toon2real: Translating Cartoon Images to Realistic Images

Labiba Kanij Rupty
labknr98@gmail.com

Mohammad Imrul Jubair
mohammadimrul.jubair@ucalgary.ca

K M Arefeen Sultan
krsultan069@gmail.com

Sayed Hossain Khan
sayedhossainkhan36@gmail.com

MD. Nahidul Islam
nahidul19967@gmail.com

Department of CSE, Ahsanullah University of Science & Technology, Bangladesh.

Abstract

In this paper, we encounter the challenge of translating cartoon images to realistic images. As the Generative Adversarial Networks have tremendous successes in image-to-image translation, its variants drew our attention to perform the task. We prepared unpaired datasets for both cartoon and realistic domains, and we discerned that CycleGAN is suitable to reach our goal—as its Cycle-Consistency loss function is fit for working with unpaired datasets. We applied this model and the results exhibit that our approach produces satisfactory realistic images from cartoon images.

1. Introduction

Movies serve as one of the most popular sources of entertainment for human beings. Cartoons, undeniably, held a large part of entertainment industry in this modern day world. While watching them, a curiosity might be prompted in our mind: *How enchanting it would be to see our favourite cartoons become realistic? What if the adventure of Chihiro from 'Sprited Away (2001)' is rendered in a real-life setup? Or the journey of Carl from 'Up (2009)'?*

We reckon the above fantasy crosses most of the cartoon lovers' minds once in a while; however, making this happen in reality is not an easy task. For instance, an upcoming movie *'The Lion King (2019)'*—remake of one of the most popular animated movies *'The Lion King (1994)'*—costs four times the original one [2, 3]. The reason is that the new movie is a live-action version of the animated one based on Computer-Generated Imagery (CGI) which is a costly task to perform [1]. Moreover, the time and the labour required to generate an image are also high.

In this paper, we consider the above subject as our research problem and we attempt to propose a time & cost effective solution. We aim to input a cartoon image and to produce its realistic version automatically. Hence, we present a technique called ***"toon2real"***—a Generative Adversarial Networks (GANs) [5] based approach—that translates cartoons to realistic images. There has been some tremendous researches on image-to-image translation using GANs such as [14, 15, 10, 7, 8]; however, to our knowledge there hasn't been any research on generating realistic images from cartoon images yet. The closest research on translating cartoon images to realistic images has been touched by Li *et al*. [9] which only covers the face generation part of the task. Besides, Tomei *et al*. translates art images to realistic domain in their work [13], where each object of an image from the cartoon is mapped with the same objects from images of realistic domain. Moreover, the CartoonGAN [4]—a motivation behind our work—converts real image to cartoon; performing contrariwise is not a solution to our problem as the detail preservation from real to cartoon is not similar for vice versa. Hence, translating cartoons to realistic images is much harder because the cartoon images are smoothed out and their details are very trivial while compared to realistic images.

In this paper, we apply a technique which is based on CycleGAN [15] to achieve desired goal. We demonstrate our results and, in addition, we compare it with the UNIT method used by Liu *et al*. [10].

2. Formulation

Our main objective is to translate images of *cartoon* domain (C) to *realistic* domain (R) by learning the mapping of C to R. To satisfy this requirement, we adopt the CycleGAN [15] method to train our models. Our overall approach is discussed below.

2.1. Data Collection

Due to the lack of paired datasets between *realistic* and *cartoon* domain, we took an approach to collect unpaired datasets for both domains. Initially, for *realistic* domain, we scraped scenery images from *Flickr* and many other sources which were tagged as *scenery, sunrise, sunset, sea, sky* &

beach and collected around $7K$ samples. Besides, for *cartoon* domain, we extracted images from various Japanese anime movies. We excluded the frames which are darker to see, and the first and last few frames—as the introductory and credits part tend to be textual in a movie. After hand-picking the appropriate images, in order to approximate with the size of the *realistic* domain, we collected images from more than 15 cartoon movies and clips. For both the domain, images were of 128×128 dimension.

2.2. Model Development

To train our datasets the CycleGAN model is exploited. In this subsection, we discuss GANs very briefly followed by two key concepts of CycleGAN: the Cycle-Consistency loss and the PatchGAN.

Generative Adversarial Networks (GANs): Two networks will simultaneously learn the probability distribution of two domains—in our case, they are the cartoon domain C and the realistic domain R—to defeat each other. The first network is a generator, $G_r : C \rightarrow R$ which learns the mapping between the distribution of domain C and R and will generate fake images r_{fake}, matched to R, whereas the second network, Discriminator D_r will learn the probability distribution of R and try to differentiate between the generated images, r_{fake} and the images r from R. GANs generate images by minimizing the differences between the original content and the newly generated distribution. The measurement of the difference is named as *adversarial loss*. Goodfellow *et al*. [5] proposed a binary cross-entropy loss function, however, we used *least squares loss* function as our *adversarial loss* as it shows more stability according to Mao *et al*. [11].

