Smoothing the Disentangled Latent Style Space for Unsupervised Image-to-Image Translation

Yahui Liu[1,3], Enver Sangineto[1], Yajing Chen[2], Linchao Bao[2], Haoxian Zhang[2],
Nicu Sebe[1], Bruno Lepri[3], Wei Wang[1*], Marco De Nadai[3*]

[1]University of Trento, Italy [2]Tencent AI Lab, China [3]Fondazione Bruno Kessler, Italy

Figure 1: Our method generates smooth interpolations within and across domains in various image-to-image translation tasks. Here, we show gender, age and smile translations from CelebA-HQ [20] and animal translations from AFHQ [10].

Abstract

Image-to-Image (I2I) multi-domain translation models are usually evaluated also using the quality of their semantic interpolation results. However, state-of-the-art models frequently show abrupt changes in the image appearance during interpolation, and usually perform poorly in interpolations across domains. In this paper, we propose a new training protocol based on three specific losses which help a translation network to learn a smooth and disentangled latent style space in which: 1) Both intra- and inter-domain interpolations correspond to gradual changes in the generated images and 2) The content of the source image is better preserved during the translation. Moreover, we propose a novel evaluation metric to properly measure the smoothness of latent style space of I2I translation models. The proposed method can be plugged in existing translation approaches, and our extensive experiments on different datasets show that it can significantly boost the quality of the generated images and the graduality of the interpolations.

1. Introduction

Translating images from one domain to another is a challenging image manipulation task that has recently drawn increasing attention in the computer vision community [9, 10, 16, 17, 26, 29, 37, 43]. A "domain" refers to a set of images sharing some distinctive visual pattern, usually called "style" (e.g., the gender or the hair color in face datasets) [10, 16, 43]. The Image-to-Image (I2I) translation task aims to change the domain-specific aspects of an image while preserving its "content" (e.g., the identity of a person or the image background) [16]. Since paired data (e.g., images of the same person with different gender) are usually not available, an important aspect of I2I translation models is the unsupervised training [43]. Moreover, it is usually desirable to synthesize the multiple appearances *modes* within the same style domain, in such a way to be able to generate *diverse* images for the same input image.

Recent work addresses the I2I translation using multi-

* These two authors contributed equally to this work. Correspondence to: wei.wang@unitn.it and denadai@fbk.eu.

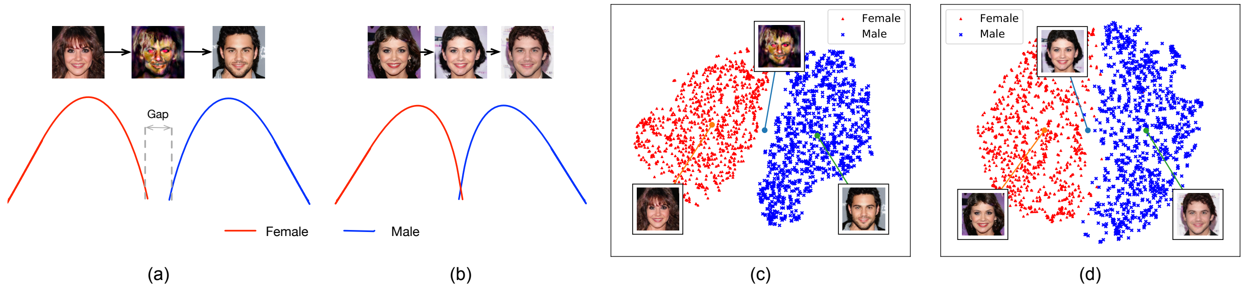

Figure 2: An illustration of the relation between smoothness and disentanglement of the style space. (a) Two well-separated distributions with a large margin in between. The intermediate area can lead to the generation of artifacts because it has not been sufficiently explored during training. (b) When the margin is reduced, the corresponding image appearance changes are smoother. (c) A t-SNE visualization of randomly sampled style codes using StarGAN v2 [10], which shows a disentangled style space but also that the inter-domain area generates images with artifacts. (d) The same visualization shows that, using our method, despite the disentanglement is preserved, the inter-domain area generates realistic images.

ple domains [9, 26, 10] and generating multi-modal outputs [26, 10]. These Multi-domain and Multi-modal Unsupervised Image-to-Image Translation (MMUIT) models are commonly evaluated based on the quality and the diversity of the generated images, including the results obtained by interpolating between two endpoints in their latent representations (e.g., see Fig. 1). However, interpolations are usually computed using only points belonging to the same domain, and most of the state-of-the-art MMUIT methods are inclined to produce artifacts or unrealistic images when tested using across-domain interpolations. This is shown in Fig. 2 (c), where, using the state-of-the-art StarGAN v2 [10], the inter-domain area in the style space frequently generates artifacts. Another common and related problem is the lack of graduality in both intra and inter domain interpolations, i.e., the generation of abrupt appearance changes corresponding to two close points in the latent space.

In this paper, we address the problem of learning a smoothed and disentangled style space for MMUIT models, which can be used for gradual and realistic image interpolations within and across domains. With "disentangled" we mean that the representations of different domains are well separated and clustered (Fig. 2), so that intra-domain interpolations correspond to only intra-domain images. With "smoothed" we mean that the semantics of the style space changes gradually and these changes correspond to small changes in the human perceptual similarity.

