

# PIXEL LEVEL DATA AUGMENTATION FOR SEMANTIC IMAGE SEGMENTATION USING GENERATIVE ADVERSARIAL NETWORKS

*Shuangting Liu<sup>†</sup>*    *Jiaqi Zhang<sup>†</sup>*    *Yuxin Chen<sup>†</sup>*    *Yifan Liu<sup>†</sup>*    *Zengchang Qin<sup>†‡</sup>*    *Tao Wan<sup>\*</sup>*

<sup>†</sup> Intelligent Computing and Machine Learning Lab, School of ASEE, Beihang University, China

<sup>‡</sup>Keep Labs, Keep Inc <sup>\*</sup> School of Biological Science and Medical Engineering, Beihang University

## ABSTRACT

Semantic segmentation is one of the basic topics in computer vision, it aims to assign semantic labels to every pixel of an image. Unbalanced semantic label distribution could have a negative influence on segmentation accuracy. In this paper, we investigate using data augmentation approach to balance the semantic label distribution in order to improve segmentation performance. We propose using generative adversarial networks (GANs) to generate realistic images for improving the performance of semantic segmentation networks. Experimental results show that the proposed method can not only improve segmentation performance on those classes with low accuracy, but also obtain 1.3% to 2.1% increase in average segmentation accuracy. It shows that this augmentation method can boost accuracy and be easily applicable to any other segmentation models.

**Index Terms**— Data augmentation, generative adversarial networks (GANs), semantic segmentation

## 1. INTRODUCTION

Semantic segmentation aims to assign semantic labels to every pixel of a given image, it is one of the basic tasks in computer vision. Recently, many deep learning models [1, 2, 3, 4] are proposed and have achieved great performance on this task. However, deep learning based segmentation is always data-hungry and needs huge amount of fine pixel-level labeled data, which are always hard to collect, not to mention the fact that manual annotations may have a huge cost. Most previous work focus on improving the structure of deep neural networks (e.g. adding more layers [5]) to improve the accuracy, yet this approach can only improve average segmentation accuracy. In some practical applications, we may need to improve the performance of some specific classes as they may contain critical information. Unbalanced label distribution is one of the reasons causing low performance in segmentation. Using data augmentation for enlarging training set can yield better results and that has reported in various literature [6, 7, 8].

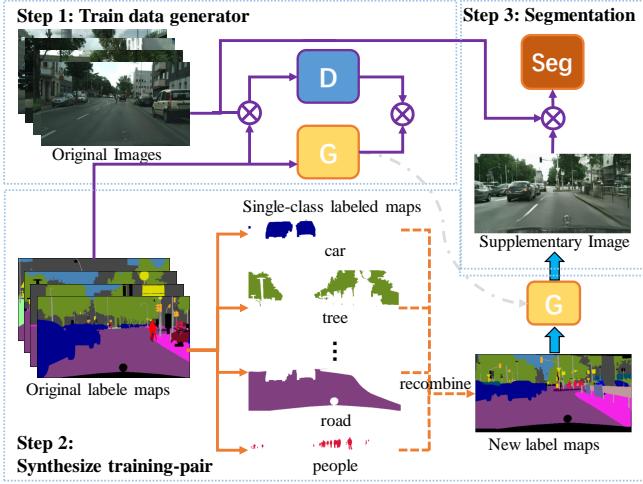
In this research, we consider using Generative Adversarial Networks (GANs) [9] for data augmentation in order to improve the segmentation accuracy. GANs are well used in computer vision and image processing [10, 11] for generating *realistic* images by learning true label distribution in a zero-sum game framework. Realistic image generation is a kind of image-to-image translation. The goal is using a semantically labeled image to generate a photographic image. Several methods have been proposed for this task including cascaded refinement networks [12], conditional GANs [13], and semi-parametric synthesis [14].

Data augmentation is simply the extension of the training data with generated data. Existing data augmentation techniques can roughly fall into two following categories: (a) geometric transformation which is computationally cheap and generic. (b) guided-augmentation or task-specific methods which using specific labels to generate image data [15]. In the case of image classification, some methods like Affine [7], elastic deformations [8], patches extraction and RGB channels intensities alteration [6] are all belong to this category. However, these methods only lead to an image-level transformation which only change depth or scale of image. They can improve the robustness of neural network, but actually no help for dividing a clear boundary between data manifolds. These methods do not improve the label distribution which is determined by higher-level features. As for the second type of data augmentation, many complex manipulated augmentation schemes have been proposed in fields such as scene text recognition [16], text localization [17], person detection [18], and emotion classification [19]. They all demonstrate the great performance on synthetic data.

