Multimodal Image Translation with Stochastic Style Representations and Mutual Information Loss

Sanghyeon Na[1]
sktkdgus98@korea.ac.kr
Seungjoo Yoo[1]
seungjooyoo@korea.ac.kr
Jaegul Choo[2]
jchoo@kaist.ac.kr

[1] Korea University
Seoul, Korea
[2] Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Korea

Abstract

Unpaired multimodal image-to-image translation is a task of converting a given image in a source domain into diverse images in a target domain. We propose two approaches to produce high-quality and diverse images. First, we propose to encode a source image conditioned on a given target style feature. It allows our model to generate higher-quality images than existing models, which are not based on this method. Second, we propose an information-theoretic loss function that effectively captures styles in an image. It allows our model to learn complex high-level styles rather than simple low-level styles, and generate perceptually diverse images. We show our proposed model achieves state-of-the-art performance through extensive experiments on various real-world datasets.

1 Introduction

Unpaired multimodal image-to-image translation (multimodal translation) is a task of translating an image belonging to one domain into diverse images in another domain without aligned data across the two domains. Multimodality means that a single image can be translated into diverse images in the other domain. For instance, the man in the first row of Fig. 1 can be converted into many different styles of women. We will define content as a factor that is invariant between domains (e.g., angle of face and gaze) and style as a varying factor within each domain (e.g., hair length, make-up, and beard). Multimodal translation is a challenging task because it should achieve high-quality and high-diversity of generated images at the same time, i.e., both high-quality but low-diversity images and high-diversity but low-quality images are not desirable. Our aim is to achieve them at the same time by focusing on two issues with the existing multimodal translation models as follows.

The first issue is that the existing multimodal translation models independently encode the source image and the target image with the two separate encoders [13, 21, 25]. In other words, one encoder extracts information only from a given image without considering what information is extracted from the other encoder. In contrast, when encoding the source

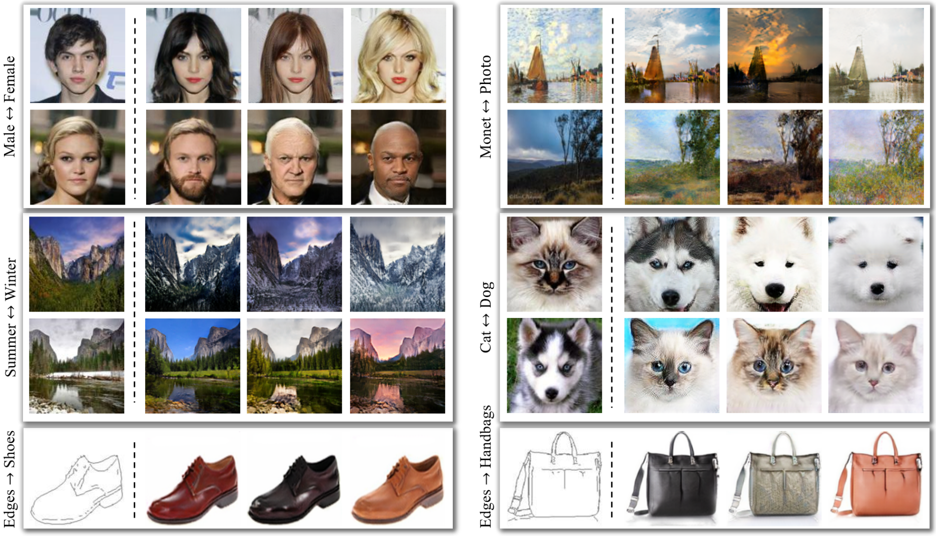


Figure 1: **Results of our proposed model (MISO).** Given the source image on the leftmost, MISO can generate high-quality and diverse images with the different high-level styles (e.g., different *hairstyle*, *age*, and *skin color* in Male↔Female) rather than manipulating the low-level styles such as overall color tone (for the low-level styles, please see the results of other models in Fig. 3). Additionally, MISO can control fine variations such as *make-up* (Male→Female), *cloud shape* (Summer→Winter), and *light reflection* (Edges→Shoes). More results are in the supplementary material.

image, our model allows the encoder to consider which target style is applied to the source image by using the concatenated source image and given target style feature as an input of the encoder. Then, the output feature is decoded to an image in the target domain. We assume that this encoder is encouraged to extract information from a source image that is not contained in the given target style feature but necessary to complete a high-quality image. We call this architecture style-conditioned encoder because our model extracts feature from the source image conditioned on the target style feature. It allows our model to generate higher-quality images than the existing models.

