Style-Guided and Disentangled Representation for Robust Image-to-Image Translation

Jaewoong Choi, Daeha Kim, Byung Cheol Song

Department of Electrical and Computer Engineering, Inha University, Incheon 22212, South Korea
chlwodnd500@naver.com, kdhht5022@gmail.com, bcsong@inha.ac.kr

Abstract

Recently, various image-to-image translation (I2I) methods have improved mode diversity and visual quality in terms of neural networks or regularization terms. However, conventional I2I methods relies on a static decision boundary and the encoded representations in those methods are entangled with each other, so they often face with 'mode collapse' phenomenon. To mitigate mode collapse, 1) we design a so-called style-guided discriminator that guides an input image to the target image style based on the strategy of flexible decision boundary. 2) Also, we make the encoded representations include independent domain attributes. Based on two ideas, this paper proposes Style-Guided and Disentangled Representation for Robust Image-to-Image Translation (SRIT). SRIT showed outstanding FID by 8%, 22.8%, and 10.1% for CelebA-HQ, AFHQ, and Yosemite datasets, respectively. The translated images of SRIT reflect the styles of target domain successfully. This indicates that SRIT shows better mode diversity than previous works.

Introduction

Generative adversarial network (GAN) (Goodfellow et al. 2014) capable of generating high-fidelity images from noise is getting a lot of attention from image synthesis (Brock, Donahue, and Simonyan 2018; Lučić et al. 2019), super-resolution (Ledig et al. 2017; Wang et al. 2018), image-to-image translation (I2I)(Zhu et al. 2017; Isola et al. 2017; Mao et al. 2019; Lee et al. 2018), and so on. Powerful image generation ability of GAN is useful for the I2I task, whose main goal is to learn the translation between different domains. The core of I2I is to translate an image via a latent vector containing generalized information in the visual domain (Choi et al. 2020).

However, since GAN is like a two-player minmax game of discriminator and generator, it cannot avoid mode collapse (Arjovsky, Chintala, and Bottou 2017). Mode refers to a type of generated images, and mode collapse happens when the generated images follow only a single mode or very few modes. Mode collapse in I2I is a phenomenon in which the translated images are biased on a single mode when translating to a different domain (Kim et al. 2017).

Recently, as an approach to mitigate mode collapse from an information-theoretic perspective, (Chen et al. 2016) tried to maximize the mutual information (MI) between the latent space and the generated image space. Also, mode seeking regularization (Mao et al. 2019) for encouraging minor modes was introduced for the I2I task. Besides, many studies to overcome mode collapse in generative models have been reported (Anoosheh et al. 2019; Jolicoeur-Martineau 2018; Choi et al. 2020).

However, the above-mentioned methods still have several limitations. (Choi et al. 2020) newly proposed a loss function for the diversity of style information, but it could not be a fundamental solution for mode collapse. Also, (Mao et al. 2019) relied solely on a regularization term, and discriminated generated images and real images based on a static decision boundary that causes mode collapse. A single decision boundary causes a critical problem that generated images are stuck in a specific mode. In particular, the static decision boundary in the I2I task makes diverse image attributes gather in a single basin (Sun, Fang, and Schwing 2020). The left side of Fig. 1 shows the limitations of the previous works.

To overcome the mode collapse problem in the I2I task, we allow a discriminator to produce a flexible decision boundary. (Jolicoeur-Martineau 2018) and (Sun, Fang, and Schwing 2020) already reported that flexible decision boundary could mitigate the phenomenon of being stuck in a single basin when translating images of a few visual attributes. Thus, inspired by the previous studies, we present a style-guided discriminator to consider diverse modes and detailed styles. Also, we borrow an extrinsic regularization term for disentanglement of encoded representations. In addition, we utilize use-specific weights based on information theory for boosting generator performance.

Main contributions of this paper are as follows:

- We propose a novel style-guided discriminator loss function that attempts to mitigate mode collapse and also disentangle style representation through flexible decision boundary. As far as we know, this is the world-first explicit discriminator based on pairwise input for a multiple domain *I2I* task.
- The discrimination plot on the right of Fig. 1 presents a way to visually confirm the degree of mode collapse relaxation and reality of each generated image. This plot

