2024 IEEE International Conference on Big Data and Smart Computing (BigComp) | 979-8-3503-7002-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIGCOMP60711.2024.00038

CartoonizeDiff: Diffusion-based Photo Cartoonization Scheme

Hwyjoon Jeon
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
hwyjeon@korea.ac.kr

Jonghwa Shim
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
indexlibrorum3822@korea.ac.kr

Hyeonwoo Kim
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
guihon12@korea.ac.kr

Eenjun Hwang
School of Electrical Engineering
Korea University
Seoul, Republic of Korea
ehwang04@korea.ac.kr

***Abstract*— Photo cartoonization seeks to create cartoon-style images from photos of real-life scenes. So far, diverse deep learning-based methods have been proposed to automate photo cartoonization. However, they tend to oversimplify high-frequency patterns, resulting in images that look like abstractions rather than a true animation style. To alleviate this problem, this paper proposes CartoonizeDiff, a new photo cartoonization method based on diffusion model and ControlNet. In the proposed method, Color Canny ControlNet and Reflect ControlNet are appended to a pretrained latent diffusion model to preserve the color, structure, and fine details of photos for better cartoonization. Through extensive experiments on animation backgrounds and real-world landscape datasets, we demonstrate that the proposed method quantitatively and qualitatively outperforms existing methods.**

Keywords—Photo Cartoonization, Controllable Diffusion Model, Diffusion Model, Generative Model

I. INTRODUCTION

Cartoons are a visual art form with a variety of subgenres. Animation, a representative subgenre of cartoons, can vividly tell stories through vibrant colors, elaborate backgrounds, special effects, and dynamic movements of characters and backgrounds. In animation, backgrounds are especially important because they can provide the background necessary for story development or effectively indicate transitions in the story. However, drawing a background is a difficult and expensive task, as it requires not only creativity and skill, but also a significant amount of time from the artist. To alleviate this, diverse photo cartoonization methods have been proposed to convert real photos into animation-style images. The traditional method is to use image editing software such as Photoshop, but it still requires a significant amount of work. Recently, many efforts have been made to automate photo cartooning using deep learning models such as GAN (Generative Adversarial Networks) [1] and Diffusion Probabilistic Model [2].

GAN-based photo cartoonization typically uses loss functions to reflect cartoon features such as clear outlines, distinct colors, and abstract representation [3]–[5]. However, this process smooths out high-frequency patterns that represent brightness, edges, and texture, creating a simplified representation of a photograph rather than a cartoon.

On the one hand, diffusion models have recently received considerable attention due to their outstanding performance in image generation. For instance, text-to-image (T2I) diffusion models combined with text encoders like CLIP [6] have demonstrated the ability to generate images that closely resemble the description provided by text. They have become very popular with the public because they can produce images of great visual quality through text prompts. Likewise, a T2I model trained on an animation dataset can generate animation images based on text. However, the limitations of text in describing all the visual details of an image are ultimately reflected in the resulting image. On the other hand, image-to-image diffusion models succeed in converting portraits into a cartoon style, but the background excluding the subject in the photo is still distorted or blurred.

To address this problem, in this paper, we propose a novel photo cartoonization scheme that preserves the color, structure, and details of photos by appending multiple ControlNets [7] to a diffusion model. The diffusion model used in the paper is the publicly available pretrained latent diffusion model known as Stable Diffusion [8]. We use two ControlNets in the cartoonization to maintain both the color and structural details of the original photo. Through various experiments, we show that the proposed method outperforms existing methods in terms of popular evaluation metrics.

The main contributions of this paper are as follows.

- We propose a new diffusion-based photo cartoonization method that effectively preserves the color and structural details of input photos.
- We demonstrate the proposed method can generate high-quality backgrounds for animation compared to existing methods.
- We verify the effectiveness of the proposed method through comparative experiments and ablation studies.

This work was partly supported by NRF (National Research Foundation of Korea) (No. NRF-2021R1A4A1031864) and NRF (No. RS-2023-00252257) grant funded by the Korean Government(MSIT).

