Appearance-preserved Portrait-to-anime Translation via Proxy-guided Domain Adaptation

Wenpeng Xiao, Cheng Xu, Jiajie Mai, Xuemiao Xu, Yue Li,
Chengze Li, Xueting Liu, and Shengfeng He, *Senior Member, IEEE*

Abstract—Converting a human portrait to anime style is a desirable but challenging problem. Existing methods fail to resolve this problem due to the large inherent gap between two domains that cannot be overcome by a simple direct mapping. For this reason, these methods struggle to preserve the appearance features in the original photo. In this paper, we discover an intermediate domain, the coser portrait (portraits of humans costuming as anime characters), that helps bridge this gap. It alleviates the learning ambiguity and loosens the mapping difficulty in a progressive manner. Specifically, we start from learning the mapping between coser and anime portraits, and present a proxy-guided domain adaptation learning scheme with three progressive adaptation stages to shift the initial model to the human portrait domain. In this way, our model can generate visually pleasant anime portraits with well-preserved appearances given the human portrait. Our model adopts a disentangled design by breaking down the translation problem into two specific subtasks of face deformation and portrait stylization. This further elevates the generation quality. Extensive experimental results show that our model can achieve visually compelling translation with better appearance preservation and perform favorably against the existing methods both qualitatively and quantitatively. *Our code and datasets are available at https://github.com/NeverGiveU/PDA-Translation.*

Index Terms—Portrait-to-anime translation, coser portrait proxy, domain adaptation.

1 Introduction

Anime is a world-wide popular art form which is extensively involved in entertainment industries. Creating an anime portrait from a real photo is not trivial as it needs careful abstraction and deformation of facial features as well as preserving the distinctive characteristics of the input portrait. This is labor-intensive and time-consuming even for professional artists.

Due to the lack of paired data, existing image-to-image translation methods [1], [2], [3] adopt an unsupervised setting to connect two domains. Notwithstanding the demonstrated success on similar domains (*e.g.*, zebra and horse), they fail in dealing with large domain gaps. In particular, human portraits show large appearance differences from the anime ones, especially for the exaggerated facial components, smooth texture, and stylish hairstyles (see the first row in Fig. 2). This huge domain discrepancy, unfortunately, incurs severe learning ambiguity during the training process and difficulty to compute local similarity other than data distributions. Therefore, domain similarity attracts all the attention of the network but lose identity characteristics in the output anime (see Fig. 1), which seriously impairs the recognizable features of generation results.

For portrait-to-anime translation, the generated anime portrait is desirable to faithfully resemble the input portrait to keep the original identity [4]. Since different anime portraits usually share similar facial patterns of large eyes, a tiny nose, and a small mouth [5], we find that there are five characteristics, namely, the hair style, hair color, face shape, expression, and skin color, that are more visually recognizable (the first row in Fig. 2). We argue that an ideal portrait-to-anime approach should be able to preserve these recognizable features of the input photo. To achieve this goal, our key idea is to introduce a proxy between the portraits and anime portraits such that the large domain gap between the above two domains can be alleviated to a smaller one, thereby eliminating the learning ambiguity during the training and facilitating the appearance-preserved translation. This proxy we identify is the *coser portrait*, *i.e.*, the portraits of costume players (humans costuming as anime characters). The coser portrait owns similar features of real-life human portraits while exhibits certain anime appearances and styles (see Fig. 2). This motivates us to take advantage of the coser portrait to aid our training, such that optimal mappings can be learned for appearance-preserved translation.

- This work is supported by Guangdong International Technology Coopertation Project (No. 2022A0505050009); the Key-Area Research and Development Program of Guangdong Province, China (No. 2020B010165004, No. 2020B010166003); National Natural Science Foundation of China (No. 61972162); Guangdong International Science and Technology Cooperation Project (No. 2021A0505030009); Guangdong Natural Science Foundation (No. 2021A1515012625); Guangzhou Basic and Applied Research Project (No. 202102021074); CCF-Tencent Open Research fund (CCF-Tencent RAGR20210114); and substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project No. UGC/FDS11/E02/21. *(Wenpeng Xiao and Cheng Xu contributed equally to this work.) (Corresponding authors: Xuemiao Xu and Shengfeng He.)*
- Wenpeng Xiao, Cheng Xu, Xuemiao Xu, Yue Li, and Shengfeng He are with the School of Computer Science and Engineering, South China University of Technology, Guangdong, China. Xuemiao Xu is also with Ministry of Education Key Laboratory of Big Data and Intelligent Robot and Guangdong, and State Key Laboratory of Subtropical Building Science, Provincial Key Lab of Computational Intelligence and Cyberspace Information. E-mail: {wpxiao.littleblack, cschengxu}@gmail.com; {xuemx, liyue, hesfe}@scut.edu.cn.
- Jiajie Mai is with King's College London, London, United Kingdom. E-mail: k20035517@kcl.ac.uk.
- Chengze Li and Xueting Liu are with Caritas Institute of Higher Education, Hong Kong SAR, China. E-mail: {czli, tliu}@cihe.edu.hk.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3228707

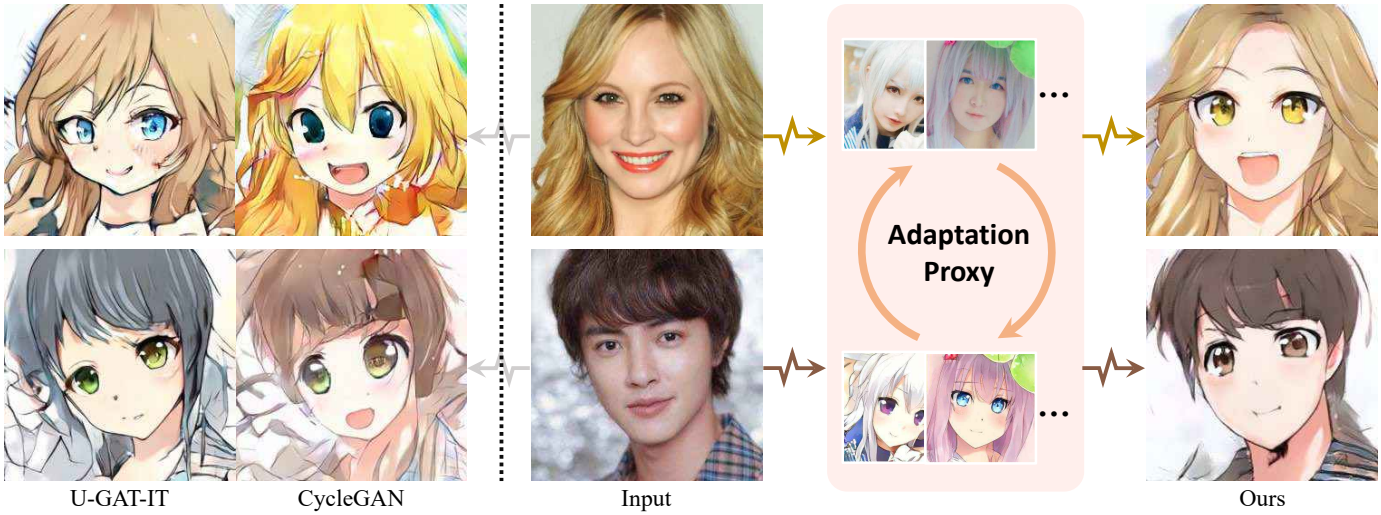


Fig. 1: The domain gap between human portrait and anime portrait is too large to translate for existing methods (left part). We introduce an intermediate translation proxy (coser portrait) with a progressive domain adaption scheme to bridge the domain gap and achieve an appearance-preserved portrait-to-anime translation (right part).

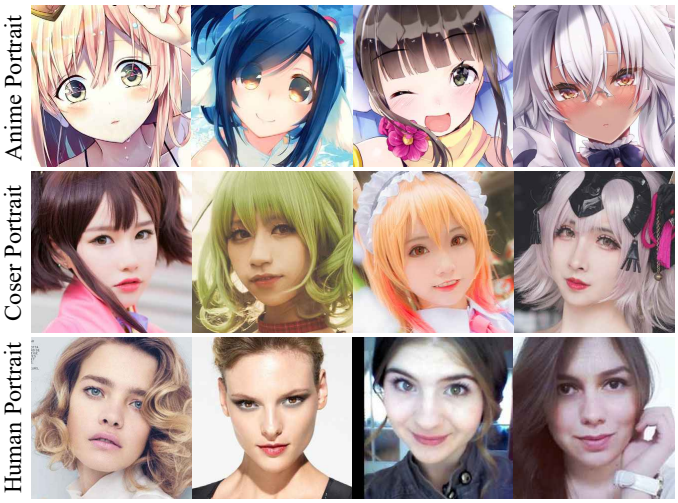


Fig. 2: Examples from different portrait domains. The three rows show the anime, the coser and the human portraits, respectively. For the coser portrait, the first two examples are of headdress-free style, while the latter two are of headdress style. For the human portrait, the first two examples are from CelebA-HQ [6] and the last two are from Selfie [3]. The coser portraits possess similar features of real-life human portraits as the human portraits and exhibit certain anime appearances and styles.

Based on the above observation, we aim to introduce the coser portraits as a proxy to guide portrait-to-anime translation with well-preserved appearance. To this end, we propose a proxy-guided domain adaptation learning scheme which consists of three progressive training stages. Concretely, in the first stage, we train our model with unpaired coser portraits and anime portraits to obtain an initial model with compelling appearance preservation capability. To further adapt the initial model to the human portrait domain, a most straightforward way is to fine-tune the model on the human portrait dataset. However, this may result in unsatisfactory adaptation results as there is a significant gap between the coser portrait and the human portrait domains. We therefore propose to achieve the adaptation in a gradual manner. To start with, we synthesize a new dataset (dubbed *Aug-Portrait*) of the intermediate domain between the coser portraits and the human portraits by fusing the coser portrait face shapes and the human facial textures together. Then, we perform our second-stage training with a combination of the Aug-Portrait and coser portrait datasets. By learning from the rich and diverse facial textures from human portraits, our model gains a better generalization capability on the human portraits. To fully adapt our model to the human portrait domain, we finally perform our third stage of training by adding the human portrait dataset into our training data. By this means, the initial model can be well adapted to the human portrait domain, enabling visual-pleasing and appearance-preserved translations on the human portraits. To further alleviate the learning difficulty, we adopt a disentangled framework with two subnetworks as our translation model, with each subnetwork focusing on face shape deformation and portrait stylization, respectively. This disentangled scheme further facilitates our model to handle large deformation and generate anime portraits with higher quality.

We evaluate our method on various human portraits for the portrait-to-anime translation task. The experimental results show that our method significantly outperforms the existing methods in generating high-quality anime portraits with well-preserved appearance features. We summarize the main contributions of our method as follows:

- We propose a proxy-guided domain adaptation learning scheme to achieve unsupervised appearance-preserved portrait-to-anime translation.
- We delve into the large domain gap problem arising from the existing portrait-to-anime datasets and propose a coser portrait dataset as a proxy bridging the gap between the human portrait and anime portrait domains.

- In comparison with the existing methods, extensive experiments demonstrate the superiority of our method in generating high-quality appearance-preserved anime portraits.

2 Related Work

2.1 Image-to-Image Translation

Image-to-image translation aims to learn a mapping from the source domain to the target domain. It has been a hot topic for years with wide applications in the field of computer vision, such as image super-resolution/restoration [7], [8], [9], colorization [10], [11], [12], style/appearance transfer [13], [14], [15], [16], *etc*. Pix2pix [17] first proposes a cGAN-based network to learn the mapping from the input to the output images, which yields plausible results on many image-to-image translation tasks such as semantic image synthesis, sketch-to-photo synthesis, *etc*. To deal with high-resolution images, pix2pixHD [18] proposes a coarse-to-fine generator and multi-scale discriminators for more stable training and fine details generation. Albeit the impressive results achieved, these methods rely on paired data for training, which is hard to obtain. To circumvent this issue, several works are proposed to solve the image-to-image translation problem with unpaired data. CycleGAN [1] proposes to learn a bilateral mapping of images between two domains via a cycle consistency constraint. UNIT [2] makes a shared latent space assumption to enable unsupervised image-to-image translation (UI2I). To achieve diverse-modality outputs for a single input image, MUNIT [19] and DRIT [20] propose to disentangle an image into domain-specific style and domain-independent content representations. By incorporating domain labels as the condition of translation, StarGAN [21] solves the multi-domain image-to-image translation with a single model only. FUNIT [22] proposes a network that performs image translation between seen classes during training and scales to the unseen classes during testing. For the anime portrait generation with no paired data, applying the aforementioned unsupervised methods simply lead to a direct domain mapping from the portrait to the anime portrait domain, which fails to cope with the large shape deformation and texture stylization simultaneously and tends to produce inferior results with unsatisfactory appearance preservation.