The main idea of our proposal is based on the hypothesis that the interpolation problems are related to the exploration of latent space areas which correspond to sparse training data. We again refer to Fig. 2 to illustrate the intuition behind this observation. Many MMUIT methods use adversarial discriminators to separate the distributions of different domains [10]. However, a side-effect of this disentanglement process is that some areas of the latent space do not correspond to real data observed during training. Consequently, when interpolating in those areas, the decoding process may lead to generating unrealistic images. We propose to solve this problem jointly using a triplet loss [35, 4] and a simplified version of the Kullback-Leibler (KL) divergence regularization [24]. The former separates the domains using a small *margin* on their relative distance, while the latter encourages the style codes to lie in a compact space. The proposed simplified KL regularization does not involve the estimation of parametric distributions [24] and it can be easily plugged in Generative Adversarial Networks (GANs) [10, 3]. On the other hand, differently from adversarial discrimination, the triplet-loss margin can *control* the inter-domain distances and help to preserve the domain disentanglement in the compact space, Finally, we also encourage the content *preservation* during the translation using a perceptual-distance based loss. Fig. 1 shows some interpolation results obtained using our method. In Sec. 6 we qualitatively and quantitatively evaluate our approach and we show that it can be plugged in different existing MMUIT methods improving their results. The last contribution of this paper concerns the proposal of the Perceptual Smoothness (PS) metric based on the perceptual similarity of the interpolated images, to quantitatively evaluate the style smoothness in MMUIT models.

The **contributions** of this paper can be summarized as follows. First, we propose a new training strategy based on three specific losses which improve the interpolation smoothness and the content preservation of different MMUIT models. Second, we propose a novel metric to fill-in the gap of previous MMUIT evaluation protocols and quantitatively measure the smoothness of the style space.

2. Related Work

Unsupervised Domain Translation. Translating images from one domain to another without paired-image supervision is a challenging task. Different constraints have been proposed to narrow down the space of feasible mappings between images. Taigman *et al.* [39] minimize the feature-level distance between the generated and the source image. Liu *et al.* [27] create a shared latent space between the domains, which encourages different images to be mapped into the same space. CycleGAN [43] uses a cycle consistency loss in which the generated image is translated back to the original domain (an approach proved to be pivotal in the field [23, 1, 32]). However, all these approaches are limited to one-to-one domain translations, thus requiring $m(m-1)$ trained models for translations with m domains. StarGAN [9] was the first single-model for *multi-domain* translation settings. The generation process is conditioned by a target domain label, input to the generator, and by a domain classifier in the discriminator. However, the I2I translation of StarGAN is deterministic, since, for a given source image and target domain, only one target image can be generated (no multi-modality).

Multi-modal and Multi-domain Translation. After the pioneering works in supervised and one-to-one image translations [44, 16, 30], the recent literature is mainly focused in multiple-domains and multi-modal translations. Both DRIT++ [26] and SMIT [34] use a noise input vector and a domain label to increase the output diversity. StarGAN v2 [10] relies on a multitask discriminator [28] to model multiple domains, a noise-to-style mapping network, and a diversity sensitive loss [30] to explore the image space better. However, qualitative results show changes of subtle "content" details (e.g., the color of the eyes, the shape of the chin or the background) while translating the image with respect to the style (e.g., the hair colour or the gender).

Although MMUIT models do not require any image-level supervision, they still require set-level supervision (i.e. domain labels for each image). Very recently, TUNIT [3] proposed a "truly unsupervised" task where the network does not need any supervision. TUNIT learns the set-level characteristics of the images (i.e., the domains), and then it learns to map the images to all the domains. We will empirically show that our method can be used with both StarGAN v2 and TUNIT, and significantly improve the interpolation smoothness with both models.

Latent-space interpolations. There is a quickly growing interest in the recent I2I translation literature with respect to latent space interpolations as a byproduct of the translation task. However, most previous works are only qualitatively evaluated, they use only intra-domain interpolations [25, 26, 34], or they require specific architectural choices. For example, DLOW [13] is a one-to-one domain translation, and RelGAN [40] uses a linear interpolation loss at training time, but it is not multi-modal. In StarGAN v2 [10], the style codes of different domains are very well disentangled, but the inter-domain interpolations show low-quality results (e.g., see Fig. 2). HomoGAN [8] learns an explicit linear interpolator between images, but the generated images have very limited diversity.

Interestingly, image interpolations are not limited to the I2I translation field. The problem is well studied in Auto-Encoders [24, 6, 5] and in GANs [2, 21, 22], where the image is encoded into the latent space without an explicit separation between content and style. For example, StyleGAN [21] and StyleGANv2 [22] show high-quality interpolations of the latent space, where the latter has been further studied to identify the emerging semantics (e.g. linear subspaces) without retraining the network [36, 18, 42]. Richardson *et al.* [33] propose to find the latent code of a real image in the pre-trained StyleGAN space. This two-stage inversion problem allows multi-modal one-to-one domain mappings and interpolations. However, these methods are not designed to keep the source-image content while changing the domain-specific appearance. Thus, they are not suitable for a typical MMUIT task.

3. Problem Formulation and Notation

Let $\boldsymbol{\mathcal{X}} = \bigcup_{k=1}^{m} \boldsymbol{\mathcal{X}}_k$ be the image set composed of m disjoint domains ($\boldsymbol{\mathcal{X}}_i \cap \boldsymbol{\mathcal{X}}_j = \emptyset, i \neq j$), where each domain $\boldsymbol{\mathcal{X}}_k$ contains images sharing the same style. The goal of a multi-domain I2I translation model is to learn a single functional $G(i,j) = \boldsymbol{\mathcal{X}}_i \rightarrow \boldsymbol{\mathcal{X}}_j$ for all possible $i, j \in \{1, 2, \cdots, m\}$. The domain identity can be represented either using a discrete domain label (e.g., i) or by means of a style code $\boldsymbol{s}$, where $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$ is a continuous vector and the set $\boldsymbol{\mathcal{S}}$ of all the styles may be either shared among all the domains or it can be partitioned in different domain-specific subsets (i.e., $\boldsymbol{\mathcal{S}} = \{\boldsymbol{\mathcal{S}}_1, \cdots, \boldsymbol{\mathcal{S}}_m\}$). In our case, we use the second solution and we denote with $\hat{\boldsymbol{x}} = G(\boldsymbol{x}, \boldsymbol{s})$ the translation operation, where $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_i$ is the *source* image (and its domain implicitly indicates the source domain i), $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}_j$ is the *target* style code and $\hat{\boldsymbol{x}} \in \boldsymbol{\mathcal{X}}_j$ is the generated image.