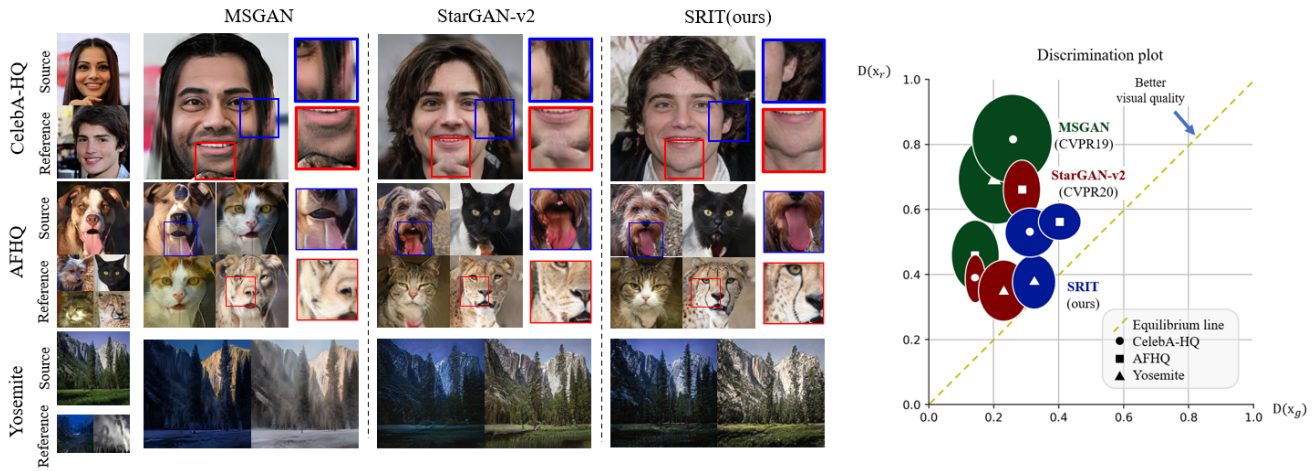

Figure 1: The left side shows the generated images of MSGAN, StarGAN and SRIT (ours) on the test split of CelebA-HQ, AFHQ, Yosemite datasets. Here, the blue and red boxes are magnified. They show that SRIT reflects the style of the reference image better than the other methods. The graph on the right shows the discrimination ability of each I2I translation model. In this plot, the closer to the equilibrium line, the closer the generative model is to the global optimum, that is, $p_{data} = p_g$ (i.e. $D(\mathbf{x}_r) = D(\mathbf{x}_g)$) (Goodfellow et al. 2014). This indicates that the generated image can be decided as a realistic image from the perspective of discriminator. The radius of each circle on the right graph indicates the standard deviation of the output of discriminator D. Refer to Experiments section for analysis of the plotted values on this graph.

can be widely used as a *performance indicator* of I2I algorithms in the future.

Related Work

Generative models for reducing mode collapse. In order to mitigate mode collapse, ether a generative model should be adaptive to input images (Goodfellow et al. 2014) or a discriminator must have the strong mode decision capability (Liu et al. 2019). So, (Mirza and Osindero 2014) and (Che et al. 2016) proposed a mode-specific generator that injects label supervision information into noise. (Mao et al. 2017) and (Arjovsky, Chintala, and Bottou 2017) proposed a loss function using f-divergence to match the distribution of the discriminator with the true label. From another point of view, (Chen et al. 2016) and (Belghazi et al. 2018) tried to alleviate mode collapse by utilizing MI between the input and output of the generator. ToDayGAN (Anoosheh et al. 2019) formed flexible decision boundary for the discriminator based on pairwise input. Since pairwise input enables a more flexible decision boundary, ToDayGAN can be an extrinsic solution that alleviates mode collapse. Inspired by the relativistic formula of (Jolicoeur-Martineau 2018), we derived a flexible decision boundary suitable for the I2I task.

Overview of image-to-image translation. As an early I2I model, Pix2Pix (Isola et al. 2017) adopted a generator based on U-Net (Ronneberger, Fischer, and Brox 2015) and a structure of adversarial learning. CycleGAN (Zhu et al. 2017) and DiscoGAN (Kim et al. 2017) introduced cycle consistency loss for stable I2I between different domains. Meanwhile, for diverse visual representations, (Mao et al. 2019) and (Yang et al. 2019) utilized diversity regularization term which plays a role of generating an image of a different mode whenever latent vectors change. Recently, I2I methods using guidance of reference images have been proposed (Choi et al. 2020; Baek et al. 2020). They achieved high visual quality by directly using the source image as an input to the generator without any encoding process in which image information may be lost.

Technical approaches for boosting visual quality. Information theory has been employed to understand I2I between different domains. For instance, MI was applied to loss functions (Chen et al. 2016; Belghazi et al. 2018) or regularization terms (Baek et al. 2020). Specifically, (Baek et al. 2020) adopted MI to maximize the dependency between style representations of samples in the same domain, thereby each class information was independently trained. While MI reflects the overall dependency of the two distributions, pointwise MI (PMI) is a metric for measuring the dependency of each point. PMI has been applied to a lot of computer vision tasks. For example, in the text supervision task (Takayama and Arase 2019), the output response sentence was determined so that PMI was maximized at the point where there was an input utterance. Since the encoded style representations from multiple domains in SRIT should be independent, PMI will contribute to SRIT for the purpose of understanding multiple domains.

Method

Let the distributions of real images and generated images be $\mathbb{P}$ and $\mathbb{Q}$, respectively. Images sampled from $\mathbb{P}$ and $\mathbb{Q}$ are represented by $\mathbf{x}_r$ and $\mathbf{x}_g$, respectively. As in Fig. 2 (a) and (b), a style encoder E and a mapping network M (Choi et al. 2020) generate style representations $\mathbf{s}_r$ and $\mathbf{s}_\mathbf{z}$ from $\mathbf{x}_r$ and $\mathbf{z}$. Here, $\mathbf{z}$ is from a random noise distribution $\mathbb{R}$: $\mathbf{s}_r = E(\mathbf{x}_r)$ and $\mathbf{s}_\mathbf{z} = M(\mathbf{z})$. In the learning stage, $\mathbf{s}_\mathbf{z}$ and $\mathbf{s}_r$ are alternately input to generator G (Choi et al.

