# Image-to-Image Translation with Conditional Adversarial Networks

Yufeng Jiang

## Abstract

*Authors investigate conditional adversrial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations. They demonstrate that this approach is effective at synthesizing photos from label maps, reconstructing objects from edge maps and colorizing images.*

## 1. Introduction

Many problems in image processing, computer graphics and computer vision can be posed as "translating" an input image into a corresponding output image. A scene may be renered as an RGB image, a gradient field, an edge map, a semanctic label map, etc. In analogy to automatic language translation, authors define automatic $image-to-image\ translation$ as the task of translating one possible representation of a scene into another, given sufficient training data (see Figure 1).

It would be highly desirable if they could instead specify only a high-level goal and then automatically learn a loss function appropriate for satisfying this goal. Fortunately, this is exactly what is done by the recnetly proposed Generative Adversarial Networks (GANs) [2, 1, 4, 5, 6]. In this paper, they explore GANs in the conditional setting. Just as GANs learn a generative model of data, conditional GANs (CGANs) learn a conditional generative model [2]. This makes cGANs suitable for image-to-image translation tasks, where they condition on an input image and generate a corresponding output image.

## 2. Method

GANs are generative models that learn a mapping from random noise vector $z$ to output image $y$, $G : z \rightarrow y$ [2]. In contrast, conditional GANs learn a mapping from observed image $x$ and random noise vector $z$, to $y$, $G : \{x, z\} \rightarrow y$. The generator $G$ is trained to produce outputs that cannot be distinguished from "real" images by an adversarially trained discriminator $D$, which is trained to do as well as possible at detecting the generator's "fake".

### 2.1. Objective

The objective of a conditional GAN can be expressed as

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] \\ + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))],\end{aligned} \quad (1)$$

where $G$ tries to minimize this objective against an adversarial $D$ that tries to maximize it, *i.e.* $G^* = arg \min_G \max_D \mathcal{L}_{cGAN}(G, D)$

To test the importance of conditioning the discriminator, they also compare to an unconditional variant in which the discriminator does not observe $x$:

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{y}[\log D(y)] \\ + \mathbb{E}_{x,z}[\log(1 - D(G(x, z)))].\end{aligned} \quad (2)$$

Previous approaches have found it beneficial to mix the GAN objective with a more traditional loss, such as L2 distance [3]. The discriminator's job remains unchanges, but the generator is tasked to not only fool the discriimator but also to be near the ground truth output in an L2 sense. They also explore this option, using L1 distance rather than L2 as L1 encourages less blurring:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[||y - G(x, z)||_1]. \quad (3)$$

Their final objective is

$$G^* = arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G). \quad (4)$$

### 2.2. Network architectures

They adapt their generator and discriminator architectures from those in [4]. Both generaotr and discriminator use modules of the form convolution-BatchNorm-ReLU. Details of the architecture are provided in the supplemental materials online, with key features discussed below.

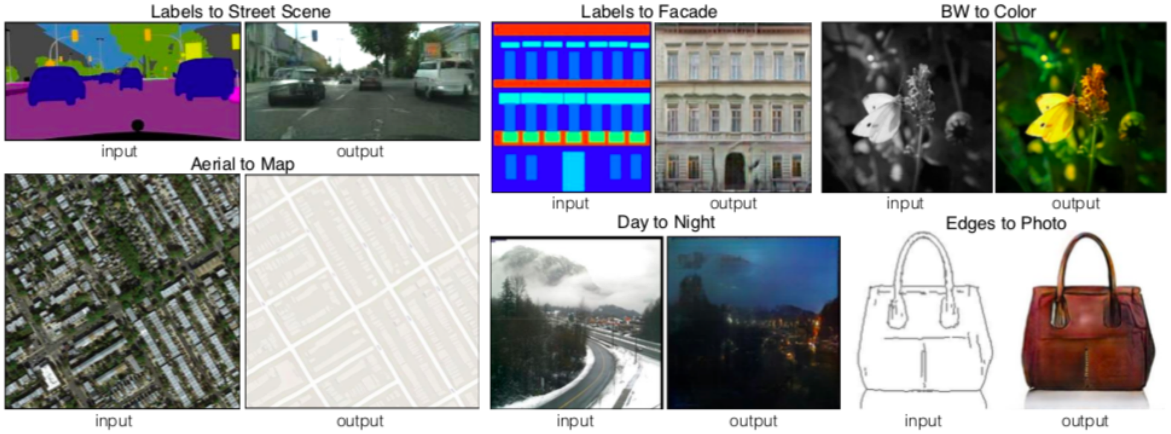

Figure 1. Many problems in image processing, graphics, and vision involve translating an input image into a corresponding output image. These problems are often treated with application-specific algorithms, even though the setting is always the same: map pixels to pixels. Conditional adversarial nets are a general-purpose solution that appears to work well on a wide variety of these problems. Here we show results of the method on several. In each case we use the same architecture and objective, and simply train on different data.

**Generator with skips:** A defining feature of image-to-image translation problems is that they map a high resolution input grid ot a high resolution output grid. To give the generator a means to circumvent the bottleneck for information like this, they add skip connections, following the general shape of a "U-Net" (see Figure 2). Specifically, they add skip connections between each layer $i$ and layer $n_i$, where $n$ is the total number of layers. Each skip connection simply concatenates all channels at layer $i$ with those at layer $n_i$. The evaluation of the generator architecture is shown at Table 1

| Loss | Per-pixel acc. | Per-class acc. | Class IOU |
|---|---|---|---|
| **Encoder-decoder (L1)** | 0.35 | 0.12 | 0.08 |
| **Encoder-decoder (L1 + cGAN)** | 0.29 | 0.09 | 0.05 |
| **U-net (L1)** | 0.48 | 0.18 | 0.13 |
| **U-net (L1 + cGAN)** | **0.55** | **0.20** | **0.14** |

Table 1. FCN-scores for different generator architectures (and objectives), evaluated on Cityscapes labelsphotos. (U-net (L1-cGAN) scores differ from those reported in other tables since batch size was 10 for this experiment and 1 for other tables, and random variation between training runs.)

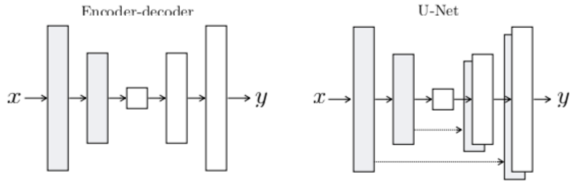

Figure 2. Two choices for the architecture of the generator. The "U-Net" is an encoder-decoder with skip connections between mirrored layers in the encoder and decoder stacks.

**Markovian discriminator PatchGAN:** It is well known that the L2 loss - and L1 prodidcues blurry results on image generation problems. GAN discriminaotr only models high-frequency structure, relying on an L1 term to force low-frequency correctness. In order to model high-frequencies, it is sufficient to restrict their attention to the structure in local image patches. Thesefore, they design a discriminator architecture which they term a $PatchGAN$ that only penalizes structure at the scale of patches. This discriminator tries to classify if each $N \times N$ patch in an image is real or fake. They run this discriminator convolutationally across the iamge, averaging all responses to provide the ultimate output of $D$.

## References

[1] E. L. Denton, S. Chintala, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015. 1

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

[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context, encoders: Feature learning by inpainting. In *CVPR*, 2016. 1

[4] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[5] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. 1

[6] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 1