Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/50/20241168

Translation from sketch to realistic photo based on CycleGAN

Xingfang Yuan

Department of Computer Science, Georgia Institute of Technology, Atlanta, 30332, United States

xyuan75@gatech.edu

Abstract. Forensic sketches serve as crucial tools for law enforcement agencies in identifying individuals of interest. However, their effectiveness can be limited due to constraints such as incomplete information and variations in interpretation by sketch artists, often rendering these sketches unrecognizable to the general public. In response to this challenge, this paper introduces an innovative approach—a CycleGAN-based image generation model. This model aims to transform monochrome forensic sketches into images with realistic colors and textures, offering an alternative visual representation that aids the public in identifying wanted individuals. The model is trained on unpaired datasets containing sketches and photographs of human faces, encompassing diverse scenarios. Through this training, it learns to generate images that closely resemble photographs captured in everyday environments. Impressively, the proposed model demonstrates rapid convergence, with both the generator and discriminator reaching optimal performance within just 500 epochs. Consequently, the generated images prove to be significantly more recognizable than the original sketches, thus enhancing the potential for successful identifications.

Keywords: Generative Adversarial Network, Image Generation, Image Translation.

1. Introduction

With the rapid development of Artificial Intelligence (AI) technologies in recent years, the current ways of solving problems and completing tasks in numerous industries are to be challenged by the new, AI-based methods [1-3]. Two of the most important tasks AI has changed so far are image recognition and image generation, which is the core topic of numerous industries. A typical problem that requires practitioners to both create and recognize images is criminal investigation, especially the use of forensic sketches. In a common situation where investigators have to locate a criminal with the description of the criminal's appearance from eyewitnesses, a conventional solution is that an artist will draw a forensic sketch of the criminal according to the description, and the police office will post this sketch on media, calling for the public to report the whereabouts of whoever they suspect to be the criminal. However, the background difference between the artist, the investigators and the public usually make the sketch quite ineffective in helping to identify the criminal, sometimes worse than the original description [4].

To solve this problem, many methods that use AI to generate pictures that work better than the traditional sketch for the public to identify the wanted person emerged in the past few years. For instance, Gunjate et al. have proposed a method using Generative Adversarial Network (GAN) combined with a U-Net Convolutional Neural Network (CNN) to transfer sketches to realistic pictures [5]. However, the design of the structure lacks incorporation of reconstruction loss, which resulted in overly general faces

DOI: 10.54254/2755-2721/50/20241168

that lack realistic details. In a similar approach, Chao et al. also used GAN with U-net in the generator network of the model to accomplish a sketch-to-photo translation. They made more and deeper changes to the basic GAN model in order to acquire more realistic and high-fidelity pictures. The model used a Patch-GAN as the discriminator and included several feature constraints. While the output of the model does resemble the sketch and has high fidelity, the model tends to oversimplify the facial features in the sketch, implying a still-high reconstruction loss [6]. Some research on the topic of face sketch to photograph have also used traditional computer vision methods like reconstructing images from Principal Component Analysis (PCA) representations [7], while others have used various GAN-based models like conditional CycleGAN to generate differently colored photos (skin tone, color of hair and eyes) from a sketch [8]. While performing well on certain aspects, most models mentioned above were only trained on datasets in which each sketch has a strictly corresponding photograph. This could be the reason for the lack of detail and realism among many of them.

In this regard, this study designed a CycleGAN-based model that trains on unpaired datasets to improve the realism in the sketch-photo translation. The proposed model consists of two generator-discriminator pairs with the first pair responsible for generating images that resemble photographs given sketches, and the second pair ensures the generated images resemble their corresponding sketches. Based on the CycleGAN structure, Resnet blocks were added in the generators to avoid vanishing and exploding gradient and applied different discriminator loss functions to fit the unpaired dataset better. The result demonstrated significant identifiability given the unpaired dataset it trains on.

2. Method

2.1. Introduction of CycleGAN

The proposed model implements CycleGAN, a preexisting model specialized in image-to-image translation [9]. Similar to the original GAN model, CycleGAN is an unsupervised approach that operates without the need for paired training data. The model consists of two types of networks, discriminators, and generators. A generator is trained to generate images that makes the discriminator incorrectly classify a generated image as real image in the training dataset, and a discriminator is trained to correctly discern the generated images from real images.