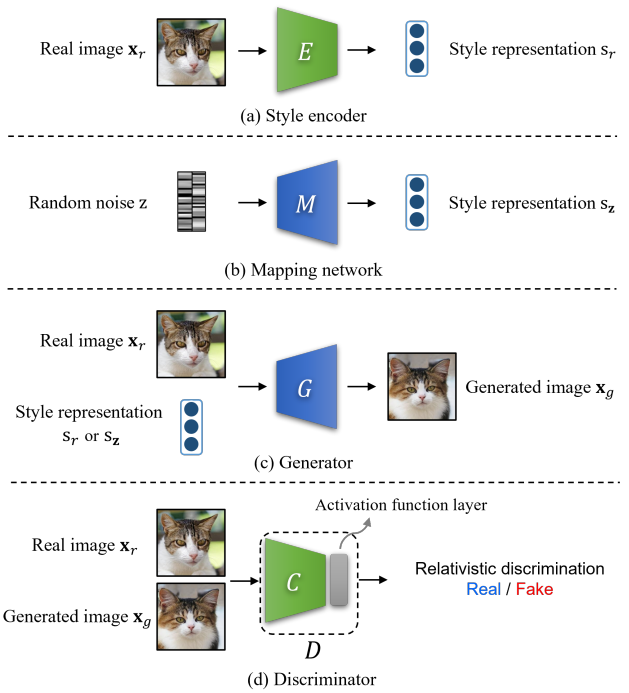

Figure 2: Components of SRIT. SRIT consists of (a) style encoder, (b) mapping network, (c) generator, and (d) discriminator.

2020) every iteration. Let translations using $\mathbf{s_z}$ and $\mathbf{s}_r$ denote latent-guided synthesis and reference-guided synthesis, respectively. D (Liu et al. 2019) is defined by $D = (f_{sig} \circ C)$ where C is a network without the last sigmoid layer f_{sig}.

Overview

Our goal is to generate an image with a rich visual representation by mitigating mode collapse. In other words, we want $\mathbf{x}_g$ (translated from $\mathbf{x}_r$) to successfully capture the style information of $\mathbf{x}_r$. As a main idea of this paper, we propose a style-guided discriminator (STGD) loss that mitigates mode collapse through flexible decision boundary. STGD loss induces a style guide effect through flexible decision boundary by generating $\mathbf{x}_g$ that reflects diverse style (i.e. mode) characteristics. In addition, we present a normalized PMI (NPMI) loss that enhances the disentanglement of the encoded style representation, and propose an importance weighting generator (IG) loss. They contribute to improving the visual representation and visual quality of translated images. The proposed SRIT is proven to provide the best quantitative performance. Also, the discrimination plot of Fig. 1 demonstrates that SRIT is closest to the global optimum (Goodfellow et al. 2014) of the generative model.

Style-Guided Discriminator

In order for diverse modes to be fully reflected in $\mathbf{x}_g$, it is desirable that D not only determines whether $\mathbf{x}_g$ is real or fake, but also considers various modes. However, since most of the I2I methods adopted a static decision boundary (Choi et al. 2020; Lee et al. 2018; Huang et al. 2018), their learning process had one convergence point ($D(x) \rightarrow 1$). As a result, as shown in the top of Fig. 3 (a), in the standard I2I method (w/o $\mathcal{L}_{STGD}$), a visually static decision boundary is formed regardless of modes. So the standard I2I method can suffer from deterioration in visual quality and diversity due to mode collapse.

To alleviate mode collapse, (Jolicoeur-Martineau 2018; Sun, Fang, and Schwing 2020) employed a specific flexible decision boundary mechanism that compares the discriminator outputs for real and fake images *in pair*. Inspired by the approach of (Jolicoeur-Martineau 2018; Sun, Fang, and Schwing 2020), we propose the STGD loss based on flexible decision boundary where relativistic discrimination is applied to the pairwise inputs of $\mathbf{x}_r$ and $\mathbf{x}_g$. The formula is as follows:

$$\begin{aligned}\mathcal{L}_{STGD} = &\mathbb{E}_{(\mathbf{x}_r,\mathbf{z})\sim(\mathbb{P},\mathbb{R})}[\log(f_{sig}(C(\mathbf{x}_r) - C(\mathbf{x}_g)))] \\ &+ \mathbb{E}_{(\mathbf{x}_r,\mathbf{z})\sim(\mathbb{P},\mathbb{R})}[\log(1 - f_{sig}(C(\mathbf{x}_g) - C(\mathbf{x}_r)))]\end{aligned} \tag{1}$$

where $\mathbf{x}_g$ is $G(\mathbf{x}_r, \mathbf{s}_r)$ or $G(\mathbf{x}_r, \mathbf{s_z})$. Note that $\mathcal{L}_{STGD}$ is defined based on pairwise inputs of a reference image $\mathbf{x}_r$ and the translated image $\mathbf{x}_g$. A given model is trained so that the difference between $C(\mathbf{x}_g)$ and $C(\mathbf{x}_r)$ is minimized through the STGD loss. As a result, the style characteristics of $\mathbf{x}_r$ are reflected and a flexible decision boundary is formed. As $C(\mathbf{x}_g)$ converges to the decision boundary of $\mathbf{x}_r$, resulting in a style-guide effect, and the translated images maintain the style characteristics of reference images as shown in Fig. 3 (b). From a generative model perspective, more realistic images are generated.

A two-point sample-based verification method that can explain the operation of Eq. 1 has been introduced in (Sun, Fang, and Schwing 2020). The two-point case showed that the relativistic formula of Eq. 1 alleviates mode collapse and is also appropriate for the pairwise input structure. The two-point case can be easily extended to the multiple-point case, and a neural network-based model can be also applied at the same time. See the claims in Sec. 3 of (Sun, Fang, and Schwing 2020) for a detailed proof.