The second issue is that the self-reconstruction (SR) loss at an image level used in most multimodal translation models [13, 21, 25] is inadequate to properly learn the styles that play an important role in multimodal translation. Before discussing the SR loss, it is necessary to consider the style. It is difficult to estimate styles from a single image accurately. For example, in a human face, hair color and skin color may look different depending on lighting conditions. This uncertainty makes it difficult to estimate accurate styles from an image. In this regard, the assumption that an image has a deterministic style feature may not be sufficient. Instead, to learn the styles involved in an image considering the uncertainty, treating style feature as a random variable may be more effective because a random variable can take into account uncertainty and variation given a particular image. However, the SR loss is simply computed on the mean value [18, 23], which may not be the best to learn the complex styles in an image considering the uncertainty. This can be critical for multimodal translation model because what styles are learned is largely related to the diversity of generated images. If the model learns simple low-level styles such as overall color tone,

	MUNIT	DRIT	MSGAN	MISO(OURS)
MI LOSS	×	×	×	√
STYLE-CONDITIONED ENCODER	×	×	×	√
LATENT VARIABLE	DV	RV	RV	RV

Table 1: **Comparisons of MISO with previous models**. The last row, named as a latent variable, shows whether each model interprets the style latent variable as a deterministic value (DV) or a random variable (RV).

the generated images will not be perceptually diverse. Thus, we adopt a loss function based on mutual information, which is similar to InfoGAN [6], that can estimate the underlying uncertainty or the variance of the styles. Through comparative experiments with SR loss, we show that the proposed loss is effective to learn complex high-level styles rather than simple low-level styles and allows our model to generate perceptually diverse images.

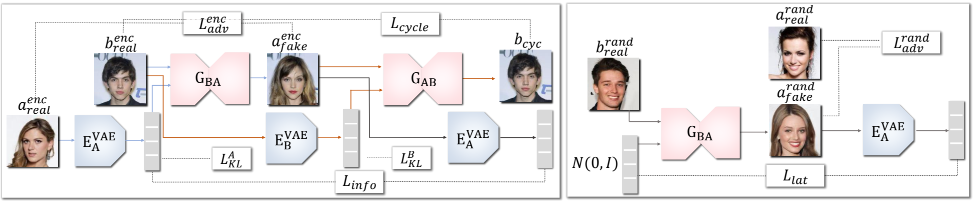
Based on these ideas, the main contributions of this work include the following:

- We model an improved framework for multimodal translation with the style-conditioned encoder. It enables our model to generate higher-quality images than the existing models.
- We introduce a loss function that improves over the self-reconstruction loss. This loss function uses mutual information to effectively learn the styles. It allows our model to learn complex high-level styles rather than low-level styles and generate perceptually diverse images.
- Using extensive experiments on numerous datasets and metrics, we show that our model produces the highest-quality yet diverse outputs in comparison to state-of-the-art multimodal translation models.

2 Related Work

2.1 Multimodal Image-to-Image Translation

Image-to-image translation has shown the remarkable performance in the various areas such as colorization [22, 28], super resolution [20, 26], and image synthesis [2, 3, 7, 8]. However, early image-to-image translation models [16, 31] can only generate a single output for a single input. BicycleGAN [32] proposed two-phase training using paired data to produce multimodal outputs. Augmented CycleGAN [1] extended this task into an unpaired setting. MUNIT [13] and DRIT [21] are the state-of-the-art frameworks for multimodal image translation, and MSGAN [25] further improved DRIT by introducing an additional regularization term to increase diversity. MUNIT, DRIT, and MSGAN follow the two-phase training of BicycleGAN, and our model also follows this two-phase training like these models. However, these models encode the source image and the target image independently, but our model introduces the style-conditioned encoder that enables our model to generate higher-quality images than the existing models. Furthermore, these models use the SR loss, but previous studies pointed out that this SR loss fails to capture detailed features [14, 19] because pixels may be blindly averaged out, resulting in blurry output. Motivated by this, we adopt a loss function based on mutual information to learn the style of an image for the multimodal translation problem. This loss shows the difference not only with MUNIT, DRIT, and MSGAN but also with BicycleGAN that utilizes image reconstruction loss that benefits from the paired dataset. Table 1 shows the differences between our model and the existing unpaired



(a) Training with the encoded style feature (b) Training with the random style feature