2375-9356/24/$31.00 ©2024 IEEE
DOI 10.1109/BigComp60711.2024.00038

The structure of this paper is as follows. In Section 2, we introduce several related works. In Section 3, we explain the proposed model in detail. Section 4 verifies the performance and effectiveness of the proposed model through comparative experiments and ablation study. Finally, Section 5 concludes this paper.

II. Related works

2.1 Photo Cartoonization

Photo cartoonization aims to convert photographs of real-world scenes into a cartoon style. Unlike other art forms, cartoons have unique characteristics such as smooth surfaces, clear edges, and highly simplified textures that differ from other artistic styles. Recently, various GAN-based photo cartoonization methods have shown great potential by capturing these cartoon characteristics well. For instance, CartoonGAN [3], an improvement on CycleGAN [9], generated clear edges and smooth shading using semantic content loss calculated through a pretrained VGG network and edge-promoting adversarial loss. Since then, various methods have been proposed to further improve the CartoonGAN structure, such as ComixGAN [10] and AnimeGAN [5]. On the other hand, Wang et al. [6] introduced a GAN-based white-box photo cartoonization framework. They broke down images into surface, structure, and texture representations and managed the output styles by adjusting the weight of each representation. Furthermore, Lee et al. [29] introduced photo cartoonization method for arbitrary style transfer. However, these GAN-based models tended to smooth out high-frequency patterns representing object textures, brightness, and details, resulting in monotonous cartoonized images. In other words, smoothing is an advantage for portraits, but a disadvantage for landscapes.

2.2 T2I diffusion models

Over the past few years, T2I diffusion models have gained great popularity in many fields through the utilization of large-scale text-image pair data and diffusion models. A. Nichol et al. [11] introduced text conditions into the diffusion model and demonstrated that classifier guidance produces better visual results. A. Ramesh et al. [12] proposed DALLE-2, a 2-stage image generation model based on CLIP, a contrastive model that effectively learns robust representations of images. It generates CLIP image embeddings from text and uses them as conditions to create images with a diffusion model. Through this process, it generates visually plausible images with a high degree of similarity to the given caption. Meanwhile, C. Saharia et al. [13] utilized a pretrained large language model to generate photorealistic images. On the other hand, a latent diffusion model called Stable Diffusion was proposed to perform denoising in the latent space of the autoencoder, effectively reducing computing resources while maintaining the quality of the generated images [8]. In addition, since its release, it has been fine-tuned using a variety of custom datasets and shared within AI model hubs such as HuggingFace [20] and CivitAI [21].

2.3 Controllable diffusion models

Controllable diffusion models were proposed to allow T2I diffusion models to tailor the generated results more effectively. This enables T2I diffusion models to generate images that match given conditions not only in the form of text, but also in various types of conditions such as canny edge, depth map, and semantic segmentation. In [14], a large diffusion model called Composer was trained from scratch to get controllability for single and multiple conditions, but the training cost was too high. On the other hand, ControlNet [7] and T2I-Adapter [15] were introduced as lightweight adapters for open-source Stable Diffusion models. By fine-tuning these lightweight adapters while freezing the parameters of original Stable Diffusion models, training costs can be significantly reduced, making them viable for the research community.

III. Proposed Method

In this section, we will first briefly review the diffusion process and ControlNet, and then present our photo cartoonization scheme.

