CUT++: Unpaired Video-to-Image Style Translation

Pengxiang Zhu, Runze Guo, Shengyang Zhou†

***Abstract*—Image translation has been a topic of wide concern. Methods including Generative Adversial Networks (GAN), neural transfer as well as diffusion has been applied to this task. Various works have proposed methods that help maintain the content consistency of the generated image and the original image. In this paper, we employ contrastive learning by sampling the features to form patches. We propose an innovative way of fusing features from different patches using attention mechanism. We also developed a more refined loss that can better capture local information from feature blocks. Empirical results have shown that our methods achieves the best results among benchmarks of different categories on a wild dataset.**

***Index Terms*—Image Style Transfer, Contrastive Learning, Attention Mechanism**

I. Introduction

When it comes to image processing and artistic creation, style transfer is a captivating technique. Style transfer aims to combine the style of one image with the content of another, resulting in unique and stunning synthesized images. In recent years, Generative Adversarial Networks (GANs) have emerged as powerful tools for achieving style transfer. GANs are machine learning models composed of a generator and a discriminator, which compete and improve through adversarial training. The generator is responsible for producing realistic images, while the discriminator attempts to differentiate between the generated and real images. Through iterative training, GANs can learn the statistical features and styles of images, enabling the generation of highly realistic and uniquely styled images.

This article aims to explore practical techniques of using GANs for style transfer. We propose the *modified CUT++ architecture* based on CUT [1]. The main framework of our pipeline is illustrated in Figure 1. For individual reflection of each member, refer to Appendix of the report.

II. Related Works

A. GAN-based Style Transfer

Generative Adversarial Networks (GANs) [2] have achieved impressive results in the field of image generation tasks. The *adversarial loss* that enforces the generated image to be identical to the real image suits the ultimate goal of image generation well. For image translation and style transfer tasks, CycleGAN [3] introduced cycle consistency loss to confine the mapping from source to domain is cycle-consistent. CUT [1] went a step further by replacing cycle-consistency with contrastive loss to improve performance and efficiency. With increasing requirement of few-shot training, [4] proposes to relax the definition of realism and proposed an adapted GAN network.

B. Neural Style Transfer

Neural style transfer intends to generate a novel image by combining the content of one image with the style of another image. It formulates image-to-image translation as an optimization problem, with objectives being both the style loss and content loss. [5] first proposed to fuse features of style images into different layers of VGG [6] features from content images. [7] incorporated perception loss for feed-forward networks. Recent works including [8] followed this framework while refining the measurement between source and target distributions to formulate novel losses.

C. Diffusion-Based Image-to-Image Translation

Over the last few years, diffusion models have shown great success in generation tasks, and have been widely applied in image-to-image translation. EGSDE [9] employed energy-guided stochastic differential equations to accomplish the inference. InST [10] proposed inversion-based style transfer model that incorporates style-transfer and text-to-image sythesis. [11] introduced phasic content fusing diffusion model and proposed a novel direction distribution consistency loss to stabilize training.

III. Method

A. Problem Formulation

Given source images $\mathcal{X} \subset \mathbb{R}^{H\times W\times C}$ and target images $\mathcal{Y} \subset \mathbb{R}^{H\times W\times C}$, we intend to derive a function f such that for each $x \in \mathcal{X}$, $f(x)$ resembles images from $\mathcal{Y}$ in style. Different from the recently popular few-shot training, we follow the traditional approach that employs a relatively large amount of training data, both from the source and target domains, due to the fact that frames of videos might not be as illustrative as a delicately formulated style image.

B. GAN-Based Framework

GAN consists of a generator network G and a discriminator network D [2], and intends to minimize the difference between generated images and real images. Formally, we have

$$\min_G \max_D V(D,G) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))] \tag{1}$$

where x is the image generated by G, $D(x) \in [0,1]$ represents the probability of x being a real image. Equation 1 states that D intends to discriminate the generated images

† All three authors are senior students from Shanghai Jiao Tong University and are of equal contribution. Shengyang Zhou is the team leader.