Figure 2: **Architecture of our proposed model**. Fig. 2a and Fig. 2b show the training phases using the encoded style feature, i.e., the style feature extracted from a_{real}^{enc}, and the random style feature sampled from $\mathcal{N}(0,I)$, respectively. The superscripts *enc* and *rand* are used to distinguish these phases. Our model jointly learns from these phases by optimizing Eq. (4) and Eq. (5). Note that the style encoder has two outputs, mean (μ) and standard deviation (σ). The grey rectangle indicates the style feature that is sampled from $\mathcal{N}(\mu, \sigma)$ except for the left grey rectangle in Fig. 2b, which is sampled from $\mathcal{N}(0,I)$. Although the sampling process from $\mathcal{N}(\mu, \sigma)$ is not described in Fig. 2 for brevity, some loss functions use μ and σ instead of the style feature, which are specified in Section 3.

multimodal translation models. In addition, although CVAE-GAN [4] has a similar structure with one of our training phases (Fig. 2a), it also uses SR loss. Recently, StarGAN-v2 [9] extended multimodal image translation task into a multi-domain setting.

2.2 Mutual Information

Mutual information is effectively used to learn deep representations [5, 12]. Measuring mutual information is usually intractable, so InfoGAN [6] introduces a lower bound on mutual information. AttGAN [10] and CFGAN [15] use mutual information to learn the styles. AttGAN defines style feature as a categorical latent feature with dimensions of a predefined number of factors and uses cross-entropy loss. CFGAN defines style feature as a continuous latent feature, which is the same with our model, and adopts the loss function proposed in InfoGAN. It means that CFGAN learns the styles by maximizing the correlation between a generated image and a *randomly sampled* latent variable. In this regard, minimizing latent reconstruction loss (Section. 3.3) also can be interpreted as maximizing mutual information, but it is not our contribution. In contrast, our loss function (Section. 3.1) learns the styles by maximizing the correlation between a generated image and a latent variable *extracted from a real image*. This loss function contributes to our model achieving state-of-the-art performance by allowing our model to learn complex high-level styles and generate perceptually diverse images.

3 Proposed Method

We introduce our unpaired multimodal image-to-image translation model called Multimodal Image translation with StOchastic style representations and mutual information loss (MISO) as well as its components. Our goal is to learn a one-to-many mapping between two domains $A \subset \mathbb{R}^{H \times W \times 3}$ and $B \subset \mathbb{R}^{H \times W \times 3}$. These two domains can be used interchangeably as a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$. One-to-many mapping between $\mathcal{S}$ and $\mathcal{T}$ can be achieved by learning a distribution $p(t|s,z)$ where $t \in \mathcal{T}$, $s \in \mathcal{S}$ and $z \sim \mathcal{N}(0,I)$. z is a random noise to achieve one-to-many mapping from given s. In other words, our model needs to learn a one-to-one mapping of $(\mathcal{S},\mathcal{Z}) \mapsto \mathcal{T}$ where $z \in \mathcal{Z}$. It is important to note that $z \sim \mathcal{N}(0,I)$ does not have any power to force a particular s to be mapped to a particular t. Therefore, we use an encoder to extract a feature z_t from t and make an arbitrary z obtain the information in z_t by minimizing the distance between $\mathcal{N}(0,I)$ and the conditional distribution of z_t given t. To this end, we can generate diverse images from a single image without the input images for the target style features by sampling from $p(t|s,z)$ where $z \sim \mathcal{N}(0,I)$.

Fig. 2 shows the training process of MISO. Though not shown in the figure, the same process is carried out with swapped domains. Our model consists of two style encoders for each domain, $E_A : A \mapsto Z_A$ and $E_B : B \mapsto Z_B$, two discriminators for each domain, $D_A : A \mapsto \mathbb{R}$ and $D_B : B \mapsto \mathbb{R}$, and two generators with encoder-decoder structure for each direction, $G_{AB} : (A, Z_B) \mapsto B$ and $G_{BA} : (B, Z_A) \mapsto A$. Unlike existing models, we do not encode the source image and the target image independently. Instead, we use the style-conditioned encoder, and our generator with the encoder-decoder structure is designed to take the source image and its target style as a condition. In Fig. 2a, b^{enc}_{real} is a source image and a^{enc}_{real} is a target image. The target style feature is extracted from a^{enc}_{real} by E^{VAE}_A, and the encoder part in G_{BA} encodes b^{enc}_{real} conditioned on the target style feature. Our style encoders are based on variational autoencoder (VAE) [17] architecture because we do not assume deterministic mappings for $E_A : A \mapsto Z_A$ and $E_B : B \mapsto Z_B$. The VAE architecture can handle the intractable true posteriors $p(z_a|a)$ and $p(z_b|b)$ via their approximate posteriors $q(z_a|a)$ and $q(z_b|b)$, which can be regarded as normal distributions, where $a \in A$, $b \in B$, $z_A \in Z_A$ and $z_B \in Z_B$.