3.1 Preliminary

Diffusion process [2] has a forward process $q(x_T|x_0)$ which is T-step Markov Chains that gradually add small amounts of noise to x_0, and a corresponding backward process $p_\theta(x_T|x_0)$ that denoise x_T to recover x_0.

$$q(x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \quad (1)$$
$$= \prod_{t=1}^{T} N\left(\sqrt{1-\beta_t}x_{t-1}; \beta_t I\right)$$
$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}, \mu_\theta(x,t), \Sigma_\theta(x_t, t)) \quad (2)$$

The hyperparameter β_t determines the noise strength of at each step. $\mu_\theta(x,t)$ and $\Sigma_\theta(x_t,t)$ are mean and standard deviation predicted by network p_θ at step t. The loss function for training can be achieved by optimizing the variational bound of negative log likelihood, expressed as (3).

$$L = E[-\log p_\theta(x_0)] \le E[-\log \frac{p_\theta(x_0:T)}{q(x_1:T|x_0)}] \quad (3)$$

ControlNet [7] is a lightweight adapter that can be attached to pretrained T2I diffusion models to reflect additional conditions. ControlNet clones the weights of the pretrained diffusion model and makes a trainable copy and locked copy as Fig. 1. Trainable copy is trained with task-specific datasets to learn conditional control, while the locked copy remains as it is to conserve the original model's capability. Using 2D feature as an example, given a feature map $x \in R^{h\times w\times c}$ with $\{h, w, c\}$ refers to height, width, and channel numbers, a neural network

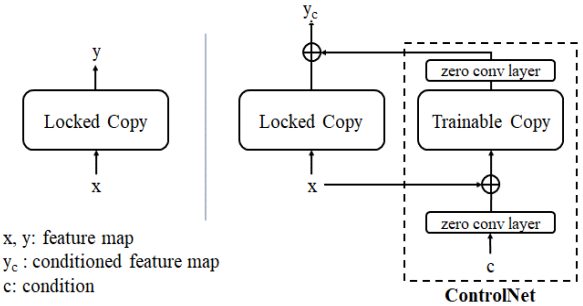

Fig. 1. ControlNet.

block $F(x; \Theta)$ transforms x into another feature map y as (4), where Θ is a set of parameters.

$$y = F(x; \theta) \tag{4}$$

The trainable copy clones all parameters in Θ and makes Θ_c. ControlNet consists of two zero-convolution layers and one trainable copy. The zero-convolution layer is a 1×1 convolution layer in which both the weight and bias are initialized as zero. Denote zero convolution operation as $Z(\cdot;\cdot)$, each layer has parameters $\{ \Theta_{z1} , \Theta_{z2} \}$. ControlNet's output y_c is expressed as:

$$y_c = F(x; \theta) + Z(F(x + Z(c; \theta_{z1}); \theta_c); \theta_{z2}) \tag{5}$$

3.2 Proposed Model

This section describes the overall structure of the proposed model. The model has a pretrained Stable Diffusion and two lightweight ControlNets-based adapters. The two adapters are Color Canny ControlNet [26] and Reflect ControlNet. The former maintains the cartoon style but cannot accurately reflect color and structural characteristics. Therefore, we propose the latter adapter to accurately reflect all the information, including the details of the photo. By leveraging these two adapters, the proposed method can preserve the color, structure, and details of the picture.

Fig. 2 shows the overall structure of the proposed photo cartoonization model. In the figure, Color Canny ControlNet is a pretrained model with color canny edges. A color canny edge is a composite image of a canny edge and color palette. However, the canny edge has limitations in detecting details in photos. Also, the color palette ignores the colors of small objects as it is represented by the averaged pixel color in the block area. Because of these characteristics, Color Canny ControlNet can create cartoon-style images, but it cannot accurately capture the color and structural properties of a photo. To alleviate this problem, we introduce Reflect ControlNet trained on the photos themselves. Because Reflect ControlNet uses the entire photo as conditional input, it can create images that are very similar to photos. By using these two ControlNet structures, we can leverage the advantages of both ControlNets and offset their disadvantages.

