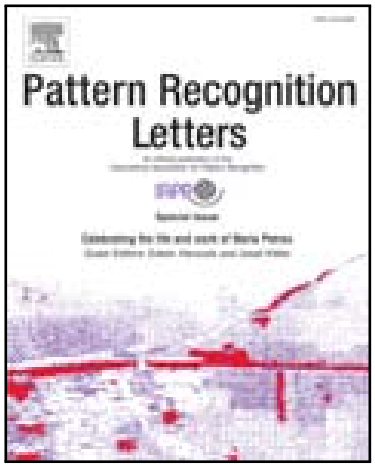
Journal Pre-proof



OccGAN : Semantic Image Augmentation for Driving Scenes

Yidong Wang, Lisha Mo, Huimin Ma, Jian Yuan

PII: S0167-8655(20)30229-4
DOI: https://doi.org/10.1016/j.patrec.2020.06.011
Reference: PATREC 7932

To appear in: *Pattern Recognition Letters*

Received date: 3 December 2019
Revised date: 16 April 2020
Accepted date: 13 June 2020

Please cite this article as: Yidong Wang, Lisha Mo, Huimin Ma, Jian Yuan, OccGAN : Semantic Image Augmentation for Driving Scenes, *Pattern Recognition Letters* (2020), doi: https://doi.org/10.1016/j.patrec.2020.06.011

Highlights

- The OccGAN structure is a semantic augmentation method on Cityscapes.
- The Rationality Module utilizes prior knowledge to implant occluders.
- The Authenticity Module ensures the plausibility by a generative adversarial network.
- Our Method improves the performance of several SOTA algorithms.

Pattern Recognition Letters
journal homepage: www.elsevier.com

OccGAN : Semantic Image Augmentation for Driving Scenes

Yidong Wang[a], Lisha Mo[a], Huimin Ma[b,**], Jian Yuan[a]

[a]*Department of Electronic Engineering, Tsinghua University, China*
[b]*School of Computer & Communication Engineering, Institute of Artificial Intelligence, University of Science & Technology Beijing, China*

ABSTRACT

Difficult images with complicated environments and occlusion have significant impacts on the performance of algorithms. They obey the long-tail distribution in the widely used datasets, which results in rare samples being overwhelmed during training. This paper presents a new approach to generate plausible occluded images with annotation as a kind of data augmentation with scenes semantics. To achieve this task, we proposed the the Occlusion-based Generative Adversarial Network (OccGAN) structure, which consists of a Rationality Module and an Authenticity Module. The Rationality Module generated preliminary occluded samples under the guidance of prior semantic knowledge. And the Authenticity Module is a generative adversarial structure to ensure the reality of the produced images. Qualitative results of the visualization process are given to verify the ablation study. Experiments on the semantic segmentation task indicate that several state-of-the-art algorithms combined with our OccGAN such as DRN, Deeplabv3+, PSPNet and ResNet-38, have boosts on IoU class scores and IoU category scores successfully.

Keyword: occlusion, GAN, semantic, augmentation, Cityscapes

1. Introduction

When the image contains abundant objects or the scene is quite complicated, sunch as MS COCO[1] and KITTI[2], deep networks are not as effective as humans in recognition [3]. Furthermore, the common property of these complicated images is occlusion. On the one hand, occlusion is arbitrary and various inherently. Several kinds of research have shown that the key parts of an object usually play an important role in recognition [4][5][6]. When the occlusion occurred at the position of key parts, the valuable information will be lost. Therefore, occlusion will lead to immense difficulty in recognition, detection, and segmentation tasks. On the other hand, deep learning tools rely on data heavily [7], while occlusion goes under a long-tail distribution in nature[8], as shown in Figure 1. Large scale datasets of nature scenes which are widely used, contain occluded images and other difficult images inevitably. The information of occluded objects is almost impossible to obtain, which means a straightforward approach cannot obtain the contour or the location box of occluded objects. In order to get annotations about occlusion, more manual work and human imagination are needed, which are laborious and costly.

Fig. 1. The long-tail distribution of the occlusion appeared in nature. We can hardly see difficult and rare occlusion in large scale image dataset. And these images lead to difficulty in recognition.

Since the long-tail distribution means that rare images are more likely to be overwhelmed by common images, direct training on a large-scale dataset cannot lead to satisfied results. Therefore, there are several data augmentation methods to assist network training[9][10][11]. Different kinds of transformation to the original image can improve the robustness of the

**Corresponding author:
e-mail: mhmpub@ustb.edu.cn (Huimin Ma)
Yidong Wang, Lisha Mo and Huimin Ma contributed equally to this work.