2.2 Anime Portrait Translation

In the past decades, anime portrait translation has witnessed impressive progress and many approaches have been proposed to address this problem. Early works [4], [23], [24] usually resort to patch-based feature matching to find the optimal candidates for further aggregation and smoothing of synthesized cartoon faces. These methods rely on predefined portrait-to-cartoon examples, which is quite labor-intensive. Furthermore, they can hardly handle with large deformation because patches are matched based on structure similarity. With the rise of deep learning, a number of CNN-based anime portrait generation methods are proposed [3], [25], [26], [27], [28], [29]. Cao *et al.* [27] propose a two-stage network for exaggerated caricature generation. However, the two stages are optimized individually without considering the mutual interactions of both stages, which may lead to sub-optimal solutions. Su *et al.* [30] address the photo-to-manga problem by translating each facial region into the manga domain with a heavy multi-GANs architecture. Wu *et al.* [26] leverage five-point facial landmarks to ensure a rational face structure for generated cartoon portraits, but it ignores the explicit deformation of two domains and thus fails to model large deformation and cannot preserve the appearance of input portraits. U-GAT-IT [3] proposes a domain attention mechanism and an Adaptive Layer Instance Normalization to achieve anime portrait generation with flexible control of shape and texture editing. Nevertheless, it struggles to achieve satisfactory translations between domains with a large gap. AniGAN [25] proposes a style-guided framework for style-controlled anime portrait generation. Despite the success of the current anime portrait generation methods, they still suffer from a large domain gap between the source and the target domains. This unfortunately leads to a high level of learning ambiguity, which causes inferior results with severe appearance changes during translation.

2.3 Image-to-Image Translation using StyleGAN

StyleGAN [31], [32] is a powerful generative model which can yield images with extraordinary visual quality. Recently, several works have demonstrated the advantages of StyleGAN in UI2I problems [33], [34], [35], [36]. The base idea is to fine-tune a source-domain model with the target-domain data to learn a new style rendering of deep layers while keeping the semantic representations of shallow layers constant. Toonify [33] performs layer swapping between the source model and the fine-tuned target model for UI2I. Similarly, UI2I-via-StyleGAN2 [34] further proposes to preserve the semantic similarity between the input and output images. To explicitly maintain the structure of the input image, Cartoon-StyleGAN [35] proposes to freeze the shadow layers of StyleGAN and apply a structure constraint during finetuning. StyleCariGAN [37] proposes a shape exaggeration blocks to progressively modulate shadow features for modeling exaggerated deformations. However, it is difficult to find an optimal trade-off between perceptual quality and semantic consistency of the generated results for the above layer-swapping based methods. To better project the real images into a robust latent space across different stylized StyleGANs and enable higher-quality translation, AgileGAN [36] designs a hierarchical Variational Autoencoder (hVAE) to map the input image to a Gaussian distribution. Nevertheless, semantic consistency is not considered to handle the large gap between human and anime. JoJoGAN [38] learns a style mapper using a single style reference image with a multi-step pipeline. However, the styles of different semantic regions (*i.e.*, hair, eyes, mouth, *etc*) of the output image are required to be consistent with the given reference, which is not applicable for our purpose of preserving the appearance of the input portrait during translation. In a nutshell, although higher-quality results can be generated compared to traditional UI2I works, the existing StyleGAN-based I2I methods still suffer from severe appearance changes and facial artifacts in portrait-to-anime translation due to the huge gap between the two domains.

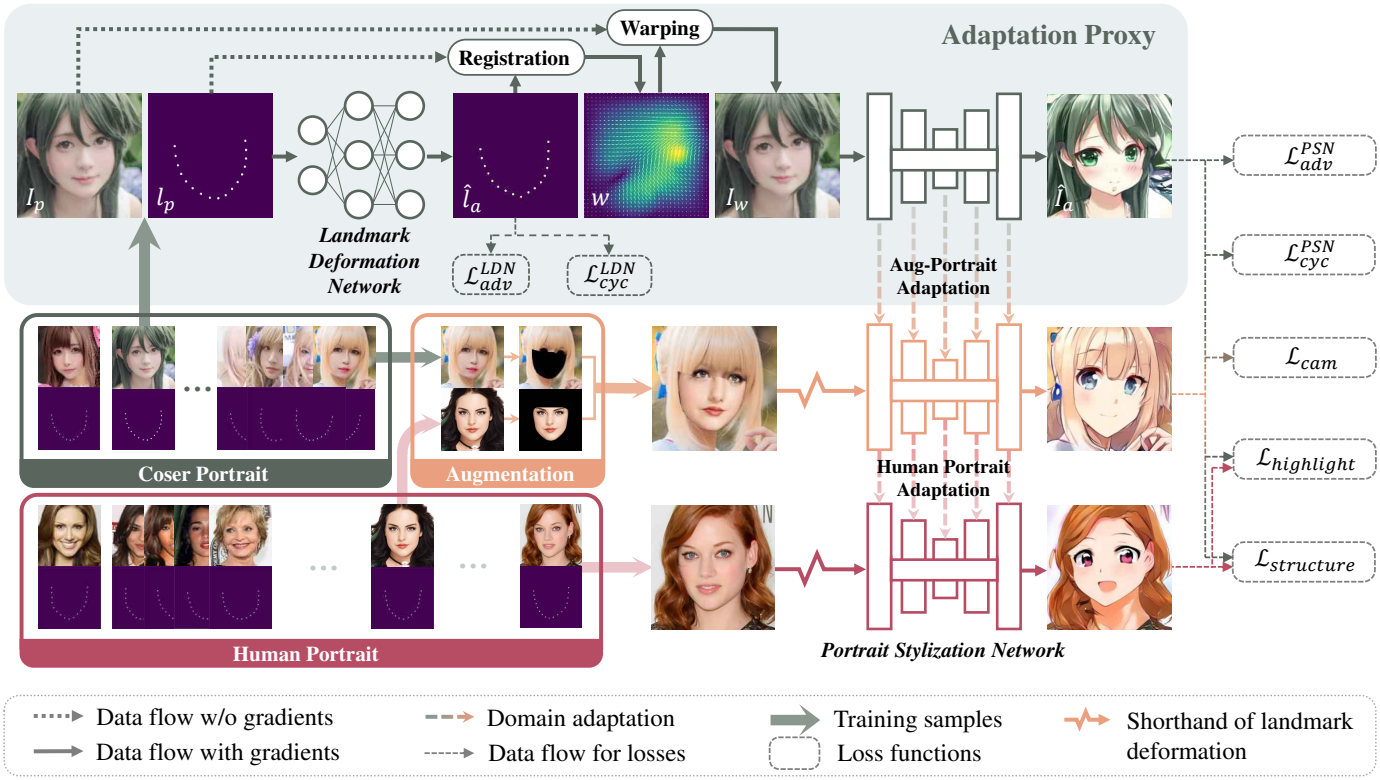


Fig. 3: Overview of our proxy-guided domain adaptation learning scheme. The model is first trained with the proposed coser portrait dataset to obtain an initial model with appearance preservation capability. Then the trained model is sequentially adapted to the Aug-Portrait and the human portrait domain. We adopt an end-to-end disentangled framework to better model the face deformation and portrait stylization during the translation. Note that the domain adaptation processes of the Landmark Deformation Network are omitted for clarity.

Furthermore, the inaccurate inversion of the input image and the inherent data bias of the pre-trained StyleGANs can lead to undesired information loss (*e.g.*, the hair textures, accessories, *etc*) that yields inferior translation results.

3 Method

Our goal is to achieve fully unsupervised appearance-preserved portrait-to-anime translation by introducing a proxy to aid the learning and shifting the appearance preservation capability inherited from the proxy domain to the human portrait domain. To accomplish this goal, we propose a new coser portrait dataset as the proxy and present a proxy-guided domain adaptation learning scheme, which consists of three progressive training steps to adapt the appearance-preserved model trained with the coser portrait to the human portrait domain in a gradual manner. An overview of our proxy-guided domain adaptation framework is shown in Fig. 3. In the following, we discuss the proxy-guided domain adaptation learning scheme and the network architecture in detail.

3.1 Initialization on the Coser Portrait Dataset

Due to the large domain gap between the human portrait and the anime portrait domains, it is extremely difficult for the model to learn correct translation patterns that are indispensable for the appearance-preserved translation. To effectively alleviate the gap between the two domains and facilitate the learning, we introduce a coser portrait dataset as a proxy to make the task easier and achieve better appearance preservation performance.

3.1.1 The Coser Portrait Dataset

Cosplay, in short for costume play, is the activity that humans are wearing costumes to represent anime characters. The cosers exhibit anime appearances and styles as well as real-life facial textures. This makes the coser portrait a good intermediate domain between the human portrait and the anime portrait domains, which can relieve the gap between them. We conduct a toy experiment to illustrate this observation more intuitively. Particularly, we first quantitatively investigate the domain gaps between the input/output domains of the widely used datasets for portrait-to-anime translation, the CelebA-HQ2Anime and Selfie2Cartoon datasets [3]. Following [39], [40], [41], we choose the FID as the metric of the gap between two domains. Specifically, we compute the FID score between the anime portraits and the CelebA-HQ datasets and the same metric between the cartoon portraits and the Selfie datasets to measure the discrepancies of different data distributions. Correspondingly, the FID score between the anime portraits and the coser portraits and the FID score between the cartoon portraits and the coser portraits are also computed for comparison. As shown in Fig. 4, the FID score between the anime (cartoon) portraits and the coser portraits is significantly smaller than those between

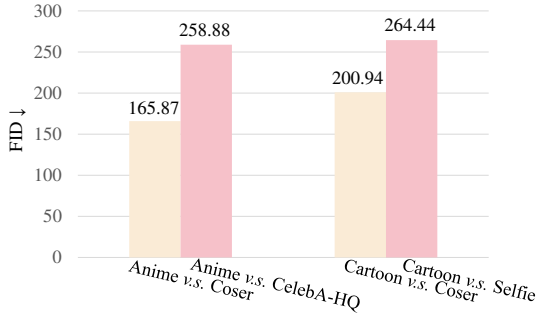


Fig. 4: FID scores between different datasets. The distribution between the anime/cartoon portraits and the coser portraits is significantly closer than those between the anime/cartoon portraits and the human portraits (with a smaller FID score compared to those of CelebA-HQ and Selfie, respectively). This makes the coser portraits a reasonable intermediate domain bridging the gap between the anime (cartoon) portraits and the human portraits.

the anime (cartoon) portraits and the CelebA-HQ (Selfie) datasets. This implies the distribution of coser portraits is significantly closer to those of the anime/cartoon portraits compared to the human portrait datasets (*e.g.*, CelebA-HQ and Selfie), making it indeed a reasonable intermediate domain that bridges the domain gap between the anime (cartoon) portraits and the human portraits.

Hence, we propose a coser portrait dataset comprising the frontal face images of cosers. These images are of more anime-like appearances compared to the human portraits in terms of the makeups, hairstyles, and facial expressions. We shall use the coser portrait dataset as the initial training set to achieve a good initialization of our model with remarkable appearance preservation capability.

