Scenimefy: Learning to Craft Anime Scene via Semi-Supervised Image-to-Image Translation

Yuxin Jiang* Liming Jiang* Shuai Yang Chen Change Loy✉
S-Lab, Nanyang Technological University
{c200203, liming002, shuai.yang, ccloy}@ntu.edu.sg

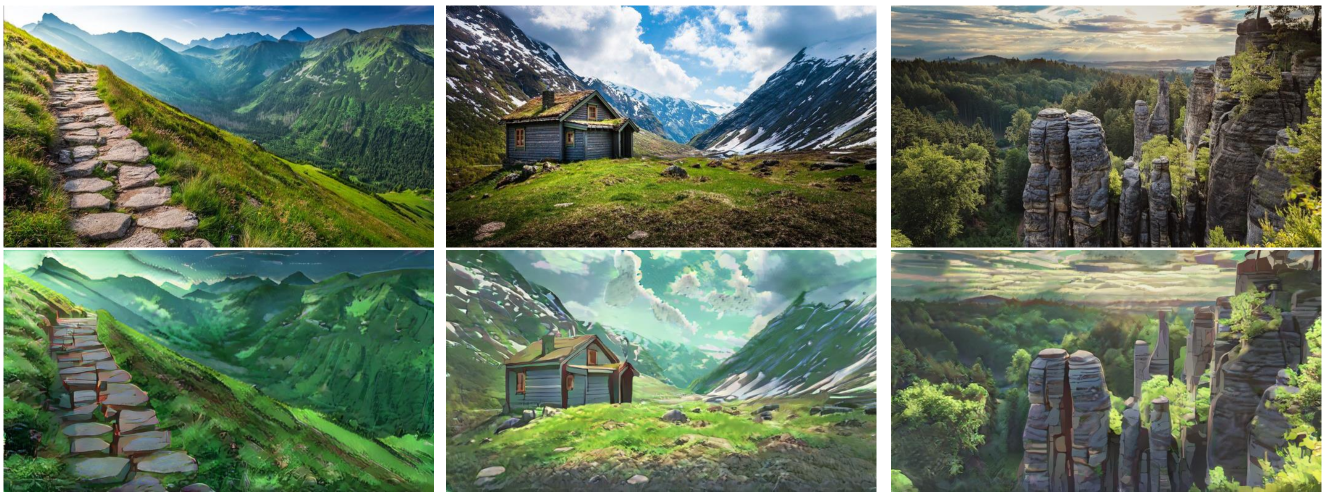


Figure 1: Examples of anime scene rendering by Scenimefy. Top row: input images; Bottom row: our translated results.

Abstract

Automatic high-quality rendering of anime scenes from complex real-world images is of significant practical value. The challenges of this task lie in the complexity of the scenes, the unique features of anime style, and the lack of high-quality datasets to bridge the domain gap. Despite promising attempts, previous efforts are still incompetent in achieving satisfactory results with consistent semantic preservation, evident stylization, and fine details. In this study, we propose Scenimefy, a novel semi-supervised image-to-image translation framework that addresses these challenges. Our approach guides the learning with structure-consistent pseudo paired data, simplifying the pure unsupervised setting. The pseudo data are derived uniquely from a semantic-constrained StyleGAN leveraging rich model priors like CLIP. We further apply segmentation-guided data selection to obtain high-quality pseudo supervision. A patch-wise contrastive style loss is introduced to improve stylization and fine details. Besides, we contribute a high-resolution anime scene dataset to facilitate future research. Our extensive experiments demonstrate the superiority of our method over state-of-the-art baselines in terms of both perceptual quality and quantitative performance. Project page: https://yuxinn-j.github.io/projects/Scenimefy.html.

1. Introduction

Crafting anime scenes requires significant artistic skill and time, making developing learning-based techniques for automatic scene stylization of unquestionable practical and commercial value. Recent advances in Generative Adversarial Networks (GANs) have led to a significant improvement in automatic stylization, but most research in this area has focused primarily on human faces [41, 33, 22, 20, 42]. Despite its high research value, generating high-quality anime scenes from complex real-world scene images remains underexplored.

Transferring real scene images into anime styles remains a formidable challenge due to several factors. 1) *The nature of a scene.* Scenes are typically composed of multiple objects with complex relationships among them, and there is an inherent hierarchy between foreground and background elements, as shown in Figure 2. 2) *The features of anime.* Anime is characterized by unique textures and intricate de-

* Equal contribution.

tails, such as the pre-designed brush strokes used in natural landscapes like grass, trees, and clouds, as illustrated in Figure 2. These textures are typically organic and hand-drawn, making their style much more difficult to mimic than the sharp edges and smooth color patches defined in previous studies [5, 36]. 3) *The domain gap and lack of data.* There is a large domain gap between real and anime scenes, and a high-quality anime scene dataset is essential in bridging this gap. However, existing datasets contain many human faces and other foreground objects, whose style is different from that of the background scene, leading to their low quality.

Unsupervised image-to-image translation [46, 21, 11, 13, 27, 14] is a typical solution for complex scene stylization without paired training data. Despite promising results, existing methods [5, 3, 36, 8] that focus on anime styles fall short in several ways. First, the absence of pixel-wise correspondence in complex scenes hinders existing methods [5, 3] from effectively performing evident texture stylization while preserving semantic content, resulting in potentially unnatural results with notable artifacts. Second, some approaches [36, 8] fall short of generating fine details of anime scenes. This is due to their handcrafted anime-specific losses or pre-extracted representations that impose edge and surface smoothness.