Our work is most similar to [19], in which GANs are used to improve classification accuracy on classes with imbalanced data. In this paper, in order to boost the performance in segmentation tasks, we explore how to use GANs to generate supplementary data with pixel-level annotation labels and balance data-distribution within the dataset. The main contributions of this paper are as follows: (1) we propose a pipeline model for data augmentation by using GANs to generate supplementary data for semantic segmentation. (2) We propose a new method for image augmentation at pixel-level. (3) We

---

Corresponding to: {zcqin,wantao}@buaa.edu.cn



**Fig. 1:** Our pipeline model for data augmentation, where semantic labels *car*, *tree* and *people* can be used to reconstruct a new label map.

improve the data distribution and increase segmentation accuracy of both specific classes and on average.

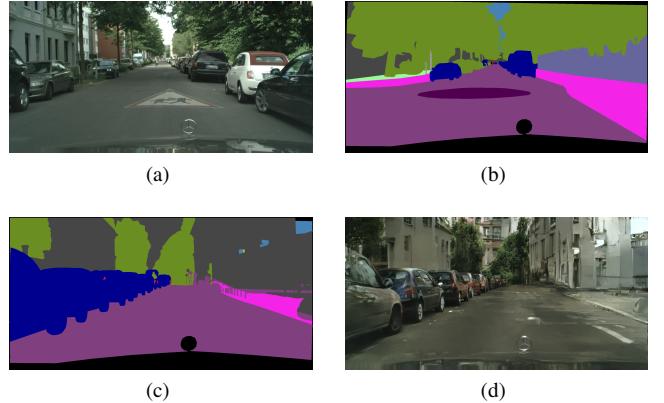
## 2. METHODOLOGY

The main idea of our method is generating supplementary (augmented) data for semantic segmentation to balance the distribution of semantic labels and improve the segmentation results. Fig. 1 schematically shows the procedure of the approach we proposed. The first step is training a GAN on original image/label pairs, it is used as a generator to transfer any human-designed semantic label maps (see Fig. 2b and 2c) to realistic images. In the second step, we use generated supplementary image to balance label distribution. Finally, we use supplementary data and original image data to train the segmentation network for better segmentation results.

### 2.1. Training Data Generator

We use the Pix2pix HD [13] model to generate realistic images given a specific semantic label map as our data generator. Real images (e.g. Fig. 2a) and their corresponding semantic label maps (e.g. Fig. 2b) from the original dataset are trained in pairs. Besides the generator  $G$ , there is a discriminator  $D$  to help completing the whole training process.  $G$  and  $D$  constitute Generative Adversarial Networks (GANs) [9]. The aim of generator  $G$  is to transfer semantic label maps to realistic images, while the discriminator  $D$  is used to distinguish real images (original images) from fake and *realistic* images generated by the generator  $G$ . We use minimax algorithm to model the strategy.

$$\min_G \max_D L_{GAN}(G, D) \quad (1)$$



**Fig. 2:** An original image (a) and its corresponding semantic label map (b). We select several semantic labels including *street*, *car*, *vegetation* and etc. to reconstruct a new semantic label map (c). Then we use GANs to generate its corresponding realistic image (d).