The process of reflecting conditional information is as follows. The output of Color Canny ControlNet denoted zero convolution layer Z' with parameters $\{ \Theta_{z1}, \Theta_{z2} \}$ and condition as c_1, can be expressed as (6).

$$y_{c_1} = F(x; \theta) + Z'(F(x + Z(c; \Theta_{z1}); \Theta_c); \Theta_{z2}) \tag{6}$$

Reflect ControlNet, denoted zero convolution layer Z'' with parameters $\{ \Theta_{z3}, \Theta_{z4} \}$ and condition as c_2, can be expressed as (7).

$$y_{c_2} = F(x; \theta) + Z''(F(x + Z(c_2; \theta_{z3}); \theta_{c_2}); \theta_{z4}) \tag{7}$$

Then, multi-ControlNet can be simply implemented by adding the outputs of each ControlNet, expressed as (8).

$$y_c = F(x; \theta) + Z'(F(x + Z(c; \theta_{z1}); \theta_c); \theta_{z2}) + Z''(F(x + Z(c_2; \theta_{z3}); \theta_{c_2}); \theta_{z4}) \tag{8}$$

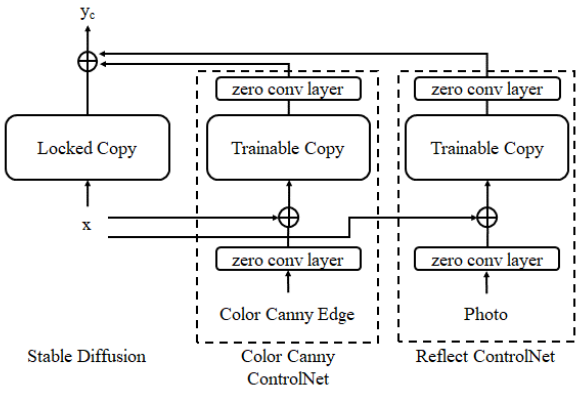

Fig. 2. CartoonizeDiff architecture.

IV. Experiments

To evaluate the effectiveness of the proposed photo cartoonization scheme, we performed extensive experiments using two public datasets. We first briefly describe the datasets used in our experiments and present training details for each ControlNet. Lastly, we present performance comparisons with other comparative methods and ablation studies.

4.1 Experimental setup

Datasets: In this paper, we used two datasets for training and one dataset for testing. For training, ControlNet requires three components: image, text, and condition image. A brief description of the datasets used is as follows.

- LAION ART [22] is a subset of the LAION-5B dataset, which includes 5B image-text pairs filtered through CLIP. Developed for research, it employs several lightweight models to predict aesthetic ratings for each image on a scale from one to ten. Color Canny ControlNet is trained with 2.6 M pairs of color canny edge processed images of LAION-ART.

- Danbooru2020 [23] is a massive collection of animated images containing over 4.2M images, each annotated with a part of the total 130 million text tags. As this dataset provides detailed descriptions of the image contents, it holds significant utility for machine learning tasks, including image recognition and generation. It's widely used in various fields, particularly in generative modeling. We used 100K pairs to train Reflect ControlNet and the images were used directly as condition images.

- LHQ1024 [24] dataset contains 90K high-quality images at 1024×1024 resolution crawled from Unsplash and Flickr and preprocessed with Mask R-CNN [27] and Inception V3 [28]. It contains a wide variety of natural and urban landscapes. We utilized 20K images for quantitative evaluation and qualitative comparison.

Performance metrics: In our experiments, we used two evaluation metrics. The first metric, Fréchet inception distance (FID) [16], is a key metric for evaluating image quality and performance. It compares the distribution between two image sets, calculated through the L2 distance of feature map

TABLE I. Comparison of FID and SSIM scores of our method and other comparative models.

Metric	Models			
	Photo	White-Box	Anime-GAN	CartoonizeDiff (Ours)
FID to animation ↓	154.26	120.39	121.84	109.37
SSIM to photo	1.0	0.845	0.828	0.605