Normalized Point-Wise Mutual Information

$\mathcal{L}_{STGD}$ in the previous section is an intrinsic solution at the image level to mitigate mode collapse. This section proposes auxiliary solution to mitigate mode collapse through extrinsic regularization for disentanglement of encoded style representation in latent space. If the encoded representation has a style attribute overlapped with other images, the visual representation of the translated image may be degraded or mode collapse may occur. Therefore, by minimizing the dependency of two style representations $\mathbf{s}_{r1}$ and $\mathbf{s}_{r2}$ encoded from different images, we make each style representation sufficiently reflect the independent characteristics of the corresponding images. Eventually, this mode collapse relaxation strategy will further improve the visual quality of the translated image.

On the other hand, since all components of the style representation affect the I2I process, we employ PMI (Van de Cruys 2011) to measure the MI of each component. PMI

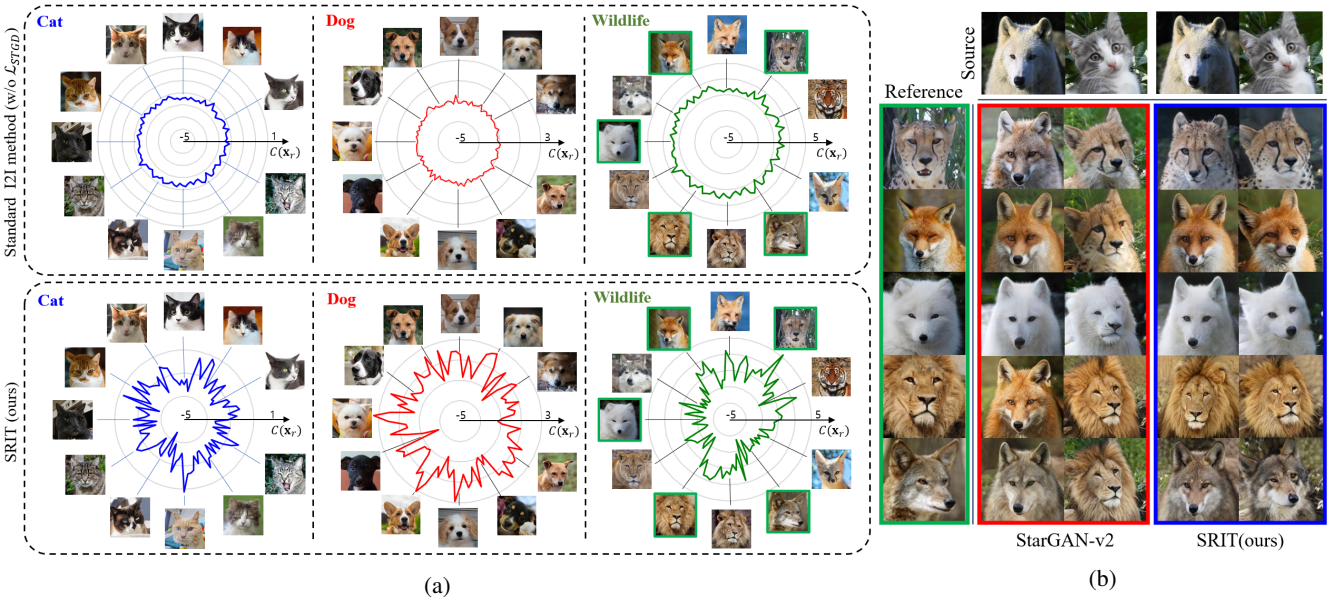

Figure 3: (a) Decision boundary according to STGD loss. Here, the decision boundaries ($C(\mathbf{x}_r)$) for 100 samples per domain of AFHQ testset were represented. In the standard I2I method (top) (Choi et al. 2018), images of various styles go through similar static decision boundaries, whereas the lower SRIT (w/ $\mathcal{L}_{STGD}$) has flexible decision boundaries according to image styles. (b) Reference-guided synthesis results for five green box images in the wildlife domain of (a). In this experiment, the results of StarGAN-v2 do not reflect the style characteristics of reference images well (see red boxes), but SRIT shows consistent style characteristics (see blue boxes). Best viewed when zoomed in.

measures the point-wise dependency of two variables and can be normalized for balance with other loss functions as in (Bouma 2009). To make the style representation of each image have an independent attribute, we define a loss function based on NPMI as follows:

$$\mathcal{L}_{NPMI} = \mathbb{E}_{(\mathbf{x}_r,\mathbf{z})\sim(\mathbb{P},\mathbb{R})}[\text{NPMI}(\mathbf{s}_{r1}, \mathbf{s}_{r2})] \tag{2}$$

where $\mathbf{s}_{r1}$ and $\mathbf{s}_{r2}$ denote style representations encoded from different real images $\mathbf{x}_{r1}$, $\mathbf{x}_{r2}$ or noises $\mathbf{z}_1$, $\mathbf{z}_2$. Since the components of the style representation encoded from each reference image through $\mathcal{L}_{NPMI}$ are independently distributed, an image with a rich visual representation can be generated (see Experiments section). **Appendix 1** qualitatively shows a disentangled representation that is independently distributed according to style.

