Knowledge-Based Systems 255 (2022) 109731

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Cross-domain image translation with a novel style-guided diversity loss design

Tingting Li [a], Huan Zhao [a,*], Jing Huang [c], Keqin Li [a,b]

[a] School of Information Science and Engineering, Hunan University, Changsha, 410082, China
[b] Department of Computer Science, State University of New York, New Paltz, NY, 12561, USA
[c] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, 411100, China

ARTICLE INFO

Article history:
Received 26 January 2022
Received in revised form 15 August 2022
Accepted 16 August 2022
Available online 22 August 2022

Keywords:
Cross-domain image-to-image (I2I) translation
Extracted style features
Generative Adversarial Networks (GAN)
Multimodal
Multiple-domain

ABSTRACT

Cross-domain image-to-image translation has made remarkable progress in recent years. It aims to map the image from the original image domain to the target domains so that the image can appear in diverse styles. Currently, existing methods are mainly based on Generative Adversarial Networks (GAN). They often employ an auxiliary encoder to extract style features from noises or reference images for the generator to translate new images. However, these approaches are usually feasible for two-domain translation and present low diversity in multi-domain translation since the extracted style features are simply served as additional input to the generator rather than fully utilized. This paper proposes a style-guided image-to-image translation (SG-I2IT) with a novel diversity regularization term named style-guided diversity loss (SD loss), making the best of the extracted style features. In our model, style features not only serve as the generator's input but also penalize the generator through the new SD loss, thus encouraging the model to capture the image styles better. The effectiveness of our method is demonstrated from two perspectives, noise-based and reference-based image translation. Qualitative and quantitative experiments validate our superiority of the proposed method against the state-of-the-art methods in terms of image quality and diversity. In addition, a user study demonstrates that the proposed method can better capture image styles and translate more realistic images.

1. Introduction

Cross-domain image-to-image (I2I) translation is one of the hottest research topics in the computer vision community, covering many basic computer vision tasks [1–4], such as image inpainting [5], style transfer [3,4], and super-resolution [6,7]. It is designed to simulate the mapping between different visual domains [8,9]. A visual *domain* refers to a group of images that share a certain common visual characteristic (such as female or male in the CelebA dataset), while each image possesses a unique *style* (such as skin tone, hair color, beard, and makeup) [10]. Presenting an image in a new style will get a new image. For example, as shown in Fig. 1, a border collie in black-and-white style can be translated into a new cat in black-striped style or black-and-white. Imagining images in different styles is an innate ability of human beings, while for a machine, it is a tough challenge. Mimicking this ability is exactly what image translation intends to learn, which makes image translation a challenging but attractive task.

Nowadays, growing efforts are taking up this challenge with Generative Adversarial Networks (GAN) [5,6,11,12]. The adversarial training of GAN has achieved excellent performance in image generation, and its introduction has promoted the rapid development of image translation [5,6,13]. For example, Hedjazi and Genc [5] proposed a multi-GAN image-to-image translation architecture to improve image inpainting that synthesizes plausible contents to fill in the missing image regions or remove unwanted objects from images. Identity-Preservation Generative Adversarial (IPGAN) is a study of photo-to-caricature translation, which generates realistic and identity-preserving caricatures from given photos [4]. Zhang et al. designed a cross-domain correspondence network (CoCosNet) for example-based image translation that translates photo-realistic images from given example images (edge maps or pose keypoints) [14]. Karras et al. [6] achieved image translation from low resolution to higher resolution based on GAN. They provided a higher quality version of the CelebA dataset (CelebA-HQ), where the image resolution is 1024 × 1024. The CelebA-HQ is widely used in image processing [5,10,14–16].

Despite impressive achievements, learning an ideal I2I translation method is still challenging for two main reasons. First, to present an object in various styles as much as possible, it is

[*] Corresponding author.
E-mail addresses: Tingting1225@hnu.edu.cn (T. Li), hzhao@hnu.edu.cn (H. Zhao), jingh@hnu.edu.cn (J. Huang), lik@newpaltz.edu (K. Li).
URL: http://www.cs.newpaltz.edu/~lik/ (K. Li).

https://doi.org/10.1016/j.knosys.2022.109731
0950-7051/© 2022 Elsevier B.V. All rights reserved.

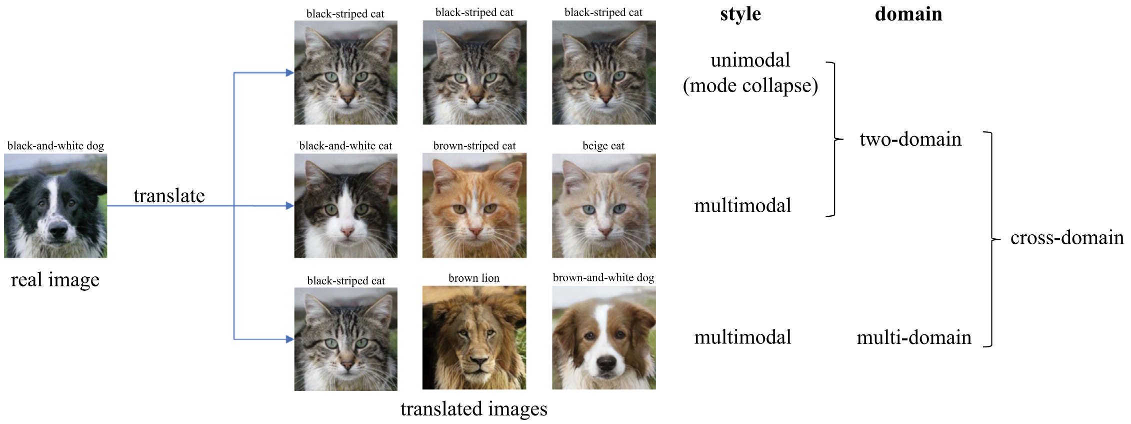


Fig. 1. Different types of cross-domain image translation. The translated images in the first row are plausible but similar, which is called unimodal image translation. Translated images in the second row are in different styles, called multimodal image translation. These image translations are both between two domains (dog → cat). The translation images in the third row have different styles and belong to different image domains (cat, wild animal, and dog), called multi-domain image translation.

required for the model to translate the object style from one to multiple. More precisely, one input will correspond to multiple possible outputs [8,17]. That is to say, such mapping from one visual domain to another is inherently *multimodal*. However, the image styles translated by current GAN-based methods are usually similar rather than diverse due to the well-known mode collapse problem of GAN [8,17,18]. Thus, increasing diversity has become a crucial issue for an ideal image translation method [3,19–21]. Second, image translation between two domains has made excellent progress [19], but extending them to multiple domains is difficult. Some widely-used datasets in the real situation often contain *multiple domains*. For example, the recent popular animal-face-high-quality (AFHQ) dataset [10] consists of three domains: cat, dog, and wildlife. Two-domain image translations are not applicable to such datasets. Therefore, multi-domain I2I translation in real scenarios is worthy of in-depth study.

To better learn the multimodal mappings, many efforts have been developed to increase the output diversity from different perspectives [19–21]. The initial attempt is based on conditional GAN (cGAN), which guides the output by adding additional conditions. Many researchers prefer adopting an auxiliary encoder to extract latent variables for the generator [22] or to explore more relationships between the original and target image domains [20]. From a new perspective, some recent efforts assume that a visual image can be decomposed into a domain-specific style feature space and a domain-invariant content space [8,19]. They enhance the image diversity by matching the domain-invariant content codes with other style feature codes. Nevertheless, these approaches mainly consider the image translation between two domains [10]. Some datasets actually often have more than two visual domains (such as the AFHQ dataset). For image datasets with K domains, these approaches need to learn $K(K-1)$ mappings to achieve the translations between different domains, thus limiting their scalability over multiple image domains [10,23].

To improve the scalability, some recent studies focusing on multi-domain image translation have been proposed. An intuitive approach is to add multiple domain-specific information [23]. But such domain information is usually fixed and limits the diversity of the output [10]. Researchers introduced an encoder to break through this limitation [10,21]. For instance, a recent Ref. [10] utilizes an encoder to extract domain codes from specific image domains as substitutes for fixed domain-specific information; The latest Ref. [21] employs multiple encoders to disentangle the given images into domain-specific and domain-invariant representations. One of the main functions of their encoders is to extract features for the generator. However, most current researchers simply use such style features as additional inputs to the generator. How to make the best of these style features for cross-domain image translation is actually still worthy of further study.

In this paper, we take full advantage of the extracted style features and design a novel style-guided regularization term for cross-domain image translation. The main contributions of this paper are summarized as follows:

- We propose a style-guided image-to-image translation (SG-I2IT) with a novel diversity loss function named style-guided diversity loss (SD Loss). It makes the best of style features to assist the generator in discovering diverse image styles.
- We theoretically analyze the limitations of the current state-of-the-art diversity loss approaches. Our method breaks through the limitations and encourages the model to capture image styles effectively to translate diverse new images.
- Extensive experiments are conducted on two wildly-used image datasets, CelebA-HQ and AFHQ. The results demonstrate our superiority against other state-of-the-art methods in terms of image quality and diversity.

Compared with the preliminary conference version of this paper, we have made the following improvements and extensions: 1. Rewrite the whole article; 2. Add Section 3.2 to theoretically analyze the limitations of the current state-of-the-art diversity loss methods and introduce the design of the proposed style-guided loss function; 3. Add diversity analysis experiments to demonstrate the significance of the proposed technique in enhancing image diversity; 4. Add a user study to support the superiority of this method.

The rest of the paper is organized as follows. Section 2 briefly introduces the related work on cross-domain image translation. Section 3 describes the details of the proposed style-guided image-to-image translation. In Section 4, we conduct a series of experiments to demonstrate the performance of the proposed method. Section 5 gives a brief conclusion of this paper and a direction for future work.

2. Related work

Generative Adversarial Networks (GANs). Recent years have witnessed the success of GANs in various artificial intelligence applications [25–27], especially in the computer vision community [28]. The core idea of GANs is to map random noise to

Table 1
Comparison of different I2I translation methods. The notations "–" and "✓" represent "no" and "yes", respectively.

Methods	Unsupervised	Unpaired	Multimodal	Multi-domain	Number of generator	Usage of style features
Pix2Pix [9]	–	–	–	–	Single	None
CycleGAN [3]	–	✓	–	–	Single	None
DiscoGAN [24]	✓	✓	–	–	Single	None
UNIT [1]	✓	✓	–	–	Multiple	Injecting
BicycleGAN [20]	–	–	✓	–	Single	Injection
StarGAN [23]	–	–	–	✓	Multiple	Injection
Augmented CycleGAN [22]	✓	✓	✓	–	Single	None
MUNIT [19]	✓	✓	✓	–	Multiple	Injection
DRIT [8]	✓	✓	✓	–	Multiple	Injection
MSGAN [17]	–	✓	✓	–	Single	Injecting
DRIT++ [21]	✓	✓	✓	✓	Multiple	Injection
StarGAN-v2 [10]	✓	✓	✓	✓	Single	Injection
Ours	✓	✓	✓	✓	**Single**	**Injecting & regularizing**

the target domain samples through the competition between a generator and a discriminator [29]. The generator is responsible for mapping the noise to the target domain samples, while the discriminator aims to distinguish the generated fake samples from the real samples. The adversarial learning between these two members makes GANs produce plausible samples that cannot be recognized. Hitherto, GAN-based methods have been widely applied to various computer vision tasks, such as target tracking [30], medical image analysis [31,32], image translation [9], etc. For image translation, Ref. [9] is the first to successfully apply conditional GANs (cGANs) to image translation, breaking through the bottleneck of traditional methods, and proposed a Pix2Pix framework, which opened up a new era for image translation. In this paper, we propose an image-to-image translation method based on GAN.

Image-to-image translation. I2I translation aims to capture the mapping from an original visual domain to target domains [9]. Many efforts have been devoted to this challenging task. The successful introduction of cGANs has made Pix2Pix the first common deep-learning framework for I2I translation [2,3,20]. Although the image quality has been improved, its training requires paired images. Focusing on unpaired I2I translation, Ref. [3] put forward a novel cycle consistency loss, encouraging the model to capture auxiliary inverse relations from the target visual domain to the original domain. Ref. [24] proposes an unsupervised I2I translation based on the GAN model named DiscoGAN, aiming to help the model discover the cross-domain relations for unpaired I2I translation. Unsupervised I2I translation (UNIT) [1] is another typical unsupervised I2I translation model. It also aims to avoid costly pairing and make a shared-latent space assumption based on the combination of GANs and Variational AutoEncoders (VAEs) [33]. Nevertheless, these early studies make a simple assumption that the mapping of I2I translation is one-to-one, which ignores the inherent multimodal characteristic of the I2I translation [19].