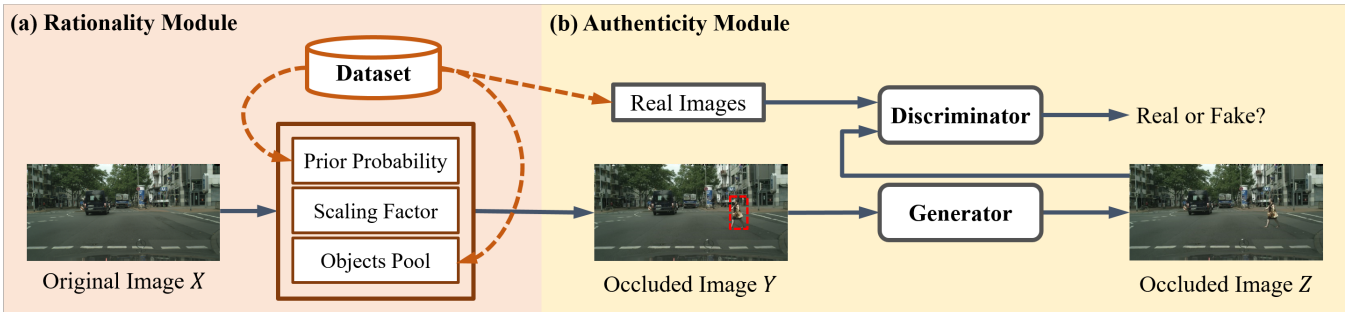


Fig. 2. Overview of the OccGAN structure. Our model has two parts: Rationality Module and Authenticity Module. The Rationality Module is highlighted by the light orange, and the Authenticity Module is highlighted by the light yellow. (a) Rationality Module: Given an original image, our rationality module utilizes the prior semantic knowledge from the dataset to complete occluders implantation; (b) Authenticity Module: In order to maintain the authenticity of occluded images, this generative adversarial network is proposed to cope with the sharp edges and the inconsistent style problems.

training model, including cropping, flipping, rotation, shifting, scaling, etc. Random Erasing[11] is also a method to provide occluded samples with gray blocks. These low-level augmentations mainly rely on the prior known invariances but lack semantic information. In other words, most traditional augmentation methods are low-level stacking of data, rather than instance-level operations that match scenes semantics. The core of this paper is to provide an augmentation method at semantic level to enlarge the number of plausible occlusion samples to balance the distribution.

Thus we propose an occlusion-based generative adversarial networks, which is abbreviated as OccGAN, to generate occlusion samples from natural images, which takes advantage of generative adversarial networks (GAN) [12] in image generation. A generative adversarial network is usually composed of two parts: a generator and a discriminator. The generator attempts to spoof the discriminator by producing samples similar to real ones, while the discriminator is trained to distinguish the generated samples. The two models are trained together to reach the Nash Equilibrium by an adversarial loss. In the seminal paper [12], the original GAN cannot control the pattern of generated samples because of random noise as input. Therefore, conditional GAN [13] is proposed to handle this problem, and many subsequent works are variants of conditional GAN. Then, Convolutional Neural Network (CNN) has also been adopted to train GAN. Proposing a set of architectural guidelines, the authors of [14] obtained a stable deep convolutional GAN. The tasks of image-to-image translation and text-to-image translation have also employed GAN [15][16][17].

A GAN with fully learning can hardly control the output images occluded or transformed. Furthermore, the generated occlusion may be as uncontrollable as noise. Although the GAN structure in A-Fast-RCNN[8] is used to generate occlusion to overcome the long-tail distribution, these occlusions occur on deep features rather than RGB images. All of the generated occlusions are modeled inside the network convolutional layers to make the network more robust. In fact, they just point out the essential region of one image. In our research, the OccGAN structure is manipulated to generate occluded images with natural objects instead of black or gray blocks.

Lee et al[18] insert object instance masks of specific classes by GAN to the semantic label maps on Cityscapes[19] dataset. Spatial and shape distributions are learned by two generators separately. Hence, the generation of RGB images can be achieved in two ways. One is to utilize other conditional GANs with semantic label maps as input[20], and the other is simple cropping and mapping. Due to the instability of GANs, the output will have blurred content or twisted edge, resulting in strange synthetic RGB images. By contrast, our approach utilizes prior knowledge to generate occlusions both on semantic label maps and RGB images, and exploits the GAN structure to ensure authenticity.