Importance Weighting

Ultimately, the closer to a point where the discriminator's outputs become equivalent, i.e., $D(\mathbf{x}_r) = D(\mathbf{x}_g)$, the more realistic the image is generated. This section presents an additional idea for accomplishing this.

Together with the image level solution $\mathcal{L}_{STGD}$ and the latent space level solution $\mathcal{L}_{NPMI}$, we additionally propose to assign use-specific weights to generator learning (Cao et al. 2020). In order to give more weights to parameter updates in proportion to the discriminator output for each image, we refer to importance sampling (Bishop 2006).

Specifically, the expected value of the known function G with respect to the real image distribution p_r may be estimated by some samples of the translated image distribution p_g. The importance weighting factor is the likelihood ratio of the probability that the generated image $\mathbf{x}_g$ has a real image distribution, i.e., $p_r(\mathbf{x}_g)$ and the probability that $\mathbf{x}_g$ has a generated image distribution, i.e., $p_g(\mathbf{x}_g)$. As a result, the weighting factor to be used for updating parameters are defined by using $\frac{p_r(\mathbf{x}_g)}{p_g(\mathbf{x}_g)}$.

Lemma 1. *The importance weighting factor, which acts as a weight when learning a generator based on the GAN's global optimal discriminator, is derived by Eq.* (3)*.*

$$\frac{p_r(\mathbf{x}_g)}{p_g(\mathbf{x}_g)} = e^{C(\mathbf{x}_g)}. \tag{3}$$

Proof. First, according to (Goodfellow et al. 2014), the ideal global optimal discriminator has the following output w.r.t. $\mathbf{x}_g$.

$$D(\mathbf{x}_g) = \frac{p_r(\mathbf{x}_g)}{p_r(\mathbf{x}_g) + p_g(\mathbf{x}_g)} \tag{4}$$

Next, since D has a sigmoid activation function layer f_{sig} at the end, it can be defined by

$$D(\mathbf{x}_g) = f_{sig} \circ C(\mathbf{x}_g) = \frac{1}{1 + e^{-C(\mathbf{x}_g)}} \tag{5}$$

Based on Eqs. (4) and (5), the importance weighting factor is derived by an exponential formula such as Eq. (3). Here, as C, i.e., D excluding the activation function layer determines that $\mathbf{x}_g$ is more realistic, a larger weight $e^{C(\mathbf{x}_g)}$ is computed. See (Cao et al. 2020) for more details. □

If the importance weighting strategy derived by Lemma 1 is applied to the loss function of the generator, the following loss function is defined.

$$\mathcal{L}_{IG} = \mathbb{E}_{(\mathbf{x}_g,\mathbf{z})\sim(\mathbb{Q},\mathbb{R})}[e^{C(\mathbf{x}_g)+\epsilon}\ \log(1 - D(\mathbf{x}_g))] \quad (6)$$

Here, we introduce an offset factor ϵ to correct the output, taking into account that C may mismatch the global optimal discriminator with an ideal zero-centered output. ϵ is a scalar factor adapted to each dataset. It is fixed at 0.8 for CelebA-HQ and AFHQ and is set to 0.7 in Yosemite. On the other hand, since the weighting factor cannot be generated by an untrained discriminator, the training phase and the boosting phase must be distinguished during the learning process. So, all networks including the discriminator are trained in the training phase. In the training phase, the importance weighting factor $e^{C(x_g)}$ is set to 1. Then, the image translation performance is boosted by minimizing $\mathcal{L}_{IG}$ in the boosting phase using the trained style-guided discriminator D. Effect of this weighting process on performance will be demonstrated in Ablation study section.

Defining Full Objective Function

The full objective function consists of the loss functions defined by Eqs. (1), (2) and (6) and the fundamental loss function $\mathcal{L}_{fund}$, which is widely used for I2I models.

First, the adversarial loss function $\mathcal{L}_{adv}$ is reconstructed based on pairwise input for the intrinsic improvement of the loss function. $\mathcal{L}_{adv}$ is defined by $\mathcal{L}_{adv} = \mathcal{L}_{STGD} + \mathcal{L}_{IG}$. Note that $\mathcal{L}_{IG}$ includes $e^{C(x_g)}$ only in the boosting phase as mentioned above.

Next, $\mathcal{L}_{fund}$ consists of three loss functions: The cycle consistency loss function $\mathcal{L}_{cyc}$ (Kim et al. 2017), which is commonly used to prevent mode collapse, and the style reconstruction loss function $\mathcal{L}_{sty}$ (Huang et al. 2018) to maintain consistency in the multiple domain I2I process, and the loss function $\mathcal{L}_{ds}$ (Choi et al. 2020) for enhancing the diversity of the translated images. Therefore, $\mathcal{L}_{fund} = \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{sty}\mathcal{L}_{sty} - \lambda_{ds}\mathcal{L}_{ds}$. Here, λ_{cyc}, λ_{sty}, and λ_{ds} are the hyper-parameters.

In addition, we experimentally observed that the mode seeking loss function $\mathcal{L}_{ms}$ proposed in (Mao et al. 2019) not only strengthens the mapping between the style representation and the translated image, but also is synergetic with $\mathcal{L}_{NPMI}$. Thus, the extrinsic regularization terms $\mathcal{L}_{NPMI}$ and $\mathcal{L}_{ms}$ are applied so that the encoded representation reflects the style attribute of domain well in the latent space. Finally, the full objective function is defined by

$$\min_{G,F,E}\max_{D} \quad \mathcal{L}_{adv} + \lambda_{NPMI}\mathcal{L}_{NPMI} + \mathcal{L}_{ms} + \mathcal{L}_{fund} \quad (7)$$

where a scale factor λ_{NPMI} was set to 0.1.