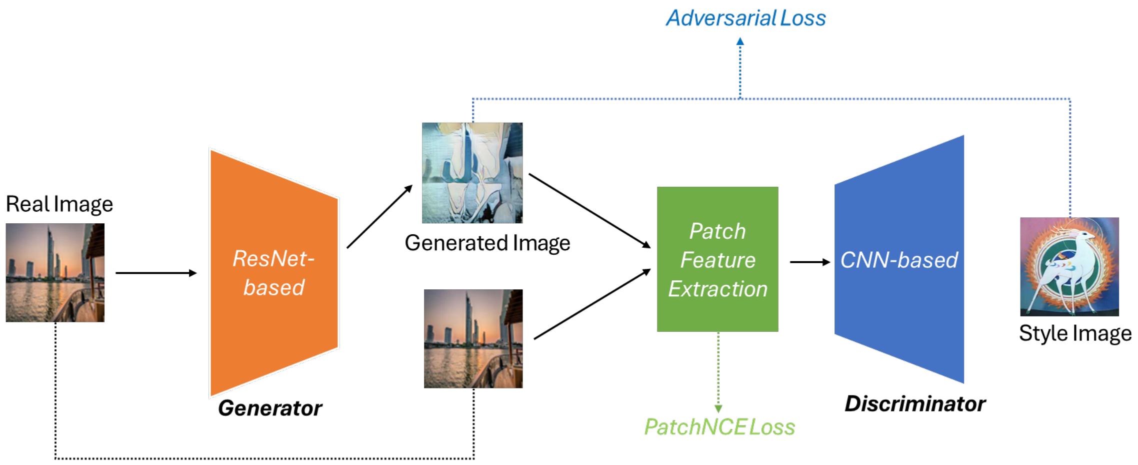

Fig. 1. Our framework for the proposed CUT++. It features a GAN-based main structure. The ResNet-based generator outputs an image while the patch feature extraction module and the CNN-based discriminator are responsible for the PatchNCE loss and adversarial loss respectively

and real images, while G does the opposite until equilibrium is reached.

CycleGAN [3] introduces two generators G and F intended to achieve $G(x) \in \mathcal{Y}$ and $F(y) \in \mathcal{X}$. A discriminator is paired with each generator to minimize the difference between the translated images and real images of that domain. As a basis, adversial loss is introduced along with GAN backbone.

$$\mathcal{L}_{\text{GAN}}(G, D, X, Y) = \mathbb{E}_y \log D(y) + \mathbb{E}_x \log(1 - D(G(x)) \quad (2)$$

In addition, cycle consistency loss is proposed to preserve the content of the original image, based on the intuition that the image generated from passing F and G consecutively should be similar to the original image, i.e. $F(G(x)) \sim x$.

CUT [1] replaces cycle consistency with maximization of mutual information based on contrastive learning [12]. The generator function G is breaked into two parts G_{enc} and G_{dec}, where G_{enc} generates the multi-layer image features and G_{dec} outputs the actual images. For L layers of feature from $G_{\text{enc}}(x)$, CUT applies a simple MLP network H_l to formulate a stack of features $\{z_l\}_L = \{H_l(G^l_{\text{enc}}(x))\}_L$. For each layer $l \in \{1, 2, \ldots, L\}$, the features z_l are discretized into patches. We denote $s \in \{1, 2, \ldots, S_l\}$ to be the index of patch at layer l. Similarly, we can decode the output image $\hat{y} = G(x)$ into $\{\hat{z}_l\}_L = \{H_l(G^l_{\text{enc}}(G(x)))\}$.

Following the terms of contrast learning, if we consider the sample $z^s_l \in \mathbb{R}^{C_l}$, then $\hat{z}^s_l$ is its only *positive pair* while all the other patches of the input $z_l^{S\backslash s}$ are all its the *negative pairs*. The loss between a single tuple $(z^s_l, \hat{z}^s_l, z_l^{S\backslash s})$ is formulated as follows in the form of cross-entropy loss with $v = z^s_l$, $v^+ = \hat{z}^s_l$, $v^- = z_l^{S\backslash s}$ and $\tau = 0.07$.

$$\mathcal{L}(v, v^+, v^-) = -\log \left[\frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{n=1}^{N} \exp(v \cdot v_n^-/\tau)} \right] \quad (3)$$

And the proposed *PatchNCE* loss that constrains the generated picture to reserve the original contents simply sums $\mathcal{L}$ over all s and l.