The MMUIT task is an extension of the above description in which:

a. *Training is unsupervised.* This is crucial when collecting paired images is time consuming or impossible.

b. *The source content is preserved.* A translated image $\hat{\boldsymbol{x}} = G(\boldsymbol{x}, \boldsymbol{s})$ should preserve domain-invariant characteristics (commonly called "content") and change only the domain-specific properties of the source image $\boldsymbol{x}$. For example, in male $\leftrightarrow$ female translations, $\hat{\boldsymbol{x}}$ should keep the pose and the identity of $\boldsymbol{x}$, while changing other aspects to look like a female or a male.

c. *The output is multi-modal.* Most I2I translations methods are deterministic, since, at inference time, they can produce only *one* translated image $\hat{\boldsymbol{x}}$ given a source image $\boldsymbol{x}$ and a target domain j. However, in many practical applications, it is desirable that the appearance of

$\hat{\boldsymbol{x}}$ depends also on some random factor, in such a way to be able to produce different plausible translations.

There are mainly two mechanisms that can be used to obtain a specific style code $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}_j$. The first option is to sample a random vector (e.g., $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$) and then use an MLP to transform $\boldsymbol{z}$ into a style code: $\boldsymbol{s} = M(\boldsymbol{z}, j)$ [21], where j is the domain label. The second option is based on extracting the code from a reference image ($\boldsymbol{x}' \in \boldsymbol{\mathcal{X}}_j$) by means of an encoder: $\boldsymbol{s} = E(\boldsymbol{x}')$. In our case, we use both of them.

4. Method

Fig. 2 shows the main intuition behind our method. A style space in which different domains are well separated (i.e., disentangled) may not be sufficient to guarantee smooth inter-domain interpolations. When the domain-specific distributions are too far apart from each other, this may lead to what we call "training gaps", i.e., portions of the space that are not populated with training samples. Consequently, at training time, the network has not observed samples in those regions, and, at inference time, it may misbehave when sampling in those regions (e.g., producing image artifacts). Moreover, a non-compact style space may create intra-domain "training gaps", leading to the generation of non-realistic images when drawing style codes in these areas. Thus, we argue that smoothness is related to reducing these training gaps and compacting the latent space.

Note that the commonly adopted domain loss [9] or the multitask adversarial discriminators [10, 28] might result in domain distributions far apart from each other to facilitate the discriminative task. In order to reduce these training gaps, the domain distributions are expected to be pulled closer while keeping the disentanglement. To achieve these goals, we propose two training losses, described below. First, we use a triplet loss [35] to guarantee the separability of the style codes in different domains. The advantage of the triplet loss is that, using a small margin, the disentanglement of different domains in the latent space can be preserved. Meanwhile, it is convenient to control the inter-domain distance by adjusting the margin. However, our empirical results show that the triplet loss alone is insufficient to reduce the training gaps. For this reason, we propose to compact style space using a second loss.

We propose to use the Kullback-Leibler (KL) divergence with respect to an a priori Gaussian distribution to make the style space compact. This choice is inspired by the regularization adopted in Variational AutoEncoders (VAEs) [24]. In VAEs, an encoder network is trained to estimate the parameters of a multivariate Gaussian given a single (real) input example. However, in our case, a style code $\boldsymbol{s}$ can be either real (using the encoder E, see Sec. 3) or randomly sampled (using M, Sec. 3), and training an additional encoder to estimate the distribution parameters may be hard and not necessary. For this reason, we propose to simplify the KL divergence using a sample-based ℓ_2 regularization.

Finally, as mentioned in Sec. 3, another important aspect of the MMUIT task is content preservation. To this aim, we propose to use a third loss, based on the idea that the content of an image should be domain-independent (see Sec. 3) and that the similarity of two images with respect to the content can be estimated using a "perceptual distance". The latter is computed using a network pre-trained to simulate the human perceptual similarity [41].

In Sec. 4.1 we provide the details of these three losses. Note that our proposed losses can be applied to different I2I translation architectures which have an explicit style space (e.g., a style encoder E, see Sec. 3), possibly jointly with other losses. In Sec. 4.2 we show a specific implementation case, which we used in our experiments and which is inspired to StarGAN v2 [10]. In the Supplementary Material we show another implementation case based on TUNIT [3].

4.1. Modeling the Style Space

Smoothing and disentangling the style space. We propose to use a triplet loss, which is largely used in metric learning [35, 38, 14, 7], to preserve the domain disentanglement:

$$\mathcal{L}_{tri} = \mathbb{E}_{(\boldsymbol{s}_a, \boldsymbol{s}_p, \boldsymbol{s}_n) \sim \boldsymbol{\mathcal{S}}}[\max(||\boldsymbol{s}_a - \boldsymbol{s}_p)|| - ||\boldsymbol{s}_a - \boldsymbol{s}_n|| + \alpha, 0)], \tag{1}$$

where α is a constant margin and $\boldsymbol{s}_a$ and $\boldsymbol{s}_p$ (i. e., the *anchor* and the *positive*, adopting the common terminology of the triplet loss [35]) are style codes extracted from the same domain (e.g., $\boldsymbol{s}_a, \boldsymbol{s}_p \in \boldsymbol{\mathcal{S}}_i$), while the *negative* $\boldsymbol{s}_n$ is extracted from a different domain ($\boldsymbol{s}_n \in \boldsymbol{\mathcal{S}}_j, j \neq i$). These style codes are obtained by sampling real images and using the encoder. In more detail, we randomly pick two images from the same domain i ($\boldsymbol{x}_a, \boldsymbol{x}_p \in \boldsymbol{\mathcal{X}}_i$), a third image from another, randomly chosen, domain j ($\boldsymbol{x}_n \in \boldsymbol{\mathcal{X}}_j, j \neq i$), and then we get the style codes using $\boldsymbol{s}_k = E(\boldsymbol{x}_k), k \in \{a, p, n\}$. Using Eq. (1), the network learns to cluster style codes of the same domain. Meanwhile, when the style space is compact, the margin α can control and preserve the disentanglement among the resulting clusters.