Experiments

We chose StarGAN-v2 (Choi et al. 2020), MSGAN (Mao et al. 2019) and RGAN (Jolicoeur-Martineau 2018) as baselines. See **Appendix 2** for network details. Quantitative and qualitative results are shown through intensive experiments on a total of three datasets including the two datasets used in (Choi et al. 2020). In addition, numerical figures shown in the right graph of Fig. 1 are quantitatively analyzed.

Datasets

This section evaluates the proposed SRIT algorithm[1] for three popular datasets, i.e., CelebA-HQ (Karras et al. 2017), AFHQ (Choi et al. 2020), and Yosemite (summer and winter scenes) (Zhu et al. 2017). CelebA-HQ is divided into male and female, and AFHQ is divided into cat, dog and wildlife, and Yosemite is divided into summer and winter. Any information except the domain label was not used, and for fair comparison, each image resolution was resized to 256×256 as in the previous works.

Training Configurations

Implementation details. Model training is composed of the training phase and the boosting phase. In training all datasets, the batch size was set to 8, and 100K iterations in the training phase and 5K iterations in the boosting phase were repeated. Three regularization parameters λ_{cyc}, λ_{sty}, and λ_{ds} were set with reference to (Choi et al. 2020), and λ_{NPMI} and ϵ were experimentally determined, respectively. For CelebA-HQ and Yosemite datasets, $\lambda_{cyc} = 1$, $\lambda_{sty} = 1$, $\lambda_{ds} = 1$, and $\lambda_{NPMI} = 0.1$. In AFHQ, $\lambda_{cyc} = 1$, $\lambda_{sty} = 1$, $\lambda_{ds} = 2$, and $\lambda_{NPMI} = 0.1$. Also, for CelebA-HQ and AFHQ, $\epsilon = 0.8$, and $\epsilon = 0.7$ in Yosemite. For training stability, λ_{ds} decreases linearly towards zero for 10K iterations. The detailed set-up such as high-order regularization (Mescheder, Geiger, and Nowozin 2018) was set the same as the baseline model (Choi et al. 2020). See **Appendix 3**.

For detailed parameter configuration and model specifications, refer to the link below, which contains methods for generating high-resolution (e.g., 512×512 and more) images and adjusting parameters.

Evaluation metrics. We adopted FID (Fréchet inception distance) (Zhang et al. 2018) and LPIPS (Learned Perceptual Image Patch Similarity) (Heusel et al. 2017) to evaluate the visual quality and diversity of translated images. FID was computed by feature vectors obtained from Inception-V3 (Szegedy et al. 2016) pre-trained with ImageNet, and LPIPS was computed by L_1 distance between features extracted from AlexNet (Krizhevsky, Sutskever, and Hinton 2012) pre-trained with ImageNet. FID is the average for the validation set and the translated images of each dataset, and the LPIPS is the average for 10 images translated from the same source image. The lower the FID, the higher the visual quality. And, the higher the LPIPS, the higher the diversity.

Performance Comparison

For fair evaluation, we extracted the quantitative values described in the latest I2I translation literature, and for qualitative performance comparison, we employed only available milestone methods under the same conditions.

Quantitative evaluation. We compared the quantitative performances of DRIT (Lee et al. 2018), MSGAN (Mao et al. 2019), StarGAN-v2 (Choi et al. 2020), LETIT (Zhao and Chen 2021), ReMix (Cao et al. 2021), and SRIT in terms of FID and LPIPS. Table 1 shows that SRIT quantitatively outperforms other methods for all datasets. Compared to

[1]Code can be found here https://github.com/jaewoong1/SRIT_Style-guided-I2I-translation

Datasets		CelebA-HQ		AFHQ		Yosemite	
Type	Method	FID (↓)	LPIPS (↑)	FID (↓)	LPIPS (↑)	FID (↓)	LPIPS (↑)
Latent	DRIT (Lee et al. 2018)	52.1	0.178	95.6	0.326	52.3	0.106
	MSGAN (Mao et al. 2019)	33.1	0.389	61.4	0.517	49.0	0.106
	StarGAN-v2 (Choi et al. 2020)	13.7	0.452	16.2	0.45	43.5	0.196
	LETIT (Zhao and Chen 2021)	12.5	-	15.9	-	-	-
	ReMix (Cao et al. 2021)	-	-	15.2	0.491	-	-
	SRIT(ours)	**12.4**	**0.499**	**12.5**	**0.591**	**39.1**	**0.381**
Reference	DRIT (Lee et al. 2018)	53.3	0.311	114.8	0.156	78.1	0.187
	MSGAN (Mao et al. 2019)	39.6	0.312	69.8	0.375	56.7	0.214
	StarGAN-v2 (Choi et al. 2020)	23.8	0.388	19.8	0.432	43.8	0.225
	SRIT(ours)	**21.7**	**0.411**	**16.9**	**0.555**	**38.1**	**0.424**

Table 1: Quantitative comparison on latent-guided and reference-guided synthesis. The numerical figures of LETIT and ReMix were directly quoted from those papers.