3.1 Mutual Information Loss

This section introduces our proposed loss function mutual information loss (MI loss). We design MI loss that maximizes mutual information between the target style feature z_a and image generated with that feature $G_{BA}(b, z_a)$ written as $I(z_a; G_{BA}(b, z_a))$. It is difficult to directly be maximized, so we derive the lower bound to be maximized by following InfoGAN [6] as

$$\begin{aligned} I(z_a; G(b,z_a)) &= H(z_a) - H(z_a|G(b,z_a)) \\ &\geq \mathbb{E}_{z_a \sim p(z_a)}[\mathbb{E}_{a' \sim p_g(a|(b,z_a))}[\log q(z_a|a')]] + H(z_a) \\ &\geq \mathbb{E}_{z_a \sim p(z_a)}[\mathbb{E}_{a' \sim p_g(a|z_a)}[\log q(z_a|a')]] \end{aligned} \tag{1}$$

where H indicates an entropy, $q(z_a|a')$ is an approximated distribution (or the encoded latent distribution) of the intractable true posterior $p(z_a|a')$ and p_g is a distribution of the generated image $G(b, z_a)$. In Eq. (1), we consider a lower bound on a single source image b for brevity, and we can treat $H(z_a)$ as a constant by fixing the distribution of z_a. However, we have an unobserved prior of a latent variable, $p(z_a)$ that we cannot directly sample from. Our model addresses the problem that it is impossible to sampling directly from $p(z_a)$ by using the distribution $p(a)$ that we can easily sample from. That is, Eq. (1) is rewritten as

$$\begin{aligned} \int_{z_a} \mathbb{E}_{a' \sim p_g(a|z_a)}[\log q(z_a|a')] p(z_a)\, dz_a &= \int_{z_a}\int_{a} \mathbb{E}_{a' \sim p_g(a|z_a)}[\log q(z_a|a')] p(z_a, a)\, da\, dz_a \\ &= \int_{a}\int_{z_a} \mathbb{E}_{a' \sim p_g(a|z_a)}[\log q(z_a|a')] p(z_a, a)\, dz_a\, da \\ &= \mathbb{E}_{a \sim p(a)}[\mathbb{E}_{z_a \sim p(z_a|a)}[\mathbb{E}_{a' \sim p_g(a|z_a)}[\log q(z_a|a')]]]. \end{aligned} \tag{2}$$

In Eq. (2), it is still impossible to sample from the true posterior $z_a \sim p(z_a|a)$. Instead, we use $q(z_a|a)$, which is an approximation of $p(z_a|a)$. As a result, we can derive

$$I(z_a; G(b, z_a)) \geq \mathbb{E}_{a \sim p(a)}[\mathbb{E}_{z_a \sim q(z_a|a)}[\mathbb{E}_{a' \sim p_g(a|z_a)}[\log q(z_a|a')]]].$$

We assume $q(z_a|a')$ to be a normal distribution because we use a VAE-based encoder, and thus $q(z_a|a')$ can be represented as $\mathcal{N}(\mu_{out}, \sigma_{out}^2)$ where μ_{out} and σ_{out} are outputs of the encoder given a' as an input. Since $\mathcal{N}(\mu_{out}, \sigma_{out}^2)$ has a closed form of a probability density function $f(z_a|\mu_{out}, \sigma_{out}^2)$, we can represent $\log f(z_a|\mu_{out}, \sigma_{out}^2)$, and the final MI loss, $\mathcal{L}_{info}$, to minimize is defined as

$$\mathcal{L}_{info} = -\log f(z_a|\mu_{out}, \sigma_{out}^2) = \frac{1}{2}\log 2\pi\sigma_{out}^2 + \frac{(z_a - \mu_{out})^2}{2\sigma_{out}^2}. \tag{3}$$