A typical CycleGAN model is made of two discriminators and two generators. One generator generates photograph-style pictures given sketches, and a corresponding discriminator tries to distinguish these generated photograph-style pictures from real photos. Another generator generates sketches given fake pictures from the other generator, and a corresponding discriminator is responsible to distinguish this new sketch from the original sketch. The first generator-discriminator pair is trained to produce pictures that look like real photos of human faces, and the second pair makes sure the generated picture looks like the original sketch.

2.2. Resnet block in generator

The generators are deeper than the discriminators and thus more susceptible to problems like exploding or vanishing gradient and training degradation in general. In order to enhance the model's resilience against these issues, this study introduced Resnet blocks, also known as residual blocks, into each generator's architecture, positioned between the encoder and decoder. These blocks include a normalization layer, two convolutional layers, and a ReLU layer in between. Incorporating Resnet blocks enables the model to effectively learn the residual function, thereby improving its robustness.

The residual function is defined by the function representing the difference between the target mapping and the input data:

$$F(x) = y - x \tag{1}$$

Where F(x) is the residual function, and y is the desired output given x. Optimizing the residual function to 0 provides an easier way to train the model to map x to the output y and makes the model gain better accuracy with increased depth [10].

DOI: 10.54254/2755-2721/50/20241168

The discriminator loss function is defined as mean square error between the generator output and the target values 0 (not real photo) or 1 (real photo). By adopting the structure known as Least Square GAN (LSGAN) instead of the commonly used sigmoid cross-entropy function, the model gains enhanced resilience against the vanishing gradient problem. In this setup, the loss function for the fake image generator is established as the mean square error between the output and the target values, which are 0 (indicating the discriminator is not deceived) or 1 (indicating the discriminator is deceived). The loss function of the generator producing sketches from generated photo-like pictures is also called the Cycle Consistency loss. It is an L1 loss function between the generated sketch and the original sketch.

2.3. Implementation details

In this setup, the Adam optimizer was employed to train the neural networks, utilizing a learning rate of 0.0002 and setting the (beta1, beta2) parameters to (0.5, 0.999) for both the discriminators and generators. Each discriminator is comprised of 5 convolutional layers, with the final layer serving as the classification layer.

ReLu function is applied on the output of each convolutional layer except for the classification layer, and batch norm is applied on the same outputs but except for the output of the first convolutional layer. The size of input image is 128×128×3, and each layer samples the input down by a factor of 2. The classification layer is the final convolution layer that converts the last convolution output into a single number as the output. Each of the generators is made of two sub-networks: an encoder and a decoder. The encoder compresses the image with three convolutional layers. Batch norm and a ReLu function are applied on each output of the convolutional layers. The output is then put into 6 consecutive Resnet blocks which don't change the dimensions of the data block. The data is then de-convolved by three transpose convolutional layers to a 128×128×3 image as output. Batch norm and ReLu functions are applied to the output of the first two transpose convolutional layers, but the last output goes through a tanh activation function.

2.4. Dataset

This research employs the CUHK Face Sketch Database, featuring 606 meticulously hand-drawn sketches derived from frontal photographs with controlled lighting and neutral expressions. For the photographic dataset, the Human Faces dataset from Kaggle, encompassing a wide range of demographics, ethnicities, ages, and facial profiles, was selected. These datasets were intentionally chosen based on their substantial size, with the aim of maximizing the effectiveness and efficiency of the proposed architecture and its potential usage scenario. The sketch domain, being the input of the model, has uniform art style, expression, and lighting. This is congruent to many of the actual situations where forensic sketches tend to have quite similar styles. On the photograph domain, photos taken under complex situations that are much more common in real life are select to make the generated image more realistic and recognizable.

3. Results and discussion

As Figure 1 shows, the discriminator and generator loss dip significantly during the initial stages of training, falling to approximately 0.5 and 3 respectively before reaching an equilibrium after 500 epochs. However, while the losses of the picture remain relatively stable as the number of epochs increases, the quality of the output does not. Figure 2 shows typical model outputs after training for 500 epochs, which are relatively clear realistic pictures. On the other hand, Figure 3 depicts the deterioration of the quality and variety of the generated images over time, with most sharing similar facial features. This is likely due to the susceptibility to mode collapse of GANs, where the generator focuses on a limited subset of the target distribution and fails to explore its entirety.

