High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs

Wenjie Niu

August 17, 2018

Abstract

The paper present a new method for synthesizing high-resolution photo-realistic images from semantic label maps using conditional generative adversarial networks (conditional GANs). Conditional GANs have enabled a variety of applications, but the results are often limited to low-resolution and still far from realistic. In this work, it generate 2048 × 1024 visually appealing results with a novel adversarial loss, as well as new multi-scale generator and discriminator architectures.

1. Introduction

In this paper, it is discussed a new approach that produces high-resolution images from semantic label maps. This method has a wide range of applications. For example, it can be used to create synthetic training data for training visual recognition algorithms, since it is much easier to create semantic labels for desired scenarios than to generate training images. Using semantic segmentation methods, it can transform images into a semantic label domain, edit the objects in the label domain, and then transform them back to the image domain. This method also gives us new tools for higher-level image editing, e.g., adding objects to images or changing the appearance of existing objects.

To synthesize images from semantic labels, one can use the pix2pix method, an image-to-image translation framework [6] which leverages generative adversarial networks (GANs) [5] in a conditional setting. Recently, Chen and Koltun [1] suggest that adversarial training might be unstable and prone to failure for high-resolution image generation tasks. Instead, they adopt a modified perceptual loss [3, 4, 7] to synthesize images, which are high-resolution but often lack fine details and realistic textures.

2. Instance-Level Image Synthesis

It first review our baseline model pix2pix(Sec. 2.1). It then describe how authors increase the photorealism and resolution of the results with the improved objective function and network design. Next, it use additional instance-level object semantic information to further improve the image quality. Finally, it introduce an instance-level feature embedding scheme to better handle the multi-modal nature of image synthesis, which enables interactive object editing.

2.1. The pix2pix Baseline

The pix2pix method [6] is a conditional GAN framework for image-to-image translation. It consists of a generator G and a discriminator D. For this task, the objective of the generator G is to translate semantic label maps to realistic-looking images, while the discriminator D aims to distinguish real images from the translated ones. The framework operates in a supervised setting. In other words, the training dataset is given as a set of pairs of corresponding images $\{(s_i, x_i)\}$, where s_i is a semantic label map and x_i is a corresponding natural photo. Conditional GANs aim to model the conditional distribution of real images given the input semantic label maps via the following minimax game:

$$\min_G \max_D \mathcal{L}_{GAN}(G, D) \tag{1}$$

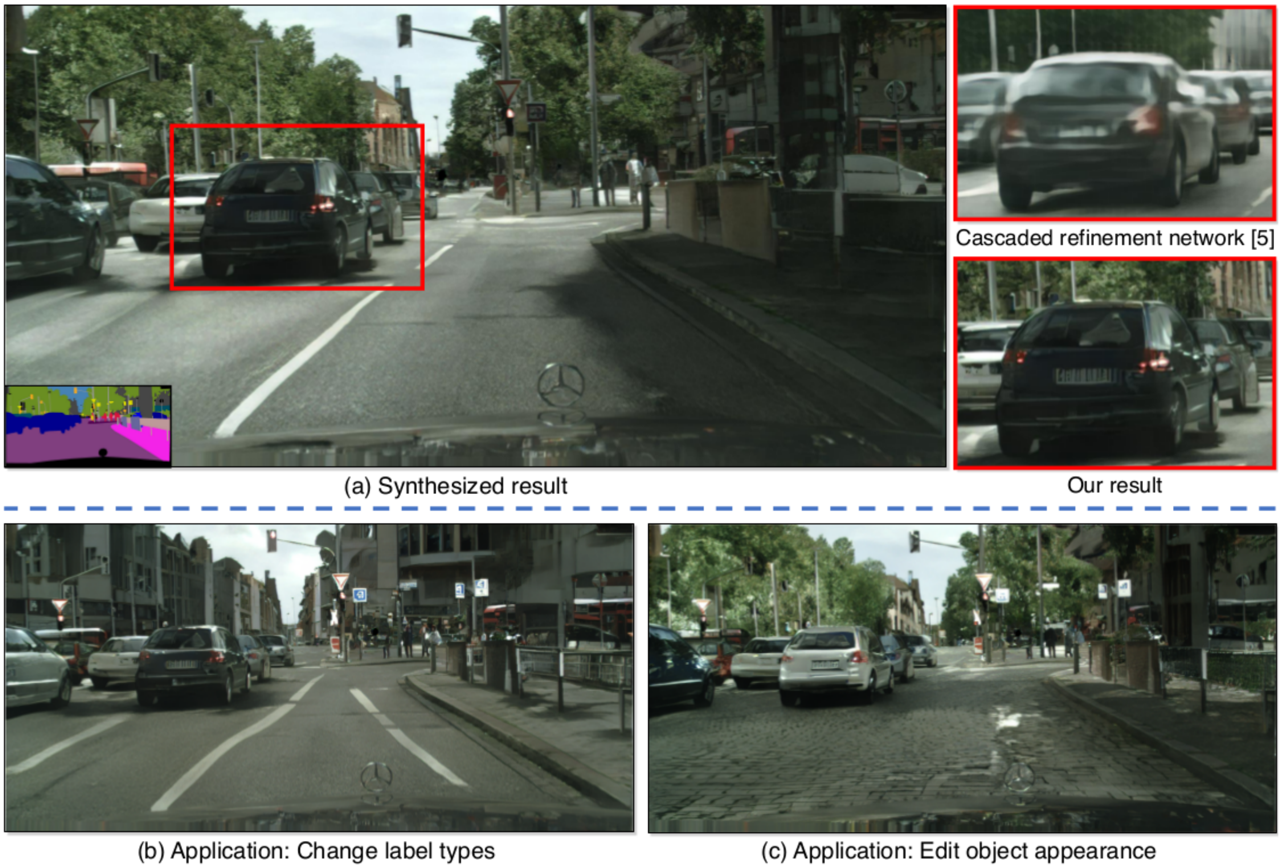
where the objective function $\mathcal{L}_{GAN}(G, D)$ is given by

$$\mathbb{E}_{(s,x)}[\log D(s, x)] + \mathbb{E}_s[\log(1 - D(s, G(s))] \tag{2}$$

The pix2pix method adopts U-Net [8] as the generator and a patch-based fully convolutional network [9] as the discriminator. The input to the discriminator is a channel-wise concatenation of the semantic label map and the corresponding image. However, the resolution of the generated images on Cityscapes [2] is up to 256 × 256. It tested directly applying the pix2pix framework to generate high-resolution images, but found the training unstable and the quality of generated images unsatisfactory.

References

[1] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 1, 2



Cascaded refinement network [5]

Our result

(a) Synthesized result

(b) Application: Change label types

(c) Application: Edit object appearance

Figure 1. It propose a generative adversarial framework for synthesizing 2048 × 1024 images from semantic label maps(lower left corner in (a)). Compared to previous work [1], the results express more natural textures and details. (b) The menthod in this paper can change labels in the original label map to create new scenes, like replacing trees with buildings. (c) The framework also allows a user to edit the appearance of individual objects in the scene, *e.g.* changing the color of a car or the texture of a road. [10]

[2] 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 *CVPR*, 2016. 1

[3] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *NIPS*, 2016. 1

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 1

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *NIPS*, 2014. 1

[6] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[7] J. Johnson, A. Alahi, and F. F. Li. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1

[8] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *ICML*, 2015. 1

[9] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE TPAMI*, 39(4):1–1, 2014. 1

[10] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 2