3.2 Adversarial Loss

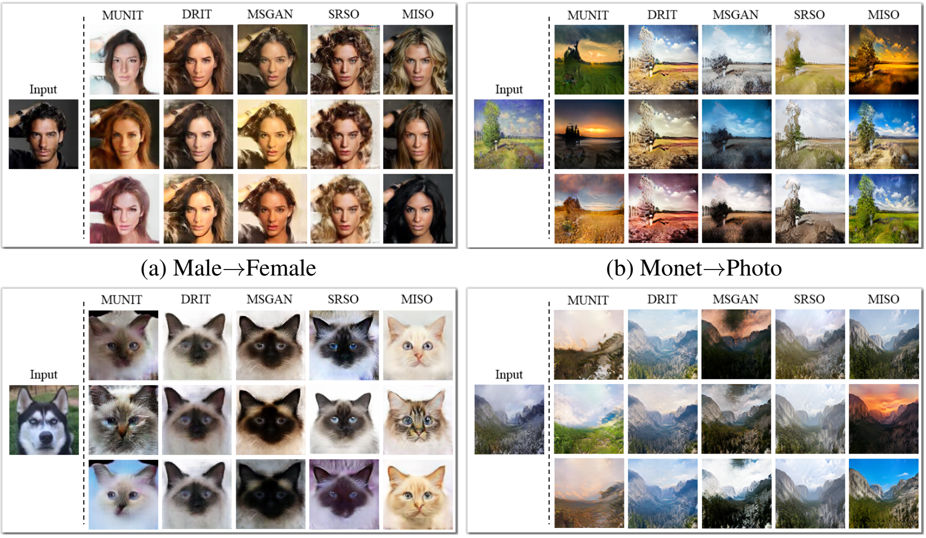
To render the generated images indistinguishable from real images, we employ the adversarial loss. In Fig. 2, both a_{fake}^{enc} generated with $z_a \sim q(z_a|a)$ and a_{fake}^{rand} generated with $z \sim \mathcal{N}(0, I)$ are needed to guarantee high-quality outputs, so we define two adversarial losses for each generated image. In other words, we define $\mathcal{L}_{adv}^{enc}$ as an adversarial loss on a_{real}^{enc} and a_{fake}^{enc}, and $\mathcal{L}_{adv}^{rand}$ on a_{real}^{rand} and a_{fake}^{rand}. Finally, the full adversarial loss can be written as $\mathcal{L}_{adv} = \mathcal{L}_{adv}^{enc} + \mathcal{L}_{adv}^{rand}$.

3.3 Other Loss functions

Cycle-Consistency Loss. It is important to preserve the contents of the source image and only change its style. To this end, the cycle-consistency loss [31] is used, i.e., $\mathcal{L}_{cyc} = \mathbb{E}_{a \sim p(a), b \sim p(b)}[||G_{AB}(\bar{a}, E_B(b)) - b||_1]$ where $\bar{a} := G_{BA}(b, E_A(a))$ corresponds to fake image of domain A, a_{fake}^{enc}, in Fig. 2a. Note that DRIT [21] and MSGAN [25] obtain style features from *generated* images which can be imperfect when compared to real images. In contrast, MISO extracts un-impaired style features $E_B(b)$ from *real* image b.

KL-Divergence Loss. At the test phase, we want to generate diverse outputs $G_{AB}(a, z)$ and $G_{BA}(b, z)$ where $z \sim \mathcal{N}(0, I)$, without requiring actual style images. The KL-divergence loss encourages the encoded latent distribution $q(z_a|a)$ and $q(z_b|b)$ to be close to $\mathcal{N}(0, I)$, i.e., $\mathcal{L}_{KL} = \mathbb{E}_{a \sim p(a)}[\mathcal{D}_{KL}(q(z_a|a)\,||\,\mathcal{N}(0, I))] + \mathbb{E}_{b \sim p(b)}[\mathcal{D}_{KL}(q(z_b|b)\,||\,\mathcal{N}(0, I))]$.

Latent Reconstruction Loss. We encourage an invertible mapping between $\mathcal{A}$ and $\mathcal{Z}$ with the latent reconstruction loss starting from $z \sim \mathcal{N}(0, I)$ which is the key part of the phase in Fig. 2b. However, L_1 loss between $z \sim \mathcal{N}(0, I)$ and $z_a \sim q(z_a|a)$ can be too strict and may bring instability to the training process. To avoid this, the latent reconstruction loss is defined as $\mathcal{L}_{lat} = \mathbb{E}_{b \sim p(b), z \sim \mathcal{N}(0, I)}[||z - \mu_A^{out}||_1]$ where μ_A^{out} is one of the outputs from $E_A(G_{BA}(b, z))$.