Unlike the idea of generating difficult occluded samples for training, some other studies devoted to image inpainting by GANs to overcome occlusion. In [21] a recurrent neural network is used to complement the occluded part of the image. A multi-scale deconvolutional network is also proposed to fix the occluded images [22]. In fact, the GAN structure is introduced to recognize occluded parts of objects [23].

In this paper, we also choose the Cityscapes dataset, which is a practical autonomous driving dataset with urban scenes. The most important thing is that we are able to establish reliable location priors because the in-vehicle cameras have a fixed viewpoint. The posture and spatial relationship of pedestrians and vehicles are relatively fixed in driving scenes. Random persons and cars are selected from natural images and are exploited to occlude the input. At the same time, inspired by the benefit of location prior in driving scene [24] and scene understanding [25][26], data statistics and spatial zooming factors could determinate where and what to occlude. Therefore, we propose the OccGAN structure to generate occlusions, which consists of a Rationality Module and an Authenticity Module. The whole structure constructs the human concept of occlusion into GAN to produce plausible occluded images guided by prior knowledge. Our method utilizes instance-level operations to achieve image semantic augmentation of the number, position and relationships of targets. Unlike the classic data augmentation results, most of which will be contrary to the real-world situation, the generated results of our OccGAN is consistent with the scene semantics.

Our main contributions are four-fold. First, we put forward a kind of instance-level semantic augmentation method OccGAN corresponding to the scenarios. Second, the Rationality Module utilizes prior knowledge, including scene understanding and

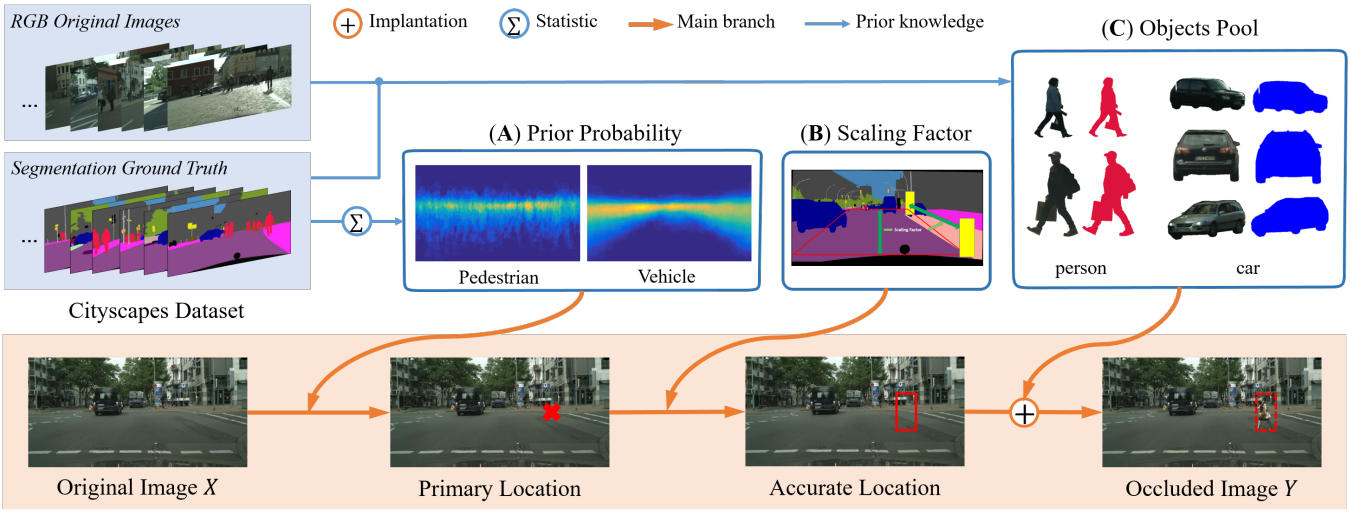


Fig. 3. Rationality Module. The prior semantic knowledge includes three parts: (A) the prior probability distribution, (B) the scaling factor and (C) the objects pool. The main branch is highlighted by the orange. Given the original image X, the module generates preliminary occluded images following three steps. The first step is to locate a point position by the prior probability distribution. Then, the second step is extending the point to a box by the scaling factor. Finally, the module selects an occluder from the objects pool to generate preliminary image Y by pixel-by-pixel implantation.

spatial relationship, to implant occluders reasonably. Third, the Authenticity Module exploits a generative adversarial structure to promote the authenticity of images. Finally, experiments on semantic segmentation task verify the effectiveness of our method and successfully improve the performance of existed state-of-the-art algorithms.