To address the challenges discussed above, we propose a novel semi-supervised image-to-image (I2I) translation pipeline, named *Scenimefy*, for producing high-quality anime-style renderings of scene images, as shown in Figure 1. Our key idea is to incorporate a new supervised training branch into the unsupervised framework using generated pseudo paired data to overcome the difficulties of unsupervised training. Specifically, we leverage the desirable properties of StyleGAN [16, 17] by fine-tuning it to generate coarse paired data between real and anime, which we call pseudo paired data. We propose a novel semantic-constrained fine-tuning strategy that leverages rich pre-trained model priors, such as CLIP [29] and VGG [31], to guide StyleGAN to capture complex scene features and alleviate overfitting. We further introduce a segmentation-guided data selection scheme to filter low-quality data. With the pseudo paired data, Scenimefy learns effective pixel-wise correspondence and generates fine details between the two domains, guided by a novel patch-wise contrastive style loss. Together with the unsupervised training branch, our semi-supervised framework seeks a desired trade-off between the faithfulness and fidelity of scene stylization.

To facilitate training, we also collected a high-quality pure anime scene dataset. We conducted comprehensive experiments that demonstrate the effectiveness of Scenimefy, surpassing state-of-the-art baselines in both perceptual quality and quantitative evaluation. In summary, our key contributions are as follows:

- We propose a novel semi-supervised image-to-image translation framework for scene stylization that generates high-quality complex anime scene images from real ones. Our framework incorporates a new patch-wise contrastive style loss to improve stylization and fine details.
- The training supervision is derived from structure-consistent pseudo paired data generated by a newly designed semantic-constrained StyleGAN fine-tuning strategy with rich pre-trained prior guidance, followed by a segmentation-guided data selection scheme.
- We collected a high-resolution anime scene dataset to facilitate future research in scene stylization.

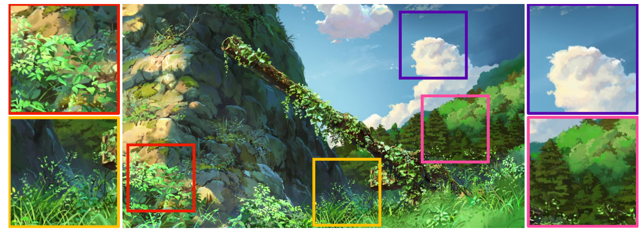


Figure 2: **Characteristics of anime scenes.** A scene frame from Shinkai's film 'Children Who Chase Lost Voices' (2011) shows the presence of hand-drawn brush strokes of grass and stones (foreground), as well as trees and clouds (background), as opposed to clear edges and flat surfaces.

2. Related Work

Image-to-image translation aims at transferring images from a source domain to a target domain. Its taxonomy can be generally grouped into two paradigms: supervised [12, 23, 28] and unsupervised [46, 21, 11, 27, 13], depending on whether paired training data are available. Pix2pix [12] is the first supervised image translation model with conditional GANs [23], and is later extended to pix2pixHD [35] for generating high-resolution images. Due to the difficulty of acquiring paired images, unsupervised models have been developed, typically based on the assumption of the cycle-consistency constraint [46]. However, the underlying bijective assumption is restrictive. Some methods [1, 27, 25, 13] have tried to break the cycle, such as the contrastive unpaired translation model (CUT) [27], which uses contrastive learning to maximize the mutual information between the local patches of input and output images. Patches are a natural unit for learning intricate anime-style textures between images and complex relations between objects within individual images, making contrastive learning-based translations useful for scene stylization. Thus, we build our unsupervised training branch with a recent contrastive learning-based translation model [14].

Domain adaptation of StyleGAN has been an active area of research, aiming to transfer knowledge of pre-trained

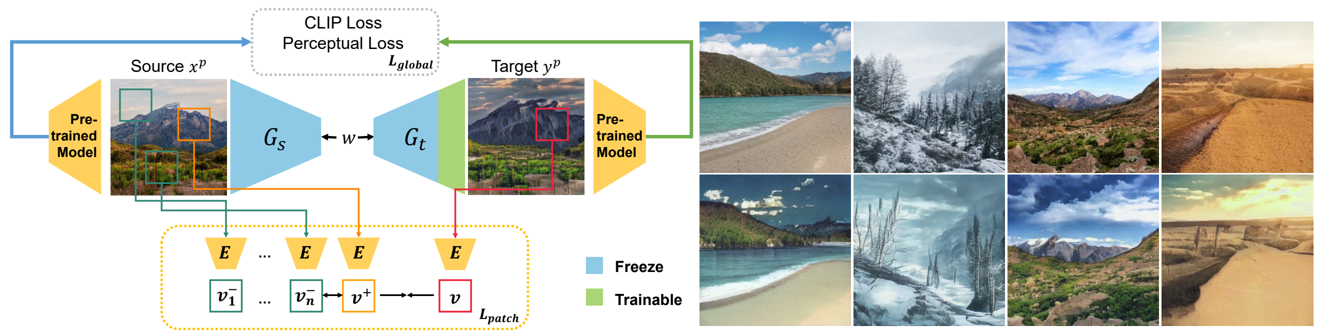


Figure 3: **Overview of our semantic-constrained fine-tuning strategy for pseudo paired data generation. Left:** We initialize a source generator G_s and a target one G_t by pre-trained on a real scene domain. G_s remains fixed throughout the process. G_t is optimized using rich pre-trained model prior (*i.e.*, CLIP, VGG) guidance with the early layers freezed to be adapted to an anime scene domain. A patch-wise contrastive loss using pre-trained CLIP embedders E is applied to better preserve local spatial details. **Right:** Examples of generated pseudo paired data after segmentation-guided data selection.

GANs to new domains. Several adaptation strategies based on fine-tuning have been proposed, including learnable parameter selection [24, 40, 37, 30], data augmentations [15, 44, 34], and regularization terms [19, 26, 45, 39]. FreezeG [2] freezes the generator's low-resolution layers to sustain the structure of the source domain. Recent studies [18, 38, 7, 47] guide attribute-level adaptation by calculating the domain gap direction in a CLIP embedding space [29]. Despite promising attempts, StyleGAN adaptation still has limitations in fixed image resolution, failure modeling of complex scenes, overfitting, and undesired semantic artifacts. In comparison, these issues are well addressed by our semantic-constrained strategy and data selection, as well as the semi-supervised framework.

Scene cartoonization. CartoonGAN [5] proposed a semantic content loss and an edge-promoting adversarial loss to retain clear edges and smooth shading. It was further extended to lightweight AnimeGAN [3] with improved anime-specific loss functions. However, their global unsupervised learning framework is unable to capture local cartoon textures. Wang et al. [36] introduced a white-box framework, decomposing images into surface, structure, and texture – to guide the cartoonization process. Nevertheless, the obtained results replace fine details with flat color blocks. Recently, Gao et al. [8] proposed a cartoon-texture-saliency-sampler (CTSS) module to better perceive and transfer cartoon textures. However, it merely learns texture abstraction and over-saturated color, limiting its ability to synthesize anime scenes with hand-drawn styles. Different from existing efforts, our work features a novel semi-supervised image-to-image translation framework that resorts to pseudo paired data guidance to simplify this task. A patch-wise constructive loss within and between images is introduced to maintain content consistency and learn local anime textures better.

3. Methodology

Our goal is to stylize natural scenes with fine-grained anime textures while preserving the underlying semantics. We formulate the proposed *Scenimefy* into a three-stage pipeline: pseudo paired data generation (Section 3.1), segmentation-guided data selection (Section 3.2), and semi-supervised image-to-image translation (Section 3.3).

3.1. Pseudo Paired Data Generation

To bridge the domain gap between the real and anime scene, paired data is beneficial to establish semantic and style correspondences to ease the standard unsupervised I2I translation. While StyleGAN [16, 17] can synthesize high-quality images, the complexity of anime scenes necessitates an elaborately designed fine-tuning strategy for StyleGAN to generate plausible pseudo paired data.

The proposed pseudo paired data generation process is illustrated in Figure 3. With a source StyleGAN G_s pre-trained on the real scene dataset, we fine-tune it on the anime scene dataset to obtain G_t. Then, we can generate paired data $\{x^p, y^p\}$ with semantic similarity from a random latent code w as $x^p = G_s(w) \in X^p$ and $y^p = G_t(w) \in Y^p$, where X^p and Y^p are the pseudo real scene domain and anime scene domain, respectively. Based on the observation that the early layers (*i.e.*, the low-resolution ones) of StyleGAN determine the structure information, we freeze the initial blocks of the generator and the initial style vectors injected to preserve the spatial layout during fine-tuning.

To better preserve category-specific objects, we propose to guide y^p to follow the semantic attributes of x^p using pre-trained model priors, VGG [31] and CLIP [29]. Specifically, we include the CLIP loss to minimize the cosine distance between the CLIP-space embeddings of the two images and use the perceptual loss [43] to constrain the overall semantics.

$$L_{global} = D_{CLIP}(x^p, y^p) + \lambda_{lpip} lpips(x^p, y^p), \quad (1)$$

where $D_{CLIP}(\cdot,\cdot)$ denotes cosine distance in the CLIP space, $lpips(\cdot,\cdot)$ is the perceptual loss, and λ_{lpips} is the loss weight.

To better maintain local spatial information and details, we incorporate a patch-wise contrastive loss (PatchNCE), inspired by CUT [27], which applies contrastive learning to the embedded features of the generator. We use the pre-trained CLIP models to extract feature embeddings instead of an additional MLP header network that may cause the potential imbalance issue during fine-tuning [18]. Specifically, we randomly crop patches in x^p and y^p and embed them with the CLIP encoder E, as shown in Figure 3. Then, we bring the positive patches closer, which are cropped at the same position, and the negative patches far apart, which are cropped from different positions. Let v denote the embedded query patch from y^p. Let v^+ and $\{v_i^-\}_{i=1}^N$ be the embedded positive patch and N negative patches from x^p, respectively. The patch-wise loss can be written as:

$$L_{patch}(v, v^+, v^-) = -\log \left[\frac{\exp(v \cdot v^+)}{\exp(v \cdot v^+) + \sum_{i=1}^N \exp(v \cdot v_i^-)} \right], \quad (2)$$

The overall loss function of this stage can be written as:

$$L_{finetune} = L^t_{GAN}(G_t, D) + \lambda_{global} L_{global} + \lambda_{patch} L_{patch}, \quad (3)$$

where L^t_{GAN} is the adversarial loss [9], and D is the StyleGAN discriminator. λ_{global} and λ_{patch} are the loss weights.

3.2. Semantic Segmentation Guided Data Selection

Through pseudo paired data generation, we obtain a synthetic paired dataset with coarse pixel-wise correspondence. However, such raw pseudo paired data still risks low quality or poor structural consistency as shown in Figure 4, suggesting a need for data filtering.

To this end, we propose a semantic segmentation guided data selection scheme to purge low-quality samples with less structural consistency. We observe that recent semantic segmentation models, such as Mask2Former [6], can generalize well to the anime domain. Such observation allows us to use Mask2Former for pseudo paired data filtering based on two elaborately designed criteria, *i.e.*, semantic consistency and semantic abundance. Specifically, we employ pixel-wise cross-entropy loss L_{BCE} as a metric to evaluate semantic consistency. The samples with a loss value higher than a threshold of 5.0 are eliminated. To enrich semantic abundance, we exclude images with only one detected category since this indicates either little semantic information or low quality. The visualizations of the retained and filtered images and their predicted masks are presented in Figure 4, and more examples of the cleaned pseudo pairs are shown

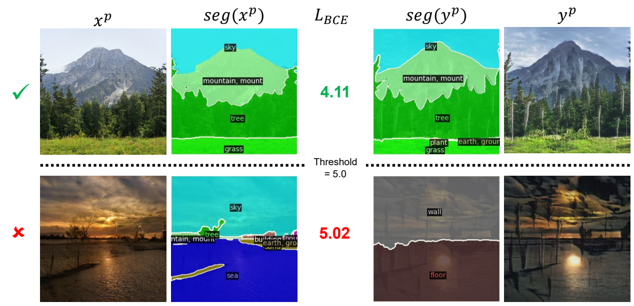


Figure 4: **Filtering examples of the segmentation-guided data selection scheme.** We automatically filter images of low quality using the pixel-wise cross-entropy loss L_{BCE}. The retained pseudo paired data (top) exhibits higher structure consistency than the discarded pair (bottom).

in Figure 3. It is observed that the remaining pseudo paired data achieves plausible quality after this stage.

3.3. Semi-Supervised Image-to-Image Translation

Our semi-supervised image-to-image translation framework (see Figure 5) consists of two branches, supervised and unsupervised. Given a set of real scene images $\{x_i\}_{i=1}^N$ in the real domain X and an anime set $\{y_j\}_{j=1}^M$ in the anime domain Y, our goal is to learn a mapping $G : X \to Y$ with the help of the pseudo paired dataset $P = \{x_i^p, y_i^p\}_{i=1}^N$. The dual-branch training procedure is detailed below.

3.3.1 Supervised Training Branch

The supervised training branch ingests the pseudo paired data to leverage the coarse pixel-wise correspondence between domain X^p and Y^p, thus facilitating the training and semantic mapping of complex scene stylization. The supervised branch is based on a conditional GAN framework [23], with the conditional adversarial loss:

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D_P) =& \mathbb{E}_{y^p, x^p}[\log D_P(y^p, x^p)] + \\ & \mathbb{E}_{x^p}[\log(1 - D_P(x^p, G(x^p))], \end{aligned} \quad (4)$$

where a patch discriminator D_P aims to distinguish between $\{(y^p, x^p)\}$ and $\{(G(x^p), x^p)\}$, and $(\cdot,\cdot)$ denotes a concatenation operation.

Different from a typical supervised I2I framework [12] that uses a reconstruction loss to impose a strong supervision, our ground truth images lie in Y^p rather than the real target Y. Instead, we introduce a novel patch-wise contrastive style loss for robust supervision. The intuition is that for a good translation, each patch within the translated image should be akin to the corresponding patch in the pseudo ground truth, rather than be identical to it. Such patches at the same location should be embedded to be closer, whereas the ones from different locations should be far away. The patch-level contrastive learning helps

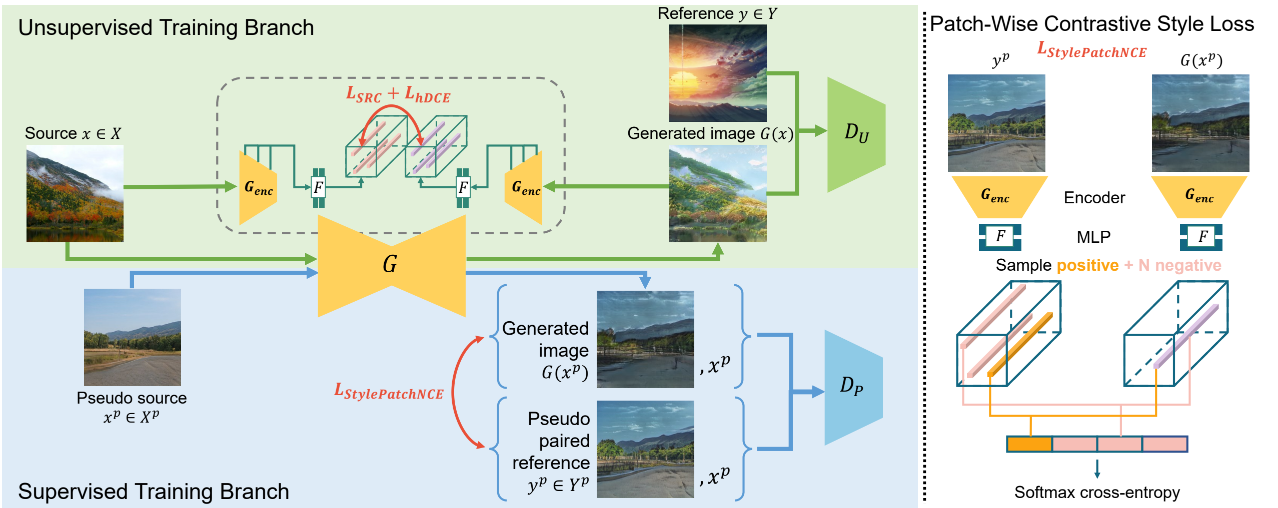


Figure 5: **Main framework of semi-supervised image-to-image translation. Left:** The proposed approach comprises two branches: unsupervised (top) and supervised (bottom). The supervised branch ingests the pseudo paired data as supervision, while the unsupervised branch learns the true target distribution using the original real and anime datasets. A novel patch-wise contrastive style loss is proposed to learn local fine details better. **Right:** Details of the patch-wise contrastive style loss.

our model learn robust local style similarity and focus on fine details. We name this loss the *StylePatchNCE* loss, as shown in Figure 5. We divide the generator G into two components, the encoder G_{enc} and the decoder G_{dec}, *i.e.*, $G(x) = G_{dec}(G_{enc}(x))$. The feature stack computed in G_{enc} is available to conduct image translation, whose element also naturally corresponds to a patch of the input image, with deeper layers representing larger patches. This feature is further passed through a two-layer trainable MLP network F, following SimCLR [4] to obtain the embedded patch feature. To precisely capture anime textures at different granularity, we select multi-scale features from a total of L layers of G_{enc}. Let $\tilde{v}_l^i$ and v_l^i be the l-th-layer embedded patch at the location i of $G(x^p)$ and y^p, respectively.

The proposed *StylePatchNCE* loss can thus be formulated as:

$$L_{StylePatchNCE}(G, F, Y^p) = \sum_{l=1}^{L} \sum_{i \neq j} L_{patch}^{style}(\tilde{v}_l^i, v_l^i, v_l^j), \quad (5)$$

L_{patch}^{style} shares the same contrastive loss form with L_{patch} in Eq. (2). The patch-level constraint enables a denser supervision of G, readily easing the training.

The training objective of the supervised branch is:

$$L_{sup} = L_{cGAN}(G, D_P) + \lambda_{style} L_{StylePatchNCE}(G, F, Y^p), \quad (6)$$

where λ_{style} is the weight for the StylePatchNCE loss.

3.3.2 Unsupervised Training Branch

The unsupervised branch directly ingests the original high-quality real dataset $\{x_i\}_{i=1}^N$ and anime dataset $\{y_j\}_{j=1}^M$ to learn the true target domain distribution. Inspired by Jung *et al*. [14], we notice the importance of tackling heterogeneous semantic relations of the image patches within a complex scene image. For instance, the patches from a mountain or a sea, and even their different parts, have diverse semantic information. Such semantic relation should be considered and preserved for plausible unsupervised scene stylization.

Accordingly, we apply the semantic relation consistency loss L_{SRC} and the hard negative contrastive loss L_{hDCE} for training [14]. L_{SRC} minimizes the Jensen-Shannon Divergence (JSD) of the in-image patch similarity distribution between x and $G(x)$, to enhance semantic consistency during translation. L_{hDCE} applies the patch-wise contrastive loss that gradually increases the discriminative difficulty of negative samples to enhance the discriminative power of the model. For the loss details, please refer to Jung *et al*. [14]. We apply both losses to the features of x and $G(x)$ extracted by G_{enc} and F, similar to our StylePatchNCE loss.

The total loss of the unsupervised branch is written as:

$$L_{unsup} = L_{GAN}(G, D_U) + \lambda_{SRC} L_{SRC} + \lambda_{hDCE} L_{hDCE}, \quad (7)$$

where L_{GAN} is the adversarial loss [9], and D_U is a standard discriminator for unsupervised training. λ_{SRC} and λ_{hDCE} are the loss weights.

3.3.3 Overall Training

The full framework of Scenimefy is thus semi-supervised, seeking a trade-off between faithfulness and fidelity of scene stylization. The full loss function is defined as:

$$L_{i2i} = L_{unsup} + \lambda_{sup} L_{sup}, \quad (8)$$

where λ_{sup} decays gradually following a cosine function as the training proceeds.

Figure 6: **Examples of our anime scene dataset.** We showcase the image samples of various anime scenes collected from nine Shinkai's films.

4. Experiments

4.1. Settings

Dataset. As our method is a semi-supervised image-to-image translation framework, the training datasets include real-world scene photos and anime scene images for the unsupervised branch, as well as a pseudo paired dataset for the supervised branch. During training, all the images are resized to a resolution of 256×256.

Real scene photos. We used 90,000 natural landscape images from the Landscapes High-Quality (LHQ) dataset [32] as our training set, and 6,656 scene images provided by the authors of CycleGAN [46] as our test set.

Anime scene photos. Our study also contributed a high-resolution (1080×1080) Shinkai-style pure anime scene dataset comprising 5,958 images. To construct this dataset, we gathered key frames from nine prominent Shinkai Mokoto films, *i.e.*, 'Weathering with You' (2019), 'Your Name' (2016), 'Children Who Chase Lost Voices' (2011), *etc*. Subsequently, we manually refined the dataset by eliminating irrelevant and low-quality images. Unlike the datasets used in previous studies [3, 36], our dataset does not contain randomly cropped images with a large portion of human portraits, which exhibit notable dissimilarities in their features when compared with the background scene. This curation was done to mitigate the potential overfitting issue during fine-tuning and narrow the domain gap. Our dataset will be made publicly available to facilitate future research in scene stylization. The example images in our dataset can be found in Figure 6.

Pseudo paired dataset. We randomly sampled 30,000 paired images with the same latent codes from the source StyleGAN generator and the fine-tuned one. We set a mild truncation trick [16] with a threshold $\gamma = 0.7$ to ameliorate data quality without sacrificing much diversity. The proposed segmentation-guided data selection (Section 3.2) was applied to ameliorate data quality.

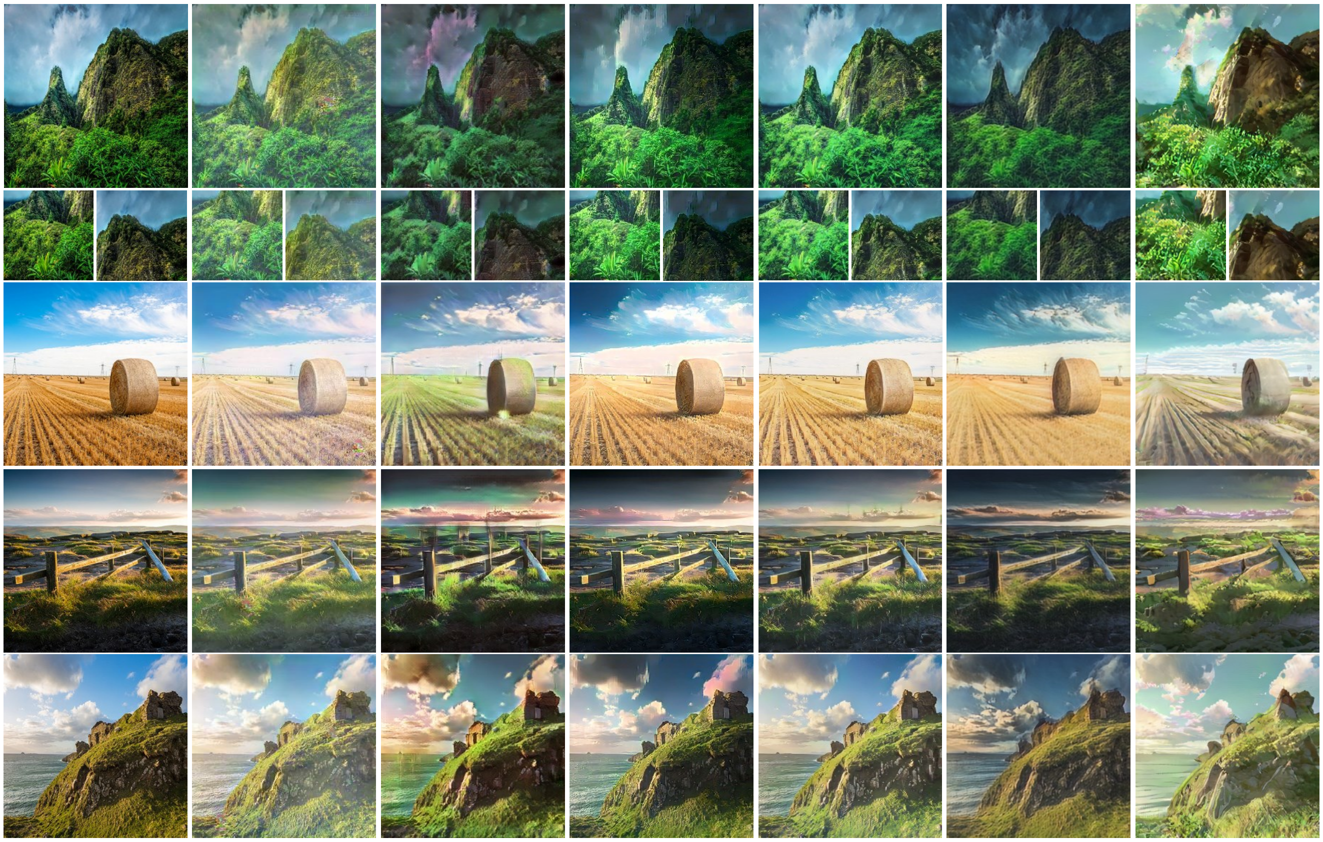
Baselines. We select five state-of-the-art baselines to make a comprehensive comparison. These baselines can be grouped into two categories: 1) representative image-to-image translation translation methods customized for scene cartoonization, *i.e.*, CartoonGAN [5], AnimeGAN [3], White-box [36], CTSS [8]; 2) and the StyleGAN-based approach, *i.e.*, VToonify [42].

Implementation details. Regarding the training process, we first train a StyleGAN2 generator on the LHQ dataset at 256×256. We then fine-tune this generator on our collected anime dataset, with the last three layers trainable and the remaining layers frozen, using the hyper-parameters $\lambda_{lpips} = 0.01$, $\lambda_{global} = 1.0$, $\lambda_{patch} = 0.05$ for 1,000 iterations. Following this, we generate 30,000 pseudo paired data using a truncation of 0.7 with the proposed data selection scheme described in Section 3.2. Our implementation of the unsupervised training branch is based on the recent image translation model [14]. Scenimefy is trained on a single NVIDIA GeForce RTX 3090 GPU for 20 epochs with $\lambda_{style} = 0.05$, $\lambda_{SRC} = 0.05$, $\lambda_{hDCE} = 0.1$, $\lambda_{sup}(t) = \cos(\frac{\pi}{40}(t-1))$, where $t = \{1, 2, .., 20\}$ is the number of training epoch. More detailed settings are provided in the *supplementary material*.

Evaluation metrics. We use Fréchet Inception Distance (FID) [10] to quantify the perceptual quality of translated images. The FID is calculated between a collection of 6,605 generated images and our introduced anime scene dataset. A lower value indicates the better image quality. In addition, we conduct a user study, where the candidates rate different methods in terms of stylization, semantic preservation, and overall translation quality. Higher scores signify better image quality.

4.2. Main Results

Qualitative comparison. Figure 7 shows the qualitative comparison with five state-of-the-art methods. Our results seek a plausible trade-off between style fidelity and semantic faithfulness, where previous methods have either focused on pursuing more content consistency at the expense of weak anime style or have degraded into an image abstraction method, leading to the loss of fine details. CartoonGAN [5] and AnimeGAN [3], design handcrafted anime-specific losses, such as the edge-smoothed loss, in attempts to manifest sharp edges, which, however, limits the stylization degree. In addition, the results of CartoonGAN are overexposed, while AnimeGAN suffers from artifacts in the sky and sea. White-box [36] and CTSS [8], generate results that look more like texture abstraction with a weak style. When tackling scene images, the StyleGAN-based style transfer method, VToonify [42] lacks both local anime texture and the global style like color. All the baseline methods are unable to fully capture the inherent anime texture features. In contrast, Scenimefy presents delicate anime features while retaining semantic consistency. For example, in the enlarged region of the first row, our model success-



(a) Source (b) CartoonGAN (c) AnimeGAN (d) White-box (e) CTSS (f) VToonify (g) Ours

Figure 7: **Qualitative comparison.** We compare our approach with five representative methods. Scenimefy (ours) produces more semantic-consistent results with rich anime-style textures compared to state-of-the-art baselines. Zoom in for details.

Table 1: **Quantitative comparison using FID.** A lower FID is better. The anime scene dataset is used as a reference.

Method	LHQ (real) [32]	CartoonGAN [5]	AnimeGAN [3]	White-box [36]	CTSS [8]	VToonify [42]	Ours
FID↓	121.807	67.200	67.739	61.973	66.729	90.578	**48.922**

Table 2: **User preference scores.** The best scores are marked in bold.

Method	CartoonGAN [5]	AnimeGAN [3]	White-box [36]	CTSS [8]	VToonify [42]	Ours
Style	0.067	0.083	0.110	0.043	0.010	**0.687**
Content	0.087	0.080	0.103	0.123	0.030	**0.577**
Overall	0.073	0.077	0.103	0.057	0.017	**0.673**

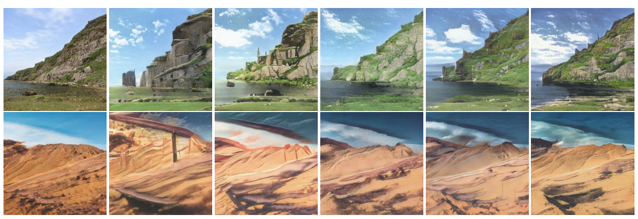
fully mimics the brush strokes of the leaves. The qualitative results indicate the effectiveness of our method, outperforming the state-of-the-art baselines in perceptual quality. More comparative results can be found in our *supplementary material*.

Quantitative results. In Table 1, we conduct a quantitative evaluation of our method against the baselines. Our approach achieves the lowest FID score, indicating that the quality of our translated results is the best, consistent with our higher visual quality. We also test the FID between the real and the anime scene datasets as a reference. The style distribution of our results is much closer to the anime domain compared with the real one.

In addition to FID, we conducted a user study with 30 subjects participating to assess the quality of anime scene rendering based on three criteria: evident anime stylization (Style), consistent semantic preservation (Content), and overall translation performance (Overall). Participants selected what they consider to be the best results from six different methods across 10 sets of images. Table 2 summarizes the average preference scores, where Scenimefy receives the best scores in all three criteria, thus further suggesting the effectiveness of our method.

(a) Source (b) w/o unsupervised (c) w/o supervised (d) w/o D_P (e) w/o $L_{StylePatchNCE}$ (f) w/o λ_{sup}decay (g) Full model

Figure 8: **Ablation study of semi-supervised I2I translation.** The effect of each key component is illustrated.



(a) Source (b) w/o freezeG (c) w/o *lpips* (d) w/o D_{CLIP} (e) w/o L_{patch} (f) Full loss

Figure 9: **Ablation study of StyleGAN fine-tuning**. The influence of each key fine-tuning technique is shown.

Table 3: L_{BCE} **of StyleGAN fine-tuning.**

Finetune	w/o freezeG	w/o *lpips*	w/o D_{CLIP}	w/o L_{patch}	Full loss
$L_{BCE} \downarrow$	4.45	4.53	4.25	4.30	**4.24**

Table 4: L_{BCE} **of semi-supervised I2I translation.**

I2I	w/o unsupervised	w/o supervised	w/o D_P	w/o $L_{StylePatchNCE}$	w/o λ_{sup}decay	Full model
$L_{BCE} \downarrow$	4.76	4.50	4.68	4.72	4.58	**4.39**

4.3. Ablation Study

StyleGAN fine-tuning. The effect of each semantic-constrained technique for fine-tuning is shown in Figure 9. Without freezing the shallow layers of the pre-trained StyleGAN (see Figure 9(b)), the spatial structure is severely altered. Ablating the global constraints (*lpips* or D_{CLIP}) from the pre-trained model prior, category-specific objects are poorly preserved, such as odd rock textures on mountains, as shown in Figure 9(c)(d). The removal of the patch-wise consistency loss in Figure 9(e) results in a loss of fine details. Applying all the pre-trained model prior constraints at both image-level and patch-level, our full method accurately generates valid pairs, transfers the anime style while maintaining the semantic structure, and has fewer artifacts.

Semi-supervised image-to-image translation. To verify the efficacy of the proposed semi-supervised image-to-image translation framework, we conduct a systematic ablation study by removing each key module independently. The visual results are shown in Figure 8. It is observed that training each branch separately results in poor outputs. Employing the supervised branch alone (Figure 8(b)) leads to low-quality results due to coarse guidance. Meanwhile, although the unsupervised branch alone successfully learns global anime style, it lacks evident local texture stylization and semantic preservation, and suffers notable artifacts, *e.g.*, translating the water into stones in Figure 8(c). Without the conditional discriminator, the results exhibit discernible local details, as depicted in Figure 8(d). Ablating the patch-wise contrastive style loss leads to noisy and weird patterns on the mountain in Figure 8(e). The coarse pseudo data may lead to negative effects, such as less evident style effects in Figure 8(f). We tackle this problem by a weight decay technique of the supervised branch. The results of our full model, as shown in Figure 8(g), exhibit superior anime rendering ability, including anime texture details, harmonious colors, and much less noise. All the modules work together to improve the overall performance of the proposed method.

Quantitative comparison. We applied L_{BCE} metric (detailed in Section 3.2, lower is better) to our ablation studies for a more comprehensive quantitative evaluation on the semantic consistency (see Table 3 and 4). In the StyleGAN fine-tuning experiments, we generate $3,000$ images with the same seeds and calculated their respective L_{BCE} loss against the nature images. For I2I translation, our test dataset comprises $6,656$ images from [46]. We achieved the best scores in both experiments, verifying the effectiveness of our design in better semantic preservation.

4.4. Further Analysis

Other anime styles. We trained our model over a different anime dataset, *i.e.*, the Hosoda Mamoru dataset, comprising $5,107$ images from [36], to validate its versatility. This dataset includes similar scenes cropped from the same movie frame, along with numerous human portraits, result-

(a) Source (b) CartoonGAN (c) AnimeGAN (d) White-box (e) CTSS (f) Ours (Hosoda) (g) Ours (Shinkai) (h) Reference

Figure 10: **Hosoda anime style.** We compare our approach with four baseline methods on the Hosoda dataset from [36]. Scenimefy (ours) exhibits a better ability to capture different anime-style textures. Zoom in for details.

ing in a significant domain gap. Thus, we used a global perceptual loss [43] between the input and output images to help maintain content consistency. The qualitative results, illustrated in Figure 10, showcase the efficacy of our method in learning and transferring desired anime styles. Notably, the grass exhibits distinct styles. In the Hosoda style, the grass texture appears more simplistic (Figure 10(f)), while in the Shinkai style, it appears more detailed and colorful (Figure 10(g)).

Anime texture transfer. As in the real anime images in Figure 12(e-g), the rock, hay bale, car and background plants (Row 3) are smooth, and only the foreground plants (Row 2) are detailed. The fence has curvy edge and smooth textures (Row 5). Our model is the only one that adheres to these key characteristics of anime scenes, while other baselines overemphasize the fidelity to the original real images so that their results look less like real anime.

Temporal consistency on videos. We simply extend our method to video stylization by transferring each single frame. The results of some representative frames are shown in Figure 11, where Scenimefy maintains smooth and coherent visual information.

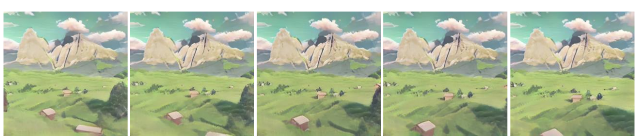
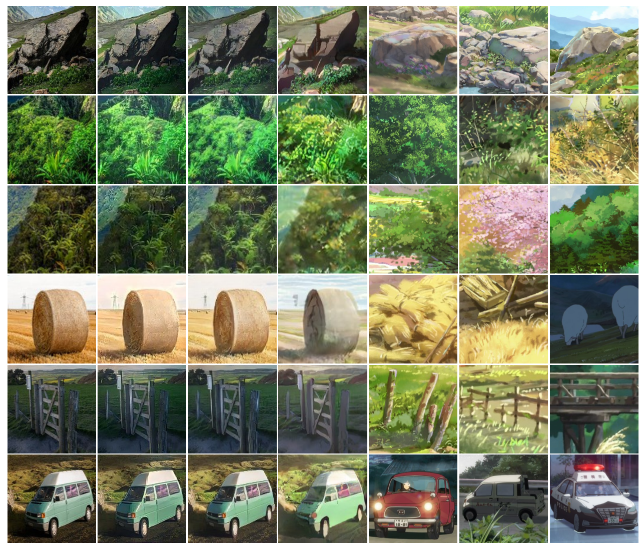


Figure 11: **Temporal consistency on videos.** The results of representative frames in video stylization.

5. Conclusion

In this paper, we proposed Scenimefy, a powerful framework for anime scene rendering, which comprises three stages: pseudo paired data generation, semantic segmentation-guided data selection, and semi-supervised image-to-image translation. Besides, we contributed a high-resolution anime scene dataset to facilitate future research in scene stylization. Our results empirically demonstrated that the use of soft pseudo paired data guidance can effectively balance style fidelity and semantic faithfulness, simplifying the pure unsupervised setting. The proposed contrastive style loss facilitates fine detail generation. Scenimefy thus outperforms state-of-the-art baselines in both perceptual quality and quantitative performance. Despite promising results, there remain a few exciting avenues for improvement, such as incorporating explicit control of stylization degree and enabling more flexible translations with user-input style. Recent breakthroughs in diffusion models have enabled remarkable image generation capability. By leveraging these advancements, we can obtain improved pseudo-paired data with enhanced details. We believe that harnessing the potential of large-scale text-to-image models may further elevate the quality of automatic anime scene rendering.



(a) Source (b) White-box (c) CTSS (d) Ours (e) Ref. 1 (f) Ref. 2 (g) Ref. 3

Figure 12: **Detailed anime texture transfer comparison.** The texture details between real anime and the generated ones by different methods.

Acknowledgment. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also supported by Singapore MOE AcRF Tier 2 (MOE-T2EP20221-0001) and the NTU NAP Grant.

References

[1] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *NeurIPS*, 2017. 2

[2] bryandlee. Freeze generator. https://github.com/bryandlee/FreezeG, 2020. 3

[3] Jie Chen, Gang Liu, and Xin Chen. AnimeGAN: A novel lightweight GAN for photo animation. In *ISICA*, 2020. 2, 3, 6, 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 5

[5] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. CartoonGAN: Generative adversarial networks for photo cartoonization. In *CVPR*, 2018. 2, 3, 6, 7

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 4

[7] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *TOG*, 2022. 3

[8] Xiang Gao, Yuqi Zhang, and Yingjie Tian. Learning to incorporate texture saliency adaptive attention to image cartoonization. In *ICML*, 2022. 2, 3, 6, 7

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 4, 5

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 6

[11] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 4

[13] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. TSIT: A simple and versatile framework for image-to-image translation. In *ECCV*, 2020. 2

[14] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *CVPR*, 2022. 2, 5, 6

[15] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 2020. 3

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3, 6

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 2, 3

[18] Gihyun Kwon and Jong Chul Ye. One-shot adaptation of GAN in just one CLIP. *arXiv preprint*, 2022. 3, 4

[19] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint*, 2020. 3

[20] Zhansheng Li, Yangyang Xu, Nanxuan Zhao, Yang Zhou, Yongtuo Liu, Dahua Lin, and Shengfeng He. Parsing-conditioned anime translation: A new dataset and method. *TOG*, 2023. 1

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *NeurIPS*, 2017. 2

[22] Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. DCT-Net: domain-calibrated translation for portrait stylization. *TOG*, 2022. 1

[23] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint*, 2014. 2, 4

[24] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning GANs. *arXiv preprint*, 2020. 3

[25] Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In *CVPR*, 2020. 2

[26] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, 2021. 3

[27] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *ECCV*, 2020. 2, 4

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[30] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv preprint*, 2020. 3

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2014. 2, 3

[32] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *CVPR*, 2021. 6, 7

[33] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. AgileGAN: stylizing portraits by inversion-consistent transfer learning. *TOG*, 2021. 1

[34] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for GAN training. *TIP*, 2021. 3

[35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, 2018. 2

[36] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *CVPR*, 2020. 2, 3, 6, 7, 8, 9

[37] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost Van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring GANs: generating images from limited data. In *ECCV*, 2018. 3

[38] Yue Wang, Ran Yi, Ying Tai, Chengjie Wang, and Lizhuang Ma. CtlGAN: Few-shot artistic portraits generation with contrastive transfer learning. *arXiv preprint*, 2022. 3

[39] Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *CVPR*, 2022. 3

[40] Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, and Bolei Zhou. One-shot generative domain adaptation. *arXiv preprint*, 2021. 3

[41] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche Master: exemplar-based high-resolution portrait style transfer. In *CVPR*, 2022. 1

[42] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. VToonify: Controllable high-resolution portrait video style transfer. *TOG*, 2022. 1, 6, 7

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3, 9

[44] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. *NeurIPS*, 2020. 3

[45] Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *CVPR*, 2022. 3

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 6, 8

[47] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the Gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *arXiv preprint*, 2021. 3