(a) Male→Female

(b) Monet→Photo

(c) Dog→Cat

(d) Winter→Summer

Figure 3: **Comparison results.** Diverse outputs obtained by using $z \sim \mathcal{N}(0, I)$. MISO generates high-quality images with diverse high-level styles. However, other models cannot preserve the contents well (MUNIT), control only low-level styles such as overall color tone (DRIT, MSGAN, SRSO), or generate artifacts (MSGAN). Furthermore, MISO controls fine variations such as *cloud shape* in Winter→Summer. The superiority of MISO over SRSO shows the effectiveness of the proposed MI loss in terms of learning high-level styles.

3.4 Full Objective Function

We can formulate the full objective to be minimized as

$$\mathcal{L}_D = -\lambda_{adv}\mathcal{L}_{adv}, \tag{4}$$

$$\mathcal{L}_{GE} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{info}\mathcal{L}_{info} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{lat}\mathcal{L}_{lat}. \tag{5}$$

4 Experiments

To demonstrate the effectiveness of our model we conduct experiments on multiple datasets with various evaluation metrics and compare with other competitive baselines. In the all experiments, we sampled the style features from $\mathcal{N}(0, I)$.

4.1 Datasets

Various translations are experimented on: **Male** ↔ **Female** on CelebA [24], **Summer (Smr)** ↔ **Winter (Wnt)** on Yosemite dataset [31], **Mnt** ↔ **Pht** on monet-to-photo dataset [31], **Cat** ↔ **Dog** on cat-to-dog dataset [21], **Edg** ↔ **Shs** on edges-to-shoes dataset [29], and **Edg** ↔ **Hdb** on edges-to-handbags dataset [27]. Although edges-to-shoes and edges-to-handbags are paired datasets, we trained our model under an *unsupervised* seeting.

	Mnt↔Pht	Yosemite	Edg→Shs	Edg→Hdb
N-Cycle	116.48	60.44	33.15	40.08
MUNIT	63.50	49.63	36.46	32.56
DRIT	57.27	45.30	56.24	42.61
MSGAN	56.04	44.67	54.15	39.75
SRSO	52.84	39.85	**24.41**	**25.08**
MISO	**51.17**	**39.62**	24.96	25.80

(a) **FID.** A low FID indicates high visual quality.

	CelebA	Mnt↔Pht	Cat↔Dog
N-Cycle	12.08	9.42	1.43
MUNIT	3.79	2.63	4.29
DRIT	6.28	11.05	9.28
MSGAN	6.79	18.95	15.00
SRSO	25.27	12.11	27.86
MISO	**45.79**	**46.84**	**42.14**

(b) **User Preference.** A high user preference score indicates high visual quality with content preservation of the source images.

Table 2: **Realism.** Evaluation on visual quality of the generated images with two metrics.

	Cat→Dog	Dog→Cat	Smr→Wnt	Wnt→Smr	Edg→Shs	Edg→Hdb
N-Cycle	0.0053	0.0020	0.0007	0.0008	0.0315	0.0644
MUNIT	0.2403	**0.1938**	0.2377	**0.2434**	0.1132	0.1794
DRIT	0.0942	0.0675	0.0954	0.0854	0.0701	0.0955
MSGAN	0.1563	0.0719	0.1489	0.1375	0.1251	0.1446
SRSO	0.0760	0.1527	0.1414	0.0850	0.1149	0.1860
MISO	**0.1694**	0.1880	**0.2434**	0.1769	**0.1383**	**0.2220**
Real	0.2209	0.2102	0.3592	0.3652	0.2755	0.3555

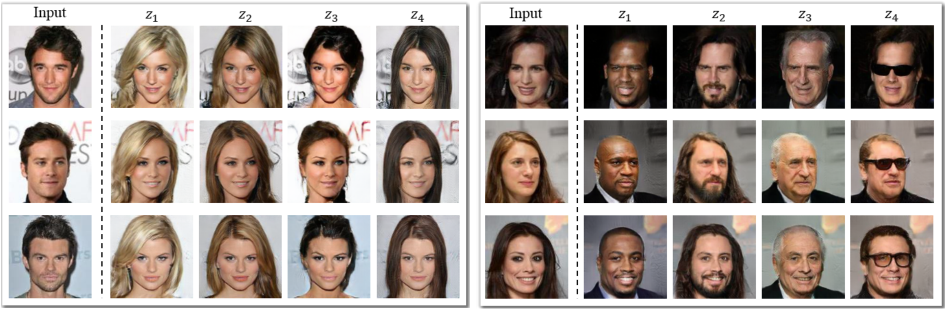
Table 3: **Diversity.** A high LPIPS distance indicates diverse outputs. However, if it is even higher than the distance of real samples, it is likely that the model made unrealistically high diversities such as artifacts and/or failed to preserve the contents of a source image.