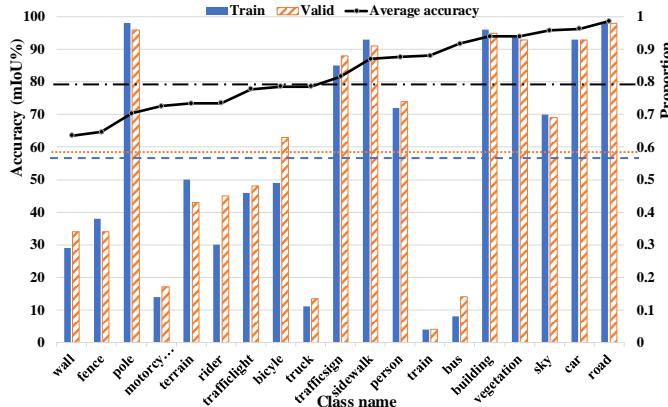
### 2.2. Synthesis of Training Data

We create a new synthesis dataset based on reconstructed label maps and generate corresponding realistic images to train semantic segmentation networks. To begin with, we separate each semantic label map in training set according to class of labels. Given a dataset of size  $N$  with  $K$  class of labels, we use  $I_1, I_2, \dots, I_N$  to represent corresponding semantic label maps of given original images. Separating those maps, we can then extract  $m (m \leq K)$  semantic labels from each map. We represent this process by:  $I \rightarrow \{L_1, L_2, \dots, L_m\}$ ,  $L_i \in \mathbf{L}$  and  $|\mathbf{L}| = K$ . Note that because one pixel only has one annotation,  $L_1 \cap L_2 \cap \dots \cap L_m = \emptyset$ . In this way, one semantic label map can be separated into  $m$  single label maps.

Then, we can reconstruct semantic label maps with these semantic labels. We arbitrarily select  $n$  semantic labels  $L_j \in \mathbf{L}$  for  $j = 1, \dots, n$ . Note that here,  $L_j$  for  $j = 1, \dots, n$  may come from different label maps, so they are not mutually exclusive. We combine them together to form a new label map (Fig. 2c). The process can be presented as following:  $\{L_1, L_2, \dots, L_n\} \rightarrow R$ . We design several specific ways to reconstruct the new label maps (details are explained in experimental study). We can change the semantic label distribution by modifying the proportion of each label in the images. With the recombined label maps, we can generate realistic images (Fig. 2d) by the data generator trained in Section 2.1.

### 2.3. Segmentation Using Data Augmentation

We use both original data and supplementary data (generated data) to train the semantic segmentation network. In the training process, we first train segmentation network with supplementary data. In this way, we can get a better initialization which contains the prior of some rare classes. Then, we use



**Fig. 3:** Label distribution analysis and model accuracy.

original dataset to fine tune the network. We use random initialization as our baseline. The comparison of two methods is discussed in experiment part. The translation of supplementary images is shown in Fig. 2.

### 3. EXPERIMENTAL STUDIES

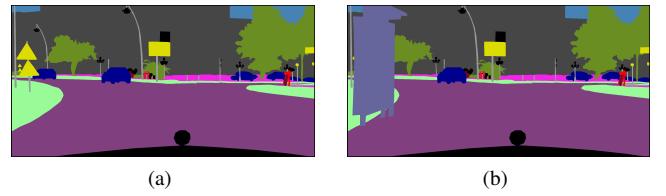
In this section, we use PSPnet [20] as the segmentation model. We compare the model performance before and after the augmentation to verify the effectiveness of our method.

#### 3.1. Analysis of Semantic Label Distribution

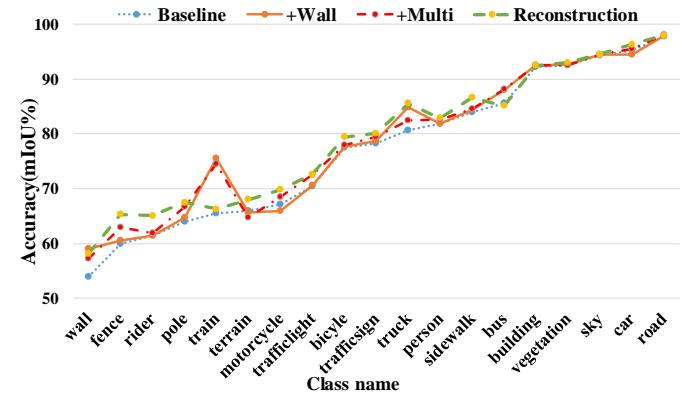
We choose Cityscapes [21] as the test dataset. This dataset records city street scenes in 50 different cities. It defines 30 visual classes (labels) for annotation, leaving 19 classes for evaluation. In our experiments, we just use 19 classes of semantic labels. We first calculate the label distribution. For each label class, the frequency of each label class appearing in the training set and the validation set is derived, and we call it as *appearance frequency*. Then we calculate the average segmentation accuracy of top 5 ranked models on Cityscapes website. Fig. 3 illustrates the correlation between the label distribution and segmentation accuracy. Comparing those classes with low appearance frequency and those have low segmentation accuracy, we find out that two groups are highly overlapped. In other word, it is possible to balance data distribution and further improve segmentation accuracy by increasing the appearance frequency on some specific classes.