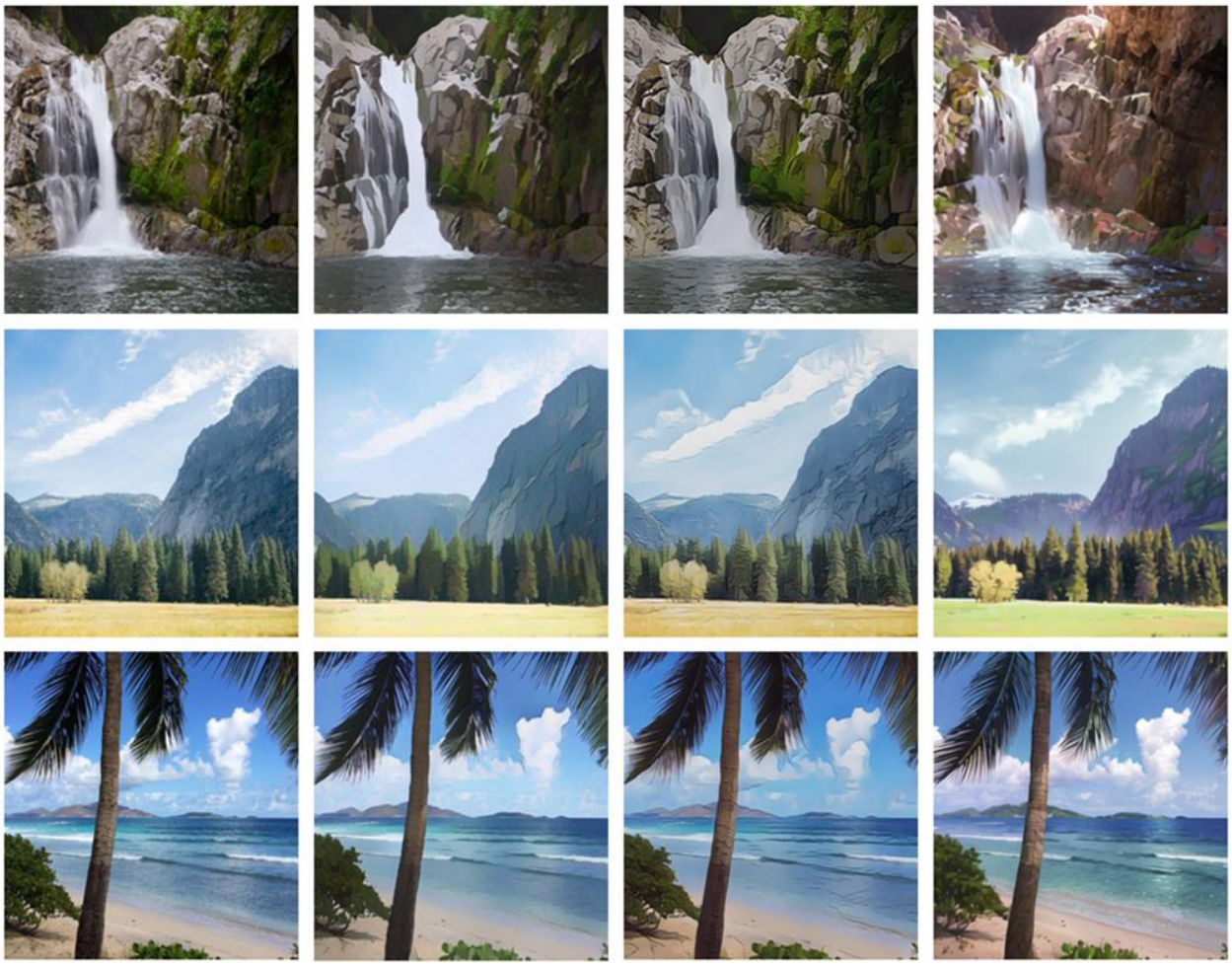

Photo White-Box Anime-GAN CartoonizeDiff (Ours)

Fig. 3. Generated image comparison with existing models. CartoonizeDiff generates better-quality animation scenery images conserving the texture and details of input photo.

distributions. A lower FID score indicates better photo cartoonization. The second metric, the Structural Similarity Index (SSIM) [25], was developed for evaluating the quality of images and videos. It measures the similarity between two images by considering factors such as brightness, contrast, and structural information. A higher SSIM score indicates a greater similarity between the two images.