Multimodal image translation. Although GAN-based methods have made profound progress in image quality [18], the inherent multi-modal characteristic of I2I translations has been hindered since the well-known mode collapse problem limits the diversity of generated images [8,17]. To improve the diversity of generated images, Encoded Multi-agent GAN (EMGAN) [34] employs multiple generators. CGAN-based methods often assist the model in discovering more image modes by introducing a regularization term [17]. For I2I translation, BicycleGAN [20] addresses the low-diversity problem through the combination of a conditional Latent Regressor GAN (cLR-GAN) and a conditional VAE-GAN (cVAE-GAN). It obtains realistic and diverse images by learning a bijective mapping between target space and latent space. Augmented CycleGAN [22] introduces an encoder to assist the model with achieving the multi-modal goal of I2I translation. Additionally, I2I Translation via Disentangled Representations (DRIT) [8] and Multimodal Unsupervised I2I Translation (MUNIT) [19] propose a new assumption that the image representation can be decomposed into different domain-specific style codes and domain-invariant content codes. They recombine domain-invariant content codes with other style codes corresponding to different domains and produce desirable and diverse outputs [8,19]. However, these methods mainly focus on the two-domain I2I translation. Some widely-used datasets in the real situation often contain more than two domains. The scalability of these two-domain translation methods is limited for datasets with more than two domains [10,23].

Multi-domain image translation. To conquer the scalability, several recent methods have been developed. In [23], the authors proposed a StarGAN framework, implementing the translations for more than two domains using target domain labels. StarGAN is a typical cGAN-based framework, but its fixed domain labels limit the diversity of the outputs. To avoid this limitation, Lee et al. abstracted domain information by introducing an image domain encoder and further extended DRIT to the multi-domain I2I translation [21]. In addition, based on the assumption that the image encoder can decompose images into style space and content space, StarGAN-v2 [10] proposes domain-specific style vectors and implements image translations over multiple domains. Nevertheless, most existing approaches simply use the extracted image style features as conditional inputs to the generator. To our knowledge, no study encourages the model to generate diverse images by regularizing the generator based on the extracted features. Therefore, we conduct a further study on making the best of the extracted image style features to enhance the diversity for multi-domain I2I translation.

Diversity maximization methods. Enhancing the diversity of samples is an unavoidable and crucial problem in many computer vision tasks [17,35,36]. For instance, to avoid tedious manual annotation work, Yang et al. proposed a multi-class active learning method for visual concept recognition [35]. They employed a similarity matrix as a diversity regularization term for the objective function to make the sampled data as diverse as possible. Liu et al. [36] designed a pair-based early active learning method for Person Re-identification (Re-ID) by introducing a pairwise diversity maximization criterion. This method can improve the diversity of selected image pairs for Re-ID tasks. However, the above approaches mainly work in the sampling stage of model training. For I2I translations, enhancing the diversity of image translation is mainly in the stage of image generation [8,15,17]. Mode-seeking GAN (MSGAN) improves the diversity of generated images by introducing a mode-seeking regularization term [17]. This regularization term works based on the conditional input of

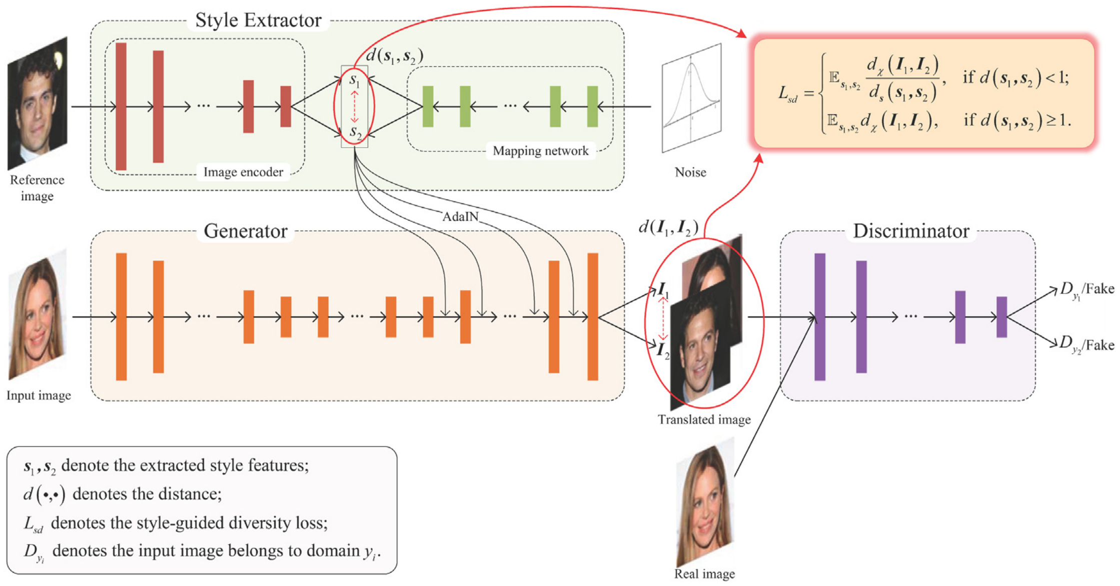

Style Extractor
$d(s_1, s_2)$
s_1
s_2
Reference image
Image encoder
Mapping network
Noise
$L_{sd} = \begin{cases} \mathbb{E}_{s_1,s_2} \frac{d_\chi(I_1, I_2)}{d_s(s_1, s_2)}, & \text{if } d(s_1,s_2) < 1; \\ \mathbb{E}_{s_1,s_2} d_\chi(I_1, I_2), & \text{if } d(s_1,s_2) \geq 1. \end{cases}$
AdaIN
$d(I_1, I_2)$
Generator
Discriminator
Input image
I_1
I_2
Translated image
D_{y_1}/Fake
D_{y_2}/Fake
Real image
s_1, s_2 denote the extracted style features;
$d(\bullet,\bullet)$ denotes the distance;
L_{sd} denotes the style-guided diversity loss;
D_{y_i} denotes the input image belongs to domain y_i.

Fig. 2. Framework overview. The framework consists of three parts: a feature extractor (a style encoder or a mapping network), a generator, and a discriminator.

the model. Inspired by MSGAN, we make full use of the extracted image style features to design a style-guided diversity regularization term for the I2I translation model, thereby improving the diversity of translation samples.

Comparing I2I translation and image harmonization. Image harmonization is crucial in image processing [37–40]. It aims to improve the generated image quality by harmonizing the appearance (e.g., color, contrast, or brightness) of the foreground to match the background image [38,40]. It can generate high-quality images that appear more realistic [38]. However, I2I translation aims to translate the image target from one original image domain to other image domains with different styles [8,10]. The styles here refer not only to objects' colors but also to some attributes [21, 23]. For example, in Fig. 1, a good image translation model can translate a black and white dog into a brown cat or a lion. Ideal image translation requires that its translated images are not only realistic but also as diverse as possible [8,15,17]. It exactly is our goal to enhance the diversity of multi-domain image translation. We compare many of the state-of-the-art I2I translation methods mentioned above, and their differences are summarized in Table 1. Among them, our method and StarGAN-v2 improve the multimodality of multi-domain image translation through a single generator. Furthermore, only our method not only takes the extracted feature styles as input to the generator but also utilizes them to propose a regularization term for the generator to encourage the model to generate more diverse images.

3. Style-guided image-to-image translation

We propose a style-guided I2I translation (SG-I2IT) with a novel style-guided diversity regularization term. The new diversity regularization term encourages our model better to explore the image space and capture image styles.

3.1. Image translation framework

Let $\mathcal{Z}$, $\mathcal{X}$, and $\mathcal{Y}$ be the spaces of the noise, image, and possible visual domain, respectively. Given a random noise $\boldsymbol{z} \in \mathcal{Z}$ or a reference image $\boldsymbol{x}' \in \mathcal{X}$ in any target visual domain $y'_i \in \mathcal{Y}$, we can extract style features $\boldsymbol{s}_i$ through the feature extractor. We aim to encourage the image translation model to make the most of the extracted style features $\boldsymbol{s}_i$ and produce diverse images reflecting the style of domain y'_i. Our method is based on a recent successful I2I translation framework, StarGAN-v2 [10]. The framework consists of the three parts shown in Fig. 2, and they are described below.

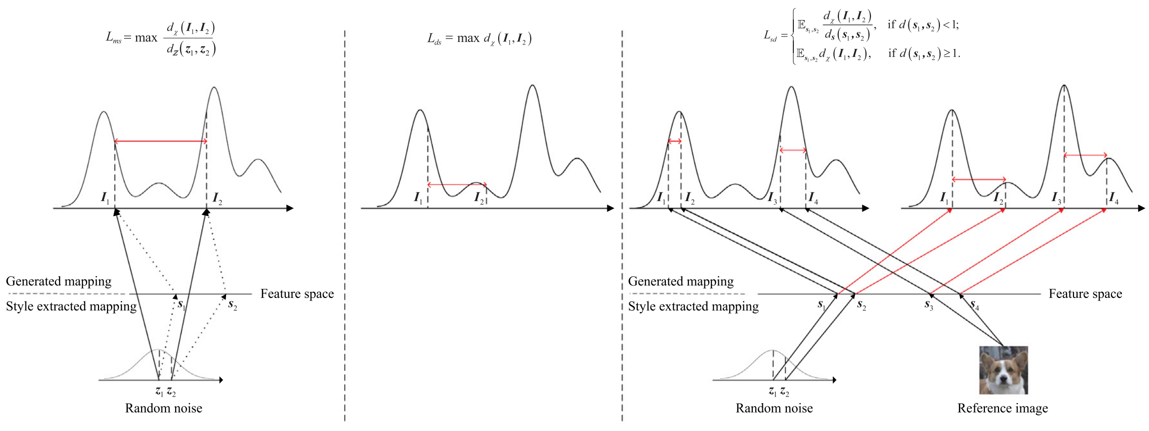
Style feature extractor (green module in Fig. 2). We extract the style features $\boldsymbol{s}_i$ through an image encoder E (the red networks) or a mapping network M (the green networks), as shown in Fig. 2. For the former, given a reference image $\boldsymbol{x}' \in \mathcal{X}$ and its domain y'_i, the encoder E extracts style features $\boldsymbol{s}_i = E_{y'_i}(\boldsymbol{x}')$ from image $\boldsymbol{x}'$. For the latter, randomly sampled a noise vector $\boldsymbol{z}$ from Gaussian distribution $N(0, 1)$, the mapping network M maps the noise vector $\boldsymbol{z}$ to style features $\boldsymbol{s}_i = M_{y'_i}(\boldsymbol{z})$ that is likely in the target domains y'_i. To make the feature extractor applicable to every domain, both E and M are designed with multiple branch outputs corresponding to different domains. Each output branch $E_{y'_i}(\cdot)$ or $M_{y'_i}(\cdot)$ provides style features $\boldsymbol{s}_i$ for a specific domain $y_i, i = 1, 2, \ldots, K$, where K refers to the number of domains. Therefore, the extractor can provide style features for all possible domains.

Generator (orange module in Fig. 2). It produces new images $\boldsymbol{I}_i = G(\boldsymbol{x}, \boldsymbol{s}_i)$ from an original image $\boldsymbol{x} \in \mathcal{X}$ with the assistance of style features $\boldsymbol{s}_i$ provided by the style feature extractor. Domain-specific style features $\boldsymbol{s}_i$ help the generator G produce images that are likely in any target domain y'_i. Thus, the model can translate images between multiple domains (such as from cats to dogs, wildlife, or cats). The generator G mainly contains two parts: downsampling blocks and upsampling blocks. Down-sampling blocks adopt the instance normalization (IN), and up-sampling blocks use the adaptive instance normalization (AdaIN). We inject the style features $\boldsymbol{s}_i$ into the generator G through the AdaIN layers [10,41]. AdaIN layers complete the fusion of the original image $\boldsymbol{x}$ and the features $\boldsymbol{s}_i$ through

$$\mathrm{AdaIN}(\boldsymbol{x}, \boldsymbol{s}_i) = \sigma(\boldsymbol{s}_i)\left(\frac{\boldsymbol{x} - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right) - \mu(\boldsymbol{s}_i), \quad (1)$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ denote mean and variance functions, respectively.