4.2 Baselines

We compared our model with several baselines. CycleGAN+Noise (denoted as **N-Cycle**) is a modified version of CycleGAN [31], which injects noise vectors to the input layer of its generator. **MUNIT** and **DRIT** are state-of-the-art multimodal translation models, and **MSGAN** is an improved version of DRIT with an additional regularization term to increase diversity. Note that Augmented CycleGAN [1] is excluded from the baselines as reproducing reasonable results is not possible. **SRSO** is our proposed model equipped with the style-conditioned encoder and the SR loss. **MISO** is the improved model of SRSO that replaces the SR loss with the proposed MI loss. In brief, MUNIT, DRIT, and MSGAN can be represented as the combination of independently encoded feature and the SR loss, SRSO as that of the style-conditioned encoder and the SR loss, and MISO as that of the style-conditioned encoder and the MI loss.

4.3 Evaluation Metric

Diversity. Learned Perceptual Image Patch Similarity (**LPIPS**) [30] measures a perceptual distance between images to mimic humans' perceptual similarity. Diversity of translated images are measured by calculating the average LPIPS distance between ten images generated from the same source image (i.e., average between $\sum_{i=1}^{10}$ LPIPS values per source image) with the average distance of real data as an upper bound. That is, the diversity of real data is measured between images with different styles and *different* contents, but the diversity of translated images is measured between images with the different styles but *same* contents.



(a) Male→Female (b) Female→Male

Figure 4: **High-level styles learned by MISO.** Images in each column are generated using the same random vector for different source images. Each random vector contains distinct high-level style (e.g., different hairstyle, skin color, age, and glasses).

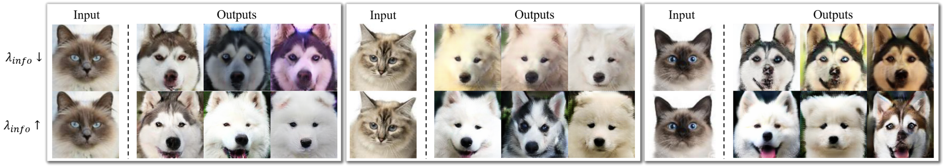


Figure 5: **Influence of MI loss.** If λ_{info} is too small ($\lambda_{info} \downarrow$), MISO generates low-level diversites (e.g., overall color tone) for a source image. By increasing λ_{info} ($\lambda_{info} \uparrow$), MISO can generate high-level diversities (e.g., species and open mouth) for a source image.

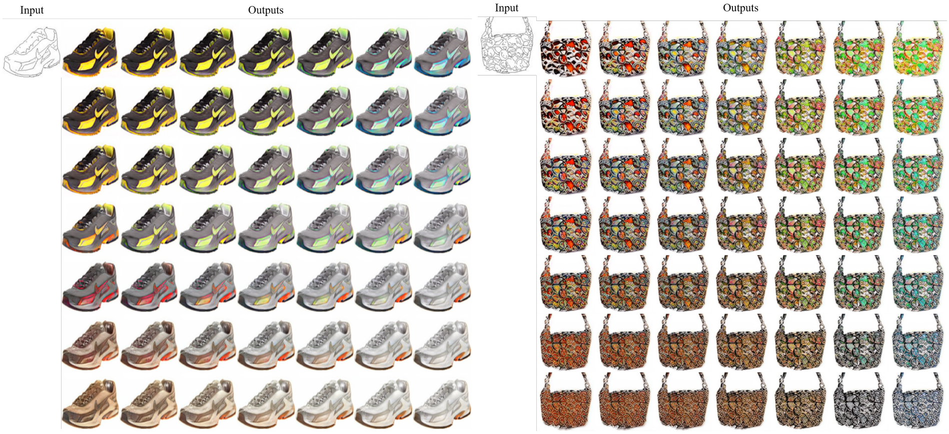
Therefore, it is difficult to exceed the distance of real data unless the model makes unrealistic diversities such as artifacts and/or fails to preserve the contents of a source image. We used all images from the test dataset as source images.

