Modified Pix2Pix Conditional Adversarial Networks for Image-to-Image Translation

Mrudula Y
Department of Computer Science and Engineering
University at Buffalo
Buffalo, NY 14260
mrudulay@buffalo.edu

Abstract

This work proposes a modified Pix2Pix cGAN for generalized image-to-image translation by adding a reconstruction loss to the objective function of the original Pix2Pix cGAN taking inspiration from StarGAN. The visual results for the task of translating an image from Day scene to a Night scene for Modified Pix2Pix are significantly better than those for Original Pix2Pix cGAN.

1 Introduction

Image-to-Image translation has become one of the hot-topics in the field of computer vision and Image processing. This is because many of the problems in this field are problems of ‘translating’ the same image of a certain domain to many different target domains, which is essentially what image-to-image translation means. As discussed in [1], the task of automatic image-to-image translation is to translate one possible representation of a scene into any of the many possible representations and that each of these translations was traditionally dealt with separate specially designed models that work only for that particular translation pair. In [1] they proposed a general-purpose conditional GAN for image-to-image translation wherein the GAN learns from a L1 loss function in addition to learning the mapping from input to output image.

An important aspect of image-to-image translation which should not be forgotten is that, at the core, the image remains same, it is the representations of that image that change. This can be understood by taking the analogy of language translation. While translating a sentence from one language to another, although the sentence looks completely different after translation but the meaning it conveys remains same. Similarly, it is important to retain the basic characteristics of the image and only change the required aspects of the image. It becomes even more moment to keep this point in mind when dealing with general purpose solutions as they are not trained for a particular domain pair and hence have more chance of producing real looking outputs that are completely different from the original image. As such a generated output is real-looking it will still succeed in fooling the discriminator but it defies the aim of image-to-image translation because it has not ‘translated’ the image but has produced an entirely new image. In this paper, the general purpose solution given in [1] has been extended to ensure that the GAN learns to retain the original aspect of the image while learning to translate it.

The technique used in this paper, for modifying the general purpose cGAN for image-to-image translation proposed in [1] has been previously introduced and used in StarGAN [2] which is a unified GAN for image-to-image translation whose effectiveness has empirically demonstrated on a facial attribute and facial expression synthesis task. Nonetheless, the technique used is just one aspect of the StarGAN architecture and as mentioned before StarGAN has been specifically tested for facial expression synthesis whereas here we are dealing with general purpose solution to image translation. The contribution of this work is to propose a modification to the general purpose cGAN for image-to-image translation and demonstrate using ablation study that the proposed modification leads to improvement in the generated translated images in terms of accuracy and sharpness.

Section 2 discusses the details of the previous works that have inspired the proposed solution in this paper. Section 3 describes the method and objective of the modified Pix2Pix Conditional Adversarial Network. Section 4 presents the experiments and ablation study results. Section 5 contains the future work.

2 Related Work

This work is an extension to the Pix2Pix Conditional GAN proposed in [1] for image-to-image translation.

Pix2Pix Conditional Adversarial Networks In this work, the authors have used conditional GANs for the purpose of image-to-image translation and have added the traditional L1 loss to the GAN objective. In the approach, the GAN not only learns the mapping from input image to output image using the adversarial loss, it also learns a loss function to train this mapping. They have used L1 distance in place of the traditionally used L2 distance, because it has been found by previous approaches that L1 loss encourages less blurry images. [1]

StarGAN This is a unified GAN model for multi-domain image-to-image translation. The StarGAN model is such that it can be trained to translate a particular image belonging to a certain domain between multiple domains for which the model has been trained. This work is similar to Pix2Pix in a way that this too proposes a single model for translating images. There are two aspects of difference between this and Pixel2Pixel – one is that in StarGAN the images can be translated bi-directionally between the multiple domains and second that at the time of training StarGAN requires data from all multiple domains one wishes to translate the images between using that particular StarGAN model.

The proposed modification to Pixel2Pixel in this work is inspired by the objective proposed for StarGAN. StarGAN talks about the importance of retaining certain characteristics of the image during translation for accurately accomplishing the task of ‘image translation’. This exercise will further help the model better classify generated images because by retaining core characteristics of an image it can develop the ability to emphasize on the discriminative components of the image.

3 Modified Pix2Pix Conditional Adversarial Network

The framework of the modified Pix2Pix Conditional Adversarial Network for generalized image-to-image translation has been described. Further in Section 4, it has been discussed why this would not work for certain image-to-image translation tasks.