Discriminator (purple module in Fig. 2). Both generated fake images $\boldsymbol{x}'$ and real images $\boldsymbol{x}$ are put into the discriminator D. The discriminator D is responsible for identifying the authenticity of


$L_{ms} = \max \frac{d_{\chi}(I_1, I_2)}{d_Z(z_1, z_2)}$
I_1
I_2
Generated mapping
Style extracted mapping
Feature space
s_1
s_2
z_1 z_2
Random noise
$L_{ds} = \max d_{\chi}(I_1, I_2)$
I_1
I_2
$L_{sd} = \begin{cases} \mathbb{E}_{s_1,s_2} \frac{d_{\chi}(I_1, I_2)}{d_s(s_1, s_2)}, & \text{if } d(s_1, s_2) < 1; \\ \mathbb{E}_{s_1,s_2} d_{\chi}(I_1, I_2), & \text{if } d(s_1, s_2) \geq 1. \end{cases}$
I_1
I_2
I_3
I_4
I_1
I_2
I_3
I_4
Generated mapping
Style extracted mapping
Feature space
s_1
s_2
s_3
s_4
z_1 z_2
Random noise
Reference image

(a) MSGAN (b) StarGAN-v2 (c) Ours

Fig. 3. The comparison of different diversity loss functions. Figs. 3(a), 3(b), and 3(c) respectively illustrate the diversity regularization terms of MSGAN, StarGAN-v2, and the proposed method. Fig. 3(a) presents the mode-seeking loss $\mathcal{L}_{ms}$ of MSGAN, maximizing the distance between two images with respect to the noises and ignoring the extracted features $\boldsymbol{s_i}$. Fig. 3(b) shows the diversity-sensitive loss $\mathcal{L}_{ds}$ of StarGAN-v2, which directly maximizes the distance between two samples. Fig. 3(c) presents the proposed style-guided diversity loss $\mathcal{L}_{sd}$, maximizing the image distance with the guidance of the extracted features $\boldsymbol{s_i}$. Black arrows represent the original mappings of Lms or Lds, and red arrows represent the mappings of the proposed method in this paper.

the input images. As shown in Fig. 2, the discriminator D is designed with multiple output branches corresponding to multiple visual domains. Each branch $D_{y_i}(\cdot)$ is a binary classifier designed to distinguish whether the input sample is a real image from the domain y_i or a fake image from the generator G corresponding to domain y_i. This multi-branch design not only encourages the model to generate plausible images but also makes the generated images cover all possible image domains as much as possible.

3.2. Diversity loss function design

A good diversity loss function can alleviate the mode collapse problem for GAN-based methods and effectively improve the diversity of generated images. The core idea of designing a diversity loss function is to encourage the model to discover more image modalities by maximizing the distance between any two generated images, $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ [10,17]. Recently, the most popular and widely-used diversity loss function is the mode-seeking loss ($\mathcal{L}_{ms}$) [10,21] proposed by Ref. [17],

$$\mathcal{L}_{ms} = \mathbb{E}\left[\frac{d_{\mathcal{X}}(\boldsymbol{I}_1, \boldsymbol{I}_2)}{d_{\mathcal{Z}}(\boldsymbol{z}_1, \boldsymbol{z}_2)}\right] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\frac{\|G(\boldsymbol{x}, \boldsymbol{z}_1) - G(\boldsymbol{x}, \boldsymbol{z}_2)\|_1}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_1}\right], \tag{2}$$

where $d_*(\cdot, \cdot)$, $\boldsymbol{z}_i$, and $G(\boldsymbol{x}, \boldsymbol{z}_i)$ respectively denote the distance measured by L_1 norm $\|\cdot\|_1$, noise vectors, and generated images. The design idea of $\mathcal{L}_{ms}$ is illustrated in Fig. 3(a); that is, the model can be encouraged to discover different image modes by maximizing the ratio of the image distance over the noise distance. As can be seen in Fig. 3(a), $\mathcal{L}_{ms}$ ignores the feature space. When it is introduced into StarGAN-v2, the slight difference in the denominator significantly increased the loss and made the model training unstable [10]. Then, Ref. [10] directly removes the denominator $\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_1$ in the original form of $\mathcal{L}_{ms}$ to maximize the distance between any two generated images and provides a diversity-sensitive loss $\mathcal{L}_{ds}$ for StarGAN-v2,

$$\begin{aligned} \mathcal{L}_{ds} =& \mathbb{E}[d_{\mathcal{X}}(\boldsymbol{I}_1, \boldsymbol{I}_2)] \\ =& \mathbb{E}_{\boldsymbol{x},\boldsymbol{s}_1,\boldsymbol{s}_2}[\|G(\boldsymbol{x}, \boldsymbol{s}_1) - G(\boldsymbol{x}, \boldsymbol{s}_2)\|_1] \\ =& \mathbb{E}_{\boldsymbol{x},y_i',\boldsymbol{z}_1,\boldsymbol{z}_2}[\|G(\boldsymbol{x}, M(\boldsymbol{z}_1)) - G(\boldsymbol{x}, M(\boldsymbol{z}_2))\|_1]. \end{aligned} \tag{3}$$

The design of $\mathcal{L}_{ds}$ is illustrated in Fig. 3(b). Obviously, $\mathcal{L}_{ds}$ does not consider the feature space, and directly removing the denominator makes $\mathcal{L}_{ds}$ fail to inherit the design idea of $\mathcal{L}_{ms}$. To this end, we attempt to find out the reason why $\mathcal{L}_{ms}$ makes the training of StarGAN-v2 unstable so that we can effectively inherit the design ideas of $\mathcal{L}_{ms}$ to design a diversity loss function utilizing style features for such multi-domain translation frameworks.

We notice that noise vectors $\boldsymbol{z}_i$ in the denominator of $\mathcal{L}_{ms}$, as shown in Fig. 2, are no longer direct inputs of the generator but are inputs of the mapping network M. M extracts style features $\boldsymbol{s}_i$ from $\boldsymbol{z}_i$ for the generator G. We respectively denote the mapping network M as a continuous function on the noise variable $\boldsymbol{z}_i$ and the generator G as a continuous function on the image $\boldsymbol{x}$ and style features $\boldsymbol{s}_i$, that is,

$$\boldsymbol{s}_i = M(\boldsymbol{z}_i) \quad \text{and} \quad \boldsymbol{I}_i = G(\boldsymbol{x}, \boldsymbol{s}_i),$$

then we get

$$\boldsymbol{I}_i = G(\boldsymbol{x}, M(\boldsymbol{z}_i)). \tag{4}$$

Eq. (4) shows that the generator G is a composite function with respect to the noise vector $\boldsymbol{z}_i$. Thus, given an input image $\boldsymbol{x}_0$ and the slight difference of noise vector, $\Delta\boldsymbol{z} = \boldsymbol{z}_2 - \boldsymbol{z}_1$, when $\Delta z \to 0$, the mode-seeking loss in StarGAN-v2 will be described as

$$\begin{aligned} \widetilde{\mathcal{L}_{ms}} =& \mathbb{E}\left[\frac{d_{\mathcal{X}}(\boldsymbol{I}_1, \boldsymbol{I}_2)}{d_{\mathcal{Z}}(\boldsymbol{z}_1, \boldsymbol{z}_2)}\right] \\ =& \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\frac{\|G(\boldsymbol{x}_0, M(\boldsymbol{z}_2)) - G(\boldsymbol{x}_0, M(\boldsymbol{z}_1))\|_1}{\|\boldsymbol{z}_2 - \boldsymbol{z}_1\|_1}\right] \\ =& \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\frac{\|G(\boldsymbol{x}_0, M(\boldsymbol{z}_1 + \Delta\boldsymbol{z})) - G(\boldsymbol{x}_0, M(\boldsymbol{z}_1))\|_1}{\|\Delta\boldsymbol{z}\|_1}\right] \\ =& \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\left\|G_{\boldsymbol{z}}'(\boldsymbol{x}_0, M(\boldsymbol{z}_1))\right\|_1\right] \\ =& \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\left\|G_{\boldsymbol{s}}'(\boldsymbol{x}_0, \boldsymbol{s}_1)\right\|_1 \cdot \left\|M_{\boldsymbol{z}}'(\boldsymbol{z}_1)\right\|_1\right], \end{aligned} \tag{5}$$

where $M_{\boldsymbol{z}}'(\boldsymbol{z}_1)$ and $G_{\boldsymbol{s}}'(\boldsymbol{x}_0, \boldsymbol{s}_1)$ respectively denote the differential (gradient) of function $M(\boldsymbol{z})$ and the partial differential (gradient) of function $G(\boldsymbol{x}, \boldsymbol{s})$ with respect to style feature $\boldsymbol{s}$. But in the general cGAN models, generated images $\boldsymbol{I}_i = G(\boldsymbol{x}, \boldsymbol{z}_i)$, and $\mathcal{L}_{ms}$ will be

$$\begin{aligned} \mathcal{L}_{ms} =& \mathbb{E}\left[\frac{d_{\mathcal{X}}(\boldsymbol{I}_1, \boldsymbol{I}_2)}{d_{\mathcal{Z}}(\boldsymbol{z}_1, \boldsymbol{z}_2)}\right] = \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\frac{\|G(\boldsymbol{x}_0, \boldsymbol{z}_2) - G(\boldsymbol{x}_0, \boldsymbol{z}_1)\|_1}{\|\boldsymbol{z}_2 - \boldsymbol{z}_1\|_1}\right] \\ =& \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\frac{\|G(\boldsymbol{x}_0, \boldsymbol{z}_1 + \Delta\boldsymbol{z}) - G(\boldsymbol{x}_0, \boldsymbol{z}_1)\|_1}{\|\Delta\boldsymbol{z}\|_1}\right] \\ =& \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\left\|G_{\boldsymbol{z}}'(\boldsymbol{x}_0, \boldsymbol{z}_1)\right\|_1\right]. \end{aligned} \tag{6}$$

By comparing Eq. (5) with Eq. (6), it can be seen that $\widetilde{\mathcal{L}_{ms}}$ in StarGAN-v2 has one more term $\left\|M'_{\boldsymbol{z}}(\boldsymbol{z}_1)\right\|_1$, that is a norm of the gradient of the mapping network. When the gradient of the generator is determined, the more enormous $\left\|M'_{\boldsymbol{z}}(\boldsymbol{z}_1)\right\|_1$ will make $\widetilde{\mathcal{L}_{ms}}$ increase more sharply, which will finally cause the unstable training of the model. Therefore, we infer that the unstable training of StarGAN-v2 with $\widetilde{\mathcal{L}_{ms}}$ results from the fact that the noise vectors $\boldsymbol{z}_i$ are no longer direct inputs of the generator.

Based on the above analysis, we propose to make full use of style features $\boldsymbol{s}_i$, conditional inputs of the generator, and design a style-guided diversity loss (SD loss) for the StarGAN-v2 framework as follows,

$$\begin{aligned}\mathcal{L}_{sd} &= \max\left\{\frac{d_{\mathcal{X}}(\boldsymbol{I}_1,\boldsymbol{I}_2)}{d_{\mathcal{S}}(\boldsymbol{s}_1,\boldsymbol{s}_2)}, d_{\mathcal{X}}(\boldsymbol{I}_1,\boldsymbol{I}_2)\right\}\\ &= \begin{cases}\mathbb{E}_{\boldsymbol{x},\boldsymbol{s}_1,\boldsymbol{s}_2}\frac{\|G(\boldsymbol{x},\boldsymbol{s}_1)-G(\boldsymbol{x},\boldsymbol{s}_2)\|_1}{\|\boldsymbol{s}_1-\boldsymbol{s}_2\|_1}, & \text{if } \|\boldsymbol{s}_1-\boldsymbol{s}_2\|_1 < 1,\\ \mathbb{E}_{\boldsymbol{x},\boldsymbol{s}_1,\boldsymbol{s}_2}\|G(\boldsymbol{x},\boldsymbol{s}_1)-G(\boldsymbol{x},\boldsymbol{s}_2)\|_1, & \text{if } \|\boldsymbol{s}_1-\boldsymbol{s}_2\|_1 \geqslant 1.\end{cases}\end{aligned} \tag{7}$$

The proposed SD loss function $\mathcal{L}_{sd}$ aims to utilize style features $\boldsymbol{s}_i$ to regularize the generator G and guide the model to discover more diverse images. When two style features are similar, that is, $\|\boldsymbol{s}_1-\boldsymbol{s}_2\|_1 < 1$, it maximizes the ratio of the distance between translated images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ over that between style features $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. When two features are far apart, that is $\|\boldsymbol{s}_1-\boldsymbol{s}_2\|_1 \geqslant 1$, then $\|\boldsymbol{I}_1-\boldsymbol{I}_2\|_1/\|\boldsymbol{s}_1-\boldsymbol{s}_2\|_1 \leqslant \|\boldsymbol{I}_1-\boldsymbol{I}_2\|_1$, so it adopts $\|\boldsymbol{I}_1-\boldsymbol{I}_2\|_1$ to maximize the distance between two images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$. As illustrated in Fig. 3(c) (black solid mappings), a mode collapse situation is likely to occur when two features $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are relatively close, and the images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ translated by them are likely to be in one mode. However, with our loss function (red solid mappings), the style feature $\boldsymbol{s}_2$ generates $\boldsymbol{I}_2$, which belongs to another undiscovered mode. This shows that the proposed diversity loss function effectively utilizes style features to maximize the distance between any two generated images, $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, thereby making it more effective to encourage the model to explore the image space to translate more meaningful images.

