Unsupervised Image-to-Image Translation on Style Transfer: A Comparison of CycleGAN and UNIT

Xuekun Wang **Tina Shen** **Sophie Tian**

Abstract

Unsupervised image-to-image translation is concerned with learning a mapping between images in two different domains using a training set of unpaired image examples. In this work, we implement and compare two such frameworks, named Cycle-Consistent Adversarial Networks (CycleGAN) and Unsupervised Image-to-Image Translation Networks (UNIT), to complete the task of style transfer. We conduct a small-scale sensitivity analysis on both algorithms and discuss our experimental findings through both quantitative and qualitative analysis. We found that UNIT outperformed CycleGAN on two style transfer datasets at the cost of computing power and training time.

1 Introduction

Image-to-image translation is a computer vision task that aims to learn a mapping between images in two domains. Many algorithms tackle this problem through unsupervised learning, where no paired examples are available. A popular area of application for such algorithms is style transfer, which converts the artistic style of an image from the source domain (e.g., a photo) to that of the target domain (e.g., a painting). In this paper, we compare the performance of two unsupervised image-to-image translation models, namely Cycle-Consistent Adversarial Networks (CycleGAN) [1], and Unsupervised Image-to-Image Translation Networks (UNIT) [2], focusing on the task of style transfer. While Liu *et al.* [2] claimed that UNIT outperformed CycleGAN in average pixel accuracy on the map dataset containing paired images [3], no comparison was done using unpaired datasets. Since both models were published in 2017, we are also curious to see why CycleGAN is more cited than UNIT (8000+ vs. 1500+ citations) if UNIT offers better performance [1, 2]. As a result, our goal is to gain a deeper understanding of both frameworks and to verify the claim of UNIT's superior performance by testing both frameworks against two unpaired style transfer datasets that UNIT did not report on: the season transfer of Yosemite images (summer2winter) and the artistic style transfer between Monet paintings and photos (monet2photo) [1]. Instead of using average pixel accuracy, we will use a metric designed for unpaired datasets and also conduct qualitative analysis for comparison.

2 Related Work

Unpaired image-to-image translation For unpaired image-to-image translation, the most effective approaches are developed based on Generative Adversarial Networks (GANs) [4]. A canonical work in this area is CycleGAN, developed by Zhu *et al.* [1], which introduced the concept of a cycle consistency loss. UNIT was released in the same year as CycleGAN, and its main difference is in making the shared latent space assumption, which implies cycle consistency [1, 2]. However, UNIT could experience unstable training due to its saddle point searching problem [2]. More recently, models such as MUNIT and DRIT extended translation models to enable many-to-many mappings between images across domains by decomposing images into a content code and a style code [5, 6]. However, MUNIT fails in tasks requiring strong geometric changes [7] and DRIT suffers from the mode collapse problem [8, 9]. Since our scope of analysis is confined to style transfer, which does not

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

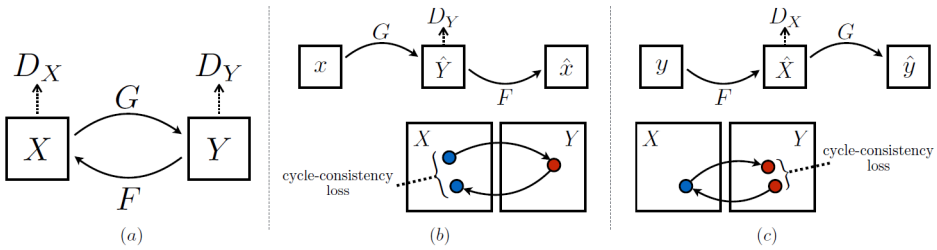

Figure 1: Model architecture of CycleGAN

require any geometric change, we decided to compare UNIT and CyleGAN instead of more recent, sophisticated models that would require more resources to train.

Style transfer Previous style transfer methods tackled example-guided style transfer, where a single image defines the target style [10, 11, 12]. Our focus, however, is on collection style transfer, in which a collection of images define the target style. For this task, conditional GANs-based image-to-image translation methods have been demonstrated to excel [13, 14, 15]. Since both CycleGAN and UNIT are built on conditional GAN-based objective functions, both models should be well-suited to perform style transfer [1, 2]. In particular, CycleGAN was demonstrated to produce natural-looking results in the style of the target domain and outperformed a classical neural style transfer method [1, 10].