3.1 Objective

The goal is to train the generator G that translates an image based on the given conditioning image. To achieve this, G is trained to translate an input noise z into an target output image y conditioned on the conditioning image x, $G : \{x, z\} \rightarrow y$. The discriminator is adversarially trained to detect images generated by generator as “fake” by taking as input the target image y and the conditioning image x, $D : \{y, x\} \rightarrow D(y|x)$. Fig. 1 illustrates the training process of the proposed approach.

Adversarial Loss. To make the generated “fake” images indistinguishable from the real one, the conditional GAN adversarial loss is adopted

$$\mathcal{L}_{adv}(G, D) = \mathbb{E}_{x,y}[\log(D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))] \tag{1}$$

where G generates an image $G(x, z)$ conditioned on the conditioning image x and taking noise z as input, while D tries to distinguish between real and fake images. The generator G tries to minimize this objective, while the discriminator D tries to maximize it.

L1 Loss. It has been seen in previous studies that combining the adversarial loss with a traditional loss like L1 or L2 has proven to be beneficial. In [1] L1 loss is considered as L1 loss has been seen to encourage less blurring when compared with L2 loss.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1] \tag{2}$$

where y is the original target image (to be generated) and G generates an image $G(x,z)$. The generator tries to minimize this loss.

Reconstruction Loss. After minimizing the adversarial loss and the L1 loss, the generator generates images which are real looking and indistinguishable by the discriminator but it is important in some aspects of image translation to maintain the basic characteristics of the original image same and only change the required aspects of the image while translating it. However, minimizing the above losses (Eqs. (1) and (2)) does not guarantee the above point. To incorporate a solution to this problem, a cycle consistency loss [2] is applied to the generator as adopted by [2].

$$\mathcal{L}_{rec}(G) = \mathbb{E}_{x,z,z'}[\|x - G(G(x,z),z')\|_1] \tag{3}$$

where G tries to reconstruct the conditioning image x, by taking as input the translated image $G(x.z)$ and noise z'. L1 norm is adopted for reconstruction loss [2]. The same generator is used twice, once to generate the fake translated image and again to reconstruct the conditioning image. It must be noted here the concept of reconstruction is not useful for all image translation tasks which will be discussed in detail in Section 4.

Full Objective. The full objective function for optimizing G and D is as follows,

$$\mathcal{L}_D = -\mathcal{L}_{adv}(G,D) \tag{5}$$

$$\mathcal{L}_G = \mathcal{L}_{adv}(G,D) + \lambda_{L1}\mathcal{L}_{L1}(G) + \lambda_{rec}\mathcal{L}_{rec}(G) \tag{6}$$

where λ_{L1} and λ_{rec} are hyper-parameters and control the relative importance of L1 loss and reconstruction loss respectively, compared to the adversarial loss. $\lambda_{L1} = 1$ and $\lambda_{rec} = 10$ have been used for the experiment.

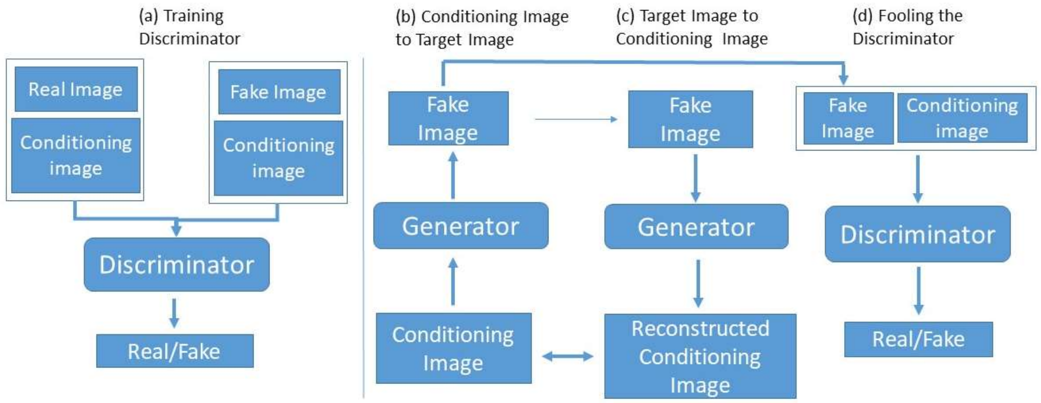

Figure 1: Overview of Modified Pix2Pix, consists of two modules Generator G and Discriminator D **(a)** D learns to distinguish between real and fake images **(b)** G takes as input the conditioning image and generates a fake image **(c)** G tries to reconstruct the conditioning image from fake image **(d)** G tries to generate images indistinguishable from the real images

3.2 Network Architecture

The generator and discriminator architecture is same as the one described in the Pix2Pix cGAN paper [1]. Below is a brief overview of the architectures of the Generator and Discriminator.