C. Optimizations based on CUT

The default setting of CUT employs StyleGAN2 [13]. We first replaced it by ResNet-based generator [14], [7]. Empirically, we discover that it can capture information about blurred and low-resolution images better.

Our optimizations are mainly focused on the feature extraction of the patches as well as the formulation of the *PatchNCE* loss. The detailed structure of this module is shown in Figure . In CUT, the patches are selected from five layers $l \in \{0, 4, 8, 12, 16\}$, and a simple MLP function $H_l(\cdot)$ maps the original features into $\mathbb{R}^{n \times F}$, where n is the number of patches and F is the feature dimension. However, CUT approaches the features of the patches individually and separately, which omits the inter-relationship between the randomly chosen patches. This might cause misinterpretation of local features and severs positional global information from the original image.

Inspired by [15] and [16] which take patches of images as inputs, we propose to add an attention layer after the original MLP. While [16] explored the semantic relationship between different patches on the original image, we intend to use attention mechanism to merge information between those patches. Considering the scale of the network G and D, we cannot use a full scale Vision Transformer (ViT) [15], but instead use a simple attention layer.

Formally, given patches $\{z^s_l\} \in \mathbb{R}^{n \times F}$, we formulate a two-headed multi-attention layer as follows, where $Q = K = V = [z^1_l, \ldots, z^s_l]$ and W_o is a linear mapping from $\mathbb{R}^{n \times 2F}$ to $\mathbb{R}^{n \times F}$.

$$\{z^s_l\}' = \text{concat}(\text{head}_1, \text{head}_2) W_o \quad (4)$$

$$\text{head}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (5)$$

The transformed patches $\{z^s_l\}'$ are then used for the calculation of *PatchNCE* loss. Originally, all samples $z^{S\backslash s}$ are considered as negative samples with equal weights.

However, when it comes to the characteristics of the patches, it is clear that the closer the patches of the image, the more

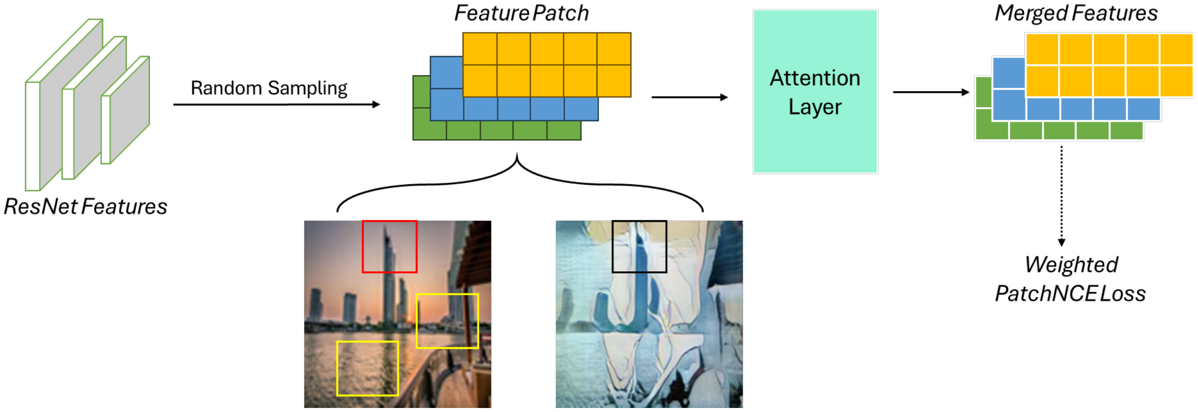

Fig. 2. The proposed structure for patch feature extraction.

similar the two should be. Therefore, we propose a method to modify the *PatchNCE* loss.