3 Methods

3.1 CycleGAN

As illustrated in Figure 1 (a), given images in two domains $\{x_i\}_{i=1}^N \in X$ and $\{y_j\}_{j=1}^M \in Y$, CycleGAN [1] consists of two generators, G and F, and two adversarial discriminators, D_Y and D_X. The generators learn the mappings $G : X \rightarrow Y$ and $F : Y \rightarrow X$. The discriminator D_X aims to distinguish images $\{x\}$ from translated images $\{F(y)\}$, and similarly D_Y aims to distinguish between $\{y\}$ and $\{G(x)\}$. For the generator G and its discriminator D_Y, the adversarial loss is: $\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))]$, where G aims to minimize this objective against an adversary D_Y that tries to maximize it, i.e., $\min_G \max_{D_Y} \mathcal{L}_{GAN}(G, D_Y, X, Y)$. Similarly, another adversarial loss for the generator F and its discriminator D_X is introduced: $\min_F \max_{D_X} \mathcal{L}_{GAN}(F, D_X, Y, X)$.

CycleGAN also proposed that the mapping functions should be cycle-consistent: when an image $x \in X$ is translated to Y, and then translated back to X, the result should look like the original image, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, as illustrated in Figure 1 (b). Similarly, $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ (Figure 1 (c)). This behavior is enforced by a cycle-consistency loss: $\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]$.

Overall, the objective of CycleGAN is to obtain generators G and F through:

$$\min_{G,F} \max_{D_X, D_Y} \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F). \quad (1)$$

3.2 UNIT

UNIT is a framework proposed by Liu *et al.* [2] based on variational autoencoders (VAEs) and GANs. The main difference between UNIT and CycleGAN is that UNIT utilizes the shared latent space assumption, shown in Figure 2 (a), which assumes that the latent representation of a pair of corresponding images from two different domains share the same latent code. As seen in Figure 2 (b), UNIT contains two encoders, E_1 and E_2; two generators, G_1 and G_2; and two adversarial discriminators, D_1 and D_2 [2]. Given two image domains $\mathcal{X}_1$ and $\mathcal{X}_2$, VAE_1 for domain $\mathcal{X}_1$ consists of the encoder-generator pair $\{E_1, G_1\}$: the encoder E_1 maps an input image $x_i \in \mathcal{X}_1$ to a latent vector z, and the generator G_1 decodes a randomly perturbed version of

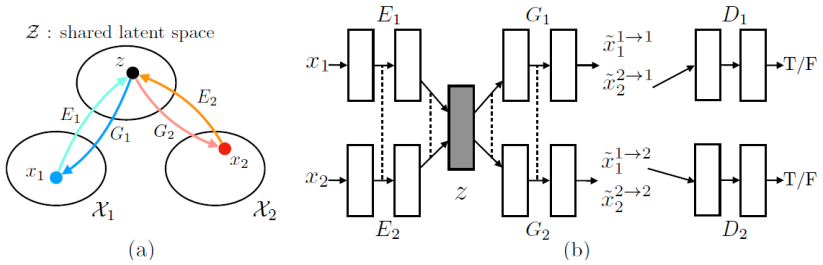

Figure 2: Model architecture of UNIT

z by adding noise from a multivariate Gaussian distribution. The VAE training ensures the reconstructed images and the original images are similar by minimizing a variational upper bound: $\mathcal{L}_{VAE_1}(E_1, G_1) = \lambda_1 KL(q_1(z_1|x_1)||p_\eta(z)) - \lambda_2 \mathbb{E}_{z_1\sim q_1(z_1|x_1)}[\log p_{G_1}(x_1|z_1)]$. The generator-discriminator pair $\{G_1, D_1\}$ is a GAN, which ensures the translated images resemble images in the target domain through a minimax objective similar to CycleGAN: $\mathcal{L}_{GAN_1}(E_2, G_1, D_1) = \lambda_0 \mathbb{E}_{x_1\sim P_{\mathcal{X}1}}[\log D_1(x_1)] + \lambda_0 \mathbb{E}_{z_2\sim q_2(z_2|x_2)}[\log(1 - D_1(G_1(z_2)))]$ [1]. Similarly, two additional losses are introduced for the second VAE and GAN: $\mathcal{L}_{VAE_2}(E_2, G_2)$ and $\mathcal{L}_{GAN_2}(E_1, G_2, D_2)$.