3.3. Training

Given an original image $\boldsymbol{x} \in \mathcal{X}$ and its domain $y \in \mathcal{Y}$, we train the proposed SG-I2IT model with our style-guided diversity loss $\mathcal{L}_{sd}$ (Eq. (7)), as well as the following three loss functions.

Adversarial loss. Given a random noise vector $\boldsymbol{z} \sim N(0,\boldsymbol{I})$ or a reference sample $\boldsymbol{x}' \in \mathcal{X}$ from a target domain $y'_i \in \mathcal{Y}$, the style feature extractor provides style features $\boldsymbol{s}_i = M_{y'_i}(\boldsymbol{z})$ or $\boldsymbol{s}_i = E_{y'_i}(\boldsymbol{x}')$ for the generator G. With the assistance of style features $\boldsymbol{s}_i$, the generator G produces plausible images $G(\boldsymbol{x},\boldsymbol{s}_i)$ to confuse the discriminator D, which in turn identifies the translated fake samples $G(\boldsymbol{x},\boldsymbol{s}_i)$ from real samples $\boldsymbol{x}$. They compete with each other via an adversarial loss [29]

$$\mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{x},y}\left[\log D_y(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x},y'_i,\boldsymbol{s}_i}\left[\log(1-D_{y'_i}(G(\boldsymbol{x},\boldsymbol{s}_i)))\right], \tag{8}$$

where $D_y(\cdot)$ refers to the output of discriminator D. Generator G minimizes $\mathcal{L}_{adv}$ to make generated images $G(\boldsymbol{x},\boldsymbol{s}_i)$ much more realistic that the discriminator cannot distinguish from real images $\boldsymbol{x}$. Discriminator maximizes $\mathcal{L}_{adv}$ to identify generated images $\boldsymbol{x}$ from fake images $G(\boldsymbol{x},\boldsymbol{s}_i)$.

Style reconstruction loss. To guarantee the generator G makes full use of style features $\boldsymbol{s}_i$ and generates new images $G(\boldsymbol{x},\boldsymbol{s}_i)$ with such styles, we utilize a style reconstruction loss [19,20]

$$\mathcal{L}_{rec} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{s}_i,y'_i}\left[\left\|E_{y'_i}(G(\boldsymbol{x},\boldsymbol{s}_i)) - \boldsymbol{s}_i\right\|_1\right]. \tag{9}$$

$E_{y'_i}(G(\boldsymbol{x},\boldsymbol{s}_i))$ refers to style features that are extracted from translated sample $G(\boldsymbol{x},\boldsymbol{s}_i)$ and correspond to the target domain y'_i. Features $\boldsymbol{s}_i$ come from the domain y'_i. Minimizing $\mathcal{L}_{rec}$ ensures that translated images $G(\boldsymbol{x},\boldsymbol{s}_i)$ better reflect the styles of the target domain, making $G(\boldsymbol{x},\boldsymbol{s}_i)$ more like the target domain image.

Cycle consistency loss. To stabilize the model training and ensure translated images $G(\boldsymbol{x},\boldsymbol{s}_i)$ preserve some characteristics of the original sample $\boldsymbol{x}$ (such as expression and posture), an L_1 norm function is employed, named the cycle consistency loss [20,22],

$$\mathcal{L}_{cyc} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{s}_i,y}\left[\left\|G(G(\boldsymbol{x},\boldsymbol{s}_i),E_y(\boldsymbol{x})) - \boldsymbol{x}\right\|_1\right]. \tag{10}$$

In Equation (10), $G(\boldsymbol{x},\boldsymbol{s}_i)$ are translated samples corresponding to target domain y'_i. Style feature $E_y(\boldsymbol{x})$ is extracted from the original sample $\boldsymbol{x}$ and corresponds to the original domain y. Hence, minimizing $\mathcal{L}_{cyc}$ helps the translated new samples $G(\boldsymbol{x},\boldsymbol{s})$ preserve some characteristics of the original image $\boldsymbol{x}$.

Formulation of Style-Guided Image Translation. With the assistance of style-guided diversity loss, the objective function of the proposed SG-I2IT model can be designed as:

$$\begin{aligned}&\min_{G,M,E}\max_{D} V(M,E,G,D)\\ =&\min_{G,M,E}\max_{D} \mathcal{L}_{adv} - \lambda_{sd}\mathcal{L}_{sd} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cyc}\mathcal{L}_{cyc},\end{aligned} \tag{11}$$

where λ_{sd}, λ_{rec}, and λ_{cyc} denote hyper-parameters, controlling the weight of each term. Terms $\mathcal{L}_{adv}$, $\mathcal{L}_{sd}$, $\mathcal{L}_{rec}$, and $\mathcal{L}_{cyc}$ correspond to Eqs. (8), (7), (9), and (10), respectively.

Considering the discriminator D and other modules containing the generator G, the mapping network M, and the encoder E as two players, the model can be viewed as a two-player game. We train the discriminator D to maximize $V(M,E,G,D)$, aiming to identify the translated fake image from the real image. The other modules aim to translate plausible images that may confuse discriminator D by minimizing $V(M,E,G,D)$. It is worth noting that the style-guided diversity loss can assist the generator in better-capturing image styles, thereby generating higher quality and more diverse images for multi-domain I2I translation.

4. Experiments

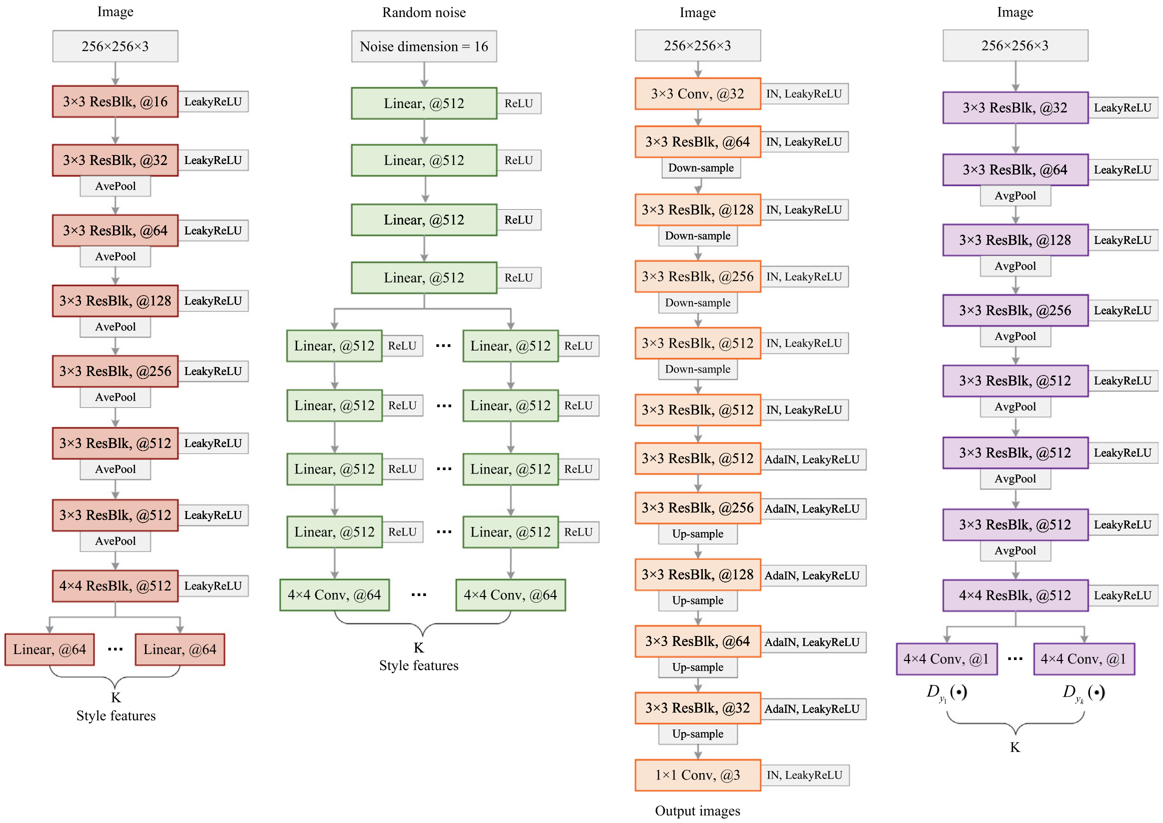
To evaluate the performance, we conduct extensive comparative experiments to compare the proposed SG-I2IT model with two state-of-the-art approaches, StarGAN-v2 [10] (in CVPR 2020) and MSGAN [17] (in CVPR 2019). Experimental details are as follows.

4.1. Datasets

We evaluate our method against baselines on two widely adopted datasets, a high-quality CelebFaces Attributes dataset (CelebA-HQ) [6] and a high-quality Animal Faces image dataset (AFHQ) [10], which follows the Ref. [10].

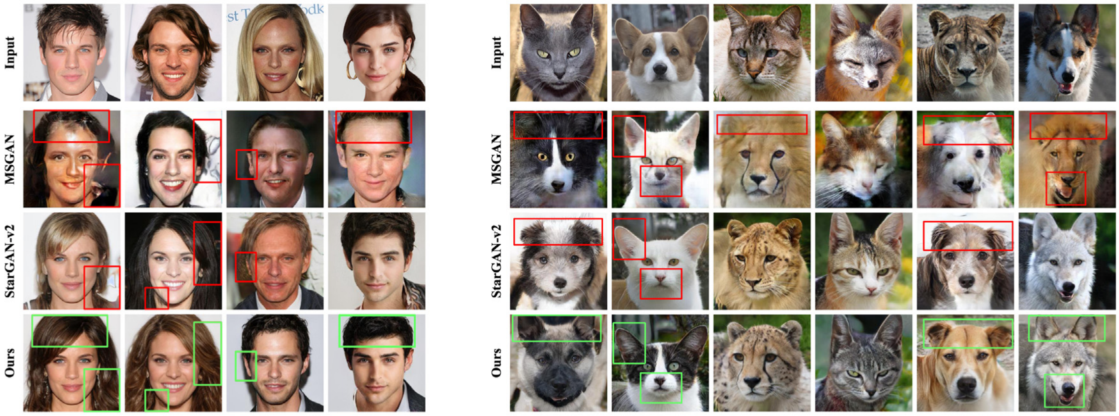
CelebA-HQ. It is a high-quality version of the CelebFaces Attributes (CelebA) dataset [6], which contains 30,000 face images with a resolution of 1024 × 1024. The images in this dataset can be split into two visual domains, male and female, according to gender. The training set contains 28,000 images, including 17,943 female images and 10057 male images, while the test set has 2,000 images, including 1,000 females and 1,000 males.

AFHQ. To further compare translation performance across multiple domains, our experiments adopt a recently released three-domain dataset AFHQ [10]. It contains 15,000 animal faces with a resolution of 512 × 512. The images in this dataset can be split into three visual domains: cat, dog, and wildlife. Each domain has 5,000 images, of which 4,500 are selected as the training set, and the remaining 500 images are used as the test set.



(a) Image encoder (b) Mapping network (c) Generator (d) Discriminator

Fig. 4. Detailed neural network architecture of four members: the image encoder, the mapping network, the generator, and the discriminator.



(a) CelebA-HQ (b) AFHQ

Fig. 5. Noise-based visual comparison on both two datasets. The other rows display the images translated by different methods using randomly sampled noise codes. In Fig. 5(a), the left two columns and the right two columns show the translated images from male to female and from female to male, respectively. In Fig. 5(b), each column from left to right displays the translations in the following order: cat-to-dog, dog-to-cat, cat-to-wildlife, wildlife-to-cat, wildlife-to-dog, and dog-to-wildlife.