In order to modify the weights of N different negative samples based on the distance between patches, we construct a coefficient matrix Coe $\in \mathbb{R}^{N\times N}$. For the given patch $z_l^{s_1}, z_l^{s_2} \in \mathbb{R}^{C_l}$, the distance D_{s_1,s_2} is defined based on l_1 norm distance of their position in the image:

$$D_{s_1,s_2} = |s_{1,x} - s_{2,x}| + |s_{1,y} - s_{2,y}| \tag{6}$$

The elements of coefficient matrix would be generated with D_{s_1,s_2} and $\lambda \in \mathbb{R}$:

$$\text{Coe}_{(s_1,s_2)} = \lambda^{D_{s_1,s_2}} \tag{7}$$

The modified *PatchNCE* loss is shown as following:

$$\text{Positive\_loss} = \exp(v \cdot v^+/\tau) \tag{8}$$

$$\text{Negative\_loss} = \sum_{n=1}^{N} \exp\left(\sum_{i=1}^{N-1} \text{Coe}_{v,v_{n,i}^-}(v \cdot v_{n,i}^-/\tau)\right) \tag{9}$$

$$\mathcal{L}(v, v^+, v^-) = -\log\left[\frac{\text{Positive\_loss}}{\text{Positive\_loss} + \text{Negative\_loss}}\right] \tag{10}$$

Finally, considering the fact that the current image of size 256 × 256 can be blurred, which affects the final visual effect. Therefore, we added a pre-trained high-resolution module proposed by [17] at the end of the pipeline.

IV. Experiments and Results

In this section, we conduct extensive experiments on two datasets `trainA` and `trainB`, where `trainA` contains 6327 landscape and portrait images and `trainB` contains 1255 frames from a Chinese cartoon. The testing is done on dataset `test` which contains 711 images of similar style to `trainA`. The dataset `trainB` is highly challenging as it is directly sampled from a part in the cartoon episode with low sampling rate, resulting in many blurred images. We have selected pioneering methods featuring GANs, neural transfers and diffusion for comparison. In some works, only one style image can be selected for training and testing, and the results from those works are tested from models trained on a randomly chosen image in `trainB`. For CUT++, we train for 200 epoches with other hyper-parameters aligned with CUT (apart from the backbone changes).

The main metrics that we have selected for evaluation are Frechet Inception Distance (FID) [18] for the quality of image generation. To measure the similarity between the input image and the generated image, we also report SSIM and PSNR. The numerical results are illustrated in Table I, while some generated images from different methods are shown in Figure 3. Our light-weight model has a size of 11.2M parameters in total.

From Table I, we can see that our proposed CUT++ achieves the best FID score among all the methods. Ablation studies show that our method attained 11.3% improvement compared with the state-of-the-art CUT model, illustrating the effectiveness and robustness of the optimizations. We can also see that the generated results reach similar performance in PSNR, which shows that the transferred images still keep the main contents. The decrease in SSIM is expected as increase in FID can inevitably lead to loss in some features in the input image due to the adversarial characteristics of the GAN backbone.

From Figure 3, we can see that our model can adequately transfer the style of the target domain. Compared with CycleGAN [18] and CUT [1], we can see that our model can still preserve most of the contents. On the other hand, we can see that the two methods based on neural style transfer has relatively inferior performance. We deem that it can be attributed to the characteristics of the dataset, that is, no image in `trainB` is generalized enough to represent the whole dataset. Therefore, it is difficult to find an image that serves as the benchmark during the inference stage.

The loss curve during the training of CUT++ is shown in Figure 4. From the figure, we can clearly see that the adversarial loss for both the generator and the discriminator the fluctuates, indicating the mutual optimization of the two networks. We can also observe the gradual decrease of the optimized *PatchNCE* loss, which illustrates better matching between the original image and the output image during the training process.