Realism. We use two metrics, **FID** [□] and **user preference** to evaluate the realism of generated images. We measured the distance between the distribution of real images and the distribution of generated images using FID. We generated 5000 images using all images from the test dataset as source images. Furthermore, we performed a user study with 30 participants. We give a random source image and its corresponding generated outputs of our model and baselines. We then ask each user which output has the best quality while maintaining the contents of the source image.

5 Results

Content Preservation. The results of MUNIT in Fig. 3 is worse than those of all the other models in maintaining contents of the source image (e.g., shoulders in the first row of Fig. 3a, trees in Fig. 3b, face angles in Fig. 3c, and mountains in Fig. 3d). This is because only MUNIT does not have a cycle-consistency loss. Note that this defect can be a reason for the high diversity scores of MUNIT in Table. 3. In contrast, MISO maintains contents of the source image while achieving the highest diversity scores for most datasets in Table. 3.

Diversity and Realism. In Figs. 3a, 3b and 3d, MUNIT generates unrealistic images compared to DRIT and MSGAN, that is consistent with the results in Table. 2 for corresponding



(a) Edges→Shoes (b) Edges→Handbags

Figure 6: **Bilinear interpolation.** It visualizes the style latent space that MISO learns.

datasets. This shows the importance of treating latent variables as a random variable. In MUNIT, the latent variables are considered deterministic, so the KL-divergence loss is inapplicable. Instead, MUNIT relies on the latent reconstruction loss that can cause problems in generating images with unseen $z \sim \mathcal{N}(0,I)$. In Fig. 3, DRIT and MSGAN only make low-level diversities among generated samples such as changing overall colors. Even SRSO, which adopts the style-conditioned encoder, only makes low-level diversities. It is consistent with the lower diversity scores of these models than that of MISO which can generate high-level diversities. However, SRSO generates high-quality images compared to DRIT and MSGAN. For example, in Fig. 3d, DRIT generates blurry images without details such as patterns on the left mountain. MSGAN generates unrealistic black spots or covers part of the image white like the third row of Fig. 3d. This is because the regularization proposed in MSGAN to generate diverse images simply increases L2 distance between generated images at pixel level. On the other hand, SRSO generates higher-quality images including details such as patterns on the left mountain without artifacts in Fig. 3d. This is because the encoder of our model can extract the information from the source image considering what information is not contained in the given target style feature to generate a high-quality image due to the style-conditioned encoder. However, SRSO only generates images with low-level diversities. By replacing SR loss of SRSO with MI loss, MISO generates high-quality images with diverse high-level styles such as daytime scene and sunset scene in Fig. 3d. Fig. 4 shows that MISO learns high-level styles, and Fig. 5 shows the importance of MI loss to learn high-level styles. As a result, MISO achieves the highest performances for most datasets in Tables. 2 and 3 by generating diverse images with high-quality. Fig. 6 visualizes the style latent space that MISO learns.

6 Conclusions

In this paper, we presented a novel framework of multimodal translation that achieves state-of-the-art performances on various datasets by utilizing the style-conditioned encoder as well as the mutual information loss function. Our model generates diverse yet high-quality images while properly preserving the contents of the source image.

Acknowledgement

This work was partially supported by Institute of Information & communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)) and the Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science and ICT (2017M3C4A7063570).

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *Proceedings of the International Conference on Machine Learning(ICML)*, 2018.

[2] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. High-resolution daytime translation without domain labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7488–7497, 2020.

[3] Hyojin Bahng, Sunghyo Chung, Seungjoo Yoo, and Jaegul Choo. Exploring unlabeled faces for novel attribute discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5821–5830, 2020.

[4] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754, 2017.

[5] Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R Devon Hjelm, and Aaron Courville. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems (NeurIPS)*, pages 2172–2180, 2016.

[7] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10639–10647, 2019.

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[10] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 2019.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems(NeurIPS)*, pages 6626–6637, 2017.

[12] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[13] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. 2018.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[15] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6089–6098, 2017.

[16] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[19] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114. IEEE, 2017.

[21] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.

[22] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5801–5810, 2020.

[23] Soochan Lee, Junsoo Ha, and Gunhee Kim. Harmonizing maximum likelihood with gans for multimodal conditional generation. *International Conference on Learning Representations(ICLR)*, 2019.

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[25] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[26] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[27] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of IEEE International Conference on Computer Vision(ICCV)*, 2015.

[28] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11283–11292, 2019.

[29] A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *Computer Vision and Pattern Recognition(CVPR)*, 2018.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[32] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.