3.2.1 Generator with skips

The generator follows the architecture of "U-net" which is a simple encoder-decoder with skip connections between each layer i and layer $n - i$ where n is the total number of layers. Skip connections concatenate all the channels in the connected layers. Skip connections are added to shuttle the low-level information shared between input and output across the net.

3.2.2 Markovian Discriminator (PatchGAN)

The used discriminator architecture is termed as *Patch*GAN – which observes the images in patches of size $N \times N$ and determines whether each of these patches is real or fake. The discriminator runs convolutionally across the whole image to get the average of all the responses as output of D. The discriminator concentrates on the high frequency structure. The low frequency crispness is taken care by the L1 loss as it has been seen that the L1 loss accurately captures them.

4 Experiments

The modified Pix2Pix cGAN method is tested on two datasets for two different tasks.

- *Architectural labels → photo*, trained on CMP Facades [5].
- *Day → night*, trained on [6].

Training Details The training and testing was done using Google Colab's Python 3 Google Compute Engine backend (GPU). For the task of *Architectural labels → photo*, 400 images, 200 epochs and batch size of 1 were used for training. For the task of *Day → night*, 600 images, 100 epochs and batch size of 3 were used for training.

4.1 Evaluation metrics

The evaluation metric that has been used to measure the quality of the generated images is per pixel accuracy, PSNR and SSIM index of the generated image against the original target image. The visual appearance of the generated images is also one of the metric though it is not quantifiable, it is the most important aspect of any image-to-image translation task.

Per-Pixel Accuracy is measured by calculating the mean squared error between the pixel values of the original and generated images. The mean square error helps to averagely measure the difference in the pixel values of the generated and original images thus quantitatively defining how closer the generated image is to the original image. Hence, the lesser the mean square error i.e. per pixel accuracy the better it is.

Peak Signal to Noise Ratio (PSNR) is also a metric to compare images and is a measure of the peak MSE error. The higher the PSNR, the better the generated image quality. [3]

Structural SIMilarity (SSIM) index is a metric that measures the similarity between two images. [4]

These metrics were measured for the *Day → night* task for each of the generated images and the corresponding original target images and an average was taken. This was performed for both the original Pix2Pix cGAN and the Modified Pix2Pix cGAN. Table 1 shows the results. Point to note here is that none of these traditional metrics for comparing images correctly capture the improvement. The visual appearance of the generated images (Fig. 2 (c)) show the improvement that modified Pix2Pix cGAN has over original Pix2Pix cGAN.

Table 1: Metric values for generated images by Original and Modified Pix2Pix cGAN compared with original target images

cGAN Models → **Metrics ↓**	**Original Pix2Pix cGAN**	**Modified Pix2Pix cGAN**
Per-Pixel Accuracy	0.002144413411081063	0.0026935997096859295
PSNR	26.766485	25.862993
SSIM index	0.97995055	0.9762408

The metric values for original Pix2Pix cGAN are better but the images generated by Modified Pix2Pix cGAN are better visually. All of the above metrics use pixel values for calculation. The reason the metric results for original Pix2Pix cGAN are better because they generate completely black images whose pixel values are closer to the pixel values of the night (darkened) target images. But, generating completely black images is not the goal. The goal is to translate the image to a night setting i.e. the basic features of the image (here, trees, road etc.) should remain intact and the required aspects of the image (sky color etc.) should change. There is a need to find some better metrics to quantitatively assess the performance of a model.

4.2 Analysis of the objective function

To know which components of the objective function are important ablation studies have been done to isolate the L1 term, the GAN term and the Reconstruction term. As the difference between Pix2Pix and modified Pix2Pix lies in the objective function, this is the best way to know the effectiveness of the modified Pix2Pix method.

Fig. 2 shows the quality of generated images with different losses i.e. when different parts of the objective function are isolated and combined.

4.3 Reason for failure in certain tasks

For the task of generating photos from architectural labels using CMP Facades dataset, the results in the Fig. 3 show that they are visually near to the labels than the original photo when Modified Pix2Pix cGAN is used. This is because the reconstruction loss penalizes the Generator if the generated image does not retain aspects of the label but it does not specify which aspects to retain. Hence, along with retaining the edges information it also retains the color of the architectural label and hence loses the original color scheme.

There is a need to find a mechanism such that we can make the Generator concentrate only on the required components of the image to retain and so that it doesn’t retain the color of the original image.