The single-modal GAN-based approach cannot generated diverse images from a single given input. Therefore, we

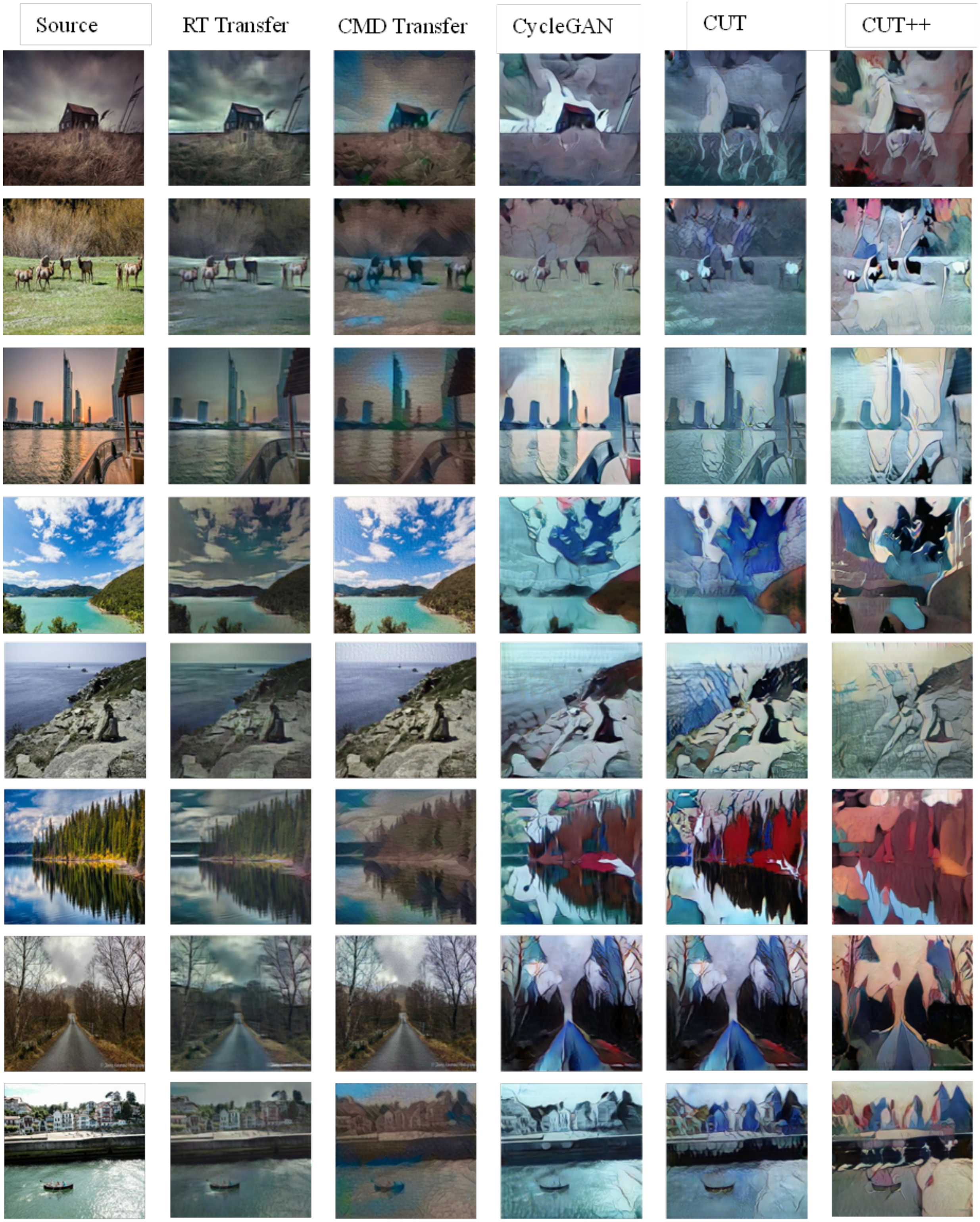

Fig. 3. The style transferring results on selected images. We can see that our proposed CUT++ carries more style features than other works, while maintaining most of the original content information. We can see that in this specific dataset, GAN-based model can achieve superior performance.