#### 3.2. Ablation Study

According to the analysis of label distribution, we propose two ways to obtain new labels: 1) Overlay a single label directly on original label maps from the dataset. 2) Totally reconstruct using labels to form an entirely new label map. To further study how the proportion of supplementary data in training data will influence the performance of semantic



**Fig. 4:** (a) Original label map and (b) the new label map by adding a semantic label *wall* (grey area on the left).

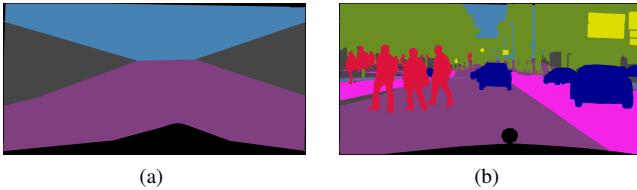


**Fig. 5:** Results comparison. Baseline: only use original Cityscapes dataset. +Wall: add label *wall* to each image. +Multi: +Wall +Fence +Pole +Traffic light +Train. Reconstruction: a new label map.

segmentation network, we conduct experiments with different proportion of augmentation data.

**Overlaying single label** We start with adding one label on original label maps to verify the effectiveness of our method. We increase the appearance frequency of those classes with low segmentation accuracy by overlaying specific single label on original label maps that do not contain this class. Fig. 4 shows the label map before and after applying our method. Taking the class *Wall* as an example, we first pick up all original label maps without the label class *Wall* from training set. We randomly choose from all wall-class label maps, and overlay one wall-class label on each original map. Furthermore, we use GANs to transfer segment images into a translated image. Finally, we use these supplementary images and original images together in different proportion to train the semantic segmentation network and calculate intersection over union (IoU) of all the classes.

We then study the effectiveness of adding more semantic labels on original label maps. We select several classes with low segmentation accuracy, and randomly choose some labels to add on. Single label of each class is overlaid on original label map. Supplementary images are also generated by GANs. In this experiment, we design one combination way that add *Wall*, *Fence*, *Pole*, *Traffic Light*, and *Train*. Results



**Fig. 6:** Reconstruction of label maps: (a) A basic label map and (b) reconstructed label map by adding various semantic labels.

**Table 1:** Experimental results of our approaches.

Method	Single Label	Multi-Label	Reconstruction
mIoU	78.65	78.82	79.41

are shown in Fig. 5 and Table 1. Adding one class of label improves mean IoU 1.3% and the IoU of the added class increases 5.0%. After adding multiple classes of labels, IoU increase significantly. Meanwhile, the mean IoU increases further up to about 1.5%.

**Reconstruction** In order to balance dataset better, we pick up segmented images from each class and combine them together to form a totally new image. We first draw a basic label map which only contains the class label *Sky*, *Road* and *Building*, which is shown in Fig. 6a. In this way we can make sure every pixel on the new label map has its label. Otherwise we may obtain an image with blank space labeled 0. Then we overlay each single label map on basic labels to form a constructed label map, as shown in Fig. 6b. We use GANs to obtain corresponding translated images. We repeat the preceding procedure for 2 times. Two different datasets are generated and input in segmentation network, and the mean IoU increases 2.0% and 2.1%, respectively. Since two results are very similar, we just show one of results in Fig. 5 and Table 1.

**Ratio of supplementary data** To figure out how the ratio of



**Fig. 8:** The result of generated an image using style transfer. (a) the original image (b) the transferred image.

supplementary data and original data will influence segmentation accuracy, we did further experiment as shown in Fig 7. Notice that when the proportion of supplementary data increasing, segmentation accuracy improves at first. But when the proportion continue goes up, the accuracy goes down. Results show that the best proportion of supplementary data in the training dataset is between 30% to 70%.