Thus, we encourage a compact space forcing an a prior Gaussian distribution on the set of all the style codes $\boldsymbol{\mathcal{S}}$:

$$\mathcal{L}_{kl} = \mathbb{E}_{\boldsymbol{s} \sim \boldsymbol{\mathcal{S}}}[\mathcal{D}_{\mathrm{KL}}(p(\boldsymbol{s}) \| \mathcal{N}(\mathbf{0}, \boldsymbol{I}))], \tag{2}$$

where $\boldsymbol{I}$ is the identity matrix, $\mathcal{D}_{\mathrm{KL}}(p\|q)$ is the Kullback-Leibler (KL) divergence and $p(\boldsymbol{s})$ is the distribution corresponding to the style code $\boldsymbol{s}$. However, $p(\boldsymbol{s})$ is unknown. In VAEs, $p(\boldsymbol{s})$ is commonly estimated assuming a Gaussian shape and using an encoder to regress the mean and the covariance-matrix parameters of each single sample-based distribution [24]. Very recently, Ghosh et al. [11] showed that, assuming the variance to be constant for all the samples, the KL divergence regularization can be simplified (up to a constant) to $\mathcal{L}_{SR}^{CV}(\boldsymbol{x}) = ||\boldsymbol{\mu}(\boldsymbol{x})||_2^2$, where "CV" stands for Constant-Variance, and $\boldsymbol{\mu}(\boldsymbol{x})$ is the mean estimated by the encoder using $\boldsymbol{x}$. In this paper we propose a further simplification based on the assumption that $\boldsymbol{\mu}(\boldsymbol{s}) = \boldsymbol{s}$ (which is reasonable if $\boldsymbol{\mu}$ is estimated using only one sample) and we eventually get the proposed *Style Regularization* (SR) loss:

$$\mathcal{L}_{SR} = \mathbb{E}_{\boldsymbol{s}\sim\boldsymbol{\mathcal{S}}}[||\boldsymbol{s}||_2^2]. \tag{3}$$

Eq. (3) penalizes samples $\boldsymbol{s}$ with a large ℓ_2 norm, so encouraging the distribution of $\boldsymbol{\mathcal{S}}$ to be a shrunk Gaussian centered on the origin. Intuitively, while the SR loss compacts the space, the triplet loss avoids a domain entanglement in the compacted region (see also the Supplementary Material). Finally, we describe below how the style-code samples are drawn in Eq. (3) ($\boldsymbol{s} \sim \boldsymbol{\mathcal{S}}$). We use a mixed strategy, including both real and randomly generated codes. More in detail, with probability 0.5, we use a real sample $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and we get: $\boldsymbol{s} = E(\boldsymbol{x})$, and, with probability 0.5, we use $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{s} = M(\boldsymbol{z}, j)$. In practice, we alternate mini-batch iterations in which we use only real samples with iterations in which we use only generated samples.

Preserving the source content. The third loss we propose aims at preserving the content in the I2I translation:

$$\mathcal{L}_{cont} = \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\mathcal{X}},\boldsymbol{s}\sim\boldsymbol{\mathcal{S}}}[\psi(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{s}))], \tag{4}$$

where $\psi(\boldsymbol{x}_1, \boldsymbol{x}_2)$ estimates the perceptual distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ using an externally pre-trained network. The rationale behind Eq. (4) is that, given a source image $\boldsymbol{x}$ belonging to domain $\boldsymbol{\mathcal{X}}_i$, for each style code $\boldsymbol{s}$, extracted from the set of *all* the domains $\boldsymbol{\mathcal{S}}$, we want to minimize the perceptual distance between $\boldsymbol{x}$ and the transformed image $G(\boldsymbol{x}, \boldsymbol{s})$. By minimizing Eq. (4), the perceptual content (extracted through $\psi(\cdot)$) is encouraged to be independent of the domain (see the definition of content preservation in Sec. 3). Although different perceptual distances can be used (e.g., the Euclidean distance on VGG features [19]), we implement $\psi(\boldsymbol{x}_1, \boldsymbol{x}_2)$ using the Learned Perceptual Image Patch Similarity (LPIPS) metric [41], which was shown to be well aligned with the human perceptual similarity [41] and it is obtained using a multi-layer representation of the two input images $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ in a pre-trained network.

The sampling procedure in the *content preserving* loss ($\mathcal{L}_{cont}$) is similar to the SR loss. First, we randomly sample $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. Then, we either sample a different reference image $\boldsymbol{x}' \in \boldsymbol{\mathcal{X}}$ and get $\boldsymbol{s} = E(\boldsymbol{x}')$, or we use $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{s} = M(\boldsymbol{z}, j)$.

We sum together the three proposed losses and we get:

$$\mathcal{L}_{smooth} = \mathcal{L}_{cont} + \lambda_{sr}\mathcal{L}_{SR} + \mathcal{L}_{tri}, \tag{5}$$

where λ_{sr} is the SR loss-specific weight.

4.2. Smoothing the Style Space of an Existing Model

The proposed $\mathcal{L}_{smooth}$ can be plugged in existing MMUIT methods which have an explicit style space, by summing it with their original objective function ($\mathcal{L}_{orig}$):

$$\mathcal{L}_{new} = \mathcal{L}_{smooth} + \mathcal{L}_{orig}. \tag{6}$$

In this subsection, we show an example in which $\mathcal{L}_{orig}$ is the original loss of the MMUIT state-of-the-art StarGAN v2 [10]. In the Supplementary Material we show another example based on TUNIT [3], which is the state of the art of fully-unsupervised image-to-image translation.