The rest of our paper is organized as follows. Section 2 introduces the whole OccGAN structure, including Rationality Module and Authenticity Module. Section 3 shows experiments validate the effectiveness of our method on semantic segmentation task. And Section 4 gives the conclusion and discussion.

2. Approach

The proposed model OccGAN takes a single RGB image as input and attempts to predict the occluders which are as plausible as possible with regard to real images. Our model is divided into two modules: one is a Rationality Module to generate occluded images by occlusion implantation, and the other is an Authenticity Module to ensure occlusion authenticity by a generative adversarial network.

2.1. Rationality Module

As shown in Figure 3, given the original image X, the Rationality Module includes three following steps to generate the preliminary occluded image Y. First of all, the implant position is randomly selected under the prior probability distribution. Secondly, we select an occluder from the objects pool and scale it by the scaling factor α. Finally, the occluder is implanted by pixel-by-pixel replacement at the previous location.

2.1.1. Occluders

Our model prefers to select real and reasoning objects in nature as occluders, instead of random gray blocks or strange unreasonable objects. The occlusion dataset should not only provide more occlusion samples but also be as realistic as possible with high-level semantic. Therefore, algorithms can have ability to waken the effects of occlusion to cope with real scenes. Furthermore, networks are expected to learn to extract the relationship between targets and occluders.

Considering that the Cityscapes dataset is a practical driving scene dataset with pixel-level annotations, it is easy to gain segmentation of occluders among all the labeled training images. For the driver, they pay more attention to moving objects, that is, pedestrians and vehicles. Therefore, we can extract segmented pedestrians and vehicles as occluders to form an objects pool. The objects pool contains more than 4000 pedestrians and 2000 cars. In addition to the selection of occluders, the rationality of occlusion mainly depends on the position and scale.

2.1.2. Prior Probability Distribution

To cope with the position issue, we introduce the prior probability distribution of the driving scene. Different categories of objects in the traffic scene have different regular probability distributions. Generally, static objects such as road, sidewalk, sky, buildings and vegetation are located in relatively fixed position. For dynamic pedestrians and vehicles, there exist specific dependencies on road. Therefore, the distribution of moving objects still has a reasonable and relatively fixed area. By adding the segmentation of training images, the prior probability distribution information of different objects can be directly obtained. Then, those probability maps will be exploited to guide the implantation of occluders.

2.1.3. Scaling Factor

To cope with the scale issue, we introduce a scale factor to simulation the depth information of occluders in three-dimension space to ensure plausibility and rationality. In the case of straight roads, it is evident to calculate the scale factor. But in the case of turning, there will be some strange results.

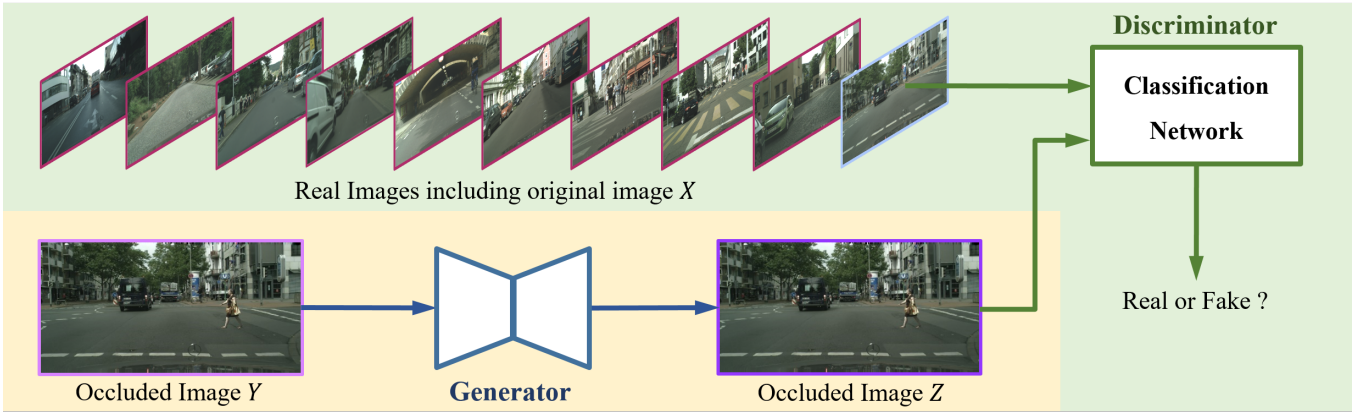


Fig. 4. Authenticity Module. This module consists of a generator and a discriminator. The generator is highlighted by the light blue background, while the discriminator is highlighted by the light green background. The generator is a fully convolutional network, and the discriminator is a 2-class classification network.