DOI: 10.54254/2755-2721/50/20241168

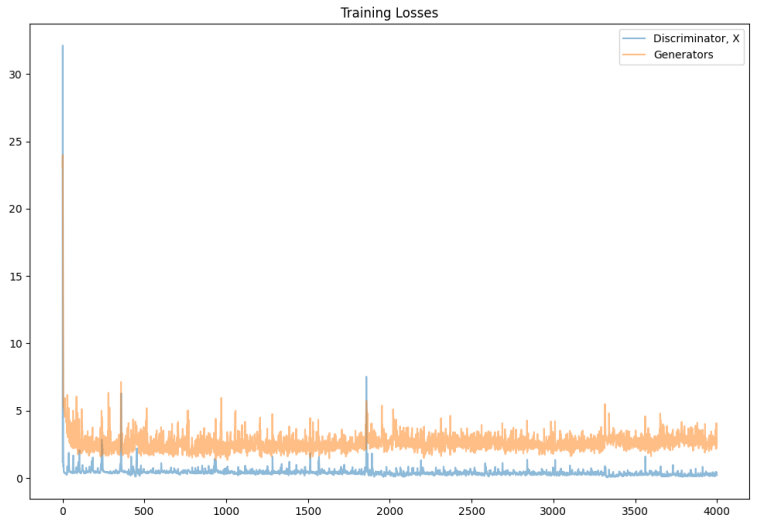

Figure 1. Training losses of the proposed model (Photo/Picture credit: Original).

Figure 2. Model output after 500 epochs (Photo/Picture credit: Original).

Figure 3. Model output after 1000 epochs (model collapse) (Photo/Picture credit: Original).

As shown in Figure 2, the output pictures are very similar to the original sketch in terms of facial structures and details. At the same time, the generated pictures have realistic and relatively natural lighting and skin tone. Considering the sketches as the baseline method used today, the generated pictures are arguably more recognizable to the public with the AI-generated details.

Likely because of the simplistic structure of the model, the model trains quite quickly. The results in Figure 2 were obtained after a mere 500 epochs. This swift training process and hence, flexibility, can be one of the benefits of implementing a single-GAN architecture like this.

The definition of the face is not clear in the output, and when the model trains for relatively large number of epochs, model collapse happens. These are probably caused by data preprocessing and selection that are not detailed and customized enough for the task. While training on unpaired datasets does yield realistic features in the output image, the Human Faces dataset may be too complex for the model to learn. Filtering out photos that are too different from the sketches in angle and lighting can be a promising direction to improve the model's performance.

4. Conclusion

This paper demonstrated the utilization of CycleGAN for sketch-to-image translation. Through modifications to the model structure, baseline loss function, hyperparameter tuning, and dataset selection, this study achieved notable enhancements over single-GAN network approaches. Model trainings with different parameters were conducted and the results were shown to be more recognizable

DOI: 10.54254/2755-2721/50/20241168

than the widely used forensic sketches. In the future, further study plans to improve this model by exploring more effective data selection and preprocessing techniques.

References

[1] Ahuja A S et al 2023 The digital metaverse: Applications in artificial intelligence, medical education, and integrative health. Integrative Medicine Research 12(1) 100917
[2] Qiu Y et al 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control 72 103323
[3] Soori M Arezoo B and Dastres R 2023 Artificial intelligence, machine learning and deep learning in advanced robotics, A review Cognitive Robotics
[4] Nikkath Bushra S and Uma Maheswari K 2021 Crime Investigation using DCGAN by Forensic Sketch-to-Face Transformation (STF)- A Review 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) Erode India pp 1343-1348
[5] Gunjate S Nakhate T Kshirsagar T and Sapat Y 2023 Sketch to Image using GAN International Journal of Innovative Science and Research Technology 8(1) pp 772–777
[6] Chao W et al 2019 High-Fidelity Face Sketch-To-Photo Synthesis Using Generative Adversarial Network 2019 IEEE International Conference on Image Processing (ICIP) Taipei Taiwan 2019 pp 4699-4703
[7] Xiaogang W and Xiaoou T 2005 Hallucinating face by eigentransformation in IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews) vol 35 no 3 pp 425-434 Aug
[8] Kazemi H F et al 2018 Unsupervised facial geometry learning for sketch to photo synthesis in 2018 International Conference of the Biometrics Special Interest Group (BIOSIG) pp 1–5
[9] Zhu J Park T Isola P and Efros A A 2017 Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks arXiv (Cornell University)
[10] He K Zhang X Ren S and Sun J 2016 Deep Residual Learning for Image Recognition 2016 IEEE Conference on Computer Vision and Pattern Recognition