In StarGAN v2, the original loss is:

$$\mathcal{L}_{orig} = \lambda_{sty}\mathcal{L}_{sty} - \lambda_{ds}\mathcal{L}_{ds} + \lambda_{cyc}\mathcal{L}_{cyc} + \mathcal{L}_{adv} \tag{7}$$

where $\lambda_{sty}, \lambda_{ds}$ and λ_{cyc} control the contribution of the *style reconstruction*, the *diversity sensitive*, and the *cycle consistency* loss, respectively.

The *style reconstruction* loss [16, 44, 10] pushes the target code ($\boldsymbol{s}$) and the code extracted from the generated image ($E(G(\boldsymbol{x}, \boldsymbol{s}))$) to be as close as possible:

$$\mathcal{L}_{sty} = \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\mathcal{X}},\boldsymbol{s}\sim\boldsymbol{\mathcal{S}}}\left[\|\boldsymbol{s} - E(G(\boldsymbol{x}, \boldsymbol{s}))\|_1\right]. \tag{8}$$

The *diversity sensitive* loss [10, 31] encourages G to produce diverse images:

$$\mathcal{L}_{ds} = \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\mathcal{X}}_i,(\boldsymbol{s}_1,\boldsymbol{s}_2)\sim\boldsymbol{\mathcal{S}}_j}\left[\|G(\boldsymbol{x}, \boldsymbol{s}_1) - G(\boldsymbol{x}, \boldsymbol{s}_2)\|_1\right]. \tag{9}$$

The *cycle consistency* [43, 9, 10] loss is used to preserve the content of the source image $\boldsymbol{x}$:

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

Finally, StarGAN v2 uses a multitask discriminator [28] D, which consists of multiple output branches. Each branch D_j learns a binary classification determining whether an image $\boldsymbol{x}$ is a real image of its dedicated domain j or a fake image. Thus, the *adversarial* loss can be formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\mathcal{X}}_i,\boldsymbol{s}\sim\boldsymbol{\mathcal{S}}_j}[\log D_i(\boldsymbol{x}) + \log(1 - D_j(G(\boldsymbol{x}, \boldsymbol{s})))] \tag{11}$$

Note that this loss encourages the separation of the domain-specific distributions without controlling the relative inter-domain distance (Sec. 4). We use it jointly with our $\mathcal{L}_{tri}$.

We refer the reader to [10] and to the Supplementary Material for additional details. In Sec. 6 we evaluate the combination of our $\mathcal{L}_{smooth}$ with StarGAN v2 (Eq. (7)), while in the Supplementary Material we show additional experiments in which $\mathcal{L}_{smooth}$ is combined with TUNIT [3].

5. Evaluation Protocols

FID. For each translation $\boldsymbol{\mathcal{X}}_i \to \boldsymbol{\mathcal{X}}_j$, we use 1,000 test images and estimate the Fréchet Inception Distance (FID) [15] using interpolation results. In more detail, for each image, we randomly sample two style codes ($\boldsymbol{s}_1 \in \boldsymbol{\mathcal{S}}_i$ and $\boldsymbol{s}_2 \in \boldsymbol{\mathcal{S}}_j$), which are linearly interpolated using 20 points. Each point (included $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$) is used to generate a translated image. The FID values are computed using the $20 \times 1,000$ outputs. A lower FID score indicates a lower discrepancy between the image quality of the real and generated images.

LPIPS. For a given domain $\boldsymbol{\mathcal{X}}_i$, we use 1,000 test images $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_i$, and, for each $\boldsymbol{x}$, we randomly generate 10 image translations in the target domain $\boldsymbol{\mathcal{X}}_j$. Then, the LPIPS [41] distances among the 10 generated images are computed. Finally, all distances are averaged. A higher LPIPS distance indicates a greater diversity among the generated images. Note that the LPIPS distance ($\psi(\boldsymbol{x}_1, \boldsymbol{x}_2)$) is computed using an *externally pre-trained* network [41], which is the same we use in Eq. (4) at training time.

FRD. For the specific case of face translations, we use a metric based on a pretrained VGGFace2 network (ϕ) [35, 7], which estimates the visual distance between two faces. Note that the identity of a person may be considered as a specific case of "content" (Sec. 3). We call this metric the Face Recognition Distance (FRD):

$$\mathrm{FRD} = \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\mathcal{X}}, \boldsymbol{s} \sim \boldsymbol{\mathcal{S}}} \left[\|\phi(\boldsymbol{x}) - \phi(G(\boldsymbol{x}, \boldsymbol{s}))\|_2^2 \right]. \tag{12}$$

PS. Karras et al. [21] recently proposed the Perceptual Path Length (PPL) to evaluate the smoothness and the disentanglement of a semantic latent space. PPL is based on measuring the LPIPS distance between close points in the style space. However, one issue with the PPL is that it can be minimized by a collapsed generator. For this reason, we alternatively propose the Perceptual Smoothness (PS) metric, which returns a normalized score in $[0, 1]$, indicating the smoothness of the style space.