TABLE I
EXPERIMENT RESULTS ON THE TEST DATASET

Method	Category	Training	FID ↓	PSNR ↑	SSIM ↑	LPIPS
RT Transfer [7]	Neural Transfer	`trainA` and `trainB`	224.28	27.94	**0.60**	-
CMD Transfer [8]	Neural Transfer	`trainA` and `trainB`	203.48	**28.24**	0.49	-
CycleGAN [3]	GAN	`trainA` and `trainB`	156.34	28.16	0.56	-
CUT-StyleGAN [1]	GAN	`trainA` and `trainB`	189.69	-	-	-
CUT-GResNet	GAN	`trainA` and `trainB`	136.61	28.16	0.47	-
CUT++ (Ours)	GAN	`trainA` and `trainB`	**120.43**	28.04	0.40	-
CUT++ with InST	GAN and Diffusion	`trainA` and `trainB`	156.91	28.04	0.40	**0.25**

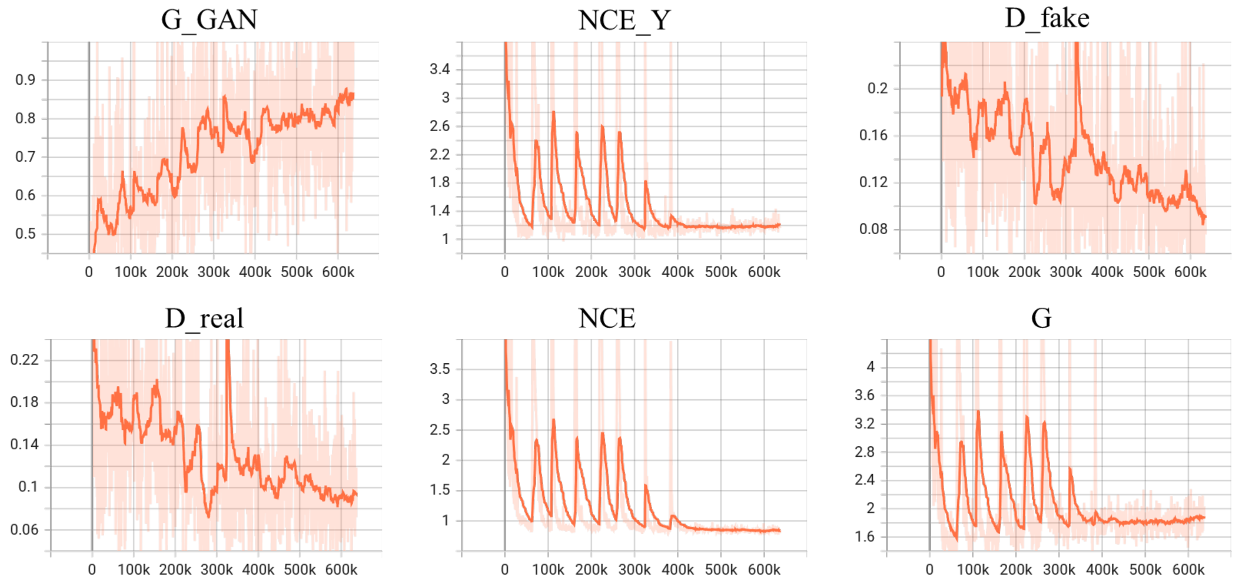

Fig. 4. The loss curve for CUT++ training during the first 50 epoches (not the full version due to time limit on GPU). G_GAN represents the main adverserial loss, while NCE and NCE_Y represents the content consistency loss. D_fake and D_real are the cross-entropy loss of the discriminator.

employ InST[10] as a post-processing module that possess the ability to generate different images given a single transferred image $\hat{y}$ from CUT++. As a post-processing module, we conduct light-weight training on a single image randomly selected from `trainB`. During the inference stage, we set strength as 0.3 to enable certain level of diversity while still preserve the overall features. By employing this post-module, we can calculate the LPIPS metric that measures the diversity of the generation, as reported in Table I. Some generated results are shown in Figure 5. From the result, we can see certain level of diversity regarding the color patches. However, empirical results illustrate the the FID values rise significantly after the modification. Therefore, we will **NOT** count it as our main model, but merely provide this method as a way of creating diversity and calculate the *optional* value of LPIPS.

V. Conclusion

To conclude, in this paper we propose CUT++ based on CUT[1] that achieves better FID results on the given dataset. We employed attention mechanism and optimized the *Patch-NCE* loss. We deem that our method can extract more features out of the style images yet still retain most features of the image features.