3.1.2 The Disentangled Framework

In order to achieve portrait-to-anime translation with high quality, we adopt a disentangled framework consisting of two subnetworks to cope with landmark deformation and portrait stylization, respectively. Different from CariGANs [27] that adopts PCA representations for landmark deformation and individually train the two sub-networks without considering the mutual interactions between them, we explicitly deform the landmarks with the first subnetwork and bridge the training of the two sub-networks via a differentiable spline interpolation operation [42] to improve the synthesis quality. Specifically, our Landmark Deformation Network (LDN) learns a bilateral mapping of facial landmarks between the portraits and the anime portraits with a pair of generator networks. With the LDN, we can warp the portrait images to an intermediate state where these warped images are of anime-like face shapes. After that, the second subnetwork, Portrait Stylization Network (PSN) focuses on bi-directional texture editing. The PSN uses another pair of generator networks to learn the style translation between the warped portraits and anime portraits. Finally, the LDN and PSN models are stacked together to build up our entire framework. During the training, the two sub-networks are optimized in an end-to-end and mutual boosting way, which effectively simplifies the overall learning difficulty and leads to more visually plausible synthesized results.

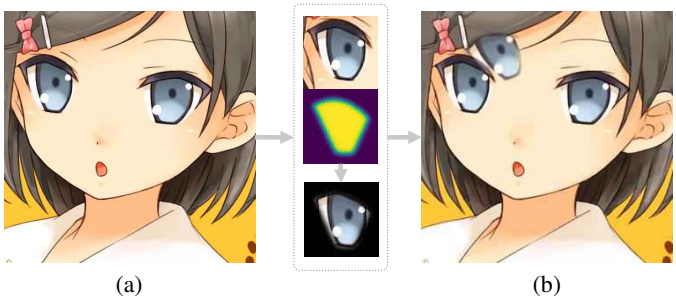


Fig. 5: Illustration of the process of abnormal face structure synthesis. We first crop the eye region based on the facial landmarks, and then randomly paste the cropped eye patch into the face region of the orginal anime portrait image (a) to produce an abnormal face structure sample (b).

Landmark Deformation Network. The Landmark Deformation Network (LDN) comprises an anime portrait landmark generator $G_{p\to a}^{LDN}$ that learns to translate the portrait landmark l_p to the anime portrait landmark $\hat{l}_a$, and a portrait landmark generator $G_{a\to p}^{LDN}$ performing the reverse translation. Here, the LDN only accounts for modeling the deformation of face shape with a 17-landmark representation since we empirically find that the patterns of other facial features are relatively homogeneous (*e.g.*, large eyes, tiny noses, and small mouths) across different anime portraits, whose deformation can be well learned by the PSN. Furthermore, involving more landmarks in LDN may inject additional learning ambiguity that leads to inappropriate deformations and therefore degrades the overall translation quality (refer to Sec. 4.5). We design the two generators of the LDN to share the same 12-layer multi-layer perceptron (MLP) structure and reinforce cycle consistency [1] during model learning. Moreover, two discriminators D_a^{LDN} and D_p^{LDN} are introduced to provide supervision signals to ensure the realism of the generated anime portrait landmarks and portrait landmarks, respectively. Here, both the two discriminators are of the same 7-layer MLP structure. With the LDN model learned, given an arbitrary input portrait I_p, we can warp it pixel-wisely with the correspondence between the portrait landmarks l_p and the predicted anime portrait landmarks $\hat{l}_a$ via a differentiable spline interpolation operation [42], yielding the warped portrait I_w. The warped portrait I_w possesses delicate anime-like face shape so that it is ready for the subsequent anime stylization stage to perform texture translation. We define two losses for optimizing the LDN on the coser and the human portrait datasets, *i.e.*, the cycle-consistency loss and adversarial loss.

Specifically, borrowing the idea from CycleGAN [1], we adopt the cycle-consistency loss to guarantee that the deformed facial landmarks of a portrait can be reconstructed to the original facial landmarks. This can effectively regularize the translation between two facial landmark domains and preserve the content of the input. The loss is defined as

$$\begin{aligned}\mathcal{L}_{cyc,p\to a}^{LDN} &= \mathbb{E}_{l_p\sim L_p}\left\|G_{a\to p}^{LDN}\left(G_{p\to a}^{LDN}\left(l_p\right)\right)-l_p\right\|_1,\\ \mathcal{L}_{cyc,a\to p}^{LDN} &= \mathbb{E}_{l_a\sim L_a}\left\|G_{p\to a}^{LDN}\left(G_{a\to p}^{LDN}\left(l_a\right)\right)-l_a\right\|_1,\end{aligned} \tag{1}$$

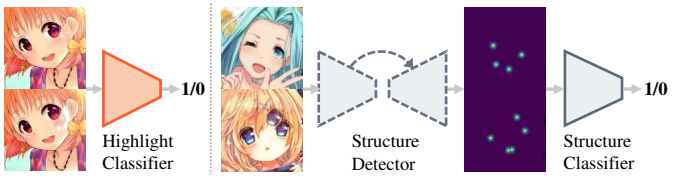


Fig. 6: The training procedures of highlight (a) and face structure (b) classifiers. The highlight classifier is trained with ground truth anime portraits (the upper part of the inputs) and synthesized anime portraits with highlight artifacts (the lower part of the inputs). The face structure classifier is trained with the corresponding face structure maps predicted by the pretrained face structure detector, given legit faces (the upper part of the inputs) and wrong faces (the lower part of the inputs) as inputs.

where l_p and l_a denote the facial landmarks of portraits and anime portrait images, respectively. And $\|\cdot\|_1$ denotes the L_1 norm.

The adversarial loss ensures the realism of the results produced by the generators by penalizing the distance between the distribution of the real and the fake data. Here, the Least Square GAN [43] is adopted for better training stability, which is defined as

$$\begin{aligned}\mathcal{L}_{adv,p\to a}^{LDN} &= \mathbb{E}_{l_a\sim L_a}\left[\left(D_a^{LDN}\left(l_a\right)\right)^2\right] \\ &\quad + \mathbb{E}_{l_p\sim L_p}\left[\left(D_a^{LDN}\left(G_{p\to a}^{LDN}\left(l_p\right)\right)-1\right)^2\right], \\ \mathcal{L}_{adv,a\to p}^{LDN} &= \mathbb{E}_{l_p\sim L_p}\left[\left(D_p^{LDN}\left(l_p\right)\right)^2\right] \\ &\quad + \mathbb{E}_{l_a\sim L_a}\left[\left(D_p^{LDN}\left(G_{a\to p}^{LDN}\left(l_a\right)\right)-1\right)^2\right].\end{aligned} \tag{2}$$

The total loss for training the LDN on the coser and the human portrait datasets is the weighted sum of the cycle-consistency loss and the adversarial loss:

$$\begin{aligned}\mathcal{L}_{LDN} &= \lambda_{cyc}^{LDN}\left(\mathcal{L}_{cyc,p\to a}^{LDN}+\mathcal{L}_{cyc,a\to p}^{LDN}\right) \\ &\quad + \lambda_{adv}^{LDN}\left(\mathcal{L}_{adv,p\to a}^{LDN}+\mathcal{L}_{adv,a\to p}^{LDN}\right),\end{aligned} \tag{3}$$

where λ_{cyc}^{LDN} and λ_{adv}^{LDN} are the weighting parameters to balance different loss items.

Portrait Stylization Network. The PSN also consists of two generators. The anime portrait generator $G_{p\to a}^{PSN}$ endeavors to produce an anime portrait conditioned on an input portrait, while the portrait generator $G_{a\to p}^{PSN}$ translates an anime portrait to its corresponding portrait counterpart. The two generators are of the same encoder-decoder structure which comprises two downsampling layers, twelve residual blocks, and two upsampling layers. In addition, two discriminators with five downsampling layers D_a^{PSN} and D_p^{PSN} are used for encouraging the visual realism of the generated images of each domain. To further enhance the visual quality of the generated results, we incorporate the domain attention module [3] on top of the sixth residual block in the generators and the last downscaling layer of the discriminators. Adaptive Layer-Instance Normalization [3] is also adopted in the residual bottlenecks for the generators.

For training the PSN on the coser portrait dataset, we apply five losses, including the cycle-consistency loss, adversarial loss, CAM loss, face structure loss, and highlight loss. Here, the cycle-consistency loss $\mathcal{L}_{cyc}^{PSN}$ and the adversarial loss $\mathcal{L}_{adv}^{PSN}$ are similarily defined as the losses in LDN to accomplish the unsupervised learning. Different from that the losses for LDN are applied to the facial landmarks, the losses for the PSN are applied to the generated images. In particular, the cycle-consistency loss $\mathcal{L}_{cyc}^{PSN}$ aims to regularize the image translation between the portrait and the anime portrait domains. And the adversarial loss $\mathcal{L}_{adv}^{PSN}$ encourages the visual realism of the generated images. Following [3], we apply the CAM loss $\mathcal{L}_{cam}$ which leads to intensively changing discriminative image regions by distinguishing two domains using the CAM approach [44], and this enables us to fully exploit the discriminative information between the source and the target domains to improve visual realism of the synthesized images.

The face structure loss is applied to suppress the distortions and ensure a reasonable structure of the generated faces. This loss estimates the likelihood of a face being distorted or with unreasonable facial features and penalizes those scenarios. For example, the loss will penalize the case that the eyes or the mouth is unreasonably placed, or the synthesis creates facial features that should not exist at all. To do so, we first pretrain a U-Net based face structure detector to predict the locations of eyes and corners of mouth from an anime-like face images and output a detection heatmap. Here, we prepare a set of normal anime portraits and synthesized abnormal anime portraits with their eyes and mouths manually labeled as the ground-truth labels for training the face structure detector. Then we train a binary anime face structure classifier to distinguish whether the heatmap is legit or not. The classifier network comprises 5 convolutional blocks followed by a global average pooling layer and two fully-connected layers. It is trained prior to the PSN model with the supervision of both legit faces and synthetic distorted and wrong faces and their ground truth labels. We show the process of abnormal face structure synthesis and the training procedure of structure classifier in Fig. 5 and Fig. 6(b) respectively. When training the PSN, we apply the pre-trained face structure classifier on the face structure maps predicted from the generated anime portraits. By forcing the corresponding face structure maps to be classified as the positive sample, the PSN has to generate results with correct face structures. Therefore, we define the face structure loss as

$$\mathcal{L}_{structure} = -\log\left(\text{sigmoid}\left(\hat{y}_{structure}\right)\right), \tag{4}$$

where $\hat{y}_{structure}$ represents the predicted probability of the face structure classifier given the face structure heatmap estimated from the generated anime portrait as input.

We observe that the PSN network tends to render unnatural highlight artifacts on the facial regions, which are inherited from the highlight patterns in the input human portrait images. To mitigate the highlight artifacts and encourage natural facial skin, we apply the highlight loss by penalizing those highlight artifacts through a pre-trained binary highlight classifier. Similar with the face structure classifier, the highlight classifier network also comprises 5 convolutional blocks followed by a global average pooling

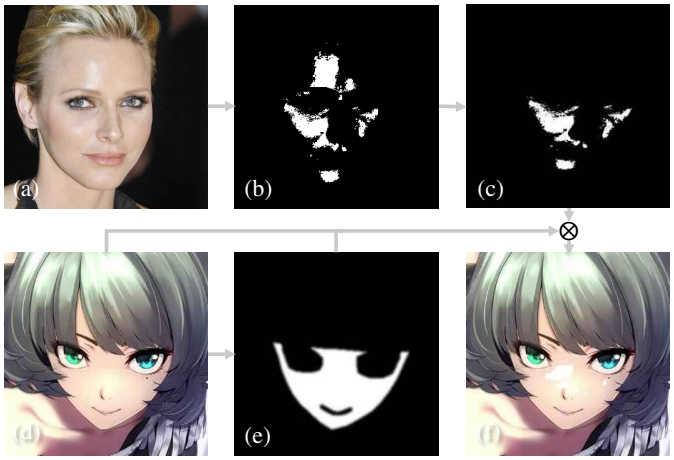


Fig. 7: The process of highlight synthesis. Given a human portrait image (a), we first extract the highlight regions (b) of the portrait and warp the highlight regions to match an arbitrary anime face (d). With the highlight regions warped as (c), we blend (c) and (d) together, with the awareness to blend within the anime facial mask (e) only, to form the final anime portrait with highlights (f).