4.2. Evaluation metrics

We evaluate translated images with two widely used evaluation metrics, including Fréchet inception distance (FID) [42] for image quality and learned perceptual image patch similarity (LPIPS) [43] for image diversity.

FID. It evaluates image quality by measuring the similarity between the translated image distribution and the real image distribution with the Fréchet distance. Image distributions are extracted from the image sets through the ImageNet pre-trained Inception-V3 [42]. If we mark the translated image distribution as $N(\mu_{ge}, \Sigma_{ge})$ and the real one as $N(\mu_{re}, \Sigma_{re})$, the FID value is calculated by

$$FID = d^2(N(\mu_{ge}, \Sigma_{ge}), N(\mu_{re}, \Sigma_{re})) \\ = \|\mu_{ge} - \mu_r\|_2^2 + Tr(\Sigma_{ge} + \Sigma_{re} - 2(\Sigma_{ge}\Sigma_{re})^{\frac{1}{2}}).$$

Thus, a lower FID value means that the translated images are more plausible and more similar to the real images.

LPIPS. We employ LPIPS to evaluate image diversity. It measures the image diversity by calculating the average perceptual

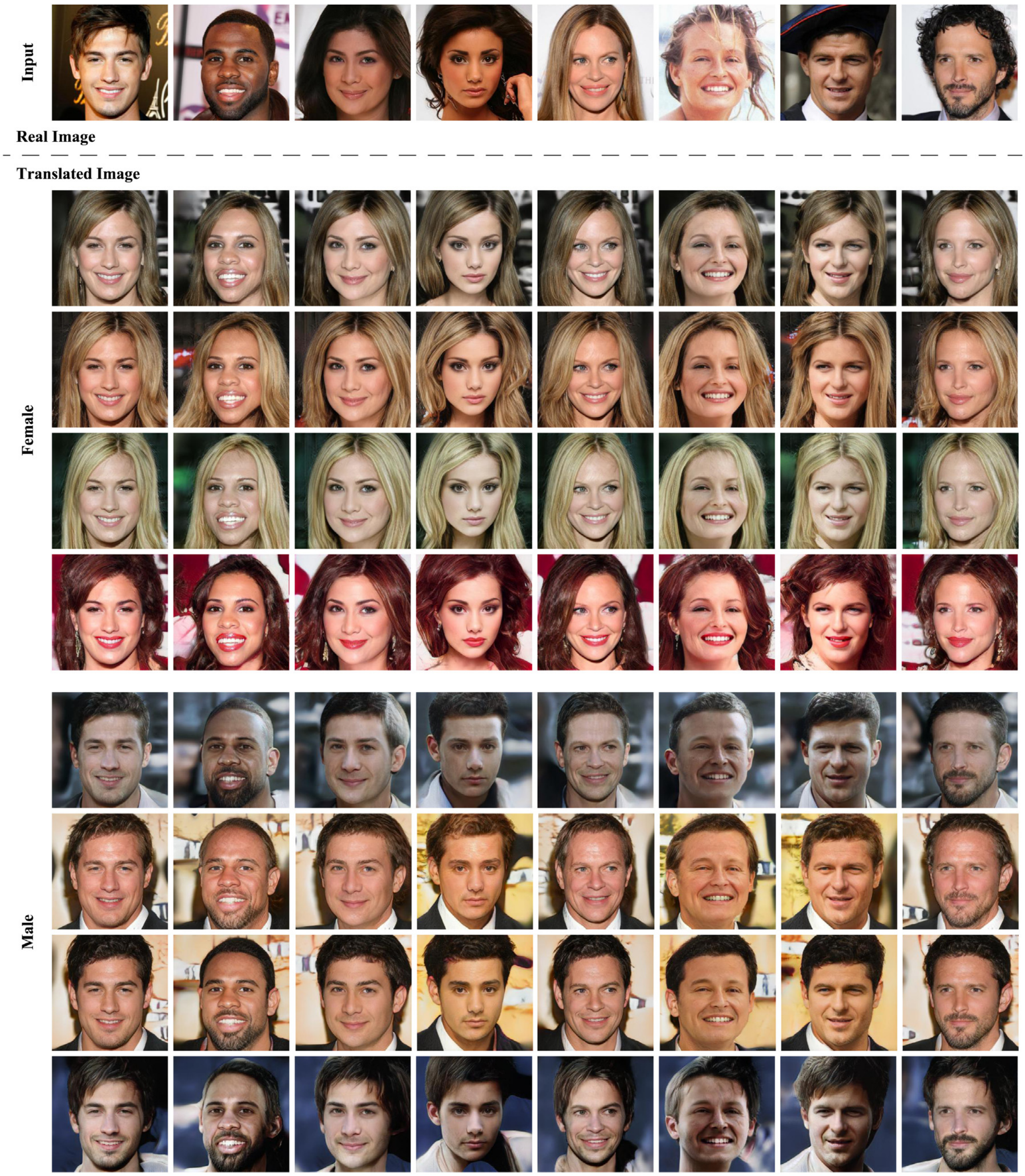


Fig. 6. The diverse samples translated by our model on CelebA-HQ dataset. Input images in the first row are real images. The rest images are our translated results based on the input images, including four rows of female translations and four rows of male translations. We preserve the identity and pose of the input images, while transforming their high-level semantics (such as makeup, hair texture, and complexion) into other diverse styles.

pairwise distances (PPDs) between all images. The pre-trained AlexNet [44] extracts features for computing the perceptual pairwise distance. PPD is an L_1 norm distance between every two features. LPIPS is the average PPD among all feature pairs. Thus, a higher LPIPS value suggests better diversity of the translated images.

4.3. Experiment setup

4.3.1. Architecture

The proposed method follows the StarGAN-v2 framework. The detailed neural network architecture is displayed in Fig. 4. For style extractor, image encoder E consists of six pre-trained residual blocks activated by Leaky ReLU functions and one fully-connected domain-specific multi-branch output layer. At the same time, the mapping network M is composed of six 512 dimensional fully-connected layers with ReLU activation functions and one fully-connected domain-specific multi-branch output layer. The generator G mainly consists of two parts: four layers of downsampling blocks normalized by IN and four layers of upsampling blocks normalized by AdaIN. Every block is activated by Leaky ReLU. The discrimination D is implemented through six pre-trained residual blocks with Leaky ReLU and a multi-branch fully connected output layer.

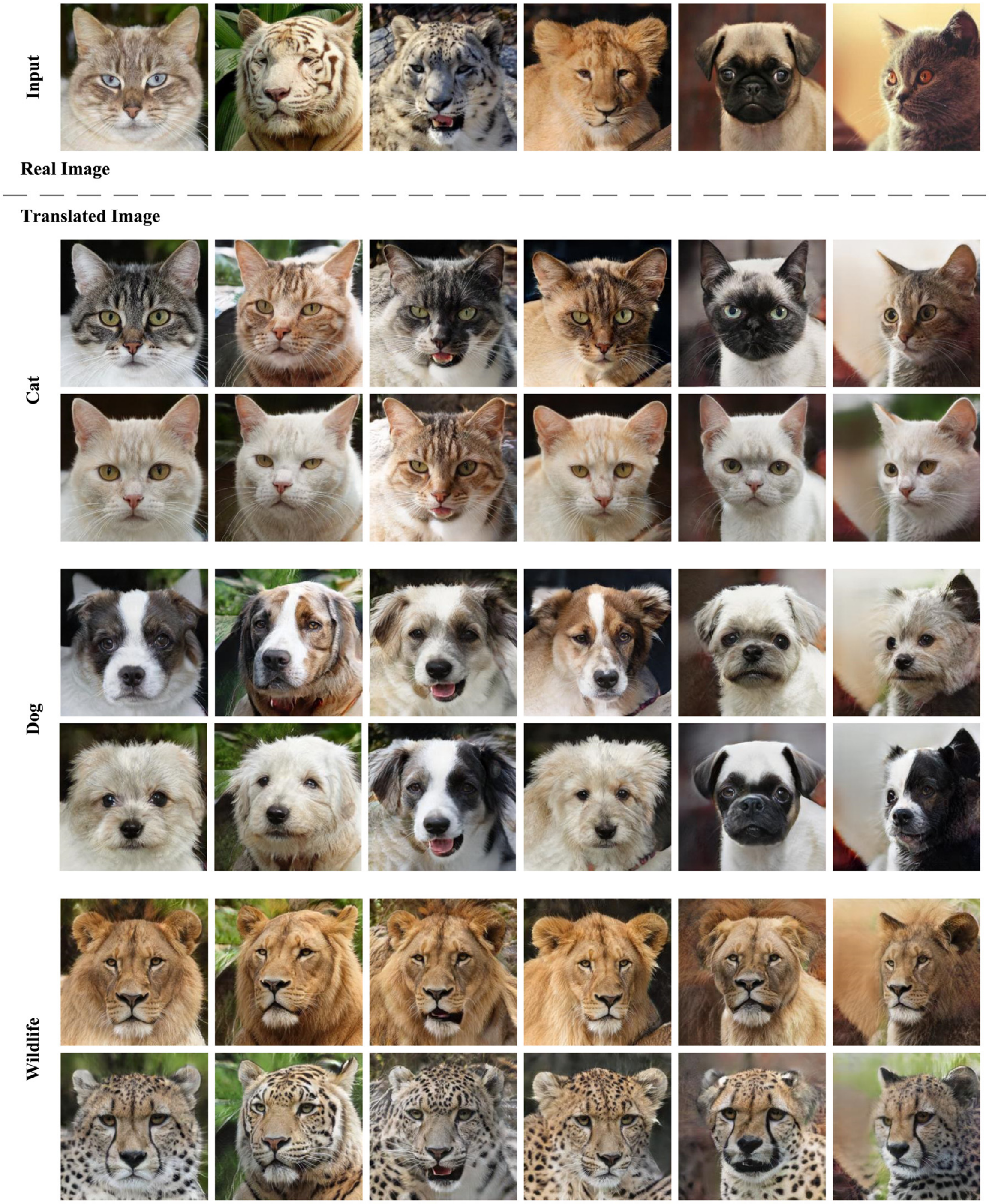


Fig. 7. The diverse samples translated by our model on AFHQ dataset. Input images in the first row are real images. The rest images are our translated results based on the input images, including two rows of cats, two rows of dogs, and two rows of wildlife. We preserve the pose and gaze of the input images, while transforming their high-level semantics (such as breed, hair texture, and color) into other diverse styles.

4.3.2. *Implementation details*

We implement the proposed model, StarGAN-v2, and MSGAN with PyTorch 1.4.0. All models are trained on an NVIDIA GeForce RTX 2080Ti GPU with 8 GB memory. For a fair comparison, all training images are resized into 256 × 256 resolution, following our baseline [10]. The maximum batch size of the model on GeForce RTX 2080Ti GPU can is 4, just half of that in Ref. [10], which causes some discrepancies in the numerical results. The iteration number is set to 100K for all models. It takes about two days to train a model with our equipment. There are three hyper-parameters, λ_{sd}, λ_{rec}, and λ_{cyc}, in our objective. These parameters are tuned by the control variable method. We choose hyper-parameter values ranging from 0.2 to 2 with a stride of 0.2, and determine hyper-parameters $\{\lambda_{sd} = 1, \lambda_{rec} = 1.2, \lambda_{cyc} = 1\}$ for CelebA-HQ and $\{\lambda_{sd} = 1, \lambda_{rec} = 0.3, \lambda_{cyc} = 0.2\}$ for AFHQ according to these empirical experiences.