To enforce the shared-latent space assumption, VAE_1 and VAE_2 have shared weights on the high level layers between $\{E_1, E_2\}$ and $\{G_1, G_2\}$. Moreover, Liu *et al.* proved that the shared latent space assumption implies cycle-consistency, and hence introduced VAE-like loss functions to model the cycle-consistency constraint: $\mathcal{L}_{CC_1}(E_1, G_1, E_2, G_2)$ and $\mathcal{L}_{CC_2}(E_2, G_2, E_1, G_1)$ [1, 2].

Overall, the objective of UNIT is to obtain generators G_1 and G_2 through:

$$\min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} \mathcal{L}_{VAE_1}(E_1, G_1) + \mathcal{L}_{GAN_1}(E_2, G_1, D_1) + \mathcal{L}_{CC_1}(E_1, G_1, E_2, G_2) + \mathcal{L}_{VAE_2}(E_2, G_2) + \mathcal{L}_{GAN_2}(E_1, G_2, D_2) + \mathcal{L}_{CC_2}(E_2, G_2, E_1, G_1). \tag{2}$$

4 Experiments

Dataset We used the summer2winter dataset containing Yosemite images and monet2photo dataset containing landscape paintings and photos, both provided by Zhu *et al.* [1]. Due to limited computing power, we cut each dataset by $\sim 60\%$ to reserve 400 images for training and 100 images for testing in each domain. All images have been preprocessed to be 256×256 pixels [1].

Model implementation and training We leveraged the official implementations of CycleGAN and UNIT [1, 2] and made modifications to significantly simplify the code, which is available in our GitHub repository. We made additional modifications to read the datasets, save model parameters, resume training from checkpoints, perform hyperparameter tuning, and evaluate the models using the same metric. Due to limitations in time and computing power, we conducted a small-scale sensitivity analysis on hyperparameters, and trained the models by varying two hyperparameters on two levels for each model and each dataset. The hyperparameter settings can be found in the Appendix (Section 7.1). Each model was trained for 200 epochs. For comparison between CycleGAN and UNIT, we selected the hyperparameters that resulted in the best average translation score.

Quantitative results Since we used unpaired datasets, we were unable to compute the per-pixel accuracy as reported by the CyclGAN and UNIT work. Frchet Inception Distance (FID) [16] was therefore used to measure the distance between the source-domain and the target-domain data distributions using the features extracted by the inception networks [17]. A lower FID score indicates higher similarity between the two domain distributions. Table 1 shows the FID scores for the best-performing CycleGAN and UNIT models on the two selected datasets. We observe that UNIT performed better at image translation (i.e., A2B and B2A) for both datasets. Although CycleGAN performed better for image reconstruction tasks (i.e., lower FID for A2B2A and B2A2B tasks), this is not the main focus of image-to-image translation and does not indicate CycleGAN's superiority.

Table 1: Comparison of FID scores between CycleGAN and UNIT. A2B, B2A are image translation tasks and A2B2A, B2A2B are image reconstruction tasks. The average translation column shows the mean of FID scores between A2B and B2A columns.

Dataset	Model	A2B	B2A	A2B2A	B2A2B	Average Translation
summer2winter	CycleGAN	132.93	121.81	72.59	92.20	127.37
	UNIT	120.85	99.51	152.03	118.16	**110.18**
monet2photo	CycleGAN	181.80	196.24	134.90	137.62	189.02
	UNIT	209.28	156.55	225.31	262.64	**182.91**

Generally, both models achieved lower FID scores on the summer2winter dataset and we believe that this is caused by Monet's fuzzy painting style which made monet2photo a more challenging dataset. In terms of training time, although UNIT performed better in FID scores for both datasets, it is much more computationally expensive. Using the same machine, the run time for UNIT was 200 seconds per epoch while CycleGAN only required 120 seconds per epoch. We observed consistent run times for both datasets, demonstrating that CycleGAN offers superior computational efficiency. Therefore, given the same training time, CycleGAN could complete 67% more epochs and could potentially achieve better performance than UNIT.

Qualitative results Figure 3 to Figure 6 in Appendix 7.2 shows the image-to-image translation results of CycleGAN and UNIT on the selected datasets. For the summer2winter dataset, although both CycleGAN and UNIT learned to color the correct portions of the images, UNIT learned to make more visible changes. It can be observed that UNIT learned the distinct colourization between two seasons by creating a brighter and greener colour scheme for summer, and a cooler, greyer colour scheme for winter. CycleGAN, on the other hand, failed to make noticeable changes for either season transfer task, corresponding to higher FID scores (Table 1). For the monet2photo dataset, UNIT did not outperform CycleGAN for translating from Monet to photo. UNIT created dark and somewhat terrifying images that are hard to interpret, while CycleGAN created more photo-realistic images. However, UNIT performed significantly better in translating from photos to Monet-style images, capturing the fuzzy style of Monet and discarding browns and earth colors from the color palette. For this task, CycleGAN failed to make significant modifications to the images, resulting in a blurry version of photos. These qualitative results correspond to the quantitative results in Table 1, where CycleGAN outperforms UNIT in the Monet to photo translation, but UNIT outperforms CycleGAN in the photo to Monet translation.