In more detail, let $\boldsymbol{s}_0$ and $\boldsymbol{s}_T$ be two codes randomly sampled from the style space, $P = (\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T)$ the sequence of the linearly interpolated points between $\boldsymbol{s}_0$ and $\boldsymbol{s}_T$, and $A = (G(\boldsymbol{x}, \boldsymbol{s}_0), \ldots, G(\boldsymbol{x}, \boldsymbol{s}_T))$ the corresponding sequence of images generated starting from a source image $\boldsymbol{x}$. We measure the degree of linear *alignment* of the generated images using:

$$\ell_{\texttt{alig}} = \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\mathcal{X}}, \boldsymbol{s}_0, \boldsymbol{s}_T \sim \boldsymbol{\mathcal{S}}} \left[\frac{\delta(\boldsymbol{x}, \boldsymbol{s}_0, \boldsymbol{s}_T)}{\sum_{t=1}^{T} \delta(\boldsymbol{x}, \boldsymbol{s}_{t-1}, \boldsymbol{s}_t)} \right] \tag{13}$$

where $\delta(\boldsymbol{x}, \mathrm{s}_1, \mathrm{s}_2) = \psi(G(\boldsymbol{x}, \mathrm{s}_1), G(\boldsymbol{x}, \mathrm{s}_2))$ and $\psi(\cdot, \cdot)$ is the LPIPS distance (modified to be a proper metric, more details in the Supplementary Material). When $\ell_{\texttt{alig}} = 1$, then the perceptual distance between $G(\boldsymbol{x}, \boldsymbol{s}_0)$ and $G(\boldsymbol{x}, \boldsymbol{s}_T)$ is equal to the sum of the perceptual distances between consecutive elements in A, thus, the images in A lie along a line in the space of $\psi(\cdot, \cdot)$ (which represents the human perceptual similarity [41]). Conversely, when $\ell_{\texttt{alig}} < 1$, then the images in A contain some visual attribute not contained in any of the endpoints. For example, transforming a short-hair male person to a short-hair girl, we may have $\ell_{\texttt{alig}} < 1$ when the images in A contain people with long hair. However, although aligned, the images in A may have a non-uniform distance, in which $\delta(\boldsymbol{x}, \boldsymbol{s}_{t-1}, \boldsymbol{s}_t)$ varies depending on t. In order to measure the *uniformity* of these distances, we use the opposite of the Gini inequality coefficient [12]:

$$\ell_{\texttt{uni}} = \mathbb{E}_{\substack{\boldsymbol{x} \sim \boldsymbol{\mathcal{X}} \\ \boldsymbol{s}_0, \boldsymbol{s}_T \sim \boldsymbol{\mathcal{S}}}} \left[1 - \frac{\sum_{i,j=1}^{T} |\delta(\boldsymbol{x}, \boldsymbol{s}_{i-1}, \boldsymbol{s}_i) - \delta(\boldsymbol{x}, \boldsymbol{s}_{j-1}, \boldsymbol{s}_j)|}{2T^2 \mu_P} \right]$$

where μ_P is the average value of $\delta(\cdot)$ computed over all the pairs of elements in $P = (\boldsymbol{s}_0, \ldots, \boldsymbol{s}_T)$. Intuitively, $\ell_{\texttt{uni}} = 1$ when an evenly-spaced linear interpolation of the style codes corresponds to constant changes in the perceived difference of the generated images, while $\ell_{\texttt{uni}} = 0$ when there is only one abrupt change in a single step. Finally, we define PS as the harmonic mean of $\ell_{\texttt{alig}}$ and $\ell_{\texttt{uni}}$:

$$\mathrm{PS} = 2 \cdot \frac{\ell_{\texttt{alig}} \cdot \ell_{\texttt{uni}}}{\ell_{\texttt{alig}} + \ell_{\texttt{uni}}} \in [0, 1]. \tag{14}$$

6. Experiments

Baselines. We compare our method with three state-of-the-art approaches: (1) StarGAN v2 [10], the state of the art for the MMUIT task; (2) HomoGAN [8]; and (3) TUNIT [3]. Moreover, as a reference for a high image quality, we also use InterFaceGAN [36], a StyleGAN-based method (trained with 1024×1024 images) which interpolates the pre-trained semantic space of StyleGAN [21] (see Sec. 2). InterFaceGAN is not designed for domain translation and for preserving the source content, but it can linearly interpolate a fixed latent space, massively trained with high-resolution images. All the baselines are tested using the original publicly available codes.

Datasets. We follow the experimental protocol of StarGAN v2 [10] and we use the CelebA-HQ [20] and the AFHQ dataset [10]. The domains are: *male-female*, *smile-no smile*, *young-non young* in CelebA-HQ; *cat*, *dog*, and *wildlife* in AFHQ. For a fair comparison, all models (except InterFaceGAN) are trained with 256×256 images. Additional details are provided in the Supplementary Material.

Settings. We test our method in two experimental settings, respectively called "unsupervised" (with only set-level annotations) and "truly unsupervised" (no annotations [3]). Correspondingly, we plug our training losses ($\mathcal{L}_{smooth}$) in the state-of-the art StarGAN v2 [10] and TUNIT [3] (see Sec. 4.1). In each setting, we plug our method in the original architecture without adding additional modules and adopting the original hyper-parameter values without tuning. We refer to the Supplementary Material for more details.

6.1. Smoothness of the Style Space

Fig. 3 shows a qualitative evaluation using the style-space interpolation between a source image and a reference

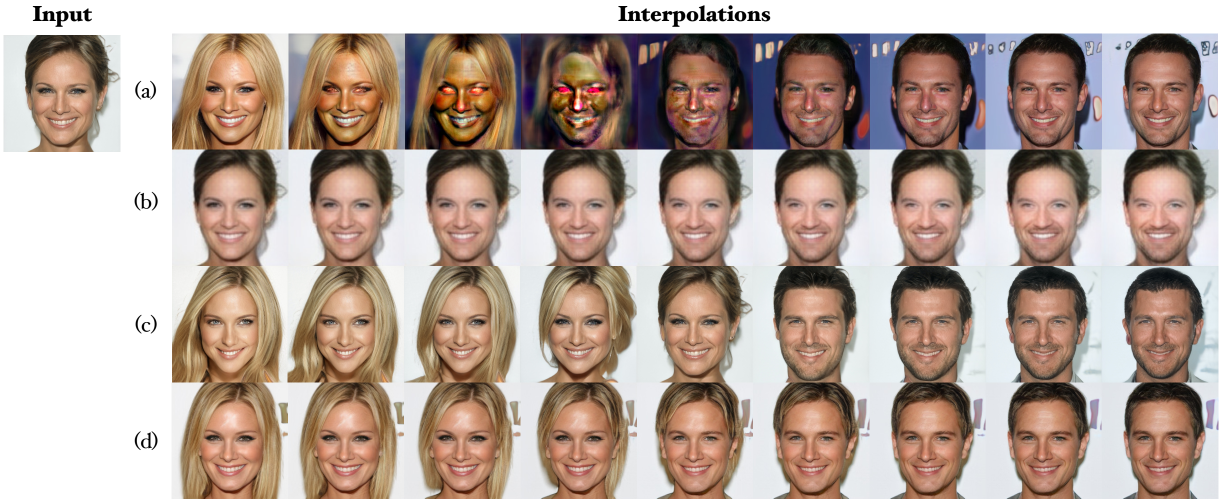

Figure 3: Inter-domain interpolation results: (a) StarGAN v2 [10], (b) HomoGAN [8], (c) InterFaceGAN [36], (d) ours. The domains correspond to genders. Our method generates smoother results while better preserving the source-person identity.

Model	**PS↑**			**FRD↓**		
	Gender	Smile	Age	Gender	Smile	Age
HomoGAN [8]	.401	.351	.389	.903	.820	.842
StarGAN v2 [10]	.272	.282	.283	1.082	.894	.882
Ours	**.504**	**.513**	**.601**	**.837**	**.625**	**.650**
InterFaceGAN [36][§]	.328	.436	.409	.884	.560	.722

Table 1: Smoothness degree and identity preservation on the CelebA-HQ dataset. [§]Trained on 1024×1024 images.

Model	**FID↓**			**LPIPS↑**		
	Gender	Smile	Age	Gender	Smile	Age
HomoGAN [8]	55.23	58.02	57.50	.010	.005	.008
StarGAN v2 [10]	48.35	29.65	26.60	**.442**	**.413**	**.407**
Ours	**23.37**	**22.21**	**23.57**	.337	.095	.128
InterFaceGAN [36][§]	13.75	12.81	12.25	.211	.115	.146

Table 2: Image quality and translation diversity on the CelebA-HQ dataset. [§]Trained on 1024×1024 images.

style. As mentioned in Sec. 1 and 4, StarGAN v2 frequently generates artifacts in inter-domain interpolations (see Fig. 3 (a)). HomoGAN results are very smooth, but they change very little the one from the other, and the model synthetizes lower quality images (Fig. 3 (b)). InterFaceGAN (Fig. 3 (c)) was trained at a higher image resolution with respect to the other models (ours included). However, compared to our method (Fig. 3 (d)), the interpolation results are less smooth, especially in the middle, while the image quality of both methods is very similar. Moreover, comparing our approach to StarGAN v2, our method better preserves the background content in all the generated images.

These results are quantitatively confirmed in Tab. 1. The PS scores show that our proposal improves the state of the art significantly, which means that it increases the smoothness of the style space in all the CelebA-HQ experiments. Note that our results are also better than InterFaceGAN, whose latent space is based on the pretrained StyleGAN [21], a very large capacity and training-intensive model. Tab. 3 and Fig. 5 show similar results also in the challenging AFHQ dataset, where there is a large inter-domain shift. In this dataset, we tested both the unsupervised and the truly unsupervised setting, observing a clear improvement of both the semantic-space smoothness and the image quality using our method.

The comparison of the qualitative results in Fig. 3 and Fig. 5 with the PS scores in Tab. 1 and Tab. 3, respectively, show that the proposed PS metric can be reliably used to evaluate MMUIT models with respect to the style-space smoothness. In the Supplementary Material we show additional evidence on the quality of the PS metrics and how domain separation can be controlled by tuning the margin value of the triplet loss.

Tab. 2 and 3 show that the improvements on the style-space smoothness and the corresponding interpolation results do not come at the expense of the image quality. Conversely, these tables show that the FID values significantly improve with our method. The LPIPS results in Tab. 2 also show that HomoGAN generates images with little diversity. However, the LPIPS scores of StarGAN v2 are higher than our method. Nevertheless, the LPIPS metric is influenced by the presence of possible artifacts in the generated images, and, thus, an increased LPIPS value is not necessarily a strength of the model. We refer to the Supplementary Material for additional qualitative and quantitative results.

Finally, we performed a user study where we asked 40 users to choose between the face translations generated by

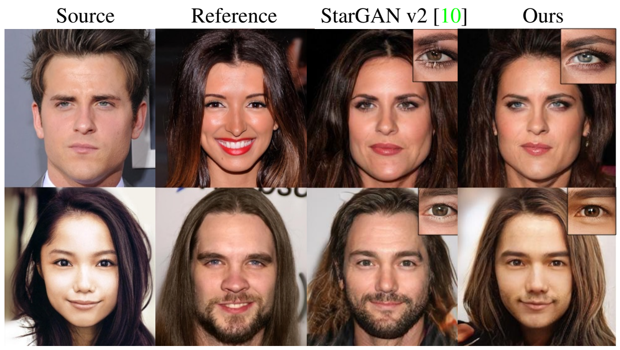

Figure 4: Content preservation using the CelebA-HQ dataset. Our method better preserves the ethnicity and identity of the source images compared to StarGAN v2.

StarGAN v2 and our method, providing 30 random image pairs to each user. In 75.8% of cases, the image generated by our model was selected as the better one, compared to StarGAN v2 (25.2%).

6.2. Identity Preservation

MMUIT models aim at translating images from one domain to another while keeping the content unchanged. While this goal is clear, the degree of content preservation is usually evaluated only qualitatively. Thus, we use the FRD (Sec. 5) and the most popular I2I translation task (face translation) to measure the content preservation of the compared models. Tab. 1 shows that our FRD is the lowest over all the methods compared on the CelebA-HQ dataset, indicating that our method better maintains the person identity of source images. Qualitatively, Fig. 4 shows that our method better preserves some distinct face characteristics (e.g., the eye color, the chin shape, or the ethnicity) of the source image while changing the style (i.e., the gender). This result also suggests that our model might be less influenced by the CelebA-HQ biases (e.g. Caucasian people). Additional experiments, with similar results, are presented in the Supplementary Material for smile and age translations.

6.3. Ablation Study

In this section, we evaluate the importance of each proposed component. Tab. 4 shows the FID, LPIPS, PS and FRD values for all the configurations, where each component is individually added to the baseline StarGAN v2, using CelebA-HQ. First, we observe that adding the $\mathcal{L}_{tri}$ loss to the baseline improves the quality, the diversity and the content preservation of the generated images. However the PS score decreases. This result suggests that better disentanglement might separate too much the styles between domains, thus decreasing the interpolation smoothness. The addition of $\mathcal{L}_{SR}$ helps improving most of the metrics but the diversity, showing that a more compact style space is a desirable property for MMUIT. As mentioned before, we note that higher diversity (LPIPS) might not be strictly related to high-quality images.

The combination of the two proposed smoothness losses dramatically improves the quality of generated images and the smoothness of the style space. This suggests that the style space should be compact and disentangled, while keeping the style clusters of different domains close to each other. Finally, $\mathcal{L}_{cont}$ further improves the FID, the PS and the FRD scores. The final configuration corresponds to our full-method and confirms that all the proposed components are helpful. We refer to the Supplementary Material Sec. B for additional analysis on the contribution of our losses.

Model	Setting	FID↓	PS↑
StarGAN v2 [10]	Unsupervised	15.64	.226
Ours		**14.67**	**.301**
TUNIT [3]	Truly Unsupervised	29.45	.443
Ours		**16.59**	**.447**

Table 3: Quantitative evaluation on the AFHQ dataset.

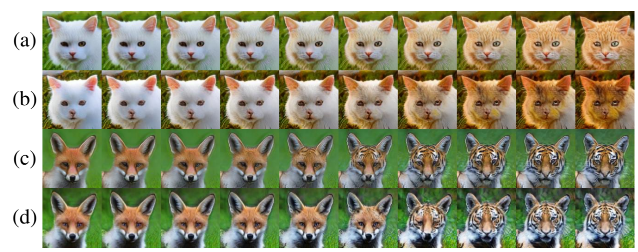

Figure 5: AFHQ dataset. (b,d) Generation results using TUNIT [3]. (a,c) TUNIT jointly with our losses.

Model	FID↓	LPIPS↑	PS↑	FRD↓
A: Baseline StarGAN v2 [10]	48.35	**.442**	.272	1.082
A + $\mathcal{L}_{tri}$	37.54	.403	.292	1.040
A + $\mathcal{L}_{SR}$	35.23	.368	.432	.912
A + $\mathcal{L}_{SR}$, $\mathcal{L}_{tri}$	24.29	.374	.501	.848
A + $\mathcal{L}_{SR}$, $\mathcal{L}_{tri}$, $\mathcal{L}_{cont}$	**23.37**	.337	**.504**	**.837**

Table 4: Ablation study on the CelebA-HQ dataset with a gender translation task.

7. Conclusion

In this paper, we proposed a new training strategy based on three specific losses which jointly improve both the smoothness of the style space and the content preservation of existing MMUIT models. We also proposed the PS metric, which specifically evaluates the style smoothness of I2I translation models. The experimental results show that our method significantly improves both the smoothness and the quality of the interpolation results and the translated images.
Acknowledgements. This work was supported by EU H2020 SPRING No.871245 and by AI4Media No.951911.

References

[1] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. *NeurIPS*, 2018. 3

[2] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *ICLR*, 2018. 3

[3] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. *arXiv preprint arXiv:2006.06500*, 2020. 2, 3, 4, 5, 6, 8

[4] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *BMVC*, 2016. 2

[5] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018. 3

[6] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *ICML*, 2018. 3

[7] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2018. 4, 6

[8] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *CVPR*, 2019. 3, 6, 7

[9] 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 *CVPR*, 2018. 1, 2, 3, 4, 5

[10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[11] Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael J. Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. In *ICLR*, 2020. 4

[12] Corrado Gini. Variabilità e mutabilità (variability and mutability). *Memorie di metodologica statistica*, 1912. 6

[13] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, 2019. 3

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017. 4

[15] 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 *NeurIPS*, 2017. 5

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1, 3, 5

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[18] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *ICLR*, 2020. 3

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 1, 6

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 4, 6, 7

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

[23] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017. 3

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3, 4

[25] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 3

[26] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10):2402–2417, 2020. 1, 2, 3

[27] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 3

[28] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 3, 4, 5

[29] Yahui Liu, Marco De Nadai, Deng Cai, Huayang Li, Xavier Alameda-Pineda, Nicu Sebe, and Bruno Lepri. Describe what to change: A text-guided unsupervised image-to-image translation approach. In *ACM MM*, 2020. 1

[30] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, 2019. 3

[31] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, 2019. 5

[32] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instance-aware image-to-image translation. In *ICLR*, 2019. 3

[33] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020. 3

[34] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. In *ICCV Workshops*, 2019. 3

[35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2, 4, 6

[36] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 3, 6, 7

[37] Aliaksandr Siarohin, Stéphane Lathuilière, Enver Sangineto, and Nicu Sebe. Appearance and Pose-Conditioned Human Image Generation using Deformable GANs. *IEEE TPAMI*, 2020. 1

[38] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. 4

[39] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017. 3

[40] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *ICCV*, 2019. 3

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4, 5, 6

[42] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020. 3

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 3, 5

[44] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 3, 5