Cycle-Consistency Loss: As we use unpaired datasets, exclusively applying adversarial loss produces images failing to match with the input data. As a result, while generating a realistic image from cartoon image, because of the void mapping, the model tends to depict random realistic images without considering the distribution of cartoon image. However, in case of *Cycle-Consistency Loss* function [15], the model learns from the most closely matched distribution of the domain. The motive of this function is that, an image generated from an input can be reconstructed back to the input again such that $x = F(G(x))$, where F and G are generators and x is the input, and thus it is able to map an image of target domain which is as close as possible to the image of input domain. In Equation 1, the loss function is showed where, m is the total number of images for each domain, and c & r are the samples from the domain C and R respectively.

$$\mathcal{L}_{cyc} = \frac{1}{m}\sum_{i=1}^{m}(F_r(G_r(c)) - c) + \frac{1}{m}\sum_{i=1}^{m}(G_r(G_c(r)) - r) \tag{1}$$

PatchGAN: As discriminators, PatchGAN model is used which was first proposed by Isola *et al*. [7]. It has a beneficial feature of working on $N \times N$ patches; it takes fewer parameters and thus decreases the computational cost. The method works best for extracting the high-frequency details of the distribution.

3. Results

We present our results followed by the comparison with UNIT [10] and our failed cases in Figure 1, 2 and 3 respectively. Our approach produces great results which is clear in Figure 1. Also, we choose the *Fréchet Inception Distance (FID)* [6] for quantitative evaluation. It shows more consistency with the human evaluation as the measurement is based on the difference between the generated and the actual dataset. We measured our output and the output of UNIT in terms of FID score where ours show the least FID score, 40.38 compared to UNIT's, which is 59.78. Here, the lower the FID score, the better the result is.

4. Discussions

In this paper, we ventured to contribute to the entertainment sector by presenting cartoons to realistic image translation. We applied CycleGAN approach which nicely adapts our unpaired datasets of cartoon and realistic domain. The experimental results show that our method performs adequately to produce desired outcomes. However, as an evolving research, it has certain limitations and we scheme to overcome those, which are shortly discussed below.

Despite producing the lowest FID score, it is still high which implies that there is still chance to improve. Besides, our technique currently works for scenery based images and fails to generate geometric structure from any cartoon image, e.g. human figure. From Figure 3, we can see that the method fails to translate a human figure as it is unaware of the semantic representation.

To improve the FID score, we have plans to train the datasets semantically, which is instead of training the entire images, we will train their segmented versions. Hence, the objects, e.g. trees, from cartoon domain will be segmented and mapped to the similar objects (trees) in the real domain [12, 13].

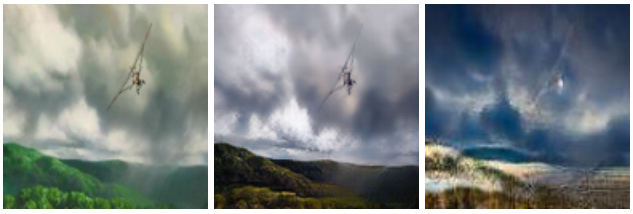
References

[1] Animation On A Budget - Pixelbox Visual Design, Jun 2015. [Online; accessed 15. Mar. 2019].

[2] Aramide-Tinubu. 'The Lion King': Is the Live-Action Film Disney's Most Expensive Movie? *Cheat Sheet*, Nov 2018.

[3] T. Bacon. Lion King's CGI Has Changed In New Trailer: Here's How It's Different. *ScreenRant*, Feb 2019.

Figure 1. ***Our results:*** The *first row* represents the input images from cartoon domain and *second row* shows the corresponding outcomes. It can be observed intuitively that the results are vivid and realistic.



(a) Input (b) CycleGAN (c) UNIT

Figure 2. ***Ours vs. UNIT:*** (b) shows the output of our proposed method which perfectly keeps the content of the input (a) and is visually sensible. It is far more realistic compared to the result of applying UNIT [10] (c) that fails to keep the content of the input.

(a) Input (b) Output

Figure 3. ***Our limitations:*** An unsuccessful case (b) of our method where it fails to translate the human figure for the input image (a).

[4] Y. Chen, Y.-K. Lai, and Y.-J. Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9465–9474, 2018.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[8] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

[9] J. Li. Twin-gan–unpaired cross-domain image translation with weight-sharing gans. *arXiv preprint arXiv:1809.00946*, 2018.

[10] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[11] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2813–2821, 2017.

[12] S. Mo, M. Cho, and J. Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018.

[13] M. Tomei, M. Cornia, L. Baraldi, and R. Cucchiara. Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation. *arXiv preprint arXiv:1811.10666*, 2018.

[14] X. Yu, X. Cai, Z. Ying, T. Li, and G. Li. Singlegan: Image-to-image translation by a single-generator network using multiple generative adversarial learning. In *Asian Conference on Computer Vision*, 2018.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.