For further work, we can try more extensive modifications of the model in an attempt to achieve better result. For example, subsequent works of CUT including [16] elaborate on the semantics of the patches.

We thank Prof. Liu and all the TAs for their dedications throughout the course.

Appendix
Individual Reflection

In this project, I am mainly responsible for formulating the direction of the research and code implementation. From this project, I have developed a deeper understanding towards the field of image style transfer from extensive literature reviews and code replications. I have also tried my best to incorporate modules to derive a better result and propose some innovations. There are several points worth mentioning during the long months of exploring.

During the project, I have also tried to make some adjustments on a neural-style transfer based work [8]. For those methods, it is common practice to incorporate the features extracted from VGG and merge them in some way. Specifically, they tend to formulate the task as an optimization problem that has the generated image as decision variables and minimizing the two losses that controls both the style and the image. In my opinion, the method itself is lightweight and intuitive, but it is also very difficult to create something new while still achieve

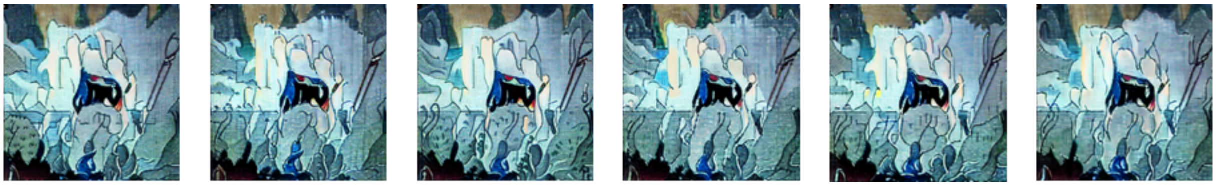

Fig. 5. Results after the InST-based post-processing module

decent results. The only possible modification that we have tried is to replace the VGG backbone with newer and heavier modules like ResNet, yet it did not attain decent results.

In the course of this project, we have explored and tried around 10 methods of various kinds. However, during the first few works, we wrongly assumed that the accuracy of the FID values is not (or just weakly) associated with the number of images. That results in some wrong judgements about the real performance of a method (*e.g.* InST). This cost us some time, but also reminds us that being especially careful about how the final metric is calculated is pivotal. Had we thought of testing the FID values using all the images in `trainB` (i.e. `mural`), we would perhaps have chosen a diffusion-based module at an earlier stage and try to work on that.

Finally, I want to comment on the metric LPIPS, which is an optional metric mentioned in the requirements of the project. Due to the limited time I have, I cannot cover all the works regarding image transfer. Yet, StarGAN(v2) is the only GAN-based method that I have read that possess the ability to generate diverse images. This pushes most groups to focus on InST and other diffusion-based methods, which are much too complicated for junior students to derive certain kind of **innovation**, neglecting what I think is the key to any of those open-answered projects.

During the main part of the report, we have stated that we employed a relatively clumsy approach to fit a heavy InST at the back of the pipeline just to meet the requirements. But obviously, this is not the best option. I have read that it is possible to integrate the original model in the StarGANv2 to generate diverse images, but in my opinion, this is still not elegant enough and there is not enough time to carry out this modification. So this is a slightly disappointing part and can be perhaps improved.

References

[1] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 319–345.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] U. Ojha, Y. Li, J. Lu, A. A. Efros, Y. J. Lee, E. Shechtman, and R. Zhang, "Few-shot image generation via cross-domain correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 743–10 752.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

[8] N. Kalischek, J. D. Wegner, and K. Schindler, "In the light of feature distributions: moment matching for neural style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9382–9391.

[9] M. Zhao, F. Bao, C. Li, and J. Zhu, "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3609–3623, 2022.

[10] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 146–10 156.

[11] T. Hu, J. Zhang, L. Liu, R. Yi, S. Kou, H. Zhu, X. Chen, Y. Wang, C. Wang, and L. Ma, "Phasic content fusing diffusion model with directional distribution consistency for few-shot model adaption," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2406–2415.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] C. Jung, G. Kwon, and J. C. Ye, "Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 260–18 269.

[17] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1905–1914.

[18] 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," *Advances in neural information processing systems*, vol. 30, 2017.