Baselines: We compared our proposed method with two representative photo cartoonization models, White-Box and Anime-GAN. For diffusion model, we take commonly used prompts from Stable Diffusion, where the positive text prompts are "best quality", "intricate details" and the negative text prompts are "low resolution", "cropped", "worst quality" and "blurry."

4.2 Quantitative Results

We evaluate the quality of the generated images using the FID score. For comparison, we used background datasets collected from Your Name [17], Weathering With You [18], and

TABLE II. Comparison of FID and SSIM using different ControlNets.

	Color Canny ControlNet	Reflect ControlNet	Color Canny + Reflect ControlNet
FID to animation ↓	106.99	120.39	109.37
SSIM to photo	0.331	0.792	0.605

(a) Photo (b) Color canny edge (c) Color Canny ControlNet (d) Reflect ControlNet (e) CartoonizeDiff (Ours)

Fig. 4. (a), (b) are conditional inputs and (c), (d), (e) are images generated by different ControlNets. For all models, the positive text prompt is "best quality, extremely detailed" and the negative text prompt is "low resolution, cropped, worst quality, blurry."

Suzume [19]. We cartoonized 20,000 images from LHQ1024 using White-Box, Anime-GAN, and CartoonizeDiff and compared their FID to animation as shown in Table 1. The table shows that CartoonizeDiff can produce images that are closer to the animation domain than other methods. Also, when considering both FID to animation and SSIM to photo, White-Box and Anime-GAN are far from the animation background and similar to photos, while our results are closer to the animation domain and relatively distant from photos.

4.3 Qualitative Results

We also performed some qualitative comparisons as shown in Fig. 3. The figure shows images generated using the proposed method and the comparative models. In the figure, we can observe that White-Box and Anime-GAN abstract elements such as trees, sea, and clouds and do not preserve the original texture of the objects. In contrast, our model successfully transformed the image into a cartoon by effectively retaining fine details.

4.4 Ablation Studies

This section describes the ablation studies that we conducted to demonstrate the effectiveness of both ControlNets in photo cartoonization. Fig. 4 shows the generated outputs of each ControlNet. Table 2 compares FID to animation and SSIM to photo of each ControlNet. Fig. 5 shows the results of changing only Reflect ControlNet's weight while the Color Canny ControlNet's weight is fixed to 1.0. Table 3 shows the FID to animation and the FID to photo under different Reflect ControlNet weights.

Color Canny ControlNet: Fig. 4 (c) shows the images generated under the conditions of Fig. 4 (b). In particular, the third row of the figure shows that when the color block of the color canny edge is set to its average color, the generated image has a different color than the original photo. In addition, Table 2 shows that Color Canny ControlNet can produce images that are close to cartoons but do not preserve the information in photos well.

TABLE III. Comparison of FID scores using different Reflect ControlNet weights. (Color Canny ControlNet weight 1.0)

ControlNet Weight	Reflect ControlNet					
	0	0.2	0.4	0.6	0.8	1.0
FID to animation ↓	106.99	109.37	121.09	121.71	121.280	121.96
FID to photo ↑	42.47	30.52	20.85	20.92	22.61	23.87

Fig. 5. Images generated using different Reflect ControlNet weights.

Reflect ControlNet: Fig. 4 (d) shows the images generated under the conditions of Fig. 4 (a). In addition, Table 2 shows that this ControlNet can create images that preserve the color, structure, and details of the input photos, but its high FID to cartoon score indicates poor photo cartoonization.

CartoonizeDiff: Fig. 4 (e) shows images generated using both Color Canny ControlNet and Reflect ControlNet. Table 2 shows that our CartoonizeDiff achieved a similar FID to animation score to Color Canny ControlNet and a relatively high SSIM to photo score. Considering the tradeoff between FID to animation and SSIM to photo, CartoonizeDiff can achieve better photo cartoonization performance than a single ControlNet.

Weight of Reflect ControlNet: We set Color Canny ControlNet to 1.0 and tried to find the optimal weight by changing the weight of Reflect ControlNet. As weight increases, the generated images retain photographic details but loses the cartoon characteristics. In particular, Table 3 shows that both FID to animation and FID to photo experience rapid changes in the weight range from 0 to 0.4. This indicates that the cartoon style is better reflected in that weight range. Therefore, we set the weight of Reflect ControlNet to 0.2, considering both FID to cartoon and FID to photo scores.

V. Conclusion

In this paper, we proposed a photo cartoonization method that preserves the colors, structure, and details of input real photos. To do that, we appended Color Canny ControlNet and Reflect ControlNet to a pretrained diffusion model. In the experiments using animation background dataset and real-world landscape dataset, we showed that the proposed method outperforms existing models quantitatively and qualitatively. We also showed in the ablation studies that combining Color Canny ControlNet and Reflect ControlNet together can achieve the best image quality.

References

[1] I. Goodfellow, et al., "Generative adversarial networks," Communications of the ACM, vol. 63, no.11, pp. 139-144, 2020.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," 34th Conference on Neural Information Processing Systems (NeurIPS 2020)

[3] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. "Cartoongan: Generative adversarial networks for photo cartoonization," In Proceedings of the IEEE conference on computer vision and pattern recognition. 9465–9474, 2018

[4] YX. Wang and J. Yu, "Learning to cartoonize using white-box cartoon representations," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8090–8099, 2020

[5] J. Chen, G. Liu, and X. Chen, "AnimeGAN: A Novel Lightweight GAN for Photo Animation," in Communications in Computer and Information Science, Springer, pp. 242–256, 2020

[6] A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” Proceedings of the 38th International Conference on Machine Learning, pp. 8748-8763, 2021.

[7] L. Zhang and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3836-3847, 2023

[8] R. Rombach, A. Blattmadnpselnn, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684-10695, 2022

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2223-2232, 2017

[10] M. Pęśko, A. Svystun, P. Andruszkiewicz, P. Rokita, and T. Trzciński, “Comixify: Transform video into a comics,” Fundamenta Informaticae, vol. 168, no. 2-4, pp. 311-333, 2019

[11] A. Nichol et al., “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,” arXiv preprint: http://arxiv.org/abs/2112.10741, 2021

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical Text-Conditional Image Generation with CLIP Latents,” arXiv preprint arXiv:2204.06125, 2022

[13] C. Saharia et al., “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” arXiv preprint: 2205.11487, 2022

[14] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, “Composer: Creative and controllable image synthesis with composable conditions,” arXiv preprint arXiv:2302.09778, 2023

[15] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie. “T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models,” arXiv preprint:2302.08453, 2023

[16] 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 Proc. 31st Int. Conf. Neural Inf. Process. Syst., pp. 6629–6640, 2017

[17] Shinkai Makoto, “Your Name,” CoMix Wave Films Inc, 2016.

[18] Shinkai Makoto, “Weathering With You,” CoMix Wave Films Inc, 2019.

[19] Shinkai Makoto, “Suzume,” CoMix Wave Films Inc, 2023

[20] Hugging Face, https://huggingface.co

[21] CivitAI, https://civitai.com/

[22] LAION/LAION-ART, https://labelbox.com/datasets/laion-art/

[23] Danbooru2020, https://www.kaggle.com/datasets/muoncollider/danbooru2020

[24] I. Skorokhodov et al., “Aligning Latent and Image Spaces to Connect the Unconnectable,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14144-14153, 2021

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

[26] Color Canny ControlNet, https://huggingface.co/ghoskno/Color-Canny-Controlnet-model

[27] K. He et al., “MASK R-CNN,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2961-2969, 2017

[28] C. Szegedy et al., “Rethinking the Inception Architecture for Computer Vision,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 2016

[29] E. Lee, H. Kim, J. Shim, E. Hwang, “Cartoon-Flow: A Flow-Based Generative Adversarial Network for Arbitrary-Style Photo Cartoonization”, Proceedings of the 30th ACM International Conference on Multimedia, pp.1241-1251, 2022