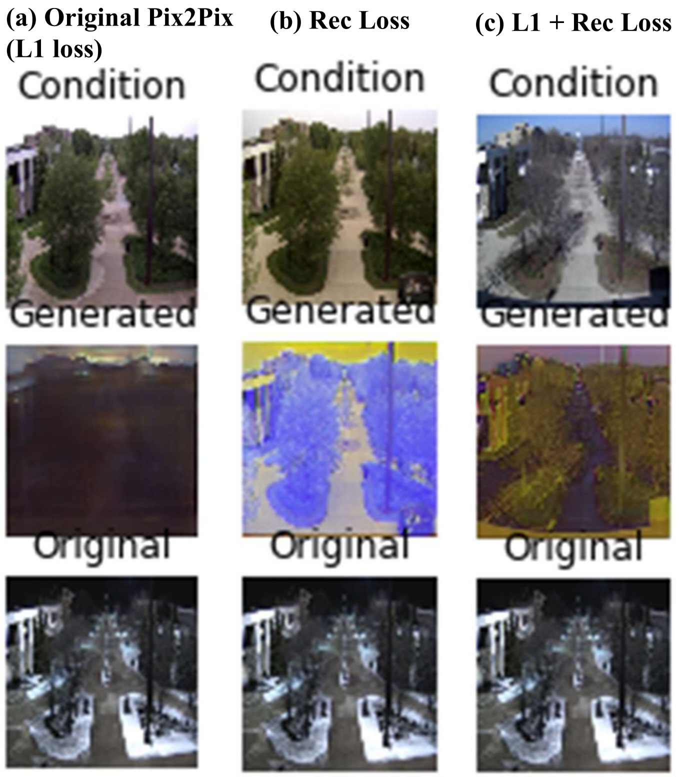

Figure 2: The middle images are the generated images by Generator, (a) image generated by orginal PixPix cGAN that has only L1 loss (b) image generated when only Reconstruction loss is applied (c) image generated by Modified Pix2Pix cGAN when L1 and Reconstruction loss is applied together

5 Future Work

The results presented in this paper are extremely preliminary but they do show that adding a reconstruction loss in the objective function does help in improving image-to-image translation.

The modified Pix2Pix model has to be tested for many other tasks such as changing black & white pictures to color, generating photos using sematic labels, generating pictures from sketches etc. As mentioned in section 2, this method has potential to build better classification models and hence it has to be tested on such a dataset to obtain quantitative results for this claim.

A mechanism needs to be researched to specify the elements on which the generator should concentrate while reconstruction. There is also a need to research image comparison metrics that will effectively capture the performance of GAN models in image generation.

As per [2], changing to using Wasserstein GANs in place of basic conditional GANs might also help improve the quality of generated images as well as bring certainty to the training process.

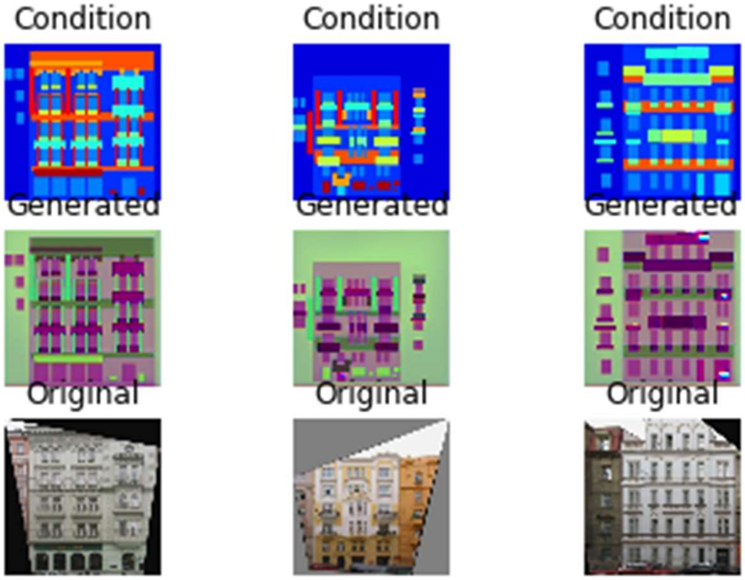

Figure 3: Modified Pix2Pix results when used for task of generating photos from architectural labels on CMP Facades dataset

Acknowledgments

I would like to thank Prof. Sargur N Srihari for teaching a very informative course and T.A Mihir Hemant Chauhan for giving an interesting project to work on. I would sincerely like to thank Nagashri Namagundlu Lakshminarayana for her guidance and constant support.

References

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

[2] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR. (2018).

[3] A. Hore and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," *2010 20th International Conference on Pattern Recognition*, Istanbul, 2010, pp. 2366-2369.
doi: 10.1109/ICPR.2010.579

[4] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing,* vol. 13, no. 4, pp. 600-612, Apr. 2004.

[5] R. S. Radim Tylecek. Spatial pattern templates for recognition of objects with regular structure. In German Conference on Pattern Recognition, 2013.

[6] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. ACM Transactions on Graphics (TOG), 33(4):149, 2014.