Datasets	CelebA-HQ			AFHQ			Yosemite		
Method	$D(\mathbf{x}_r)$	$D(\mathbf{x}_g)$	γ	$D(\mathbf{x}_r)$	$D(\mathbf{x}_g)$	γ	$D(\mathbf{x}_r)$	$D(\mathbf{x}_g)$	γ
MSGAN (Mao et al. 2019)	0.806	0.258	0.320	0.693	0.211	0.304	0.455	0.148	0.325
StarGAN-v2 (Choi et al. 2020)	0.387	0.147	0.380	0.349	0.233	0.668	0.653	0.288	0.441
SRIT(ours)	0.526	0.315	**0.598**	0.378	0.326	**0.862**	0.560	0.405	**0.723**

Table 2: The quantitative values in the discrimination plot on the right side of Fig. 1. γ is an index obtained by dividing $D(\mathbf{x}_g)$ by $D(\mathbf{x}_r)$, and the closer γ is to 1, the better performance. The results of reference-guided synthesis and latent-guided synthesis were summed. It is worth noting that SRIT has γ closest to 1 for all datasets.

(a) CelebA-HQ (b) AFHQ

Figure 4: Qualitative comparison of reference-guided synthesis on the CelebA-HQ and AFHQ datasets.

StarGAN-v2, SRIT has a marginal improvement in quantitative performance. However, as can be seen in Figs. 4 and 5, SRIT is superior to StarGAN-v2 in terms of qualitative image quality. On the other hand, Table 2 lists the quantitative values of each technique shown in the discrimination plot on the right side of Fig. 1. See **Appendix 4** for the standard deviation per model. Here, $\gamma = D(\mathbf{x}_g)/D(\mathbf{x}_r)$ becomes 1 when the model reaches the global optimum (Goodfellow et al. 2014). Thus, as γ approaches 1, it can be interpreted that the discriminator determines that the translated image is realistic. Note that SRIT has γ that is closest to 1 compared to other methods.

For the quantitative performance of the proposed method in terms of Inception Score (Salimans et al. 2016) other than FID and LPIPS, refer to **Appendix 5**.

Qualitative evaluation. Fig. 4 qualitatively compares the reference-guided synthesis results of SRIT, MSGAN, and StarGAN-v2. For CelebA-HQ, SRIT naturally reflects the style of the reference image and achieves higher visual quality than other methods (see Fig. 4 (a)). In AFHQ of Fig. 4 (b), we can see that SRIT generates an image with a richer visual representation than the others, and represents the characteristics of the reference image better than state-of-the-art (SOTA) methods. From the latent-guided synthesis results of Fig. 5, we could observe high mode diversity and rich visual representation of SRIT. Refer to **Appendix 5** for further visual examples.

Also, Fig. 6 shows the sequential images generated by traversing the encoded style representations from two reference images. In this figure, we can observe that SRIT reflects the style information of reference images better than StarGAN-v2. This indicates that SRIT is good at disentanglement of style information.

Ablation Study

This section analyzed the effect of each key technique of SRIT on overall performance. For this experiment, the quantitative performance of latent-guided synthesis was measured on the Yosemite dataset. Table 3 shows the results. (A) in Table 3 gives the performance of the baseline model (Choi et al. 2020). (B) proves that the model performance is im-

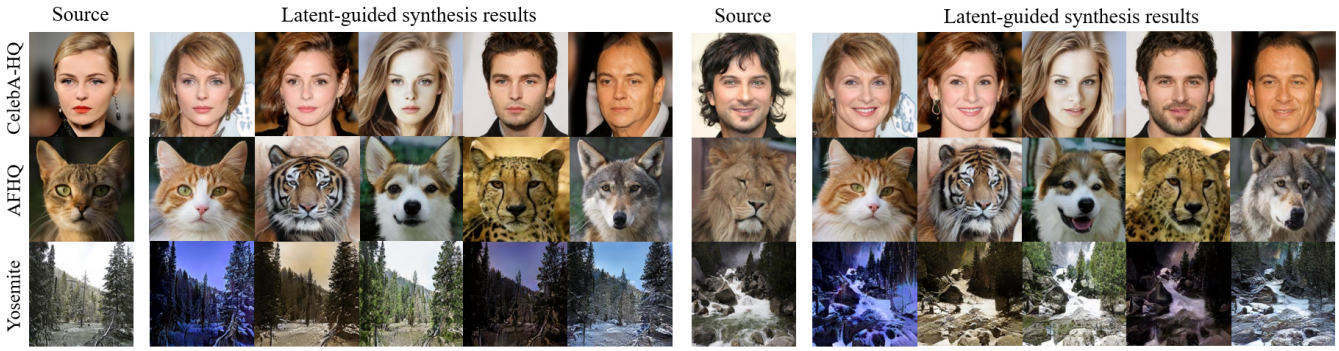

Figure 5: Latent-guided synthesis of SRIT per dataset. SRIT can generate more diverse images than the others.