Therefore, the constant k is used to adjust the factor so that objects in the distance will not be too large. The scale factor is defined as:

$$\alpha = \frac{y_{obj} - y_{road}}{w_{longest}} - k \tag{1}$$

Where y_{obj} and y_{road} represent the y-axis coordinate. y_{obj} represents the center of the object and y_{road} represents the maximum value of the road. $W_{longest}$ represents the maximum width of the pixel area occupied by the road in this image. In this paper, the constant k is set to be 0.1.

The Rationality Module can generate occluded images with high-level semantic. Note that implantation is just a pixel-by-pixel replacement. It inevitably brings problems of sharp edges and inconsistent style, which need to be addressed in the next module.

2.2. Authenticity Module

The purpose of the Authenticity Module is to ensure the authenticity of occlusion. The main problems of occlusion authenticity are the edge sharpness and style inconsistency caused by the pixel replacement. In general, the whole module is a generative adversarial network, as shown in Figure 4 consisting of a generator G and a discriminator D.

2.2.1. Generator

The input of the generator is the preliminary occluded image Y, and the output is a more plausible occluded image $Z = G(Y)$. The generator is pre-trained as an Auto-Encoder to complete a regression task. There are 8 convolutional layers in the down-sampling stage and 8 convolutional layers in the up-sampling stage in our generator. Inspired by SRGAN [27], we utilize skip-connected layers in the last three of down-sampling and the first three of up-sampling to obtain high-level semantic information. Therefore, the rich features in the earlier layers can be fully made use of to maintain authenticity. In the pre-training step, we establish an identity map on the Cityscapes dataset to fine-tune the generator firstly. Then, the generator and the discriminator will be trained together.

2.2.2. Discriminator

The discriminator is a 2-class classification network to encourage the authenticity. The real image and the fabricated image are input together to obtain two feature vectors. Then the two feature vectors are fully-connected to a regression output to distinguish which is the real one. When the discriminator is trained together with the generator, the occluded image Y is compared with m random real images including the original one X, and the predictions are weighted to be losses of the generator and the discriminator separately.

The discriminator is supposed to distinguish the forged images with the original images successfully and is trained by the dichotomous loss:

$$\mathcal{L}(X, Z) = -log\Big[D(X)\Big] - log\Big[1 - D(Z)\Big] \tag{2}$$

Let $\{I^0_{ori}, I^1_{ori}, I^2_{ori}, \dots, I^{m-1}_{ori}\}$ denote real images, where I^0_{ori} denotes the original image X. The loss function of the discriminator is defined as:

$$\mathcal{L}_D = \mathbb{E}\Big[\mathcal{L}(I_{ori}, Z)\Big] = \frac{1}{m}\sum_{k=0}^{m-1}\mathcal{L}(I^k_{ori}, Z) \tag{3}$$

Where the number of real images m is 10 in this paper. The $\mathcal{L}(I^k_{ori}, Z)$ stands as the output of the dichotomy results by the discriminator. If the occluded image is considered to be real, the value stands for 1. Otherwise, the value stands for 0. In terms of the discriminator, the weight for each pair (that is, a real image and a forged image) is equal.

2.2.3. Adversarial Loss

In terms of the generator, we first use congruent tasks to pre-train network to allow the generator to completely reconstruct the input image. Because the modification of images by our occlusion objects is very limited, it is feasible to pre-train the task through a congruent task. For the discriminator, which is a

binary classification Network, we directly use the preliminary generated occlusion images and real images to constitute the database to pre-train it. The two networks are trained together after pre-training. The discriminator updates itself in once cycle by its own loss $\mathcal{L}_D$, while it gives the loss $\mathcal{L}_{D-G}$ back to the generator. The adversarial loss function is defined as:

$$\mathcal{L}_{D-G} = w^0 \mathcal{L}(I^0_{ori}, Z) + \frac{w^*}{m-1} \sum_{k=1}^{m-1} \mathcal{L}(I^k_{ori}, Z) \qquad (4)$$
$$s.t. \quad w^0 + w^* = 1$$
$$0 \le w^0, w^* \le 1$$

Where the number of real images m is 10 in this paper, and the $\mathcal{L}(I^k_{ori}, Z)$ stands as the output of the dichotomy results by the discriminator. The pair of the original image and the forged image takes part in the weight w^0, and the other pairs share the remaining weight w^* equally. In this paper, we set $w^0 = w^* = 0.5$.