4.4. *Experimental results*

Since the extracted style features $\boldsymbol{s}_i$, as shown in Section 3.1, are extracted from random noises $\boldsymbol{z}$ or from reference images $\boldsymbol{x}$,

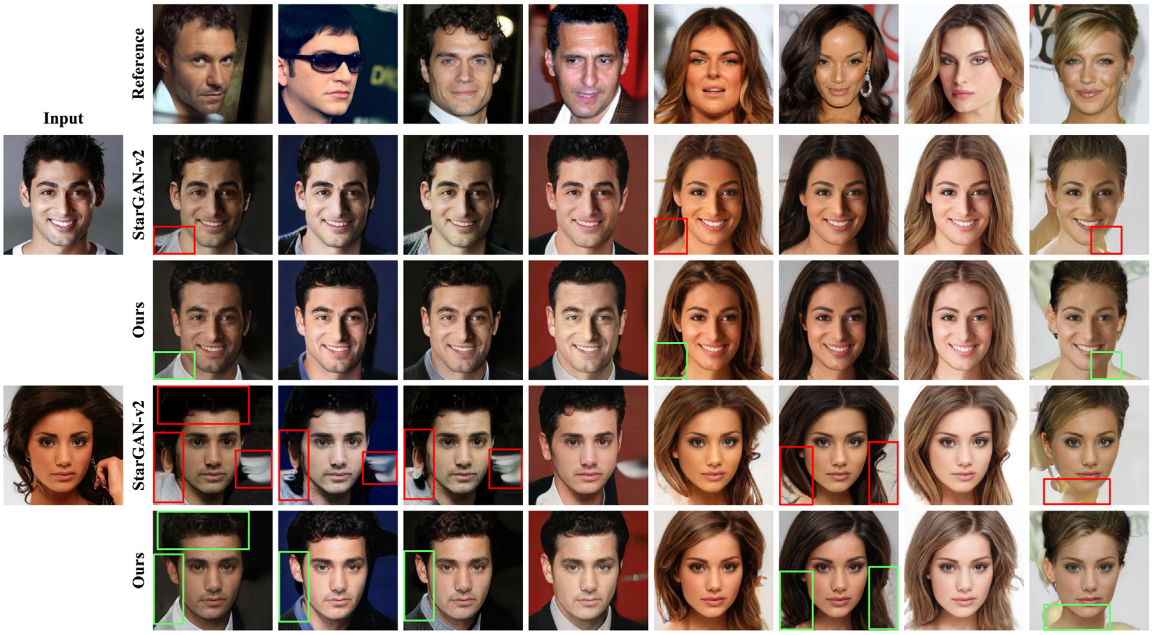


Fig. 8. Reference-based visual comparison on CelebA-HQ dataset. Input images and reference images in the leftmost column and the uppermost row are real images. Each method translates input images to the target domains, reflecting the styles of the reference images. The even-numbered rows display the translated images of StarGAN-v2, and the remaining rows present the translated images of our method.

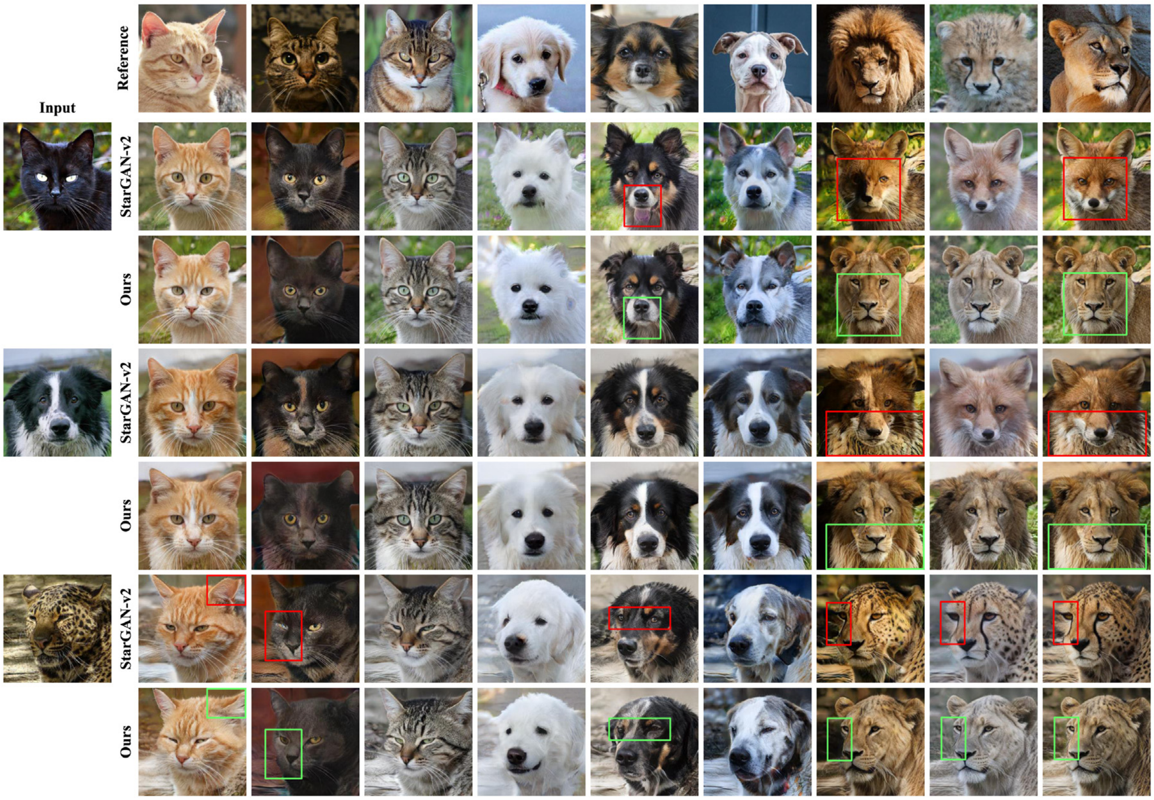


Fig. 9. Reference-based visual comparison on AFHQ dataset. The input images in the leftmost column and reference images in the uppermost row are real images. Each method translates the input image to the target domains, reflecting the styles of the reference images. The even rows present the images generated by the StarGAN-v2 model, and the rest rows show the images generated by our method.

we evaluate the proposed method from two perspectives: noise-based translation and reference-based translation. In addition, a user evaluation of the user preference between the proposed method and the state-of-the-art approaches is provided.

4.4.1. Noise-based results

For the qualitative comparison of noise-based translation, Fig. 5 shows some translation images of the proposed method and two competing approaches, MSGAN and StarGAN-v2. Each

Table 2
Quantitative evaluation of noise-based translations on CelebA-HQ and AFHQ datasets with FID and LPIPS. The lower the FID, the higher the quality of the translations. The higher the LPIPS, the more diverse the translations.

Methods	CelebA-HQ		AFHQ	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
Real images	14.8	–	12.9	–
MUNIT [19]	31.4	0.363	41.5	0.511
DRIT [8]	52.1	0.178	95.6	0.326
MSGAN [17]	33.1	0.389	61.4	**0.517**
StarGAN-v2 [10]	20.6	0.369	27.2	0.441
Ours	**19.4**	**0.463**	**25.9**	0.497

Table 3
Quantitative evaluation of reference-based translations on CelebA-HQ and AFHQ datasets with FID and LPIPS.

Methods	CelebA-HQ		AFHQ	
	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
Real images	14.8	–	12.9	–
MUNIT [19]	107.1	0.176	223.9	0.199
DRIT [8]	53.3	0.311	114.8	0.156
MSGAN [17]	39.6	0.312	69.8	0.375
StarGAN-v2 [10]	24.1	0.335	32.1	0.370
Ours	**21.7**	**0.363**	**30.3**	**0.392**

method provides several translated images in rows that are translated from different input images (in the first row) with the features extracted from random noise. Fig. 5(a) presents the visual results on the CelebA-HQ dataset. Obviously, the images translated by MSGAN do not have smooth and complete contours, especially facial shapes and hairstyles. Its generated images have many artifacts, which means that MSGAN cannot generate realistic face images. Compared with StarGAN-v2, the images translated by our method have more clear contours, such as more refined hairstyles and more clear backgrounds.

Fig. 5(b) presents a visual comparison of translated images on the AFHQ dataset. AFHQ is a more challenging dataset, for its three domains bring relatively significant differences. For the AFHQ dataset, MSGAN could learn some recognizable features of animals, such as ears, eyes, and noses. But these features are not correctly combined to form an animal with a realistic anatomical structure. The StarGAN-v2 method could generate relatively realistic images. Still, in comparison, our model not only generates realistic images but also preserves the contours of input images more accurately, such as ears and chins. The visual results on the CelebA-HQ dataset and the AFHQ dataset both show that our method could translate realistic images with higher quality than other methods.

Figs. 6 and 7 provide some image samples translated by our method on the CelebA-HQ and AFHQ datasets, respectively. In Fig. 6, we preserve the identity and pose of the input images and translate them into diverse appearances with different styles, such as their hairstyle, hair color, complexion, etc. In Fig. 7, we reserve the pose and gaze of input images and translate the input images into three visual domains (cat, dog, and wildlife). Each image can be translated into diverse appearances with different styles, such as their breed, hair texture, and color. The diverse plausible translations of our method on the two datasets demonstrate that our method can effectively make the images appear in various styles with high quality.

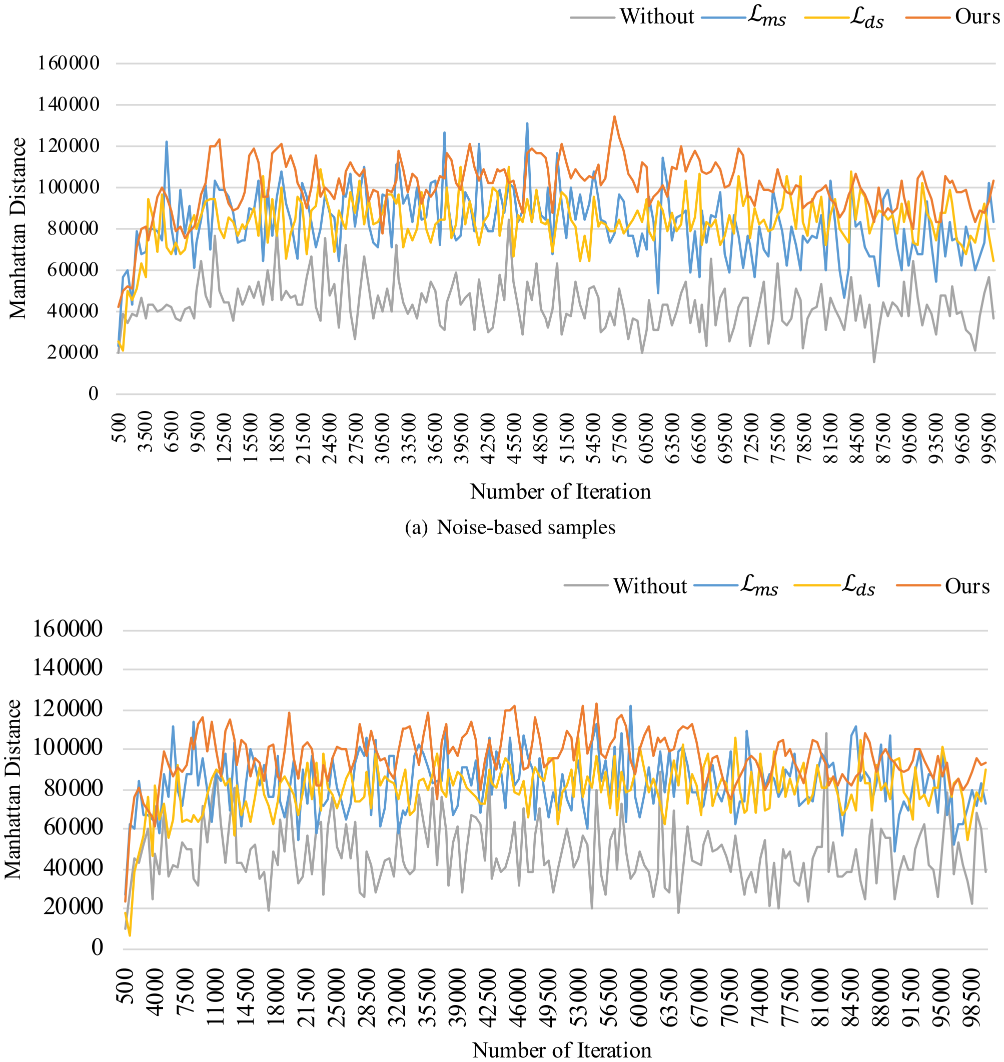
Quantitative evaluation results of noise-based translations are provided in Table 2. As shown in the table, our method is quantitatively compared with four image translation models, MUNIT, DRIT, MSGAN, and Star-GAN-v2. For quality evaluation, our method obtains the lowest FID scores among all methods on both two datasets. The best FID results validate that images translated by our method are the most realistic, demonstrating that our method performs the best in image quality compared to other methods. For diversity evaluation, the LPIPS score of our method on CelebA-HQ dataset is the best among all methods, which means that our method is superior to other methods in image diversity on CelebA-HQ dataset. On AFHQ dataset, our method gets a higher LPIPS score than StarGAN-v2 and DRIT, indicating that our method outperforms StarGAN and DRIT in image diversity. Regarding the high LPIPS scores of MSGAN and MUNIT on the AFHQ dataset, we speculate from the FID scores and visual comparisons that this is due to their inability to effectively reflect image features resulting in large differences between the generated images. The quantitative comparison results confirm that our method significantly outperforms other methods concerning the quality and diversity of the translated samples.