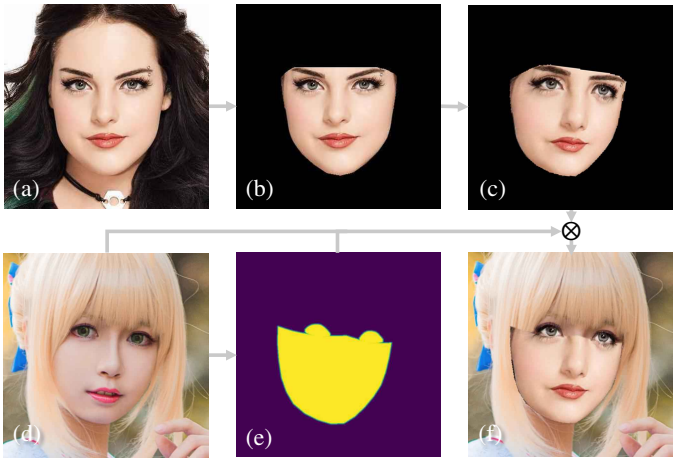


Fig. 8: The synthesis pipeline of the Aug-Portrait dataset. We first detect the facial landmarks of the human portrait (a) and the coser portrait (d) images by using the dlib [45]. Then the human portrait face is cropped and warped into the shape of the coser portrait face based on facial landmarks of the two faces via a spline interpolation operation. Finally, the warped human portrait face (c) is pasted to the coser portrait face according to the coser portrait facial mask (e) to produce an augmented portrait (f). Here, (e) is obtained via curve fitting with facial landmarks of (d).

layer and two fully-connected layers. To train the classifier, we prepare a dataset containing both real anime portraits and synthetic anime portraits with highlight artifacts. The training procedure of the highlight classifier is shown in Fig. 6(a). To imitate these facial highlight artifacts, we first use the level tool in Adobe Photoshop to obtain the highlight regions in a human portrait and warp the regions to a randomly chosen anime face via triangulation affine transformation. Note that here the highlights on the forehead region are excluded before warping. Then we blend the warped facial highlights to the anime face based on the facial mask of the anime portrait via a pixel-wise maximum operation to create the final synthetic anime portraits with highlight artifacts. The synthesis process is shown in Fig. 7. Similar with the face structure loss, we also apply the pre-trained highlight classifier to enforce the generated anime portraits to be classified as positive sample during the training, which can urge the PSN to generate results with no highlights. The highlight loss is defined as

$$\mathcal{L}_{highlight} = -\log\left(\text{sigmoid}\left(\hat{y}_{highlight}\right)\right), \tag{5}$$

where $\hat{y}_{highlight}$ represents the predicted probability of the highlight classifier given the generated anime portrait as input.

Consequently, the total loss for training the PSN on the coser portrait dataset is the weighted sum of all the above-mentioned five losses:

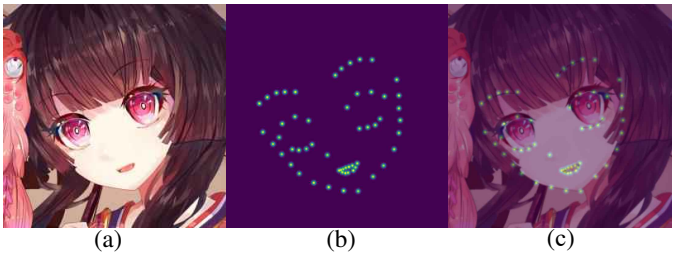
$$\begin{aligned}\mathcal{L}_{PSN} = &\lambda_{cyc}^{PSN}\left(\mathcal{L}_{cyc,p\to a}^{PSN} + \mathcal{L}_{cyc,a\to p}^{PSN}\right) \\ &+ \lambda_{adv}^{PSN}\left(\mathcal{L}_{adv,p\to a}^{PSN} + \mathcal{L}_{adv,a\to p}^{PSN}\right) + \lambda_{cam}\mathcal{L}_{cam} \\ &+ \lambda_{structure}\mathcal{L}_{structure} + \lambda_{highlight}\mathcal{L}_{highlight},\end{aligned} \tag{6}$$

where λ_{cyc}^{PSN}, λ_{adv}^{PSN}, λ_{cam}, $\lambda_{structure}$, and $\lambda_{highlight}$ are the weighting parameters to balance different loss items.

3.2 Domain Adaptation to Aug-Portrait

Since there exists a significant difference between the distributions of the coser portrait and the human portrait domains, directly adapting the initial model trained with coser portrait to the human portrait domain may lead to severe learning complexity, which results in inferior adaptation results.

To address this issue, we propose to first adapt the model to an intermediate domain closer to the human portrait domain, so that the adapted model can be subsequently transferred to the human portrait domain more easily. We construct this intermediate domain by synthesizing portraits through fusing the coser portrait face shapes and the human facial textures together to form a new Aug-Portrait dataset. In particular, as shown in Fig. 8, we adopt a face warping strategy to warp the face of a human portrait to the shape of a coser portrait face image. More specifically, we first detect the facial landmarks of the human portrait and the coser portrait images by using the dlib face detector [45]. Then the human portrait face is cropped and warped into the shape of the coser portrait face based on the facial landmarks of two faces via a spline interpolation operation [42]. Finally, the warped human portrait face is pasted to the coser portrait face according to the coser portrait facial mask to produce a fused result. The Aug-Portrait dataset features coser-like styles, *e.g.*, unrealistic hairstyles, but still exhibits a wider variety of facial textures from the human portraits, which not only bridges the gap between the coser portrait and human portrait domains, but also greatly improves the diversity and richness of trainining data, leading to better generalization capability of our model. To achieve the adaptation, we train the initial model on a combined training data of the coser portrait and the Aug-Portrait



(a) (b) (c)

Fig. 9: The manually labeled 56-point facial landmarks of the anime portrait. 17, 10, 16, 1, and 12 points are used for marking the shape of the face, eyebrows, eyes, nose, and mouth, respectively. (a) The anime portrait. (b) Facial landmarks of (a). (c) The anime portrait overlaid with landmarks.

datasets. This mixing training strategy enables the model to learn more generalizable representation that is closer to the human portrait domain while maintaining the appearance preservation capability. During the training on Aug-Portrait, we applied the same optimization objectives (Eq. 6 and Eq. 3) as used for the training on the coser portraits.

3.3 Domain Adaptation to Human Portrait

Once the adaptation to the Aug-Portrait domain is performed, our model can well cope with diverse facial patterns from the human portraits. However, the complex hair styles and colors in human portraits are still challenging since they are not involved in our training. To further generalize our model to the exact domain of human portraits, we incorporate the human portrait dataset into our mixing training data for the third-stage of training. Here, we apply the same optimization objectives of Eq. 6 and Eq. 3 on both the coser portraits and the aug-portraits. For the training on the human portraits, we only adopt the face structure loss (Eq. 4) and highlight loss (Eq. 5). The reason behind this is that performing adversarial training on the human portraits may push the generated results of human portraits to be as close as possible to the real anime portrait domain. For example, the model may tend to synthesize anime portraits with thick bangs even if such characteristic does not appear in the input human portraits. This can heavily impair the appearance preservation performance of the model on human portraits.

4 Experiments

In this section, we conduct extensive experiments to demonstrate the superiority of our method over the existing methods, especially for appearance preservation during translation. Both quantitative and qualitative evaluations are involved and in-depth ablation studies are also conducted to verify the effectiveness of our main contributions.

4.1 Datasets

To validate the effectiveness of our method, we conduct our evaluations on two unpaired portrait-to-anime datasets: CelebA-HQ2Anime and Selfie2Cartoon [3]. Specifically, the CelebA-HQ2Anime datasets contains two subsets of the CelebA-HQ and the Anime datasets for the human portrait and anime portrait domains respectively, while the Selfie2Cartoon dataset [3] contains the Selfie and the Cartoon datasets. To facilitate the model learning and elevate the generation quality on the human portraits, we collect a coser portrait dataset as the proxy to guide the domain adaptation learning. Moreover, the Aug-Portrait dataset is also synthesized by fusing the coser and the human portraits. Note that the Anime portrait, Cartoon, coser portrait, and Aug-Portrait datasets are used only for training, while the CelebA-HQ, Selfie datasets are used as the human portraits for both training and testing.

CelebA-HQ contains 30,000 human portraits from CelebA [46], among which, we randomly select 1400 images for training and other 1380 images for testing.

Anime contains 1400 anime portraits randomly selected from the Danbooru2020 dataset [47]. We manually label 56 landmarks for each anime portrait. In particular, 17, 10, 16, 1, and 12 landmarks are used for marking the shape of the face, eyebrows, eyes, nose, and mouth, respectively. Here, the 17 landmarks of the face shapes are used as the anime landmark domain data for training the LDN. 16 landmarks of eyes are used to extract the eye region for abnormal face structure synthesis, and the centers of each eye and corners of the mouth are used to train the anime face structure classifier for computing $\mathcal{L}_{structure}$. An example of anime portrait landmarks is shown in Fig. 9.

Selfie [3] contains 3500 selfies with diverse portrait styles from real-world scenarios. We conduct evaluations on Selfie to verify the generalization capability of our method on more realistic scenarios. We randomly choose 2276 and 1024 images for training and testing, respectively.

Cartoon [3] contains 3500 celluloid cartoon images. Similar with the Anime dataset, we label each image with 56 landmarks. We randomly select 2476 images for training.

Coser Portrait contains 1400 high-resolution Asian coser portraits collected from the Internet, which can be generally categorized into two styles (*i.e.*, the headdress-free and the headdress). Following the setting of the previous work [3], all the collected samples are female. As a preprocessing procedure, each raw image is first fed to a pretrained anime face detector [48] for detecting and cropping the head portraits. We show a gallery of Coser Portrait in the second row of Fig. 2.

Aug-Portrait is synthesized on the fly during the training process, whose scale increases linearly to the training iterations. Here, we perform the synthesis with a probability of 0.5 in each iteration. To synthesize the augmented portraits, we first randomly sample a batch of human portrait images (*e.g.*, CelebA-HQ or Selfie) and coser portrait images, and then warp the face regions of the human portraits to the coser portraits. Here, the face region is extracted based on the detected facial landmarks. It is worth noting that since both the human and the coser portraits in our training set are mostly near-frontal with few occlusions, the pose discrepancy between the human portraits and the coser portraits can be well addressed by the landmark alignment and a warping operation, which makes the warped results generally reasonable for learning a robust model. Additionally, although artifacts inevitably exist in the fusion boundary of the synthesized images, the model can still

TABLE 1: Quantitative comparison with existing methods.

Method	LPIPS ↓		Hue-HistD ↓	
	CelebA-HQ	Selfie	CelebA-HQ	Selfie
UNIT [2]	0.5670	0.6780	0.4535	0.5172
MUNIT [19]	0.5597	0.7029	0.4902	0.5579
CycleGAN [1]	0.5101	0.5303	0.3335	0.4024
U-GAT-IT [3]	0.5487	0.6179	0.4017	0.5352
Toonify [33]	0.4229	0.6860	0.4058	0.4748
Cartoon-StyleGAN [35]	0.6174	0.6822	0.4206	0.4658
UI2I-via-StyleGAN2 [34]	0.6199	0.6789	0.3671	0.4772
Ours	**0.4079**	**0.5137**	**0.1760**	**0.3889**

produce visually pleasant anime portraits via adversarial learning to match the real artifact-free anime domain. The synthesis process is shown in Fig. 8.

In our experiments, all the facial landmarks are detected by the dlib [45] except for the Anime and the Cartoon datasets and all the images are resized to 256×256. We showcase the examples of the anime portrait, coser portrait, and human portrait in Fig. 2.

4.2 Implementation Details

Our network is implemented on the PyTorch framework [49] on a NVIDIA RTX 3090 GPU with 24 GB memory. We train the full model in an end-to-end manner in all three stages of our training except for the first-stage training on the coser portraits, where the LDN and the PSN are first trained individually to ensure a good initialization. Specifically, for each stage, the model is trained for 500k, 250k, and 250k iterations, respectively. The λ_{cyc}^{LDN} and λ_{adv}^{LDN} are empirically set as 100 and 1 for the LDN, and the λ_{cyc}^{PSN}, λ_{adv}^{PSN}, λ_{cam}, $\lambda_{structure}$, and $\lambda_{highlight}$ are set as 10, 1, 1000, 1, and 1 for the PSN. Furthermore, we use the Adam optimizer [50] with a batch size of 1. During the first stage of training, the learning rates of both LDN and PSN are fixed to 1e-4. For the following two stages, the learning rates for training the LDN and the PSN are first initialized to 1e-5 and 1e-4, respectively. Then the learning rates of both the networks are linearly decayed to zero as the training proceeds.

4.3 Metrics

For portrait-to-anime translation, it is crucial to preserve the most recognizable appearances of the input portrait [34], whereas the existing methods suffer from unpredictable appearance changes (*e.g.*, the hairstyle, the skin and hair color, *etc*). On the other hand, previous works usually adopt the Frechet Inception Distance (FID) [51] to evaluate the generation quality. However, we argue that FID is unsuitable for indicating the quality of appearance-preserved portrait-to-anime translation for two reasons: 1) FID simply measures the similarity between the overall distributions of the generated and real anime portraits without considering the appearance consistency between the input human portrait and its generated anime portrait; 2) the generated appearance-preserved anime portrait shares consistent semantics with the input human portrait, which has significantly different distributions from those in the real anime portraits. We give the corresponding examples and discussions to elaborate on this problem in the supplementary material due to limited space in the main paper. To provide a more comprehensive assessment of the appearance-preserved portrait-to-anime translation, in this paper, we identify three key aspects that need to be taken into account. In particular, we consider the cartoonity, structure consistency, and color consistency of the generated anime portraits.

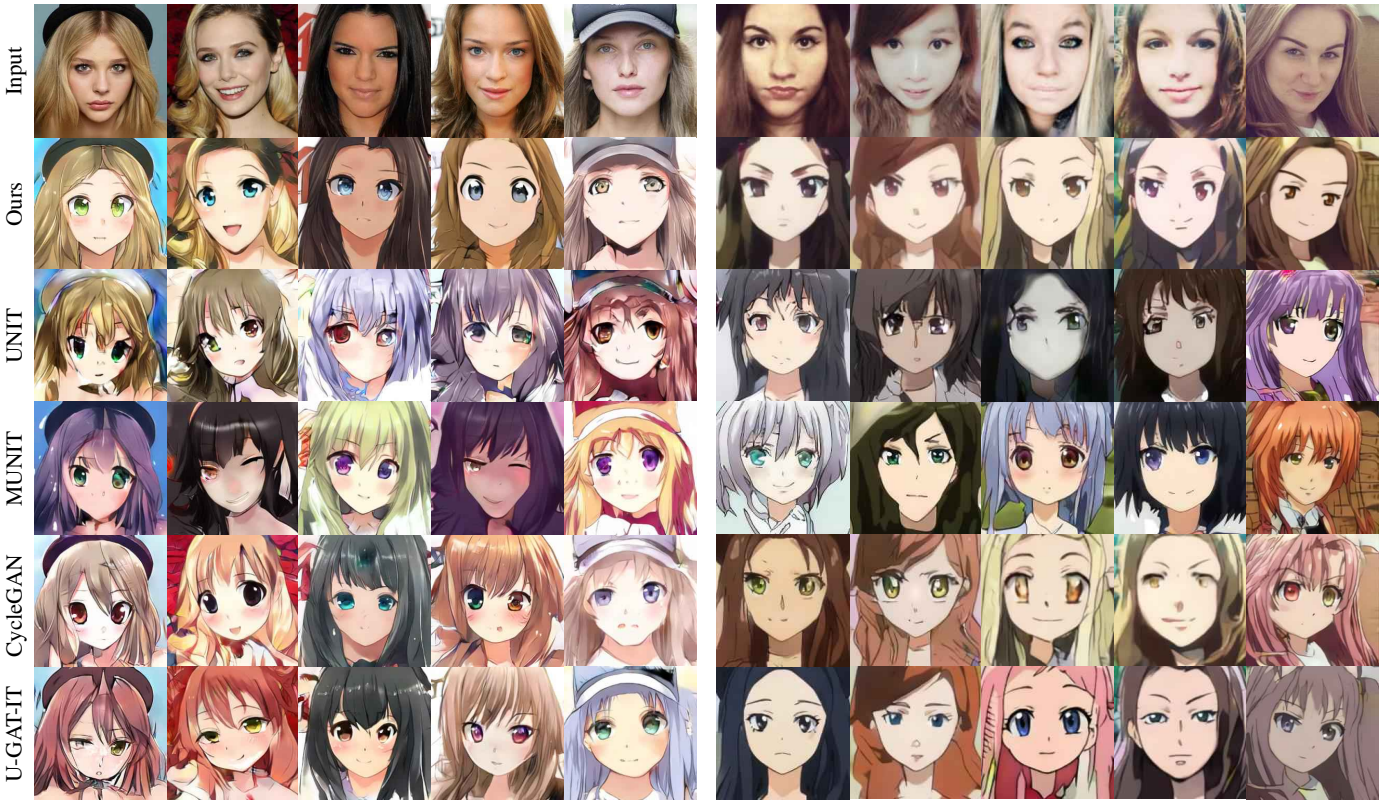
• **Cartoonity**: The generated anime portraits should possess visually plausible anime appearances and styles (*e.g.*, unrealistic hairstyles, smooth facial textures, big and sparkling eyes, abstract nose and mouth, stylized face shape, *etc*) that are as close as possible to real anime portraits. As mentioned above, the commonly used FID is not feasible to reflect the generation quality of the appearance-preserved anime portraits. We therefore conduct a user study to evaluate the cartoonity of the generated anime portraits.

• **Structure Consistency**: The overall structures of each semantic region (*e.g.*, the hairstyles) in the generated anime portrait should be consistent with the input human portrait. We therefore propose to measure the semantic similarity between the input portrait and the translated anime portrait to evaluate the structure preservation performance of the generation. Following Kwong *et al.* [34], we compute the LPIPS [52] distance between the input human portrait and its translated anime portrait. A lower LPIPS distance indicates that the structures of semantic regions of the generated anime portrait are more consistent with the input.

• **Color Consistency**: The color of the key semantics (*e.g.*, hair and skin color) in the generated anime portraits should be consistent with the input human portrait. To evaluate the color consistency, we measure the color distribution similarity in HSV space between the input portrait and the translated anime portrait. Specifically, we compute the cosine distance between the Hue histogram vectors (Hue-HistD) of the input and output portraits.

4.4 Comparison with the Existing Methods

We compare with four unsupervised I2I methods, including CycleGAN [1], UNIT [2], MUNIT [19], and U-GAT-IT [3], and three StyleGAN-based methods, including Toonify [33], Cartoon-StyleGAN [35], and UI2I-via-StyleGAN2 [34] both quantitatively and qualitatively. For the evaluation on CelebA-HQ2Anime, we progressively train our full model on {*Coser Portrait+Aug-Portrait+CelebA-HQ*, *Anime*}, and all the competitors are trained or fine-tuned on {*CelebA-HQ*, *Anime*}. We evaluate all the methods on the same testing set of *CelebA-HQ*. Similarly, for the evaluation on Selfie2Cartoon, we progressively train our full model on {*Coser Portrait+Aug-Portrait+Selfie*, *Cartoon*}, and all the competitors are trained or fine-tuned on {*Selfie*, *Cartoon*}. All the methods are evaluated on the same testing set of *Selfie*. In addition, we also conduct comparison experiments on CelebA-HQ2Anime by training the compared methods with both the CelebA-HQ and our proposed coser portrait datasets. This serves the purpose to demonstrate our proxy-guided domain adaptation learning scheme can indeed effectively utilize the coser portraits to assist the appearance-preserved anime portraits generation compared to the existing works. The corresponding results and discussions are provided in the supplemental materials due to the limited space of the main paper.



(a) Results on CelebA-HQ Portraits (b) Results on Selfie Portraits

Fig. 10: Qualitative comparison with the UI2I methods.

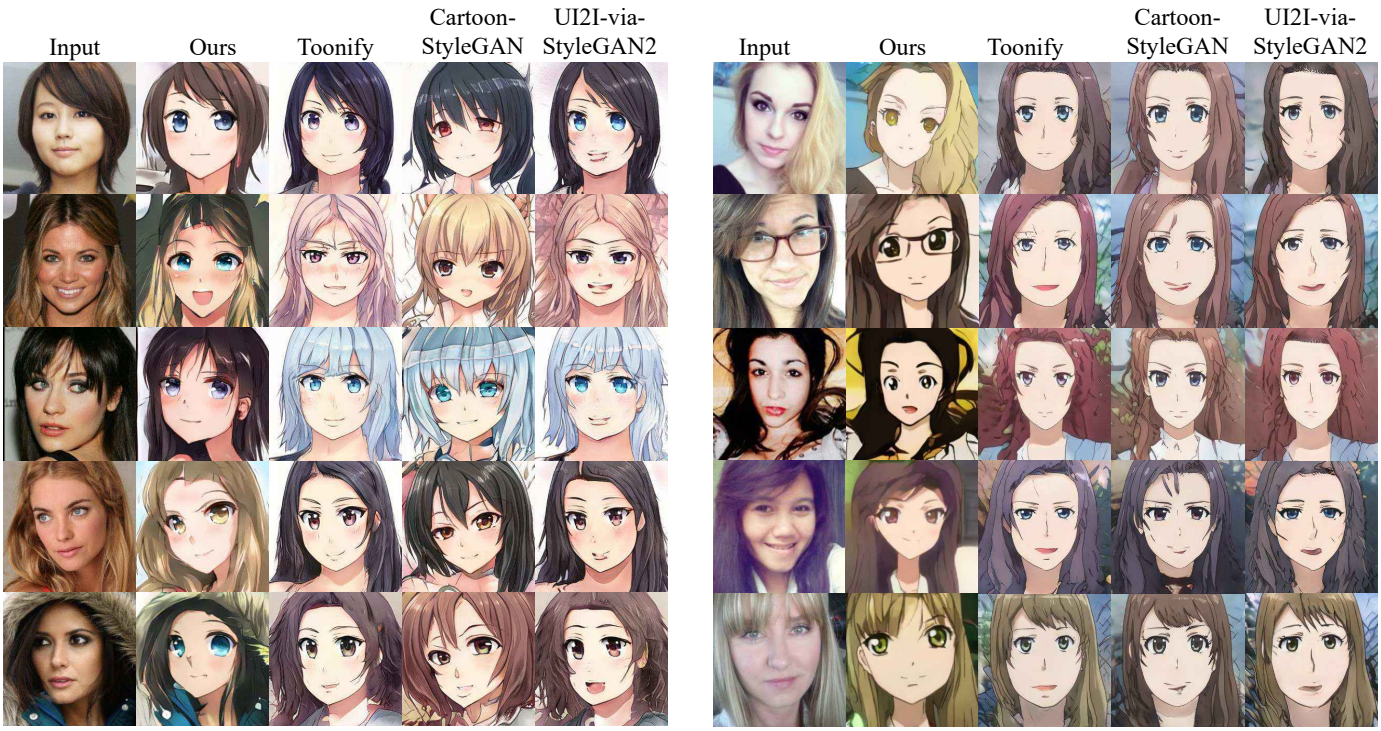
4.4.1 Quantitative Evaluation

We first compare our method with the existing works in terms of the structure consistency and color consistency quantitatively. Here, all the evaluations are conducted on the entire testing sets. The results are shown in Table. 1. We can observe that our method obtains the lowest LPIPS distance, demonstrating that our method can better preserve discriminative semantics of the translated portraits. Furthermore, our method achieves the lowest Hue-HistD score, which indicates that our method can achieve better color consistency compared with other competitors. Considering both the LPIPS and the Hue-HistD scores, our method has a superior appearance preservation capability over the other methods during portrait-to-anime translation.

4.4.2 Qualitative Evaluation

Apart from the quantitative analysis, we also evaluate the visual quality of our synthesized images. The qualitative comparisons between our approach and the other competitors are shown in Figs. 10 and 11. Notably, our method can produce more visual-pleasing and appearance-preserved anime portraits on both the CelebA-HQ and Selfie datasets. This mainly attributes to the involvement of the coser portrait dataset and our proxy-guided domain adaptation learning scheme, which effectively encourage the model to learn the correct translation patterns and improve the appearance preservation capability. In contrast, all the compared UI2I methods suffer from appearance changes due to the learning ambiguity caused by the huge gap between the two domains, and undesired face distortions also appear in the generation results. For example, UNIT suffers from noticeable facial features distortions (*e.g.*, the third row in Figs. 10(a) and (b)), and CycleGAN is prone to generate unnatural facial patterns (*e.g.*, inconsistent eye color in the fourth example of the fifth row in Fig. 10(a) and the last example of the fifth row in Fig. 10(b)). MUNIT tends to produce artifacts like messy textures on CelebA-HQ (the fourth row in Fig. 10(a)). Although U-GAT-IT can produce reasonable results in some sense, all the compared UI2I methods fail to preserve the appearance of the input portrait. In particular, look at the hair region and we can see that all the four UI2I methods tend to generate anime portraits with thick bangs that are not present in the input portraits, and the hair color also changes dramatically. Likewise, as shown in Fig. 11, despite the remarkable visual quality of the results generated by the StyleGAN-based methods, they also suffer from severe appearance changes. In particular, all the three StyleGAN-based methods undergo significant hair color or shape changes, making the generated portraits less recognizable. Furthermore, they also fail to preserve some special patterns due to the inherent data bias of the pre-trained StyleGAN generator (*e.g.*, the hat in the fifth row in Fig. 11(a), the glasses and the complex hair styles in the second and third rows in Fig. 11(b)). We provide more qualitative results on the CelebA-HQ and the Selfie datasets in the supplementary materials.

To demonstrate the generalization capability of our model in realistic scenarios, we also test our model on the in-the-wild portraits collected from the Internet. The qualitative results are shown in Fig. 12, it can be observed that our



(a) Results on CelebA-HQ Portraits (b) Results on Selfie Portraits

Fig. 11: Qualitative comparison with the StyleGAN-based UI2I methods.

TABLE 2: The average ranking scores (the smaller, the better) on perceptual cartoonity and the preference percentage of users (the higher, the better) on appearance preservation of different methods.

Method	Perceptual Cartoonity ↓		Appearance Preservation ↑	
	CelebA-HQ	Selfie	CelebA-HQ	Selfie
UNIT [2]	7.38	6.09	1.84	1.52
MUNIT [19]	4.56	5.36	2.40	3.48
CycleGAN [1]	4.34	4.76	4.60	7.08
U-GAT-IT [3]	4.20	4.00	4.92	5.04
Toonify [33]	5.39	5.75	3.28	4.40
Cartoon-StyleGAN [35]	2.69	**1.91**	3.40	5.72
UI2I-via-StyleGAN2 [34]	5.26	6.22	2.08	3.08
Ours	**2.18**	1.99	**77.48**	**69.68**

TABLE 3: MMD scores between the landmarks of different datasets. The MMD score between the warped landmarks by the LDN and the landmarks of anime portraits is significantly smaller than that between the landmarks of human portraits and anime portraits, demonstrating the LDN indeed deforms a human portrait face to a more anime-like face shape, which effectively eases the overall translation task and elevates the generation quality.

Metric	Human. *v.s.* Anime	Warped. *v.s.* Anime
MMD	1.4992	0.3459

model can produce compelling anime portraits given input portraits with different expressions, ages, races, genders, *etc*. Moreover, our method can even cope with portraits beyond the human portrait domain (the last example in Fig. 12). This verifies that our model possesses a strong generalization capability for various scenarios, which is due to that our proposed coser portrait guided domain adaptation learning scheme enables the model to learn generalizable representations while maintaining the appearance of the input portraits. We provide more qualitative results of in-the-wild portraits in the supplementary materials.

For a more comprehensive evaluation, we conduct a user study to evaluate the quality of the generated anime portraits in terms of cartoonity and appearance preservation. Specifically, given a human portrait and its translated anime portrait images by different methods, the participants are asked to answer two questions: (1) please rank the generated anime portraits of different methods based on the perceptual cartoonity (*e.g.*, visually plausible anime appearances and styles like unrealistic hairstyles, smooth facial textures, big and sparkling eyes, abstract nose and mouth, stylized face shape, *etc*); and (2) please choose the generated anime portrait with the best appearance consistency (*e.g.*, hair style, hair and skin color, face expression, *etc*) with the input portrait. We randomly choose 50 CelebA-HQ portraits and 50 Selfie portraits from the testing sets for evaluation. Totally, 50 users participate in the study and the statistics are presented in Table. 2. As can be seen, our method surpasses all the competitors by a large margin in terms of appearance preservation on both CelebA-HQ and Selfie datasets. For the perceptual cartoonity, our method also achieve the best ranking score on the CelebA-HQ dataset and obtain comparable ranking score with the best-ranking method, Cartoon-StyleGAN [35], on the Selfie dataset. This demonstrates the superiority of our method in generating visual-pleasing and appearance-preserved anime portraits.

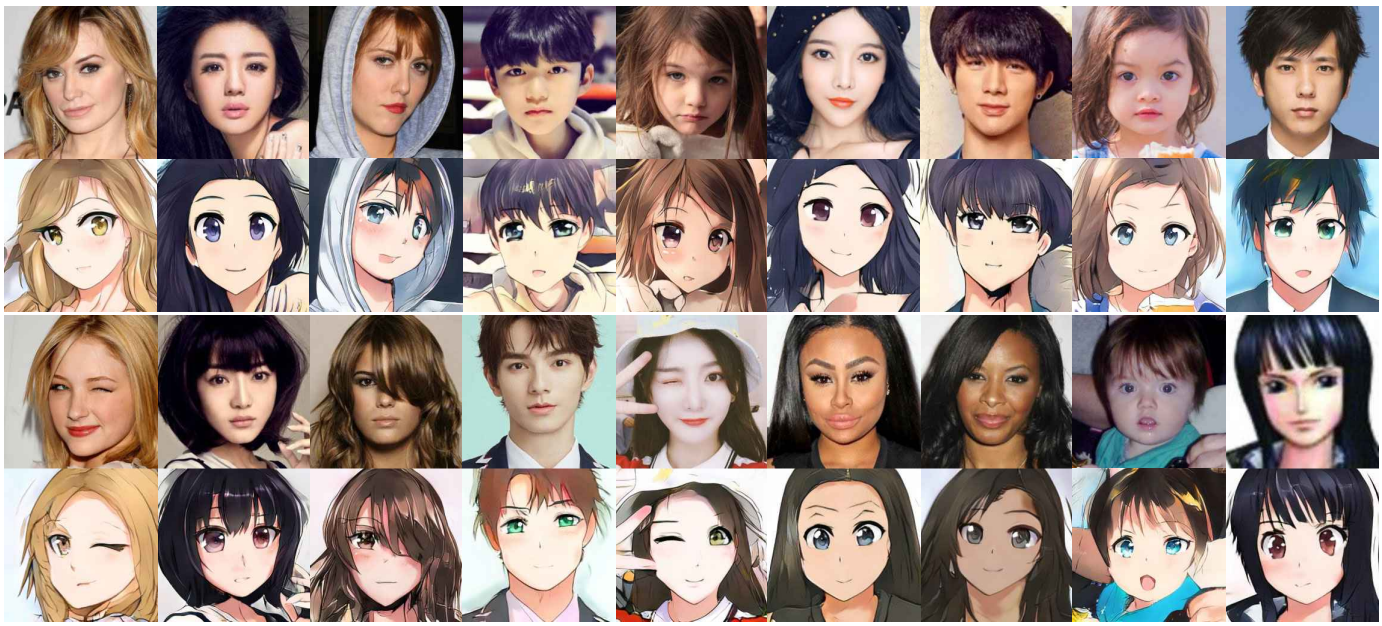


Fig. 12: Visual results of the in-the-wild portraits. Our model can produce compelling anime portraits given input portraits with different expressions, ages, races, genders, *etc*. The first and third rows are the input and the rest are the corresponding output.

4.5 Ablation Study

Here we analyze the efficacy of the key components of our proposed method. Specifically, we investigate the model variants with different training strategies, network architectures, and training objectives. All the ablations are conducted on the CelebA-HQ datasets.

We first investigate the efficacy of our proposed coser portrait dataset and the progressive proxy-guided domain adaptation learning scheme. Specifically, we consider different configurations as follows:

- **Full**: the model trained with the proposed three-stage proxy-guided domain adaptation learning scheme.
- **Baseline**: the model trained with only human portraits.
- **Coser**: the model trained with only coser portraits.
- **Coser**+**Aug**: the model trained with coser portraits and adapted to the Aug-Portrait only.
- **Coser**+**Human**: the model trained with coser portraits and directly adapted to the human portraits.
- **Single Stage**: the model trained with a combination of the coser portraits, the Aug-Portrait, and the human portraits in a single stage only.

As shown in Fig. 13, if we simply adopt the human portraits to train our model (Baseline), significant appearance changes may appear in the generated results. This is because the huge domain gap between the human portrait and the anime portrait domains may incur severe learning ambiguity, which can mislead the model to capture the erroneous translation patterns. By performing the training on the coser portraits (Coser), the generated results can preserve most distinctive appearances of the input portraits. This demonstrates that the coser portrait indeed effectively bridges the domain gaps between the input/output domains and facilitates the correct learning. Performing domain adaptation to the Aug-Portrait (Coser+Aug) can improve the facial patterns as the model is able to learn more diverse and rich facial textures from the human portraits. The generated results may suffer from unnatural facial textures and artifacts when directly adapting the model to the human portrait domain (Coser+Human). The important facial features (*e.g.*, the eyes) cannot be preserved when simply training the model with a mixture of the three portrait datasets in one stage (Single Stage). The reason behind this may be that a single stage of training with three datasets from different domains can lead to increasing learning ambiguity which hinders the network to find the optimal. The results with best visual quality and appearance preservation can be obtained when the domain adaptations to Aug-Portrait and human portraits (Full) are both considered, demonstrating the effectiveness of our full proxy-guided domain adaptation learning scheme.

We also verify the effectiveness of the disentangled framework. Here, we compare our disentangled model (LDN and PSN) with a single mapping model (PSN only). As shown in Fig. 14, a single mapping fails to model large face deformation during translation and may yield unnatural and distorted face contours in generated anime portraits (w/o LDN). Particularly, the face shapes of the samples shown in the second row in Fig. 14 are stiffly inherited from the input portraits without appropriate deformation. These human-like face shapes further result in distorted patterns at the junction of the face and neck in the translated anime portraits. This proves that a single mapping can increase the burden of the network in finding the optimal mapping of portrait-to-anime translation with both face deformation and portrait stylization, and hence incurs unsatisfactory results. Instead, our framework can generate results with more natural face structures and vivid textures by dealing

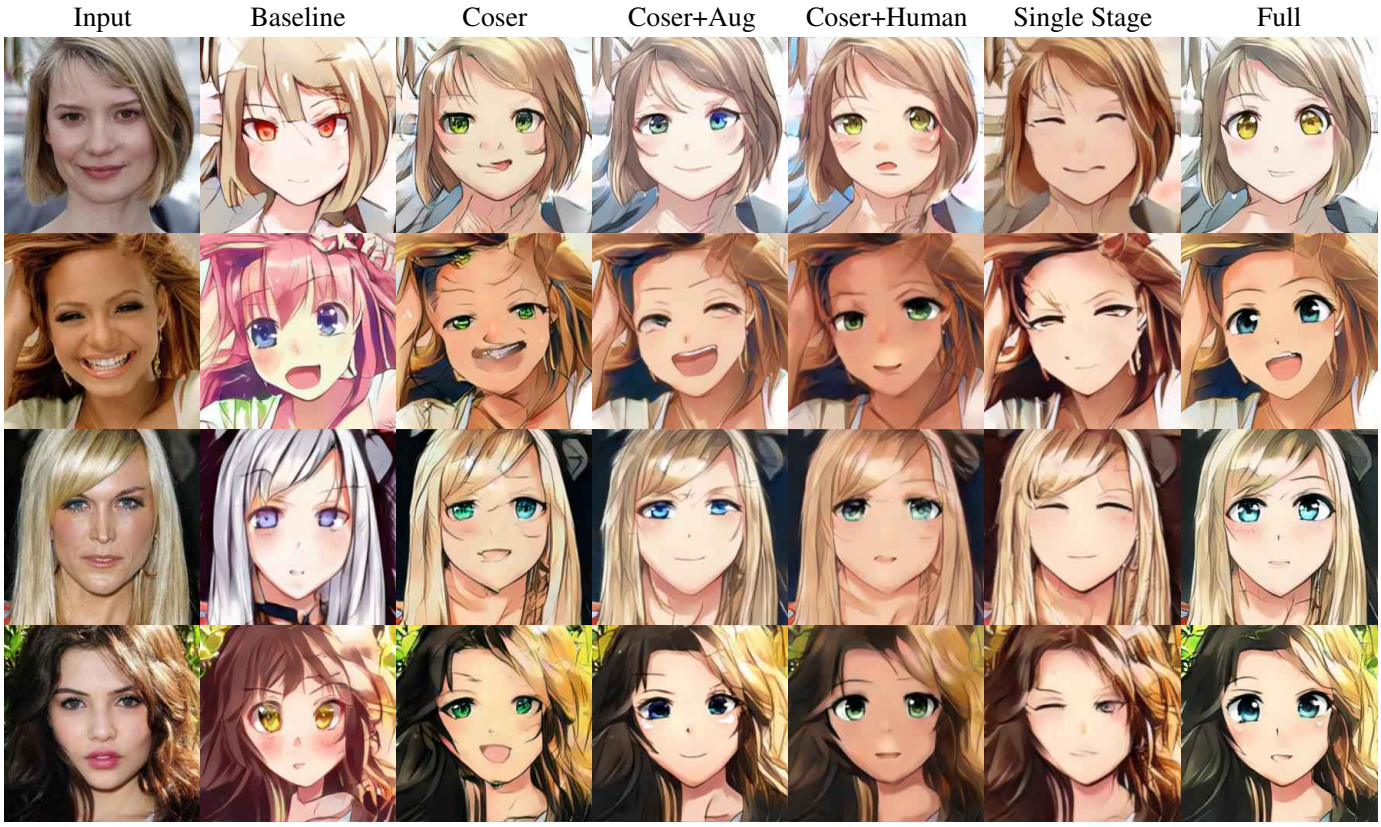


Fig. 13: Visual comparison of variants with different training strategies.

Fig. 14: Visual comparison of variants with different architectures. The model fails to model reasonable face deformation and produces distorted patterns at the junction of the face and neck when replacing the disentangled networks with a single mapping network (the second row).

Fig. 15: Visual comparison of different components involved for landmark deformation in the LDN. The second and the third rows show the results of the whole face deformation (56 landmarks), and the face shape deformation (17 landmarks), respectively.

with the face deformation and portrait stylization in a divide-and-conquer manner. To provide a more intuitive understanding of the role of the LDN, we compared the Maximum Mean Discrepancy (MMD) [53] between the landmarks of human portraits and anime portraits with that between the deformed landmarks by the LDN and the landmarks of anime portraits in Table. 3. We can see that the MMD between the landmarks of anime portraits and the deformed landmarks is much smaller compared to that between the landmarks of anime and human portraits. This indicates that the LDN indeed narrows the face shape gap between the human portrait and anime portrait domains by explicitly transforming the face shape of the human portrait to a more anime-like one, and this effectively facilitates the

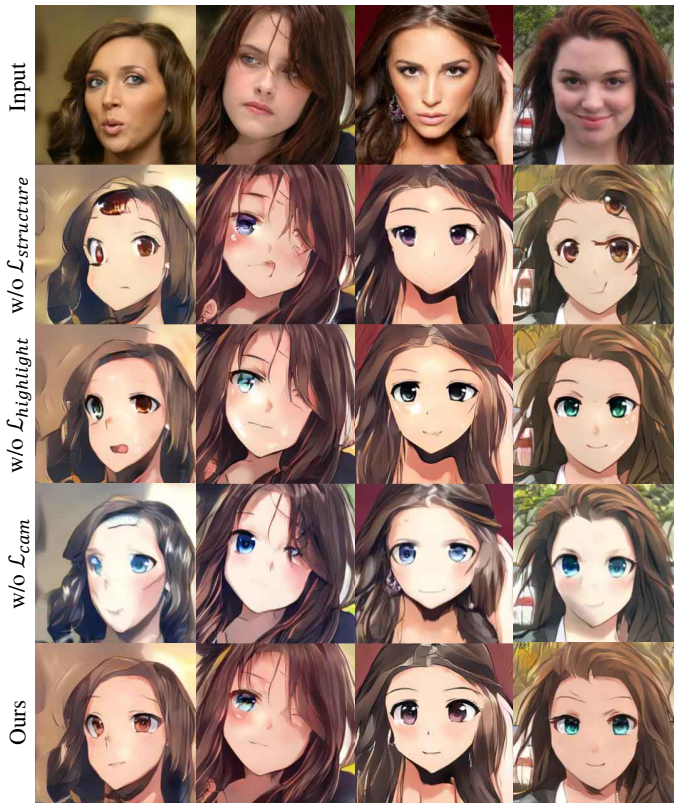


Fig. 16: Visual comparison of variants with different training objectives.

portrait-to-anime translation with large face deformation.

Apart from the LDN, the landmarks used for modeling the face deformation is also a key factor affecting the overall deformation and synthesis quality. Here we conduct an experiment to examine the rationale of our 17-landmark representation for LDN, which accounts for the deformation of the face shapes. Specifically, we replace the 17-landmark representation with 56 landmarks of the whole face for face deformation modeling. The results are shown in Fig. 15. Unfortunately, we can see that the full landmark configuration yields unsatisfactory results with significant distortions or inappropriate patterns in the face regions (the second row in Fig. 15). This may be due to that incorporating more landmarks may inject additional learning ambiguity to the LDN and give rise to inappropriate face deformation. In contrast, the results become far more visually compelling when considering only the explicit deformation of the face shapes with 17 landmarks (the third row in Fig. 15). This is because the patterns of some facial features are relatively homogeneous (*e.g.*, large eyes, tiny noses, and thin mouths) across different anime portraits, whose deformation can be well learned by the PSN. Considering all the above, we choose the 17-landmark representation as our default setting of face deformation modeling for LDN.

To further investigate the effectiveness of different losses, we depict the results of model variants with different losses in Fig. 16. As can be seen, the model tends to generate results with wrong facial structures (*e.g.*, three eyes) when $\mathcal{L}_{structure}$ is removed and noticeable highlight appears in the generated anime portraits when $\mathcal{L}_{highlight}$ is omitted. If discarding the CAM loss, the results suffer from unnatural facial textures and colors. Conversely, with all the above losses considered, our full model not only yields visual compelling translated results, but also considerably maintains the distinctive appearances of the input portraits.

TABLE 4: Quantitative results of ablation study on different variants of our model.

Models	LPIPS↓	Hue-HistD↓
w/o LDN	0.4088	0.1955
Full Landmarks	0.4576	0.2242
w/o $\mathcal{L}_{structure}$	0.4251	0.1858
w/o $\mathcal{L}_{highlight}$	0.4319	0.1851
w/o $\mathcal{L}_{cam}$	0.4112	0.1838
Baseline	0.5380	0.4099
Coser	0.4518	0.2313
Coser+Aug	0.4387	0.1861
Coser+Human	0.4089	0.2184
Single Stage	0.4317	0.1924
Full (Ours)	**0.4079**	**0.1760**

Additionally, we report the quantitative performances of different variants in Table. 4. We can observe that our full model outperforms all the variants in terms of LPIPS and Hue-HistD scores, demonstrating that our model possesses better appearance preservation capability. In particular, when simply adopting the human portraits for training (Baseline), the LPIPS and the Hue-HistD scores increase drastically, as the semantic content cannot be well preserved during the translation. By training on the coser portraits (Coser), the two metrics are improved by a large margin. The proressive domain adaptations to the Aug-Portrait and the human portraits can further boost the quantitative performance of the generated anime portraits. In comparsion with the progressive training, simply performing one-stage training with a mixture of the three datasets can yield inferior performance in the quantitative metrics, demonstrating the necessity and effectiveness of our progressive domain adaptation learning scheme.

On the other hand, a single mapping model (w/o LDN) yields inferior results with higher LPIPS and Hue-HistD scores compared to a disentangled model. Incorporating the full landmarks into the LDN can introduce unstable deformations in the generated results, and this further leads to a significant degradation in both LPIPS and Hue-HistD scores. The adopted three losses also effectively elevate the quality of the generated anime portraits by eliminating abnormal facial structures, highlights, and improving facial textures respectively, which contributes to notable gains in both metrics. In summary, all of our main proposals indeed boost both the visual quality and quantitative performance of the generated anime portraits, demonstrating the superiority of our method in visual-pleasing anime portrait translation with well preserved appearance.

5 Limitations and Discussion

Although our method can generate visually pleasant results with superior appearance consistency, we observe failure cases when the input human portraits present extreme occlusions, illuminations, expressions, poses, or aging patterns (Fig. 17(a)-(e)). This is because such extreme patterns are not involved in our training data. Also, due to the

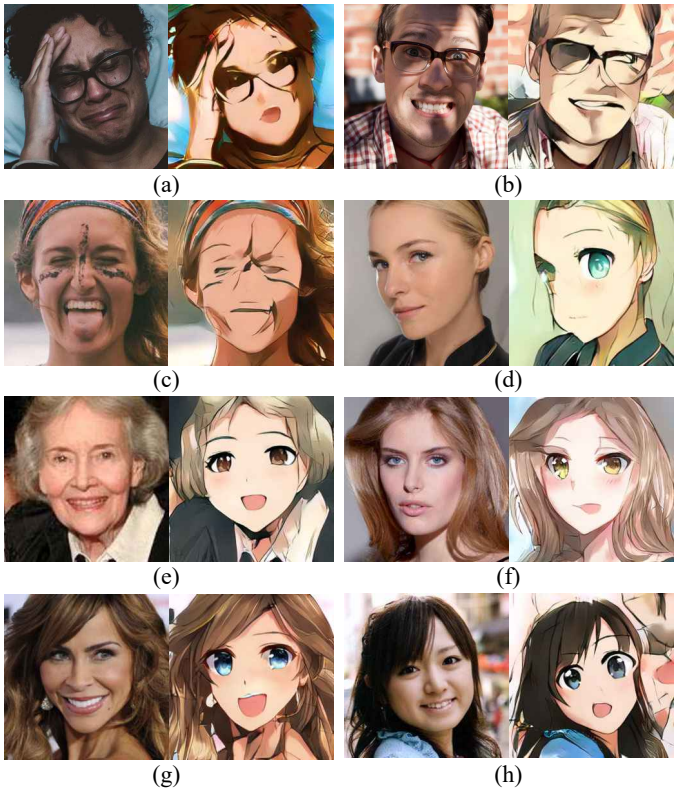


Fig. 17: Failure cases of our model. Although our method can generate appearance-preserved anime portraits, several failure cases are still observed when the input human portraits present extreme occlusions (a), illuminations (b), and expressions (c). Also, facial features (*e.g.*, eyes, nose, *etc*) and face contours under large poses cannot be well translated (d). In addition, albeit our model is capable of dealing with portraits of children (Fig. 12), face aging patterns (*e.g.*, wrinkles, age spots, *etc*) of old people cannot be properly retained (e). Undesired artifacts (*i.e.*, asymmetry eyes, mouth (f), distorted necks (g), and messy backgrounds (h)) may sometimes appear due to the inherent learning ambiguity in unsupervised learning. Samples of (a)-(c) are from the non-photorealistic rendering benchmark NPRportrait 1.0 [54], while the rest are from CelebA-HQ [6].

inherent learning ambiguity arising from the unsupervised learning scheme, the generated anime portraits may sometimes suffer from undesired artifacts (*e.g.*, distorted facial structures or necks and messy background in (Fig. 17(f)-(h)). This may be alleviated by applying explicit geometry constraints during translation. Furthermore, despite the fact that the translation quality between the human and the anime portraits can be greatly elevated with the aid of the coser portraits, it may not be feasible to use coser portraits to bridge domain gaps between human portraits and portraits of other styles (*e.g.*, caricature, sketch, *etc*). Nevertheless, our idea still sheds light on a possible proxy-guided domain adaptation solution for other tasks by effectively narrowing the gap between the source and the target domains and easing the entire learning process.

6 Conclusions

In this paper, we propose a proxy-guided domain adaptation learning framework to achieve appearance-preserved portrait-to-anime translation. We first introduce the coser portrait as a proxy bridging the large domain gap between the human portrait and anime portrait domains, which allows for better learning of mappings and yields an initial model with impressive appearance preservation performance. Based on the proxy, a progressive domain adaptation learning scheme with three training stages is proposed to transfer the initial model to the human portrait domain in a gradual manner. By this means, our model can not only generalize fairly well to the human portraits but also retain the substantial appearance preservation capability inherited from the coser portrait domain. In addition, we adopt a disentangled network to cope with the large face deformation modeling and portrait stylization in a divide-and-conquer manner. This further improves the generation quality. Extensive quantitative and qualitative experiments verify that our method can generate high-quality anime portrait images with well preserved appearance. Furthermore, it shows a considerable generalization capability on portraits with different expressions, ages, races, and genders, demonstrating its potential applicability for real scenarios.

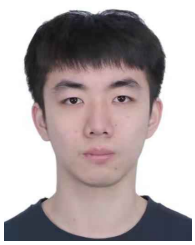
References

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.
[2] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NeurIPS*, 2017, pp. 700–708.
[3] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *ICLR*, 2020.
[4] Y. Zhang, W. Dong, C. Ma, X. Mei, K. Li, F. Huang, B.-G. Hu, and O. Deussen, "Data-driven synthesis of cartoon faces using different styles," *IEEE Trans. Image Proc.*, pp. 464–478, 2016.
[5] B. Li, Y. Zhu, Y. Wang, C.-W. Lin, B. Ghanem, and L. Shen, "Anigan: Style-guided generative adversarial networks for unsupervised anime face generation," *IEEE Trans. Multimedia*, pp. 1–1, 2021.
[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.
[7] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pat. Ana. & Mach. Int.*, pp. 295–307, 2015.
[8] Y. Wei, S. Gu, Y. Li, R. Timofte, L. Jin, and H. Song, "Unsupervised real-world image super resolution via domain-distance aware training," in *CVPR*, 2021, pp. 13 385–13 394.
[9] C. Xu, W. Qu, X. Xu, and X. Liu, "Multi-scale flow-based occluding effect and content separation for cartoon animations," *IEEE Trans. Vis. & Comp. Graphics*, pp. 1–1, 2022.
[10] M. Shi, J.-Q. Zhang, S.-Y. Chen, L. Gao, Y. Lai, and F.-L. Zhang, "Reference-based deep line art video colorization," *IEEE Trans. Vis. & Comp. Graphics*, 2022.
[11] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016, pp. 649–666.
[12] C. Furusawa, K. Hiroshiba, K. Ogaki, and Y. Odagiri, "Comicolorization: semi-automatic manga colorization," in *SIGGRAPH Asia*, 2017, pp. 1–4.
[13] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, 2016, pp. 2414–2423.
[14] J. Cheng, A. Jaiswal, Y. Wu, P. Natarajan, and P. Natarajan, "Style-aware normalized loss for improving arbitrary style transfer," in *CVPR*, 2021, pp. 134–143.
[15] Y. Shu, R. Yi, M. Xia, Z. Ye, W. Zhao, Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Gan-based multi-style photo cartoonization," *IEEE Trans. Vis. & Comp. Graphics*, 2021.

[16] C. Xu, Z. Chen, J. Mai, X. Xu, and S. He, "Pose and attribute consistent person image synthesis," *ACM Trans. Multimedia Comput. Commun. Appl.*, 2022.
[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.
[18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR*, 2018, pp. 8798–8807.
[19] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 172–189.
[20] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, 2018, pp. 35–51.
[21] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, pp. 8789–8797.
[22] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *CVPR*, 2019, pp. 10 551–10 560.
[23] H. Li, G. Liu, and K. N. Ngan, "Guided face cartoon synthesis," *IEEE Trans. Multimedia*, pp. 1230–1239, 2011.
[24] Y. Zhang, W. Dong, O. Deussen, F. Huang, K. Li, and B.-G. Hu, "Data-driven face cartoon stylization," in *SIGGRAPH Asia*, 2014, pp. 1–4.
[25] B. Li, Y. Zhu, Y. Wang, C.-W. Lin, B. Ghanem, and L. Shen, "Anigan: Style-guided generative adversarial networks for unsupervised anime face generation," *IEEE Trans. Multimedia*, pp. 1–1, 2021.
[26] R. Wu, X. Gu, X. Tao, X. Shen, Y.-W. Tai *et al.*, "Landmark assisted cyclegan for cartoon face generation," *arXiv preprint arXiv:1907.01424*, 2019.
[27] K. Cao, J. Liao, and L. Yuan, "Carigans: Unpaired photo-to-caricature translation," *ACM Trans. on Graphics*, 2018.
[28] F. Han, S. Ye, M. He, M. Chai, and J. Liao, "Exemplar-based 3d portrait stylization," *IEEE Trans. Vis. & Comp. Graphics*, 2021.
[29] Z. Ye, M. Xia, Y. Sun, R. Yi, M. Yu, J. Zhang, Y.-K. Lai, and Y.-J. Liu, "3d-carigan: an end-to-end solution to 3d caricature generation from normal face photos," *IEEE Trans. Vis. & Comp. Graphics*, 2021.
[30] H. Su, J. Niu, X. Liu, Q. Li, J. Cui, and J. Wan, "Mangagan: Unpaired photo-to-manga translation based on the methodology of manga drawing," in *AAAI*, 2021, pp. 2611–2619.
[31] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.
[32] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *arXiv preprint arXiv:2006.06676*, 2020.
[33] J. N. Pinkney and D. Adler, "Resolution dependent gan interpolation for controllable image synthesis between domains," *arXiv preprint arXiv:2010.05334*, 2020.
[34] S. Kwong, J. Huang, and J. Liao, "Unsupervised image-to-image translation via pre-trained stylegan2 network," *IEEE Trans. Multimedia*, 2021.
[35] J. Back, "Fine-tuning stylegan2 for cartoon face generation," *arXiv preprint arXiv:2106.12445*, 2021.
[36] G. Song, L. Luo, J. Liu, W.-C. Ma, C. Lai, C. Zheng, and T.-J. Cham, "Agilegan: stylizing portraits by inversion-consistent transfer learning," *ACM Trans. on Graphics*, vol. 40, no. 4, pp. 1–13, 2021.
[37] W. Jang, G. Ju, Y. Jung, J. Yang, X. Tong, and S. Lee, "Stylecarigan: caricature generation via stylegan feature map modulation," *ACM Trans. on Graphics*, vol. 40, no. 4, pp. 1–16, 2021.
[38] M. J. Chong and D. Forsyth, "Jojogan: One shot face stylization," *arXiv preprint arXiv:2112.11641*, 2021.
[39] X. Sun, Y. Hou, W. Deng, H. Li, and L. Zheng, "Ranking models in unlabeled new environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 761–11 771.
[40] Y. Yao, L. Zheng, X. Yang, M. Naphade, and T. Gedeon, "Simulating content consistent vehicle datasets with attribute descent," in *European Conference on Computer Vision*. Springer, 2020, pp. 775–791.
[41] G. Fang, Y. Bao, J. Song, X. Wang, D. Xie, C. Shen, and M. Song, "Mosaicking to distill: Knowledge distillation from out-of-domain data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 920–11 932, 2021.
[42] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, "Synthesizing normalized faces from facial identity features," in *CVPR*, 2017, pp. 3703–3712.
[43] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017, pp. 2794–2802.
[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.
[45] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, pp. 1755–1758, 2009.
[46] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015, pp. 3730–3738.
[47] Anonymous, D. community, and G. Branwen, "Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset," https://www.gwern.net/Danbooru2020, 2021.
[48] Nagadomi, "lbpcascade_animeface," https://github.com/nagadomi/lbpcascade_animeface, 2014.
[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[51] 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 *NeurIPS*, 2017, p. 6629–6640.
[52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.
[53] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, pp. 723–773, 2012.
[54] P. L. Rosin, Y.-K. Lai, D. Mould, R. Yi, I. Berger, L. Doyle, S. Lee, C. Li, Y.-J. Liu, A. Semmo *et al.*, "Nprportrait 1.0: A three-level benchmark for non-photorealistic rendering of portraits," *Computational Visual Media*, vol. 8, no. 3, pp. 445–465, 2022.

Wenpeng Xiao received the bachelor's degree in computer science and technology from South China University of Technology in 2020. He is currently working toward the master's degree in the School of Computer Science and Engineering, South China University of Technology. His research interests include image processing, deep learning for artificial arts, and the applications of computer vision.

Cheng Xu received his B.Eng. degree in Software Engineering from Xiangtan University, China, in 2016. He is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing, and deep learning.



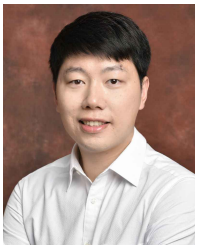
Jiajie Mai received his master's degree in Mobile Personal Communication from King's College London in 2021. He recently researches machine learning and mainly concentrates on medical artificial intelligence.

Xuemiao Xu received her B.S. and M.S. degrees in Computer Science and Engineering from South China University of Technology in 2002 and 2005 respectively, and Ph.D. degree in Computer Science and Engineering from The Chinese University of Hong Kong in 2009. She is currently a professor in the School of Computer Science and Engineering, South China University of Technology. Her research interests include object detection, tracking, recognition, and image, video understanding and synthesis, particularly their applications in the intelligent transportation.

Yue Li received the PhD degree in electronic engineering from Tsinghua University, China, in 2008. Since 2009, she is working with the School of Computer Science and Engineering in South China University of Technology. Her research interests include machine learning, data mining, and computational thinking. She is one member of 2018 Google China CS Education Program Advisory Group, and also one member of 2019 Google China High Education Program Advisory Group.

Chengze Li received his B.Eng. degree from the University of Science and Technology of China in 2013, and Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2020. He is currently an assistant professor in the School of Computing and Information Sciences, Caritas Institute of Higher Education, with research interests in 2D non-photorealistic media analysis and processing, computational photography, and computer graphics.

Xueting Liu received her B.Eng. degree from Tsinghua University and Ph.D. degree from The Chinese University of Hong Kong in 2009 and 2014 respectively. She is currently an Assistant Professor in the School of Computing and Information Sciences, Caritas Institute of Higher Education. Her research interests include computer graphics, computer vision, machine learning, computational manga and anime, and non-photorealistic rendering.



Shengfeng He (Senior Member, IEEE) is an associate professor in the School of Computer Science and Engineering, South China University of Technology. He obtained B.Sc. and M.Sc. degrees from Macau University of Science and Technology in 2009 and 2011 respectively, and a Ph.D. degree from City University of Hong Kong in 2015. His research interests include computer vision, image processing, computer graphics, and deep learning. He serves on the editorial board of Neurocomputing.