### 3.3. Model Comparisons and Analysis

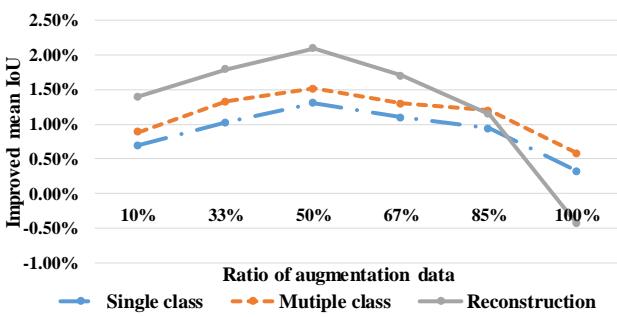
We compare each of our data augmentation methods to traditional data augmentation methods using rotation and zooming in and a state-of-art data augmentation method using style transfer. The result of using style transfer is shown in Fig. 8. Results of each method are shown in Table 2. According to the table, all of our methods have a better performance than traditional augmentation method, and our method outperforms state-of-art methods when using reconstructed data as supplementary data. Also, reconstructing separate classes can obtain highest segmentation accuracy among all our methods. This is because that it generates supplementary data with more variety and better balance the dataset. The results show the effectiveness of our method.

**Table 2:** Comparisons to other data augmentation models.

Method	Tradition	Style Transfer	Reconstruction
mIoU/Improve	77.31	79.10/+1.79	79.41/+2.1

## 4. CONCLUSIONS

In this paper, we explored how data augmentation method can be used to improve the performance of semantic segmentation. We proposed an augmentation method to generate supplementary data by using GANs. By adding generated label maps and images to original images as supplementary data, we can improve the diversity of data and balance the semantic label distribution. Comparing to other approaches in experiments, we found that the best way to implement our approach was using the *Reconstruction* method and setting the proportion of supplementary data as around 50%. The results shown that mean accuracy of a specific class can increase up to 5.5% and the average segmentation accuracy can increase 2%.



**Fig. 7:** Influence of supplementary ratio in training set.

## 5. REFERENCES

- [1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [2] Kaiming He, Georgia Gkioxari, and Piotr Dollar, “Mask r-cnn,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, “Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” in *IEEE transactions on pattern analysis and machine intelligence(TPAMI)*, 2018.
- [4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, “Rethinking atrous convolution for semantic image segmentation,” in *IEEE transactions on pattern analysis and machine intelligence(TPAMI)*, 2017.
- [5] George Papandreou, Florian Schroff, Hartwig Adam, Liang-Chieh Chen, Yukun Zhu, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” 2018.
- [6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems(NIPS)*, 2012.
- [7] Dan C. Cirean, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jrgen Schmidhuber, “High-performance neural networks for visual object classification,” *Computer Science*, 2011.
- [8] Patrice Y. Simard, Dave Steinkraus, and John C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in *International Conference on Document Analysis and Recognition*, 2003.
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems(NIPS)*, 2014.
- [10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, “Image-to-image translation with conditional adversarial networks,” *arXiv preprint arXiv:1611.07004*, 2016.
- [11] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” *arXiv preprint arXiv:1609.04802*, 2016.
- [12] Qifeng Chen and Vladlen Koltun, “Photographic image synthesis with cascaded refinement networks,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [13] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” *arXiv:1711.11585 [cs.CV]*, 2017.
- [14] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun, “Semi-parametric image synthesis,” *arXiv preprint arXiv:1804.10992v1*, 2018.
- [15] Mandar Dixit, Roland Kwitt, Marc Niethammer, and Nuno Vasconcelos, “Aga: Attribute guided augmentation,” *arXiv:1612.02559*, 2016.
- [16] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, “Synthetic data and artificial neural networks for natural scene text recognition,” 2014.
- [17] Andrew Zisserman, Ankush Gupta, Andrea Vedaldi, “Synthetic data for text localisation in natural images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [18] Jie Yu, Dirk Farin, Christof Krger, and Bernt Schiele, “Improving person detection using synthetic training data,” in *Image Processing (ICIP)*, 2010.
- [19] Xinyue Zhu, Yifan Liu, Zengchang Qin, and Jiahong Li, “Emotion classification with data augmentation using generative adversarial networks,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD)*, 2018.
- [20] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, “Pyramid scene parsing network,” in *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.
- [21] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.