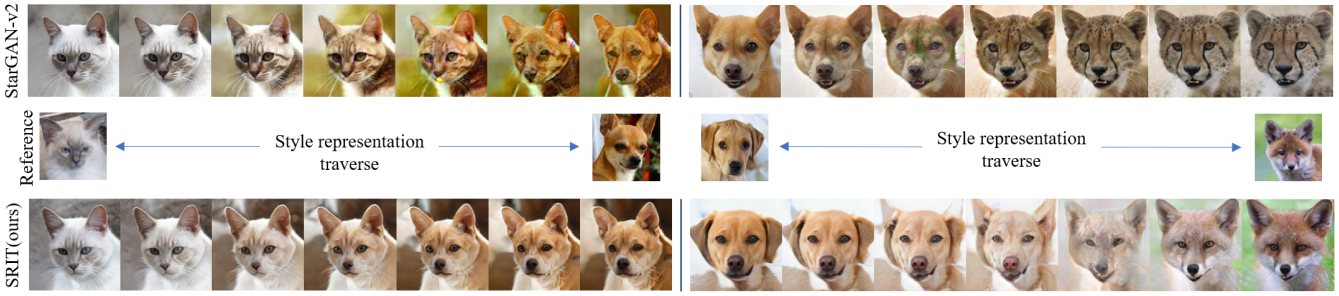

Figure 6: Transition of translated images along the style representation traverse. This experiment shows how the translated images change while varying the ratio of style representation of two reference images in the AFHQ dataset.

	MS	NPMI	STGD	IG	FID	LPIPS
(A)					43.5	0.196
(B)	✓				42.5	0.248
(C)	✓	✓			41.1	0.318
(D)	✓	✓	✓		39.4	0.372
(E)	✓	✓	✓	✓	**39.1**	**0.381**

Table 3: Performance analysis of various configurations of latent-guided synthesis for Yosemite dataset. This table shows the quantitative performance change whenever each loss of the proposed method is applied to StarGAN-v2 one-by-one. MS stands for mode seeking regularization.

proved when mode seeking (MS) regularization (Mao et al. 2019) is applied as a loss function. Although not shown in Table 3, in the case of CelebA-HQ and AFHQ, we could observe a tendency for the FID to deteriorate by about 10-20%. This is due to unstable learning caused by large gradients of MS regularization (Choi et al. 2020).

When the NPMI loss function (that allows the encoded style representation to have an independent component) and the MS regularization term (that allows different encoded style representations to generate each mode) are used together, synergy is exerted and overall performance is improved as in (C) in Table 3. The same tendency was observed in CelebA-HQ and AFHQ datasets. Continuing, a noticeable performance improvement was observed whenever the STGD loss function (D) and the IG loss function (E) were sequentially added to this.

Conclusion

This paper proposes the SRIT that can generate images of improved visual quality and visual representation by mitigating the mode collapse problem in the I2I task. By adopting a strategy that independently guides the style of the target domain, SRIT could preserve style characteristics of the reference image or the target domain better than the SOTA algorithms. By effectively encoding the style representation of reference images, SRIT can generate an image with more realistic and richer representation compared to SOTA techniques. In other words, SRIT can significantly improve the I2I performance by using a style guide, an intrinsic objective function for boosting visual quality, and an extrinsic objective function for independent representation. In the future, we will study a method to encode a meaningful representation even for a small number of complex images.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01389, Artificial Intelligence Convergence Research Center(Inha University)) and partly supported by IITP grant funded by the Korea government (MSIT) (No.2021-0-02068, AI Innovation Hub).

References

Anoosheh, A.; Sattler, T.; Timofte, R.; Pollefeys, M.; and Van Gool, L. 2019. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, 5958–5964. IEEE.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Baek, K.; Choi, Y.; Uh, Y.; Yoo, J.; and Shim, H. 2020. Rethinking the truly unsupervised image-to-image translation. *arXiv preprint arXiv:2006.06500*.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 31–40.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Cao, J.; Hou, L.; Yang, M.-H.; He, R.; and Sun, Z. 2021. ReMix: Towards Image-to-Image Translation with Limited Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15018–15027.

Cao, J.; Huang, H.; Li, Y.; He, R.; and Sun, Z. 2020. Informative sample mining network for multi-domain image-to-image translation. In *European Conference on Computer Vision*, 404–419. Springer.

Che, T.; Li, Y.; Jacob, A. P.; Bengio, Y.; and Li, W. 2016. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8188–8197.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, 172–189.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jolicoeur-Martineau, A. 2018. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, 1857–1865. PMLR.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, 35–51.

Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; and Kautz, J. 2019. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10551–10560.

Lučić, M.; Tschannen, M.; Ritter, M.; Zhai, X.; Bachem, O.; and Gelly, S. 2019. High-fidelity image generation with fewer labels. In *International Conference on Machine Learning*, 4183–4192. PMLR.

Mao, Q.; Lee, H.-Y.; Tseng, H.-Y.; Ma, S.; and Yang, M.-H. 2019. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1429–1437.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Mescheder, L.; Geiger, A.; and Nowozin, S. 2018. Which training methods for GANs do actually converge? In *International conference on machine learning*, 3481–3490. PMLR.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29: 2234–2242.

Sun, R.; Fang, T.; and Schwing, A. 2020. Towards a Better Global Loss Landscape of GANs. *Advances in Neural Information Processing Systems*, 33.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Takayama, J.; and Arase, Y. 2019. Relevant and informative response generation using pointwise mutual information. In *Proceedings of the First Workshop on NLP for Conversational AI*, 133–138.

Van de Cruys, T. 2011. Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 16–20.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Yang, D.; Hong, S.; Jang, Y.; Zhao, T.; and Lee, H. 2019. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, Y.; and Chen, C. 2021. Unpaired image-to-image translation via latent energy transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16418–16427.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.