3. Experiments

OccGAN architecture can produce occluded images quickly and plausibly. Moreover, due to the semantic synthesis, these occluded images have fine annotations of occluded objects. Our method can be regarded as a new kind of augmentation tool to produce a large number of occluded images based on original real images to obtain sufficient training samples. Besides, our method can be combined with traditional augmentation methods to improve several state-of-the-art algorithms. In order to prove the effectiveness of our augmentation approach OccGAN, a series of experiments were conducted as follows.

Experiments are conducted on Cityscapes dataset [19]. There are 2975 images for training, 500 images for validation and 1525 images for testing with fine annotations. Therefore, the task are determined to be semantic segmentation and four state-of-the-art approaches are considered as baselines: DRN [28], DeepLabv3+ [29], PSPNet [30] and ResNet-38 [31]. Noting that the data augmentation methods of these baselines are different. For the evalution of pixel-level semantic labeling task, the metric scores IoU Class and IoU Category, which are consistent with official standards, are used to assess the performance. There are two semantic granularities provided by Cityscapes dataset, i.e. 19 classes and 7 categories, to calculate the mean IoU of whole pairs of prediction and ground truth.

3.1. Qualitative Results of the OccGAN

First of all, a qualitative analysis is performed to demonstrate the process by which our OccGAN gradually generate occluded images. As Figure 5 shows, there are three sets of demos (a, b, c). Those demos mainly include the intermediate result after Rationality Module, the final results after the Authenticity Module and the detailed comparison of occluder regions. The occluder region after the Rationality Module is enclosed with a yellow bounding box, while the same region after the Authenticity Module is enclosed with an orange bounding box.

Table 1. 1x and 4x Extra Images in Training

	IoU Class	IoU Category
DRN	82.8	91.8
DRN (Ours ×1)	83.1	91.8
DRN (Ours ×4)	83.2	92.0
DeepLabv3+	82.1	92.0
DeepLabv3+ (Ours ×1)	82.7	91.8
DeepLabv3+ (Ours ×4)	82.6	92.1
PSPNet	81.2	91.2
PSPNet (Ours ×1)	81.8	91.3
PSPNet (Ours ×4)	82.2	91.4
ResNet-38	80.6	91.0
ResNet-38 (Ours ×1)	81.2	91.5
ResNet-38 (Ours ×4)	81.4	91.9

Table 2. 1x Extra Images in Training

	#Cities=6		#Cities=12		#Cities=18	
	IoU_{cls}	IoU_{cat}	IoU_{cls}	IoU_{cat}	IoU_{cls}	IoU_{cat}
DRN	70.4	82.2	77.8	89.4	82.8	91.8
DRN (Ours)	72.5	84.7	78.6	89.8	83.2	92.0
DeepLabv3+	71.1	85.5	77.7	88.4	82.1	92.0
DeepLabv3+ (Ours)	72.1	87.5	78.4	89.3	82.6	92.1
PSPNet	69.9	83.1	75.8	88.7	81.2	91.2
PSPNet (Ours)	71.9	86.3	76.7	89.4	82.2	91.4
ResNet-38	69.7	82.9	75.3	87.2	80.6	91.0
ResNet-38 (Ours)	72.2	84.8	76.5	88.0	81.4	91.9

As shown in the comparison of the first demo, the Authenticity Module promotes the consistency of image style, making the occluder more natural and harmonious with the environment. In the other two demos, it is explicit that the preliminary occluded images suffer from sharp edge and contour noise, resulting the images look strange and fake. With the Authenticity Module, the generated images become more realistic and reasonable.

3.2. 1× and 4× Extra Images in Training

In order to implement our OccGAN as data augmentation, only the training set can be used as original images to generate occluded samples, while the testing is on the validation set. For each baseline, our approach combines its own traditional augmentation to train the model.

We design experiments with 1× extra data and 4× extra data to verify the effectiveness of the quality of generated images. As shown in Table 1, our approach has improvement on all four baselines. It is effective for our OccGAN to improve algorithms performance by providing more occlusion samples. Nevertheless, the improvement is limited, which also indicates that there exists information redundancy and excessive data is unnecessary.