Sensitivity analysis In our experiment, we varied the hyperparameter values to analyse their impact on the translation task performance (i.e., average translation FID score). Specifically for UNIT, setting the KL terms to 0.1 resulted in the best performance for the summer2winter dataset, while setting the terms to 0.01 gave a slight advantage for the monet2photo dataset. Overall, we confirm the claim made by Liu *et al.* [2] that setting the KL terms to 0.1 would result in good performance consistently. For the hyperparameter settings of CycleGAN, setting the learning rate to 0.0002 and the pooling size to 50 gave consistent good performance, which correspond to the default values recommended by Zhu *et al.* [1]. Increasing the learning rate or the pooling size would decreases the model performance.

5 Conclusion

In this paper, we implemented two unsupervised image-to-image translation models, CycleGAN and UNIT, and compared their performance on two style transfer datasets. From the experimental results, UNIT was able to outcompete CycleGAN quantitatively and qualitatively most of the time for this task. We attribute UNIT's performance to its shared latent space assumption which enhances the framework by modeling the latent space in a probabilistic way. However, we also came to understand why CycleGAN is more referred to than UNIT in the literature: the CycleGAN framework is more intuitive to understand, easier to implement and faster to train due to its simpler network structure, and requires less effort to tune (i.e., less hyperparameters). Overall, UNIT and CycleGAN completed the unsupervised image translation tasks using single-modal outputs under cycle-consistency constraints. Based on their unique contributions, further research work is able to expand image-translation tasks into different aspects such as modelling multi-modal outputs or even multi-domain image translation.

6 Attributions

For this project, all members of the team contributed equally. In particular, Tina implemented the UNIT model, Xuekun implemented the CycleGAN model, and Sophie conducted sensitivity analysis for both models. We also pair-programmed to ensure each step is correct. The final report and all other components of the project received equal contributions from each member of the team.

References

[1] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[2] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[5] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *CoRR*, abs/1804.04732, 2018.

[6] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. *CoRR*, abs/1808.00948, 2018.

[7] Taewon Kang and Kwang Hee Lee. Unsupervised image-to-image translation with self-attention networks. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 102–108. IEEE, 2020.

[8] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019.

[9] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. DRIT++: diverse image-to-image translation via disentangled representations. *CoRR*, abs/1905.01270, 2019.

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[11] Leon A. Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Preserving color in neural artistic style transfer. *CoRR*, abs/1606.05897, 2016.

[12] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *CoRR*, abs/1603.03417, 2016.

[13] Taewon Kang and Kwang Hee Lee. Unsupervised image-to-image translation with self-attention networks. *CoRR*, abs/1901.08242, 2019.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

7 Appendix

7.1 Model Hyperparameters

Table 2 shows the hyperparameters tuned for CycleGAN for both datasets. We decided to tune pool size to understand whether a larger pool size helps to stabilize the model training better, and to tune the initial learning rate to understand its effect on model convergence.

Table 3 shows the hyperparameters tuned for UNIT for both datasets. We decided to tune the two KL weights to investigate if setting a weight of 0.1 is the optimal choice as claimed by Liu *et al.* [2]. Similar to CycleGAN, we also tuned the initial learning.

Table 2: Hyperparameters tuned for CycleGAN

Pool Size	Initial Learning Rate
50	0.0002
100	0.0002
50	0.0005
100	0.0005

Table 3: Hyperparameters tuned for UNIT

KL Weights	Initial Learning Rate
0.1	0.0001
0.01	0.0001
0.1	0.0005
0.01	0.0005

7.2 Output Images from Test Sets

Figure 4 to Figure 6 below shows the output images for each translation task in the test set. In each figure, the first row shows the input figures from the source domain, and the second and third row show the translation results from the best-performing CycleGAN and UNIT models, respectively.

Figure 3: Summer Yosemite → winter Yosemite image translation results

Figure 4: Winter Yosemite → summer Yosemite image translation results

Figure 5: Monet → photo image translation results

Figure 6: Photo → Monet image translation results