4.4.2. Reference-based results

For the visual comparison of reference-based translations, the proposed method is mainly compared with StarGAN-v2, as the above visual results already show that the translation images of MSGAN are blurry and have many artifacts. Fig. 8 provides some translated images of these two methods on CelebA-HQ dataset. Except for the first row, each row of translated images is produced from an input image (in the first column) referring to different reference images (in the first row). For a normal image without other interference (such as the male image in the first column), both methods can effectively translate realistic images based on reference images. But for the original images with interference (such as the female images with hands in the first column), it is difficult for StarGAN-v2 to overcome the interference information (hands) to generate realistic images. Its translated images have a lot of artifacts. Our method can effectively overcome the interference to translate more realistic images. Artifacts in the background of generated images are also much less. Our method could effectively overcome the interference to translate realistic images.

Fig. 9 presents some reference-based translated samples on AFHQ dataset. The images translated by both methods are realistic. But the proposed method performs better on some details, such as animal face contours, marked in Fig. 8 (poor in red and better in green). The visual comparisons on the AFHQ dataset also demonstrate the superiority of our method in translated image quality. However, in most image translation models, including the model proposed in this paper, the background of the generated image has some artifacts. We conjecture that it is caused by the fact that the translation model mainly focuses on the style transfer of the image target and ignores the processing of the image background.

Table 3 displays the FID and LPIPS results of reference-based samples translated by different methods, including MUNIT, DRIT, MSGAN, StarGAN-v2, and the proposed model. The experiments are performed on two datasets, CelebA-HQ and AFHQ. For quality comparison, our method obtains the significantly lowest FID values on both datasets. Our FID values are 2.4 and 1.8 points on the CelebA-HQ and AFHQ datasets, respectively, which are lower than the state-of-the-art method StarGAN-v2. The lowest FID scores indicate that our method performs best among all approaches in terms of image quality. For diversity evaluation, the proposed method also obtains the best LPIPS values on both two datasets, indicating that, for reference-based translation, our method also performs the best among all methods in terms of image diversity. The best quantitative comparison results in Table 3 demonstrate that the proposed method outperforms other methods with regard to image quality and diversity for reference-based translation.



(a) Noise-based samples

(b) Reference-based samples

Fig. 10. Manhattan distance of image samples trained on AFHQ dataset. Different colors represent the results obtained by the model with different diversity regularization terms: orange represents the results of the model with our style-guided diversity regularization term; yellow represents the results of the diversity-sensitive loss function; blue represents the results of the model using mode-seeking regularization; gray represents the results obtained by the model without any diversity function.

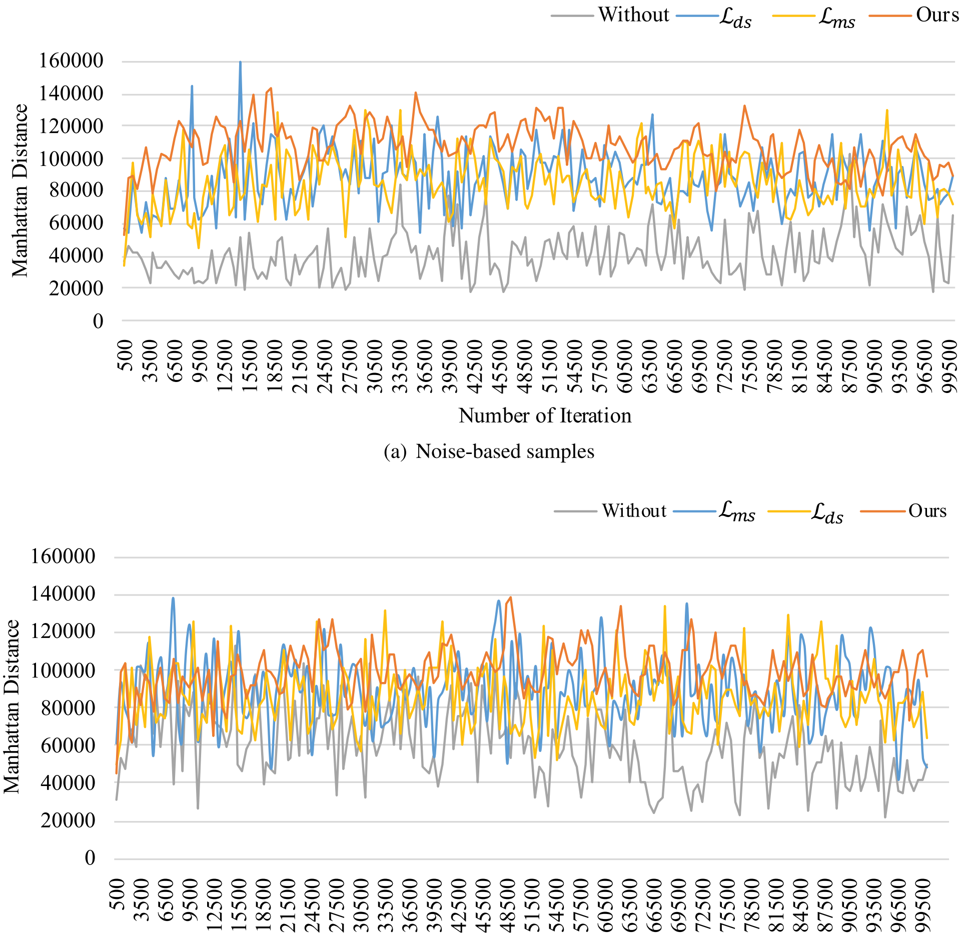
4.4.3. Diversity analysis

Currently, many I2I translation models enhance the diversity of translated images by adopting a regularization term that maximizes the distance between images [10,17]. Our proposed method is a novel technique to maximize image distance, thus enhancing image diversity for multi-domain I2I translation. Therefore, we demonstrate the effectiveness of our technique in enhancing image diversity by computing image distances. Such image distance is calculated by the Manhattan distance [45]. The proposed style-guided diversity regularization term $\mathcal{L}_{sd}$ is mainly compared to the diversity-sensitive loss $\mathcal{L}_{ds}$ [10] and mode-seeking regularization term $\mathcal{L}_{ms}$ [17]. Additionally, we also calculated the image distance obtained by the model without any diversity loss. For a fair comparison, we conduct all experiments on the StarGAN model. They are compared on the AFHQ and CelebA-HQ datasets, respectively. Translated images are all color images with 256 × 256 pixels. We calculated the average of Manhattan distances for six pairs of translated images. All models are trained 100,000 iterations and output a distance result every 500 iterations.

Table 4
Pixel-level image distances on the AFHQ dataset. 'S.D.' denotes the standard deviation of the image distance. The best results are highlighted in bold. The second best results are underlined.

Methods	Noise-based		Reference-based	
	Mean ↑	S.D. ↓	Mean ↑	S.D. ↓
Without	0.221	0.067	0.242	0.079
$\mathcal{L}_{ms}$	0.421	0.082	0.420	0.075
$\mathcal{L}_{ds}$	0.425	**0.051**	0.404	0.066
Ours	**0.507**	0.065	**0.489**	**0.064**

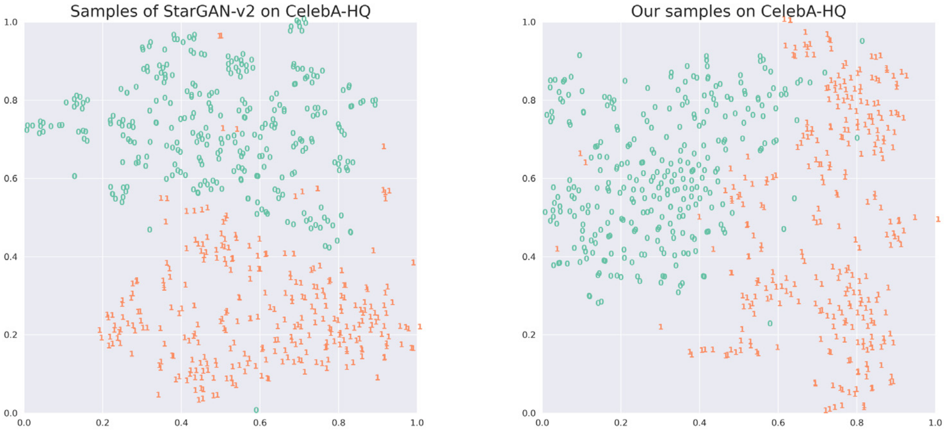
Fig. 10 presents the generated image distances from the model trained on the AFHQ dataset. Figs. 10(a) and 10(b) show the image distance results of the model guided by noises and reference images, respectively. In the figures, the red, yellow, and blue solid lines are higher than the gray ones. It shows that, compared with the model without diversity loss, diversity loss functions can effectively help the model generate images that are different from each other, thereby increasing the diversity of samples. Additionally, the red line is slightly higher than the yellow line and the blue line overall, indicating that the diversity regularization term $\mathcal{L}_{sd}$ is better than mode-seeking regularization term $\mathcal{L}_{ms}$



(a) Noise-based samples

(b) Reference-based samples

Fig. 11. Manhattan distance of image samples trained on CelebA-HQ dataset. Different colors represent the results obtained by the model with different diversity regularization terms: orange represents the results of the model with our style-guided diversity regularization term; yellow represents the results of the diversity-sensitive loss function; blue represents the results of the model using mode-seeking regularization, and gray represents the results obtained by the model without any diversity function.



(a) Translated samples of StarGAN-v2

(b) Translated samples of the proposed method

Fig. 12. Visualization of the feature space of images generated on the dataset CelebA-HQ via t-SNE. Each data point represents an image. “0” represents a female image; “1” represents a male image. There are 300 image samples per class.

Table 5
Pixel-level image distances on the CelebA-HQ dataset. 'S.D.' denotes the standard deviation of the image distance. The best results are highlighted in bold. The second best results are underlined.

Methods	Noise-based		Reference-based	
	Mean ↑	S.D. ↓	Mean ↑	S.D. ↓
Without	0.210	0.075	0.305	0.101
$\mathcal{L}_{ms}$	0.447	0.090	0.454	0.092
$\mathcal{L}_{ds}$	0.435	0.086	0.434	0.088
Ours	**0.536**	**0.071**	**0.502**	**0.076**

and diversity-sensitive loss function $\mathcal{L}_{ds}$. That is to say, $\mathcal{L}_{sd}$ can more effectively encourage the model to enhance the diversity of generated images.

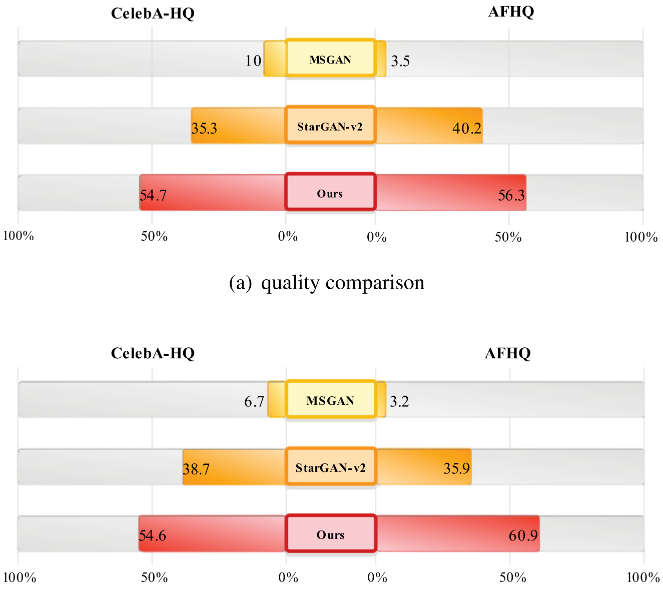
Table 4 gives the quantified results of translated image distance on the AFHQ dataset. These results are pixel-level image distances. The larger the image distance, the better the image diversity. The comparison of the mean values shows that the proposed style-guided diversity loss function achieves the maximum image distance among all methods. It shows that, among all approaches, the proposed method performs the best in encouraging the model to increase image diversity. Additionally, comparing the standard deviation (S.D.) of image distance, the smaller S.D., the more stable the model training. In noise-based image translation, $\mathcal{L}_{ds}$ achieves the minimum variance, and $\mathcal{L}_{sd}$ is the second best. In reference-based image translation, our method achieves the minimum variance. Overall, $\mathcal{L}_{sd}$ can more effectively encourage the model to translate diverse images without reducing the stability of the model.