3.3. Semantic Augmentation Replace Real Images Partially

As we all know, the GAN structure can produce images upon augmentation. If the OccGAN is able to replace real images to

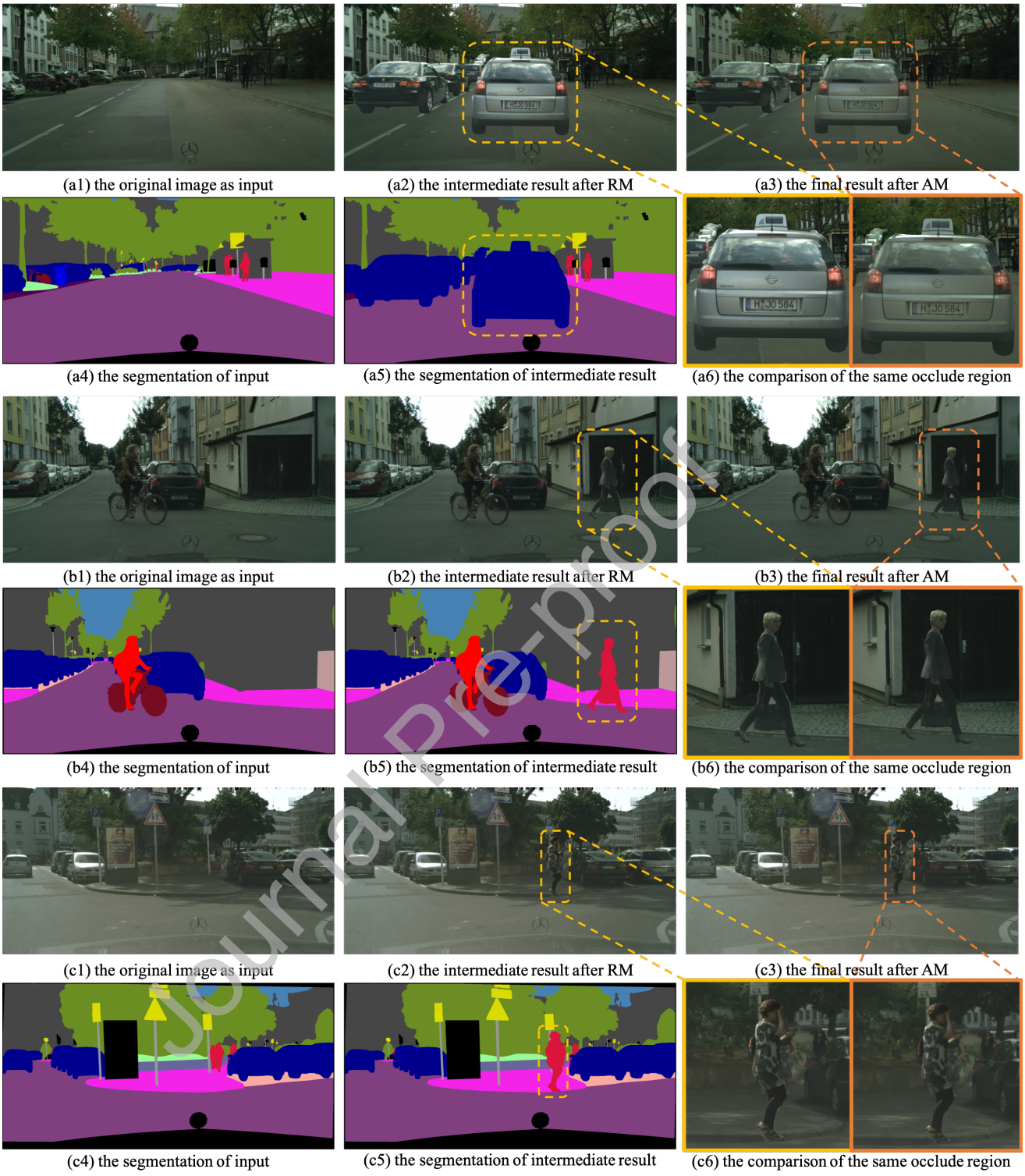


Fig. 5. Qualitative Results of the OccGAN. There are three demos above. For each demo, the results in the first row are the original RGB image, the intermediate result and the final generated image in order. In the second row, there is the segmentation of the original image, the segmentation of the intermediate result and the detailed comparison of occluder. Here the yellow bounding boxes represent the occluder region just after Rationality Module, and the orange bounding boxes represent the same region after the Authenticity Module.

reduce the demand of annotations, the cost to train a deep learning network will be lower. We design comparative experiments by gradually reducing the number of cities in real data.

Taking into account regional differences and sample balance, our manual division of cities in the dataset ensures that selected subsets cover different countries and regions. The number of

Table 3. 1x Extra Images in Training (#Cities=12)

	IoU Class	IoU Category
DRN	77.8	89.4
DRN(only ours)	77.6	89.3
DeepLabv3+	77.7	88.4
DeepLabv3+ (only ours)	78.1	88.8
PSPNet	75.8	88.7
PSPNet (only ours)	75.7	88.7
ResNet-38	75.3	87.2
ResNet-38 (only ours)	75.4	87.1

cities in the training set is 18. Then, every six cities are divided into one batch, and one batch and two batches are selected as subsets, respectively.

