

SPECIAL ISSUE PAPER

Multi-Style Cartoonization: Leveraging Multiple Datasets with GANs

Jianlu Cai¹ | Frederick W. B. Li² | Fangzhe Nan¹ | Bailin Yang¹

¹Department of Computer Science, Zhejiang Gongshang University, Hangzhou, China

²Department of Computer Science, University of Durham, Durham, United Kingdom

Correspondence

Bailin Yang, Department of Computer Science, Zhejiang Gongshang University
Email: ybl@zjgsu.edu.cn

Funding Information

This research was supported by the Zhejiang Provincial Natural Science Foundation of China (Grant No. LD24F020003) and the National Natural Science Foundation of China (Grant No. 62172366).

Abstract

Scene cartoonization aims to convert photos into stylized cartoons. While GANs can generate high-quality images, previous methods focus on individual images or single styles, ignoring relationships between datasets. We propose a novel multi-style scene cartoonization GAN that leverages multiple cartoon datasets jointly. Our main technical contribution is a multi-branch style encoder that disentangles representations to model styles as distributions over entire datasets rather than images. Combined with a multi-task discriminator and perceptual losses optimizing across collections, our model achieves state-of-the-art diverse stylization while preserving semantics. Experiments demonstrate that by learning from inter-dataset relationships, our method translates photos into cartoon images with improved realism and abstraction fidelity compared to prior arts, without iterative re-training for new styles.

KEY WORDS

Generative Adversarial Network, Multi-Style Transfer, Photo Cartoonization

1 | INTRODUCTION

Cartoons stylize real concepts through techniques such as simplified shapes and exaggerated features. Popular styles include comics, animation and caricature. Traditionally, manual cartoon generation relies heavily on specialized artistic expertise and skills. It also demands significant time investment and human resources. These practical constraints limit amateur creation and general applications.

Typically, cartoons demonstrate rich visual diversity. As shown in Figure 2, early anime exhibits clear edges and smooth colors (a), while popular 3D animations emphasize realistic lighting and textures (b)¹. Traditional Chinese ink paintings possess a hierarchical structure, commonly using lines solely for outlining edges (c). Such painting technique variation means authors may depict identical subjects through differentiated styles. This stylistic range presents challenges for computationally generating cartoons. Previous work applies image style transfer to assist non-experts translating photos into assorted cartoon aesthetics^{2,3,4,5,6,7,8}. However, existing approaches are limited in focusing solely on distinguishing real images from individual cartoon datasets, neglecting intrinsic differences between domains. As a result, prior methods lack the flexibility required for multi-style cartoonization, failing to represent the full spectrum of representational diversity.

To address prior limitations, we propose a simple yet effective multi-style generative adversarial network (GAN) model controlling generation through style encodings, as depicted in Figure 1. Unlike previous work extracting encodings from individual images, we directly map latent encodings to style encodings. This obtains encodings better aligned with overall dataset styles rather than any single image. A novel multi-branch discriminator encourages different encodings to yield diverse stylistic outputs. Leveraging multiple cartoon domains aids more characteristic result synthesis. Finally, we employ the XDoG operator⁹ to extract structural information, calculating MS-SSIM¹⁰ on such features for content evaluation along with FID¹¹ for style similarity assessment. This provides a more comprehensive quality metric. By disentangling style representations from whole collections, our method flexibly synthesizes outputs spanning multiple cartoon aesthetics, addressing limitations of prior



FIGURE 1 Our Unpaired Multi-style Cartoon Image Transfer Method converts (a) Input Photos into cartoon images with specific cartoon styles, namely (b) Ink style, (c) Nezha style, and (d) Spirited style.

single-domain/image-level approaches. Extensive experiments validate our approach achieves improved content preservation and stylistic diversity for multi-style cartoonization. Our main contributions include:

- We propose a GAN-based model for flexible multi-domain cartoonization, leveraging relationships between diverse datasets.
- A novel style encoder extracts disentangled representations from entire collections rather than images, addressing prior reliance on input styles.
- We evaluate translation quality using FID¹¹ for style similarity along with XDoG⁹-based MS-SSIM¹⁰ for structural retention, providing a more comprehensive quantitative assessment.

2 | RELATED WORK

Let \mathcal{X}_P denote real photos and \mathcal{C}_c denote specific cartoon styles $c = 1, \dots, n$ (Figure 3). Our goal is transforming a source photo $x_p \in \mathcal{X}_P$ into a target cartoon $c_c \in \mathcal{C}_c$ without strict pairings between datasets. We address this problem through a model

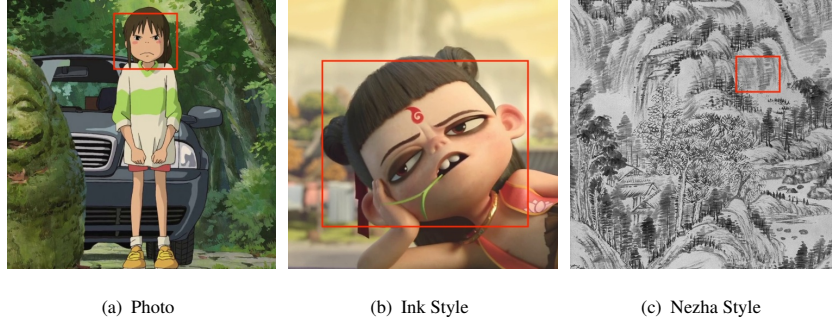


FIGURE 2 Images with different cartoon styles have their own cartoon characteristics.

leveraging relationships across C_c to flexibly synthesize diverse outputs spanning multiple cartoon aesthetics, including ink and animation styles. The following section discusses related efforts in single-style translation and generation from unpaired data.



FIGURE 3 Relationship between photo collection and various cartoon-style collection.

GAN-based image-to-image translation: GANs¹² enable tasks such as image translation^{2,13,14}, generation^{15,16,17} and style transfer¹⁸. Image translation maps source $X \rightarrow$ target Y domains. CycleGAN² introduced unsupervised arbitrary domain translation via cyclic mappings. However, it only suits simple styles. While U-GAT-IT¹⁹ uses self-attention in GANs to improve quality and generalization, it produces incorrect results for significantly different domains. Prior works solely focus on individual rather than relationships between diverse datasets, lacking flexibility to fully capture cartoonization styles through a single model. We address this by explicitly modeling inter-dataset representations.

Recent multi-domain translation models achieved impressive results. StarGAN²⁰ concatenates conditional domain information to the generator input, alongside a domain classifier to aid the discriminator. However, it focuses only on distinguishing individual rather than relationships between styles. StarGANv2²¹ introduced a style encoder and mapping network, believing each image possesses unique encodings. The encoder extracts while the mapping network generates encodings for diversity. However, with significant content changes like cartoonization, per-image encodings prone to distorting or losing content by overfitting individual rather than global styles. Conversely, our method leverages disentangled encodings across datasets for retaining semantic flexibility during translation.

Finally, Hneg-SRC²² applies contrastive learning on semantic correlations for translation across domains using CUT²³ as the generator. However, with large semantic mismatches, it can struggle preserving input details. Specifically, focusing training on semantically aligned rather than fully disentangled representations risks the generator only matching output styles to domains without fully retaining input content.

Advancements in Cartoon Style Synthesis: Chen et al.⁵ proposed a GAN framework leveraging semantic and edge-promoting losses to enhance edges while preserving content. AnimeGAN⁶ improved upon this using group convolutions and grayscale losses for richer color cartoonization. Additional works analyzed real vs cartoon attributes using a white-box representation⁷, leveraged multiple decoders/discriminators⁸, introduced guided image translation using Content-Concept Inversion (CCI) and Content-Concept Fusion (CCF)²⁴, combined pseudo paired data²⁵, or incorporated texture and color controllers²⁶. However, existing approaches concentrate solely on individual cartoon domains, failing to capture representational diversity and requiring separate costly training for each style.

To address the aforementioned issues, we propose extracting style encodings from an image set to assist in the image translation task. This approach makes our style encodings independent of the style information of individual images, focusing instead on the entire image set. It enhances the accuracy, generalization, and diversity of image translation tasks.

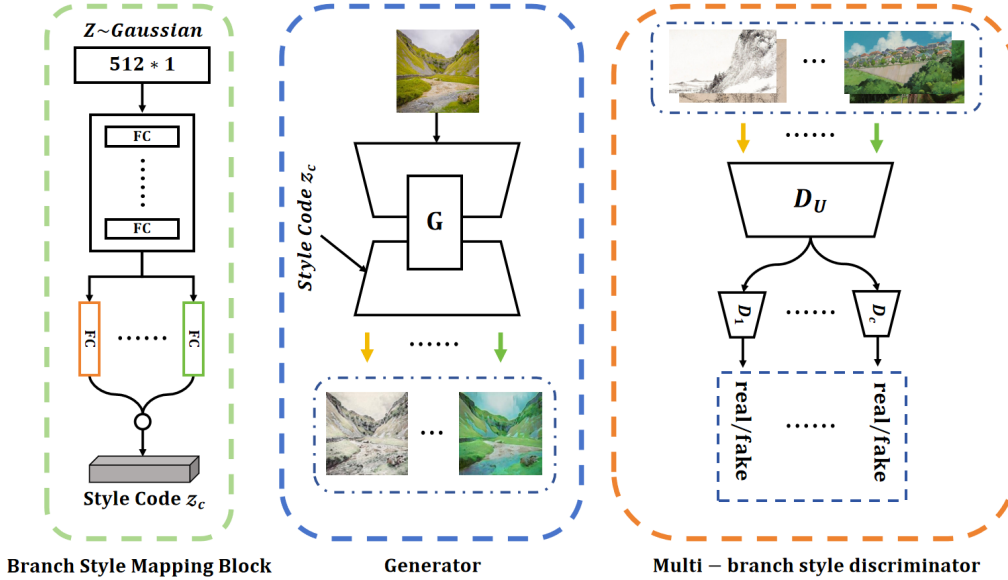


FIGURE 4 The architecture of our network model, illustrating each part and how it works.

3 | OUR METHOD

3.1 | Network Architecture

Our novel GAN architecture for multi-style cartoonization is illustrated in Figure 4. The generator consists of an encoder-decoder backbone with a multi-branch style mapping module to disentangle latents from multiple datasets for conditional synthesis. The discriminator extracts shared features and then splits into style-specific branches to discern dataset distributions, providing stronger supervision. Leveraging content, style, and adversarial losses to exploit inter-dataset relationships, this framework generates diverse cartoon outputs while preserving semantics, overcoming the reliance on single images or styles in prior approaches.

3.1.1 | Generator Architecture

The generator takes as input a source image $x_p \in \mathcal{X}_p$ and style code \mathcal{Z}_c sampled from the branch style mapping module, where \mathcal{Z}_c represents the style code for category c . It employs an encoder-decoder structure, where the encoder maps x_p to a high-dimensional latent space, encoding semantic content. The style code \mathcal{Z}_c is injected into the decoding process via AdaPoLIN¹⁹, which normalizes feature map statistics to match those of the style code at each decoder layer. Moreover, an MLP module maps \mathcal{Z}_c to mean and variance, and the decoder minimizes the distance between image features at different scales and these statistics, realizing style translation while preserving content from x_p . The AdaPoLIN operation is defined as:

$$f(x_i, \gamma_i, \beta_i) = \gamma_i \text{Conv}\left(\frac{x_i - \mu_I(x_i)}{\sigma_I(x_i)}, \frac{x_i - \mu_L(x_i)}{\sigma_L(x_i)}\right) + \beta_i \quad (1)$$

where x_i refers to each input feature map extracted from the source image x_p . (γ_i, β_i) denotes the mean and variance of the specific style code Z_c mapped from the MLP module. AdaPoLIN normalizes each input feature map x_i using the mean $\mu_I(x_i)$, $\mu_L(x_i)$ and variance $\sigma_I(x_i)$, $\sigma_L(x_i)$ of the content and style respectively. Hence, the generated cartoon image \mathcal{Y} can be expressed as:

$$\mathcal{Y} = G(x_p, Z_c) \quad (2)$$

where G is the generator function that combines the encoded content x_p with the injected style code Z_c of a specific style dataset to synthesize the output image \mathcal{Y} .

3.1.2 | Multi-branch style mapping module

Many existing methods extract style codes from single images using AdaIN^{27,19} or pretrained VGG networks²⁸, focusing on individual styles over dataset distributions. While some generative models leverage mapping networks for disentanglement^{15,16,17}, they operate on solitary references. However, representing styles with individual images fails to capture manifold textures in cartoon collections.

We propose a novel multi-branch style mapping module to address limitations of prior approaches encoding from single data points. Rather than extracting styles from specific references, a nonlinear mapping network \mathcal{F} maps a latent code \mathcal{Z} to a common style space \mathcal{W} , correlating relationships between datasets at a higher level. Each branch \mathcal{F}_c then projects \mathcal{W} to codes Z_c capturing whole collection distributions, unlike prior methods encoding solitary images. By disentangling representations across manifold datasets using mapping networks, our approach generates diverse stylizations without re-training, advancing the state-of-the-art. Our technique leverages powerful disentangled representations of multiple manifolds jointly rather than focusing on localized points^{15,16,17,27,19,28}, providing a flexible yet robust solution. The multi-branch style mapping module is mathematically defined as:

$$\begin{aligned} \mathcal{W} &= F(\mathcal{Z}) \\ Z_c &= \mathcal{F}_c(\mathcal{W}) \end{aligned} \quad (3)$$

where \mathcal{Z} is latent noise input randomly sampled from a Gaussian distribution. The mapping network \mathcal{F} projects \mathcal{Z} to a common style space \mathcal{W} , capturing correlations between datasets at a higher level. Each branch style encoder \mathcal{F}_c then maps the common code \mathcal{W} to a specific style code Z_c representing the style manifold of collection c . By mapping the input twice through \mathcal{F} and \mathcal{F}_c , our model effectively disentangles style representations from multiple manifolds without focusing on localized data points, enabling generation of diverse cartoon styles.

3.1.3 | Multi-branch style discriminator

Previous work⁸ developed multi-style generation with multiple discriminators, but each focused only on the gap between one cartoon set and real photos. This ignores relationships between various cartoon styles that are important for high-quality translation.

We propose a novel multi-branch discriminator to address this limitation. Ours has a shared component \mathcal{D}_U that captures image features, and multi-branches \mathcal{D}_c that separately determine if features match specific styles. Unlike prior single discriminators, our model jointly learns patterns within and between multiple real and fake domains. The branches provide stronger learning than individual discrimination by characterizing similarities within styles and differences between styles at local and global scales. This takes advantage of relationships ignored previously. Our discriminator leverages the connections between styles that a single discriminator does not, leading to more natural depictions of diverse yet coherent cartoon worlds through combined modeling of manifolds.

Our discriminator receives images $x(x \in (\mathcal{X}_p | \mathcal{C}_c))$ and their categories $c=(1, \dots, n)$. It determines if images match category c . The shared component \mathcal{D}_U extracts important features from all inputs. These features are then identified by \mathcal{D}_c . The discriminator is formulated as:

$$\mathcal{D}(x) = \mathcal{D}_c(\mathcal{D}_U(x)) \quad (4)$$

Since the discriminator learns not only specific domain distributions from \mathcal{D}_c but also cross-collection distributions from the shared \mathcal{D}_U , \mathcal{D} can better distinguish cartoon sets. This improves the quality and accuracy of generated images \mathcal{G} .

3.2 | Loss function

To produce higher quality images and stabilize network training, we employ the LSGAN objective²⁹. Given source image x_p , style codes $\mathcal{Z}_c (c = 1, \dots, n)$, and corresponding cartoon images $c_c \in \mathcal{C}_c (c = 1, \dots, n)$, our loss function contains:

- An adversarial term \mathcal{L}_{adv} ensuring generated images are indistinguishable from real data, guiding the model.
- A content term \mathcal{L}_{con} preserving semantic content from the input x_p .
- A style term \mathcal{L}_{sty} enforcing consistency with prescribed texture representations.

Previous methods focused separately on aspects like realism or style. Our combined loss leverages relationships between content, texture and authenticity through an integrated formulation. This holistic optimization outperforms more disjointed objectives, leading to controllable photo-realistic stylization.

Adversarial loss: We apply adversarial loss to both our generator and discriminator to make generated images indistinguishable from real cartoon images. The generator aims to translate source photos x_p into cartoon-like images $\mathcal{G}(x_p, \mathcal{Z}_c)$ matching style code \mathcal{Z}_c , while the discriminator judges whether inputs are real or generated. Compared to prior works, our discriminator also considers inter-dataset relations by distinguishing between multiple cartoon styles instead of just classifying real vs. fake. We adopt LSGAN²⁹ for more stable training:

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}^{\mathcal{D}} + \mathcal{L}_{adv}^{\mathcal{G}} \quad (5)$$

$$\mathcal{L}_{adv}^{\mathcal{D}} = \mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_C \quad (6)$$

$$\mathcal{L}_{adv}^{\mathcal{G}} = \mathbb{E}_{c_c \sim \mathcal{C}_c} [(\mathcal{D}(c_c) - 1)]^2 \quad (7)$$

where \mathcal{L}_A , \mathcal{L}_B , and \mathcal{L}_C are defined as follows:

$$\mathcal{L}_A = \mathbb{E}_{c_c \sim \mathcal{C}_c} [2(\mathcal{D}(c_c) - 1)]^2,$$

$$\mathcal{L}_B = \mathbb{E}_{x_p \sim \mathcal{X}_p} [\mathcal{D}(\mathcal{G}(x_p, \mathcal{Z}_c))]^2,$$

$$\mathcal{L}_C = \mathbb{E}_{c_{other} \sim \mathcal{C}_c} [\mathcal{D}(c_{other})]^2,$$

Specifically, $c_{other} \in \mathcal{C}_c (c = 1, \dots, n)$ and $c_{other} \neq c_c$ refer to cartoon images from other style datasets, allowing us to consider inter-relations that were ignored in prior works focusing only on distinguishing real from fake images.

Content Loss: To preserve semantic content during translation, we extract perceptual features from generated and real images using a pre-trained VGG-19 network²⁸. Compared to prior works extracting features from only single images, we leverage entire datasets to learn richer content representations capturing relationships between styles. The content loss function encourages generated images $G(x_p, \mathcal{Z}_c)$ to match content of source photos x_p :

$$\mathcal{L}_{con} = \mathbb{E}_{x_p \sim \mathcal{X}_p} [\|VGG(G(x_p, \mathcal{Z}_c)) - VGG(x_p)\|] \quad (8)$$

where we extract features from the "conv4-4" layer of VGG-19 to focus on high-level semantic content while discarding low-level details unrelated to semantics. By jointly optimizing perceptual similarities across datasets via this loss, our method better preserves input contents during diverse stylization compared to methods operating on single images.

Style Loss: To transfer textures and brushstroke styles, we use the gram matrix³⁰ to calculate style similarities based on deep feature correlations. Compared to prior works focusing on single image styles, we model styles as distributions over entire datasets to encode more diverse, dataset-level representations. The style loss constrains generated images $G(x_p, \mathcal{Z}_c)$ to match real cartoon statistics.

$$\mathcal{L}_{sty} = \mathbb{E}_{x_p \sim \mathcal{X}_p, \mathbb{E}_{c_c \sim \mathcal{C}_c}} [\|\text{gram}(VGG(G(x_p, \mathcal{Z}_c))) - \text{gram}(VGG(c_c))\|] \quad (9)$$

By jointly optimizing correlations across datasets, our method learns richer texture representations beyond individual styles. This enables generalization to new photo inputs with a variety of cartoon visual trends compared to prior single-image based methods.

Total loss function: The total objective comprises generator \mathcal{L}_{gen} and discriminator \mathcal{L}_{dis} losses:

$$\mathcal{L}_{gen} = \lambda_{adv} \mathcal{L}_{adv}^{\mathcal{G}} + \lambda_{con} \mathcal{L}_{con} + \lambda_{sty} \mathcal{L}_{sty} \quad (10)$$

$$\mathcal{L}_{dis} = \lambda_{adv} \mathcal{L}_{adv}^{\mathcal{D}} \quad (11)$$

Compared to methods optimizing losses on individual images, we combine dataset-level adversarial, content and style losses within our unified GAN framework. By minimizing \mathcal{L}_{gen} and \mathcal{L}_{dis} alternately for generator \mathcal{G} and discriminator \mathcal{D} , our method trains all components jointly using relationships across diverse cartoon datasets. This allows us to better capture complex, multi-modal dependencies between photo contents and cartoon textures than prior single-domain methods.

4 | EXPERIMENTAL RESULTS

In this section, we will first describe the dataset, experimental details, and the qualitative and quantitative analysis used for evaluation. Subsequently, we will verify the effectiveness of the proposed method in scene cartoonization. Finally, we will illustrate our advantages over other methods through quantitative and qualitative comparisons.

Dataset: The dataset contains one collection of 6,656 real scene photographs from Flickr for training, with the remaining 790 images for testing. Additionally, there are three collections of cartoon images sourced from movies and illustrations. All images were resized uniformly to 256×256 pixels to serve as the input-output pairs required to train our proposed unsupervised image translation framework. Further, the inclusion of multiple cartoon styles allows leveraging their inter-relations and enhances the model’s ability to generate diverse outputs.

Cartoon images were obtained by sampling movie frames at 0.5 second intervals, filtering adjacent duplicates with PSNR>16 and SSIM<0.7 thresholds. This preprocessing ensures sufficient variance between frames to train the method’s novel multi-branch encoding scheme, which aims to model overall style distributions rather than individual images. In our experiments, we leveraged relationships between multiple cartoon style datasets to enhance stylization diversity, addressing limitations of prior single-dataset methods^{5,6,7}. Specifically, we employed the following four publicly available style datasets:

- The first dataset comprises 1388 Chinese ink paintings, featuring variances in depiction abstraction suited for the model’s spatial style modulation.
- The second "Nezha: Birth of the Demon Child" animation dataset contains 6825 images, providing photorealistic textures beneficial for the generator’s content-style disentanglement.
- The third "Spirited Away" collection includes 9132 images with clear lines and colors, facilitating the style discriminator’s manifold learning across modalities.

We additionally incorporated the 1752-image "Hayao" and 1284-image "Paprika" styles from AnimeGANv2⁶ to demonstrate generalization. These expanded the modeled style relations, addressing single-dataset limitations.

These datasets covering diverse visual themes trained our proposed method’s novel multi-style encoder and discriminator, leveraging cross-set style relationships for enhanced stylization effects, validating our approach.

Experimental details: Our method is implemented in PyTorch, utilizing two NVIDIA 3060 GPUs with 12GB memory. The detailed model architecture has been described in Section 3. The training batch size is set to $N = 8$, and the type of cartoon-style dataset used is $n = 3$. The hyperparameters are set to $\lambda_{adv} = 5$, $\lambda_{con} = 7$, $\lambda_{sty} = 1$. We optimize our model using the Adam optimizer³¹ with a learning rate of 1×10^{-4} , $\beta_1 = 0.5$, $\beta_2 = 0.999$ for around 100 iterations.

Qualitative and Quantitative Analysis: We conducted qualitative and quantitative analyses on three different style datasets. We compared our approach with seven advanced image translation methods, including CycleGAN², U-GAT-IT⁶, White-Box⁷, CartoonGAN⁵, MSCartoonGAN⁸, StarGAN²⁰, StarGANv2²¹, and HnegSRC²².

In our quantitative experiments, we decided to create an evaluation metric termed the "Structuro-Style Measure." This metric combines FID (Fréchet Inception Distance) with MS-SSIM (Mean Structural Similarity Index) and the XDoG algorithm. We opted for a combination of FID and MS-SSIM. FID gauges the data distribution difference between generated and real images, proving valuable for assessing image diversity and quality. MS-SSIM measures structural similarity across multiple scales, crucial for tasks emphasizing image structure. Our approach also introduced the XDoG algorithm, which further extracts image content information while disregarding color-related details, thereby enhancing the comprehensiveness of our evaluation. The combined use of these metrics contributes to a better understanding and explanation of the effectiveness and performance of our approach.

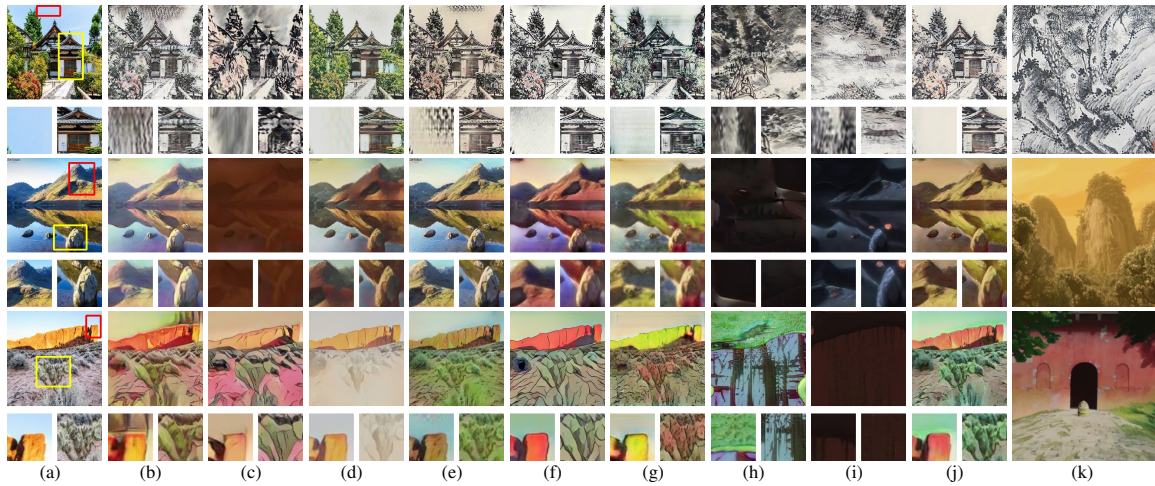


FIGURE 5 (a) Input Photo, (b) CycleGAN, (c) U-GAT-IT, (d) White-Box, (e) CartoonGAN, (f) MSCartoonGAN, (g) StarGAN, (h) StarGANv2, (i) HengSRC, and (j) Ours, where (k) is Representative image of cartoon style. Comparison of our method with advanced scene animation methods on three different style datasets (image details are enlarged). Quantitative evaluation of results, with the first row in Ink style, the second row in *Nezha* style, and the third row in *Spirited* style.

4.1 | Qualitative comparison

The comparison between our method and previous methods is shown in Figure 5, where we qualitatively evaluate them on content preservation and style transfer ability. CycleGAN² learns the mapping between two domains using cycle consistency, but may lose content or produce artifacts when translating between divergent scene and cartoon domains. U-GAT-IT⁶ uses attention and AdaLIN to control generation but can introduce distortions or modify colors excessively, sometimes destroying content. White-Box⁷ produces different results by adjusting hyperparameters, but performance varies significantly with them and inconsistent effects arise using the same hyperparameters across styles. CartoonGAN⁵ guides clearer edge generation using an edge loss, suiting datasets with obvious edges, but results on less-defined styles resemble inputs. MSCartoonGAN⁸ appears visually nearest to ours but still experiences content loss. StarGAN²⁰ and StarGANv2²¹ are multimodal frameworks, yet the large mismatch in scene and cartoon content leads to severe content loss. HnegSRC²² relies on semantic similarity but may ignore content with great semantic differences between domains. In contrast, our method produces more detailed results while maintaining content virtually unchanged and better matching target styles in texture and color based on datasets, validating leveraging multiple dataset relationships.

In addition, we have conducted stylistic tests on more scenarios. As shown in Figure 6, *Ink* Style shows a good effect on scene, animal and face data. Although our training data does not include facial and animal photos, the model still can generate high-quality images. The model makes good use of lighting to show the hierarchy of hair and retain the basic content of the image. *Nezha* Style better remains the style of the original collections, and can show the haziness of the animation collections in various scenes. And the images generated under *Spirited* Style have clearer lines and more obvious color blocks, which is similar to that of the style of cartoon collections.

4.2 | Quantitative comparison

As shown in Table 1, our method generally outperforms previous single-dataset methods in terms of FID, indicating its ability to generate more stylized results. However, FID alone cannot assess changes in image content. Therefore, in addition to FID, we also employ the XDoG and MS-SSIM metrics to evaluate structural content consistency. Our method gains better performance in maintaining image content compared to previous works. Finally, we propose using the SSM (Structuro-Style Measure) as a comprehensive evaluation metric.

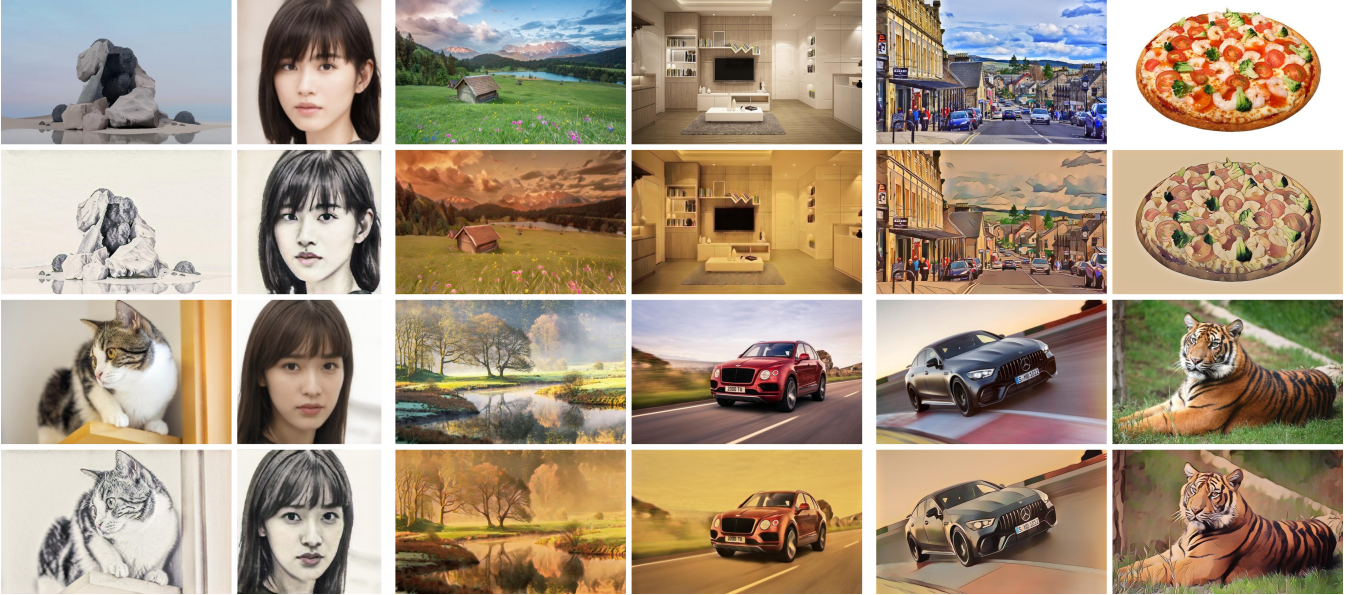


FIGURE 6 Stylization test in more scenarios.

TABLE 1 SSM scores (lower are better) comparison on Ink, Nezha, and Spirited datasets.

Method	$XDoG+MS-SSIM(content)\uparrow$			Avg	$FID(style)\downarrow$			Avg	SSM
	<i>Ink</i>	<i>Nezha</i>	<i>Spirited</i>		<i>Ink</i>	<i>Nezha</i>	<i>Spirited</i>		
CycleGAN	0.660	0.652	0.573	0.628	129.08	105.45	126.11	120.21	129.44
U-GAT-IT	0.510	0.481	0.487	0.492	147.74	115.51	128.81	130.69	162.12
White-Box	0.770	0.579	0.554	0.634	147.42	108.84	129.26	128.51	137.60
CartoonGAN	0.882	0.823	0.786	0.830	118.19	95.92	132.76	115.62	118.03
MSCartoonGAN	0.506	0.537	0.443	0.495	95.17	90.80	92.39	92.79	116.97
StarGAN	0.542	0.664	0.646	0.617	97.88	113.83	135.24	115.65	125.50
StarGANv2	0.015	0.027	0.023	0.023	58.54	90.33	117.80	88.89	2.54e7
HnegSRC	0.017	0.417	0.098	0.177	70.22	85.59	90.98	80.26	2105.05
ours	0.731	0.871	0.745	0.782	83.01	90.74	113.80	95.85	100.10

SSM balances style similarity by FID and content similarity using MS-SSIM. By penalizing content changes and using XDoG extraction, SSM addresses significant content loss better. This avoids reliance on style similarity alone for misleading assessment. SSM thus enables more comprehensive generation quality evaluation.

4.3 | Ablation experiment

The Structuro-Style Measure(SSM): Prior works proposed using FID to evaluate stylistic similarity between generated and reference images based on analyzing feature distributions in terms of mean, covariance and distances, with lower FID indicating higher similarity. Thus, FID quantitatively assesses generated image authenticity in style. However, relying solely on FID to evaluate translation quality has limitations. Since FID focuses only on feature matching, it may report favorable scores even when significant content is lost. As shown in Figure 7, both StarGANv2 and HnegSRC achieved excellent FID but disregarded input content, focusing only on the target domain. In such cases, FID alone does not facilitate reasonable assessment.

To address this, we introduced the Mean Structural Similarity Index metric (MS-SSIM) to evaluate content consistency. While FID captures style similarity, MS-SSIM assesses preservation of semantic details. By combining FID and MS-SSIM,

our proposed SSM enables a more comprehensive evaluation of both style imitation and content preservation for robust quality assessment of image translation models.

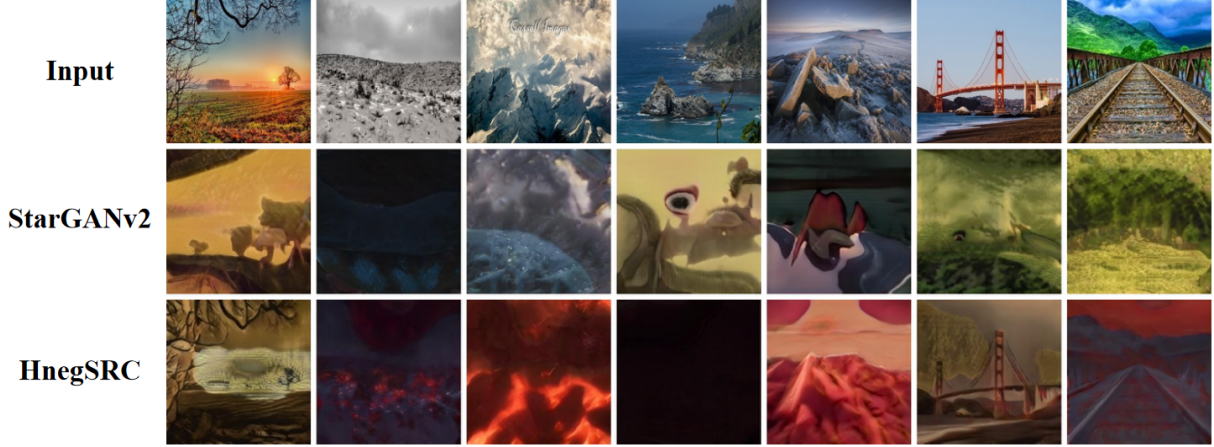


FIGURE 7

Randomly selected high FID-generated results from StarGANv2 and HnegSRC, comparing them with the input images.

We propose a comprehensive evaluation metric called the Structuro-Style Measure (SSM). SSM balances style similarity and content preservation by combining FID and MS-SSIM scores. The SSM calculation is:

$$SSM(S_{ty}, S_{tr}) = S_{ty} + S_{ty}^{-\ln(S_{tr})} \quad (12)$$

where S_{ty} represents the degree of image stylization, measured by FID between generated (\mathcal{Y}) and target (\mathcal{C}_c) images. S_{tr} represents content similarity, quantified by MS-SSIM between \mathcal{Y} and input (\mathcal{X}_p) images. Specifically:

$$\begin{aligned} S_{ty} &= FID(\mathcal{Y}, \mathcal{C}_c) \\ S_{tr} &= MSSSIM(\mathcal{Y}, \mathcal{X}_p) \end{aligned} \quad (13)$$

By combining S_{ty} and S_{tr} , SSM facilitates a comprehensive evaluation of image translation quality in terms of both style imitation and content consistency. By computing FID and MS-SSIM and incorporating them with an exponential term, the objective is to apply a heightened penalty in cases where there is a substantial loss of content in the generated images. In situations where MS-SSIM is equal, SSM varies with fluctuations in FID. Conversely, when FID is equal, SSM introduces an exponential penalty term as MS-SSIM decreases.

While MS-SSIM can effectively assess content similarity in series of images with added noise or blurring, relying solely on it also has limitations. Specifically, MS-SSIM faces challenges with scenarios involving significant content differences, such as completely inconsistent grayscale images. As illustrated in Figure 8, MS-SSIM may struggle to precisely capture structural similarity in such cases.

To address this, we propose preprocessing generated images with the XDoG edge extraction algorithm before MS-SSIM calculation. As shown, when there are large variations in content, directly utilizing MS-SSIM can produce imprecise structural similarity measurements. However, by first extracting edge maps from images via XDoG, contextual information unrelated to structure is removed. This proven preprocessing approach helps reduce error rates when employing MS-SSIM, overcoming challenges posed by dissimilar contents to more effectively evaluate preservation of semantic content integrity during translation.

Loss Function Ablation Experiment: We conducted ablation studies to analyze the impact of different loss components on image generation quality, as shown in Figure 9. Excluding the adversarial loss $\mathbb{E}_{c_{other} \sim \mathcal{C}_c}$ enhances texture but results in inaccurate descriptions, such as overstressed internal cloud lines. Removing the additional style loss \mathcal{L}_{sty} weakens line textures. Omitting both losses during adversarial training substantially reduces the model’s ability to capture proper textures and colors, resulting in a noticeable decline in output fidelity. These results validate the importance of each loss term for optimizing realistic stylization. The complete loss formulation enables holistic translation with balanced texture stylization and content consistency, demonstrating its effectiveness.










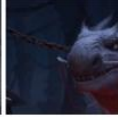

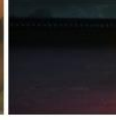

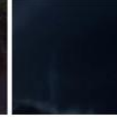
Image1							
Image2							
MSSSIM	0.586	0.521	0.531	0.559	0.523	0.502	0.662
XDOG+MSSSIM	0.018	0.000	0.021	0.046	0.000	0.085	0.162

FIGURE 8

Structural similarity between two different images was calculated by MS-SSIM with and without XDoG preprocessing.

TABLE 2 Ablation studies on loss functions.

Method	FID↓			Avg
	<i>Ink</i>	<i>Nezha</i>	<i>Spirited</i>	
w/o \mathcal{L}_{sty} & w/o $\mathbb{E}_{c_{other} \sim \mathcal{C}_c}$	116.26	107.57	161.33	128.38
w/o \mathcal{L}_{sty}	84.20	101.56	121.32	102.36
w/o $\mathbb{E}_{c_{other} \sim \mathcal{C}_c}$	90.19	118.69	133.90	107.59
Total Loss	83.01	90.74	113.80	95.85

TABLE 3 Ablation studies on dataset usage.

Method	FID↓			Avg
	<i>Ink</i>	<i>Nezha</i>	<i>Spirited</i>	
single dataset	94.02	100.94	125.30	106.75
two datasets	106.87	100.47	121.13	109.49
three datasets	83.01	90.74	113.80	95.85

In contrast to prior works, Table 2 demonstrates that our approach significantly improves stylization performance and generated output aesthetics through leveraging multiple datasets for comparison and introducing global style losses. Quantitative indices reveal the significance of each component in the cartoonization process. Specifically, without the multi-branch discriminative loss and corresponding structure, cartoonization effects weaken while distortions amplify. Employing additional style losses aids the model in better fitting stylization across diverse datasets. Combining various data sources proves effective as it notably bolsters the degree of input cartoonization while enhancing generation quality. Overall, the results validate that utilizing multiple reference manifolds beneficially fosters artistic stylization abilities compared to single-dataset schemes.

Evaluating the Effect of Dataset Quantity on Generation Quality: We conducted ablation studies to analyze the effect of varying the number of reference style datasets (one, two, three) on output fidelity. As reported in Table 3, FID scores were used to evaluate stylization under each configuration. When two datasets were utilized, results were averaged across all combinations of mixed training. The experiments revealed that while two datasets can slightly reduce stylization for very dissimilar styles like Ink, incorporating three manifolds generally enhances the overall stylization level attained. This validates that leveraging a greater variety of reference styles through multi-dataset training empowers models with a more comprehensive understanding of different art media. The findings thus provide insights into adapting the dataset quantity to optimize generation of diverse artistic textures and patterns.

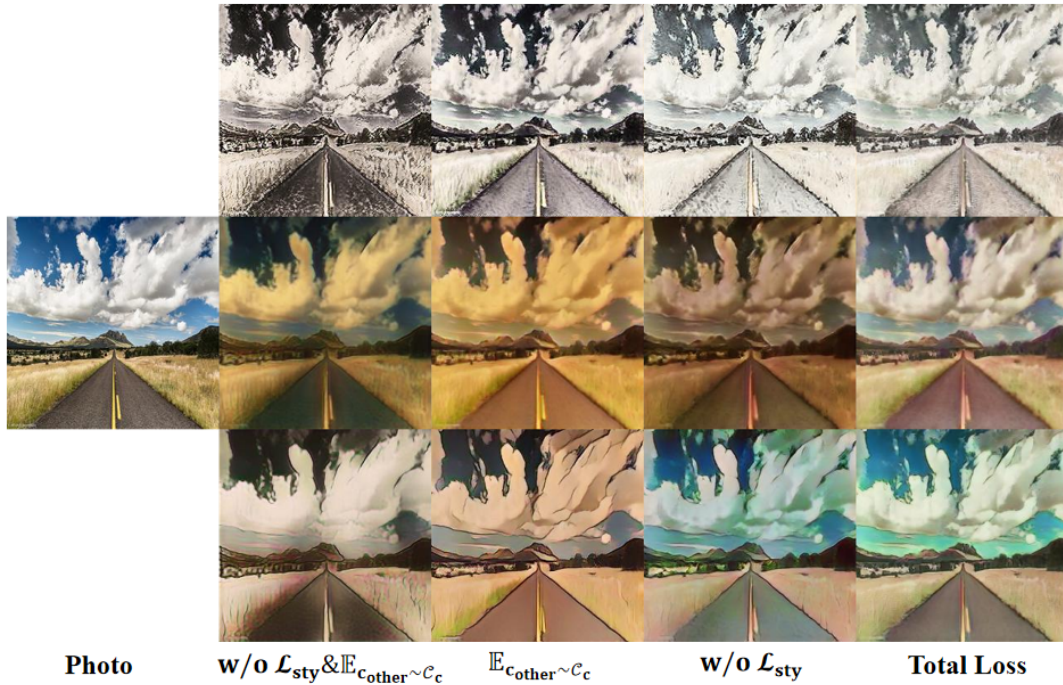


FIGURE 9 Qualitative ablation experiment on components of our loss function, including multi-branch loss and gram loss.

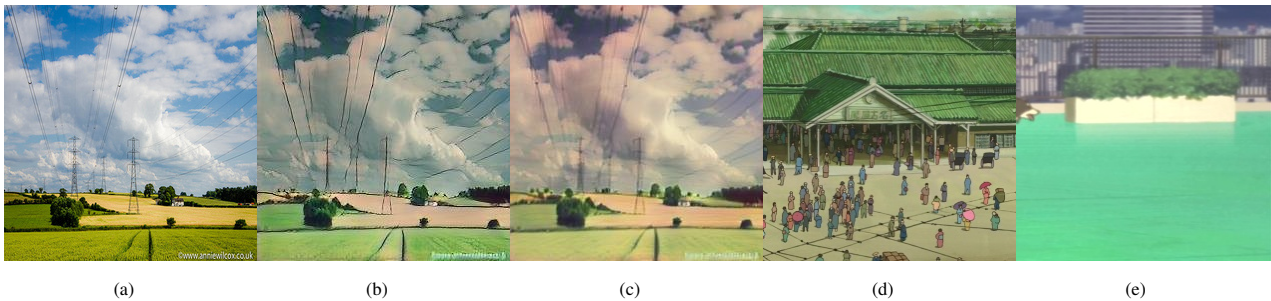


FIGURE 10 Generation samples when extending our model to two added animation styles. (a) Input photo. (b,c) Stylization results in Hayao and Paprika styles, respectively. (d,e) Further stylized outputs in Hayao and Paprika styles, respectively.

4.4 | Extension to New Styles

During experiments, we observed combining a pre-trained generator with initialized style mappings yielded satisfactory reconstruction, showing the model’s stylization abilities can be extended by retaining fixed generator weights and solely training additional discriminators and encoders on new styles.

Figure 10 demonstrates this extension’s effectiveness. Leveraging the pre-trained parameters, expanded training was conducted on encoders and discriminators for the Hayao and Paprika animation datasets. Results show the generator has acquired the capacity to mimic both styles credibly. Stylized outputs in the Hayao style exhibit well-defined edges, while those in the Paprika style showcase smoother textures and vivid colors. This validates that our approach can flexibly generate diverse artistic depictions by learning new styles independently through the style-specific branches.

5 | CONCLUSIONS

This work presents a GAN framework for photorealistic cartoonization across artistic manifolds. A novelty is a multi-style mapping and multi-branch discriminator addressing translation into distinct styles. The mapping module learns optimized style encodings, enabling fine-grained translation control. The multi-discriminative scheme orientates transfer through comparison to diverse cartoon attributes. Experiments validate leveraging manifold information augments generative abilities. Additionally, SSM is proposed to balance FID and MS-SSIM scores, facilitating examination of consistency and style emulation, two critical objectives. Overall, the model and evaluation constitute a significant step towards high-fidelity cartoon outputs preserving semantic content through multi-attribute style guidance. The approach captures richer artistic expressions and paves the way for future multi-target translation work.

ACKNOWLEDGMENTS

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020003. The authors also acknowledge the support from the NSFC (National Natural Science Foundation of China) under Grant No. 62172366.

References

1. Gurney J. *Color and light: A guide for the realist painter*. 2. Andrews McMeel Publishing . 2010.
2. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ; 2017: 2223–2232.
3. Ahn N, Kwon P, Back J, Hong K, Kim S. Interactive Cartoonization with Controllable Perceptual Factors. In: ; 2023: 16827–16835.
4. Liu MY, Breuel T, Kautz J. Unsupervised image-to-image translation networks. *Advances in neural information processing systems* 2017; 30.
5. Chen Y, Lai YK, Liu YJ. Cartoongan: Generative adversarial networks for photo cartoonization. In: ; 2018: 9465–9474.
6. Chen J, Liu G, Chen X. AnimeGAN: a novel lightweight GAN for photo animation. In: Springer. ; 2020: 242–256.
7. Wang X, Yu J. Learning to cartoonize using white-box cartoon representations. In: ; 2020: 8090–8099.
8. Shu Y, Yi R, Xia M, et al. Gan-based multi-style photo cartoonization. *IEEE Transactions on Visualization and computer graphics* 2021; 28(10): 3376–3390.
9. Winnemöller H, Kyprianidis JE, Olsen SC. XDoG: An eXtended difference-of-Gaussians compendium including advanced image stylization. *Computers & Graphics* 2012; 36(6): 740–753.
10. Wang Z, Simoncelli EP, Bovik AC. Multiscale structural similarity for image quality assessment. In: . 2. Ieee. ; 2003: 1398–1402.
11. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* 2014.
12. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. *Communications of the ACM* 2020; 63(11): 139–144.
13. Mirza M, Osindero S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* 2014.
14. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: ; 2016: 2818–2826.
15. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: ; 2019: 4401–4410.
16. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan. In: ; 2020: 8110–8119.
17. Karras T, Aittala M, Laine S, et al. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 2021; 34: 852–863.
18. Cai Q, Ma M, Wang C, others . Image neural style transfer: A review. *Computers and Electrical Engineering* 2023; 108: 108723.
19. Kim J, Kim M, Kang H, Lee K. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830* 2019.
20. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: ; 2018: 8789–8797.
21. Choi Y, Uh Y, Yoo J, Ha JW. Stargan v2: Diverse image synthesis for multiple domains. In: ; 2020: 8188–8197.

22. Jung C, Kwon G, Ye JC. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In: ; 2022: 18260–18269.
23. Park T, Efros AA, Zhang R, Zhu JY. Contrastive learning for unpaired image-to-image translation. In: Springer. ; 2020: 319–345.
24. Cheng B, Liu Z, Peng Y, al. e. General image-to-image translation with one-shot image guidance. In: ; 2023: 22736–22746.
25. Jiang Y, Jiang L, Yang S, others . Scenimefy: Learning to Craft Anime Scene via Semi-Supervised Image-to-Image Translation. In: ; 2023: 7357–7367.
26. Ahn N, Kwon P, Back J, Hong K, Kim S. Interactive Cartoonization with Controllable Perceptual Factors. In: ; 2023: 16827–16835.
27. Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. In: ; 2017: 1501–1510.
28. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 2017; 30.
29. Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S. Least squares generative adversarial networks. In: ; 2017: 2794–2802.
30. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: ; 2016: 2414–2423.
31. Kinga D, Adam JB, others . A method for stochastic optimization. In: . 5. San Diego, California;. ; 2015: 6.

AUTHOR BIOGRAPHY



Cai Jianlu is currently pursuing his studies at Zhejiang Gongshang University. His main research interests lie in the field of computer vision, particularly in image style transfer. He explores how techniques in artificial intelligence programming, computer animation, and computer graphics can be leveraged to enhance the effects of image stylization.



Frederick W. B. Li received a B.A. and an M.Phil. degree from Hong Kong Polytechnic University, and a Ph.D. degree from the City University of Hong Kong. He is currently an Associate Professor at Durham University, researching computer graphics, deep learning, collaborative virtual environments, and educational technologies. He is also an Editorial Board Member of *Virtual Reality & Intelligent Hardware*. He chaired conferences such as ISVC and ICWL.



Fangzhe Nan, a doctoral candidate at Zhejiang Gongshang University, received the B.S. and master's degrees from Xinjiang University, Urumqi, China, in 2017 and 2020, respectively. Her research interests include computer vision, video and image processing, and point cloud compression.



Bai-Lin Yang received his Bachelor's and Ph.D. degrees from Dept. Computer Science, Hangzhou Dianzi University in 2003 and Zhejiang University in 2007, respectively. Now, he is a faculty member of Zhejiang Gongshang University. Yang's research interests include web graphics, realtime rendering and mobile game.



Citation on deposit:

Ryan, G. A. (2024). Receptive Ecumenism in a Synodal Catholic Church. In J. A. Berry, & V. Coman (Eds.), *Living Tradition: Continuity and Change as Challenges to Churches and Theologies*. Leipzig: Evangelische Verlagsanstalt

For final citation and metadata, visit Durham Research Online URL:

<https://durham-repository.worktribe.com/output/2408024>

Copyright statement: This accepted manuscript is licensed under the Creative Commons Attribution 4.0 licence. <https://creativecommons.org/licenses/by/4.0/>