Fig. 11 shows the image distances of the model trained on the CelebA-HQ dataset. Figs. 11(a) and 11(b) show the image distance results of noise-based image translation and reference-based image translation, respectively. Comparing these two subgraphs, it is evident that diversity loss functions are more effective in improving image diversity for noise-based image translation. Among all diversity regularization terms, our method performs the best. The quantified image distance results on the CelebA-HQ dataset are shown in Table 5. The table shows that the proposed style-guided diversity regularizer term obtains the most significant mean distances in noise-based and reference-based image translation, 0.536 and 0.502, respectively. They are 0.089 and 0.048 higher than the mean distances of the second best method $\mathcal{L}_{ms}$, respectively. It shows that the proposed diversity regularization term still outperforms other methods in enhancing image diversity on the CelebA-HQ dataset.

Furthermore, we visualize the feature distribution of translated images via t-SNE. Fig. 12 presents the feature space distributions of the generated images on the CelebA-HQ dataset. The CelebA-HQ dataset consists of two domains: female and male. For each image domain, 300 generated samples are randomly selected for experiments. Sub Fig. 12(a) shows the feature distribution of the images generated by the StarGAN-v2 model. Sub Fig. 12(b) visualizes the feature distribution of the images generated by the proposed method. In sub Fig. 12(a), multiple samples are clustered into groups, while this phenomenon is rare in sub Fig. 12(b). Overall, the sample distribution in sub Fig. 12(b) is more spread out than that in sub Fig. 12(a). It shows that the proposed method can indeed effectively increase the image distance.

4.4.4. User study

Besides the qualitative and quantitative comparisons, we also evaluate the performance of our method through a user study on CelebA-HQ and AFHQ datasets. There are 40 users taking part in our study. Everyone is provided with 15 groups of translated images. Every group of images contains three shuffled translated images coming from different models. This study assesses translated images from two aspects: quality and style reflection. For quality comparison, users evaluate each image's realism through specific questions about the facial features, contours, and background. For style reflection comparison, users mainly evaluate through questions related to the similarity between the style of translated images and that of real images. They are also asked to select the highest quality image and the best one reflecting the style of the reference image from each group of images.



(a) quality comparison

(b) style reflection comparison

Fig. 13. User Votes (%) on the most preferred method concerning the visual quality and style reflection. Our method outperforms the baselines with obvious advantages in both aspects.

Fig. 13 presents the vote results of the user study on CelebA-HQ and AFHQ datasets regarding image quality comparison and style reflection comparison. The vote results show that our method is approved by most users. On the one hand, our translations are obviously considered to outperform other baselines not only in quality but also in style reflection. On the other hand, comparing the results on two datasets, our method is perceived to perform better than baselines on the complex three-domain dataset AFHQ. The results show that the proposed method can better extract styles and render them into input images to translate more realistic and diverse images, compared with other baselines.

5. Conclusion

We have proposed a multi-domain I2I translation method SG-I2IT with a novel loss function named style-guided diverse loss. The proposed loss takes full advantage of extracted style features to maximize the distance between different translated images. As a result, the model can efficiently capture various image styles. Extensive experiments have been conducted on a two-domain dataset, CelebA-HQ, and a three-domain dataset, AFHQ. The results demonstrate that the proposed method performs better than two state-of-the-art methods, MSGAN and StarGAN-v2, in terms of image quality and diversity. In the user study, the proposed method is also approved by most users, which shows that the proposed method can capture image style more effectively. In the future, we will attempt to design a specific feature extractor for image translation to achieve the combination of noise-based and reference-based methods. Besides, there are some artifacts in the images translated by many I2I translation methods. Addressing the artifacts of translation samples will also be an essential task in our future work.

CRediT authorship contribution statement

Tingting Li: Methodology, Software, Data curation, Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Huan Zhao:** Funding acquisition, Writing – review & editing, Validation, Supervision. **Jing Huang:** Conceptualization, Resources, Data curation, Investigation, Writing – original draft, Writing – review & editing. **Keqin Li:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2020YFB1713400 and the Special Project of Foshan Science and Technology Innovation Team, China under Grant FS0AA-KJ919-4402-0069.

References

[1] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems, NeurIPS, Long Beach, CA, USA, 2017, pp. 700–708.

[2] A. Ul Hassan, H. Ahmed, J. Choi, Unpaired font family synthesis using conditional generative adversarial networks, Knowl.-Based Syst. 229 (2021) 107304.

[3] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, Venice,Italy, 2017, pp. 2242–2251.

[4] L. Yan, W. Zheng, C. Gou, F.-Y. Wang, IPGAN: Identity-preservation generative adversarial network for unsupervised photo-to-caricature translation, Knowl.-Based Syst. 241 (2022) 108223.

[5] M.A. Hedjazi, Y. Genc, Efficient texture-aware multi-GAN for image inpainting, Knowl.-Based Syst. 217 (2021) 106789.

[6] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: Proceedings of the International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, 2018, pp. 1–12.

[7] P. Li, S. Tu, L. Xu, GAN flexible lmser for super-resolution, in: Proceedings of the ACM International Conference on Multimedia, MM, Nice, France, 2019, pp. 756–764.

[8] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, M.-H. Yang, Diverse image-to-image translation via disentangled representations, in: Proceedings of the European Conference on Computer Vision, ECCV, Munich, Germany, 2018, pp. 36–52.

[9] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, 2017, pp. 1125–1134.

[10] Y. Choi, Y. Uh, J. Yoo, J. Ha, StarGAN v2: Diverse image synthesis for multiple domains, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Seattle, WA, USA, 2020, pp. 8185–8194.

[11] M. Shao, Y. Zhang, Y. Fan, W. Zuo, D. Meng, IIT-GAT: Instance-level image transformation via unsupervised generative attention networks with disentangled representations, Knowl.-Based Syst. 225 (2021) 107122.

[12] Y. Xiao, H. Zhao, T. Li, Learning class-aligned and generalized domain-invariant representations for speech emotion recognition, IEEE Trans. Emerg. Top. Comput. Intell. 4 (4) (2020) 480–489.

[13] Y. Xiao, W. Lei, L. Lu, X. Chang, X. Zheng, X. Chen, CS-GAN: Cross-structure generative adversarial networks for Chinese calligraphy translation, Knowl.-Based Syst. 229 (2021) 107334.

[14] P. Zhang, B. Zhang, D. Chen, L. Yuan, F. Wen, Cross-domain correspondence learning for exemplar-based image translation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Seattle, WA, USA, 2020, pp. 5142–5152.

[15] T. Li, H. Zhao, S. Wang, J. Huang, Style-guided image-to-image translation for multiple domains, in: Proceedings of the ICMR 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, MMPT, Taipei, Taiwan, 2021, pp. 28–36.

[16] J. Wu, Z. Huang, J. Thoma, D. Acharya, L. Van Gool, Wasserstein divergence for GANs, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, Cham, 2018, pp. 673–688.

[17] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, M.-H. Yang, Mode seeking generative adversarial networks for diverse image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA, 2019, pp. 1429–1437.

[18] L. Zhang, L. Zhao, High-quality face image generation using particle swarm optimization-based generative adversarial networks, Future Gener. Comput. Syst. 122 (2021) 98–104.

[19] X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Proceedings of the European Conference on Computer Vision, ECCV, Munich, Germany, 2018, pp. 179–196.

[20] J.-Y. Zhu, R. Zhang, D. Pathak, T. Dar-rell, A.A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, in: Advances in Neural Information Processing Systems, NeurIPS, Long Beach, CA, USA, 2017, pp. 465–476.

[21] H.Y. Lee, H.Y. Tseng, Q. Mao, J.B. Huang, M.H. Yang, DRIT++: Diverse image-to-image translation via disentangled representations, Int. J. Comput. Vis. 128 (10) (2020) 2402–2417.

[22] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, A.C. Courville, Augmented CycleGAN: Learning many-to-many mappings from unpaired data, in: Proceedings of the International Conference on Machine Learning, ICML, Stockholmsmässan, Stockholm, Sweden, 2018, pp. 195–204.

[23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Salt Lake City, UT, USA, 2018, pp. 8789–8797.

[24] T. Kim, M. Cha, H. Kim, J. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: Proceedings of the International Conference on Machine Learning, ICML, Sydney, NSW, Australia, 2017, pp. 1857–1865.

[25] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, B. Schuller, Speech-based diagnosis of autism spectrum condition by generative adversarial network representations, in: Proceedings of the International Conference on Digital Health, ICDH, London, United Kingdom, 2017, pp. 53–57.

[26] Y. Yang, G. Xie, J. Wang, J. Zhou, Z. Xia, R. Li, Intrusion detection for in-vehicle network by using single GAN in connected vehicles, J. Circuits Syst. Comput. 30 (01) (2021) 2150007.

[27] M. Chen, X. Shi, Y. Zhang, D. Wu, M. Guizani, Deep feature learning for medical image analysis with convolutional autoencoder neural network, IEEE Trans. Big Data 7 (4) (2021) 750–758.

[28] J. Fan, S. Wang, P. Yang, Y. Yang, Multi-view facial expression recognition based o n multitask learning and generative adversarial network, in: Proceedings of the IEEE International Conference on Industrial Informatics, INDIN, China, 2020, pp. 573–578.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, NeurIPS, Montréal, Canada, 2014, pp. 2672–2680.

[30] P. Wang, H. Fu, X. Li, J. Guo, Z. Lv, R. Di, Multi-feature fusion tracking algorithm based on generative compression network, Future Gener. Comput. Syst. 124 (2021) 206–214.

[31] J. Wang, R. Li, R. Li, K. Li, H. Zeng, G. Xie, L. Liu, Adversarial de-noising of electrocardiogram, Neurocomputing 349 (2019) 212–224.

[32] Y. Yang, F. Nan, P. Yang, Q. Meng, Y. Xie, D. Zhang, K. Muhammad, GAN-based semi-supervised learning approach for clinical decision support in health-IoT platform, IEEE Access 7 (2019) 8048–8057.

[33] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, in: Proceedings of the International Conference on Learning Representations, ICLR, Banff, Canada, 2014, pp. 1–9.

[34] H. Zhao, T. Li, Y. Xiao, Y. Wang, Improving multi-agent generative adversarial nets with variational latent representation, Entropy 22 (9) (2020) 1055.

[35] Y. Yang, Z. Ma, F. Nie, X. Chang, A.G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, Int. J. Comput. Vis. 113 (2) (2015) 113–127.

[36] W. Liu, X. Chang, L. Chen, D. Phung, X. Zhang, Y. Yang, A.G. Hauptmann, Pair-based uncertainty and diversity promoting early active learning for person re-identification, ACM Trans. Intell. Syst. Technol. 11 (2) (2020) 21:1–21:15.

[37] J.M.J. Valanarasu, H. Zhang, J. Zhang, Y. Wang, Z. Lin, J. Echevarria, Y. Ma, Z. Wei, K. Sunkavalli, V. Patel, Interactive portrait harmonization, 2022, pp. 1–16, arXiv Preprint, arXiv:2203.08216.

[38] Y. Jiang, H. Zhang, J. Zhang, Y. Wang, Z.L. Lin, K. Sunkavalli, S. Chen, S. Amirghodsi, S. Kong, Z. Wang, SSH: A self-supervised framework for image harmonization, in: IEEE/CVF International Conference on Computer Vision, ICCV, Montreal, QC, Canada, October 10-17, 2021, pp. 4812–4821.

[39] X. Cun, C.-M. Pun, Improving the harmony of the composite image by spatial-separated attention module, IEEE Trans. Image Process. 29 (2020) 4759–4771.

[40] W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, L. Zhang, DoveNet: Deep image harmonization via domain verification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Seattle, WA, USA, 2020, pp. 8391–8400.

[41] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Long Beach, CA, USA, 2019, pp. 4396–4405.

[42] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local Nash equilibrium, in: Advances in Neural Information Processing Systems, NeurIPS, Long Beach, CA, USA, 2017, pp. 6626–6637.

[43] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Salt Lake City, UT, USA, 2018, pp. 586–595.

[44] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.

[45] V.H. Kamble, M.P. Dale, Machine learning approach for longitudinal face recognition of children, in: Machine Learning for Biometrics, 2022, pp. 1–27.