Comparison experiments are performed by training with 1× extra images generated by our OccGAN. The results are shown in Table 2. In terms of the IoU Class score, our method improves the performance by at least 1.0, 0.7, and 0.4 points on three subsets, respectively. Moreover, for the IoU Category score, our method improves by at least 2.0, 0.4, and 0.1 points, respectively. It is reasonable that the performance will decrease as the real data reduces. Our method could resist a certain degree of data loss. Our method still cannot provide unknown information, that is, data cannot be produced without foundation.

3.4. Semantic Augmentation Replace Real Images Completely

Furthermore, when training the network, we use the semantic generated images to replace the real data entirely. Our experiment is conducted on the subset of 12 cities with 1× extra images generated by the OccGAN for training. As Table 3 shows, there are only minor differences in the results of different training data. The generated data by the OccGAN is qualified compared to the original training data. The results indicate that our OccGAN can replace real data to a certain extent.

4. Conclusion

This paper proposes an OccGAN structure as a kind of augmentation method to generate plausible occluded images on autonomous driving dataset. Under the guidance of prior knowledge, the Rationality Module of OccGAN generates preliminary occluded images by the prior probability distribution and scale factor. Furthermore, the Authenticity Module handles the problem of edge sharpness and style inconsistency by a generative adversarial structure. At the same time, we can obtain the precise pixel-level annotations of the occluded objects as valid data augmentation. Experiments on the semantic segmentation task show that generated images by our OccGAN are almost indistinguishable from real images in the cases of many severe occlusion. Combined with other traditional augmentation, our augmentation method has improved the performance of several state-of-the-art algorithms successfully, which makes sense in training process of machine learning.

In the abundant autonomous driving dataset, which contains various weather, customs and traffic signs, more rational cues, more complex prior models and more accurate cost functions can be designed to enhance our method. Nevertheless, in general object detection datasets, the long-tail distribution of occlusion is still an important issue in complicated scenarios. To cope with this issue, the relationship between objects needs to be explored to establish prior knowledge, such that our OccGAN will be extended to more application scenes.

Acknowledgments

This work was supported by the National Key R&D Plan (No.2016YFB0100901), the National Natural Science Foundation of China (No.61773231, No.61673237) and the Beijing Municipal Science & Technology Project (No.Z191100007419001).

References

[1] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, pp. 740–755, 2014.

[2] A. Geiger, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[4] Z. Zhao, H. Ma, and X. Chen, "Semantic parts based top-down pyramid for action recognition," *Pattern Recognition Letters*, vol. 84, pp. 134–141, 2016.

[5] Z. Zhao, H. Ma, and S. You, "Single image action recognition using semantic body part actions," 2016.

[6] Z. Zhao, H. Ma, and X. Chen, *Generalized symmetric pair model for action classification in still images*. Elsevier Science Inc., 2017.

[7] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," 2017.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *neural information processing systems*, vol. 141, no. 5, pp. 1097–1105, 2012.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[11] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *International Conference on Neural Information Processing Systems*, pp. 2672–2680, 2014.

[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Computer Science*, pp. 2672–2680, 2014.

[14] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Computer Science*, 2015.

[15] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016.

[16] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017.

[17] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," pp. 1060–1069, 2016.

[18] D. Lee, M. Liu, M. Yang, S. Liu, J. Gu, and J. Kautz, "Context-aware synthesis and placement of object instances," pp. 10393–10403, 2018.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," pp. 8798–8807, 2018.
[21] A. V. D. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," 2016.
[22] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," 2016.
[23] K. Ehsani, R. Mottaghi, and A. Farhadi, "Segan: Segmenting and generating the invisible," 2017.
[24] X. Li, H. Ma, X. Wang, and X. Zhang, "Traffic light recognition for complex scene with fusion detections," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 199–208, 2017.
[25] X. Li, H. Ma, and X. Wang, "Feature proposal model on multidimensional data clustering and its application," *Pattern Recognition Letters*, vol. 112, pp. 41–48, 2018.
[26] X. Li, H. Ma, X. Wang, and K. Zhang, "Saliency detection via alternative optimization adaptive influence matrix model," *Pattern Recognition Letters*, vol. 101, pp. 29–36, 2018.
[27] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv: Computer Vision and Pattern Recognition*, 2016.
[28] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.
[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.
[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.
[31] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition." arXiv:1611.10080, 2016.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: