GMM-UNIT: UNSUPERVISED MULTI-DOMAIN AND MULTI-MODAL IMAGE-TO-IMAGE TRANSLATION VIA ATTRIBUTE GAUSSIAN MIXTURE MODELLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised image-to-image translation aims to learn a mapping between several visual domains by using unpaired training pairs. Recent studies have shown remarkable success in image-to-image translation for multiple domains but they suffer from two main limitations: they are either built from several *two-domain mappings* that are required to be learned independently and/or they generate low-diversity results, a phenomenon known as *model collapse*. To overcome these limitations, we propose a method named GMM-UNIT based on a content-attribute disentangled representation, where the attribute space is fitted with a GMM. Each GMM component represents a domain, and this simple assumption has two prominent advantages. First, the dimension of the attribute space does not grow linearly with the number of domains, as it is the case in the literature. Second, the continuous domain encoding allows for interpolation between domains and for extrapolation to unseen domains. Additionally, we show how GMM-UNIT can be constrained down to different methods in the literature, meaning that GMM-UNIT is a unifying framework for unsupervised image-to-image translation.

1 INTRODUCTION

Translating images from one domain into another is a challenging task that has significant influence on many real-world applications where data are expensive, or impossible to obtain and to annotate. Image-to-Image translation models have indeed been used to increase the resolution of images (Dong et al., 2014), fill missing parts (Pathak et al., 2016), transfer styles (Gatys et al., 2016), synthesize new images from labels (Liu et al., 2017), and help domain adaptation (Bousmalis et al., 2017; Murez et al., 2018). In many of these scenarios, it is desirable to have a model mapping one image to multiple domains, while providing visual diversity (i.e. a day scene $\leftrightarrow$ night scene in different seasons). However, the existing models can either map an image to *multiple* stochastic results in a single domain, or map in the same model *multiple* domains in a deterministic fashion. In other words, most of the methods in the literature are either *multi*-domain or *multi*-modal.

Several reasons have hampered a stochastic translation of images to multiple domains. On the one hand, most of the Generative Adversarial Network (GAN) models assume a deterministic mapping (Choi et al., 2018; Pumarola et al., 2018; Zhu et al., 2017a), thus failing at modelling the correct distribution of the data (Huang et al., 2018). On the other hand, approaches based on Variational Auto-Encoders (VAEs) usually assume a shared and common zero-mean unit-variance normally distributed space (Huang et al., 2018; Zhu et al., 2017b), limiting to two-domain translations.

In this paper, we propose a novel image-to-image translation model that disentangles the visual content from the domain attributes. The attribute latent space is assumed to follow a Gaussian mixture model (GMM), thus naming the method: GMM-UNIT (see Figure 1). This simple assumption allows four key properties: *mode-diversity* thanks to the stochastic nature of the probabilistic latent model, *multi-domain translation* since the domains are represented as clusters in the same attribute spaces, *scalability* because the domain-attribute duality allows modeling a very large number of domains without increasing the dimensionality of the attribute space, and *few/zero-shot generation* since the continuity of the attribute representation allows interpolating between domains and extrapolating to unseen domains with very few or almost no observed data from these domains. The code and models will be made publicly available.

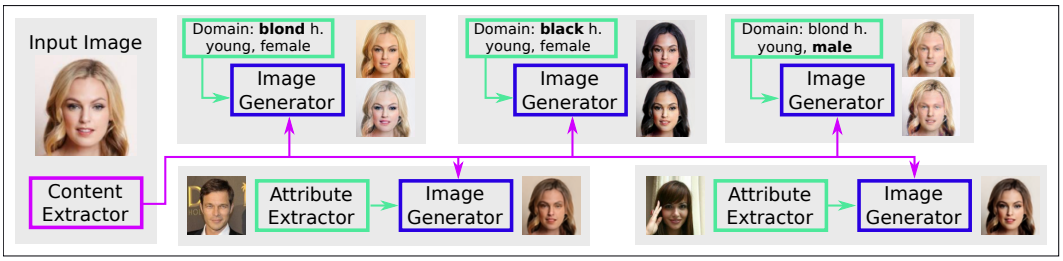

Figure 1: GMM-UNIT working principle. The content is extracted from the input image (left, purple box), while the attribute (turquoise box) can be either sampled (top images) or extracted from a reference image (bottom images). Either way, the generator (blue box) is trained to output realistic images belonging to the domain encoded in the attribute vector. This is possible thanks to the disentangled attribute-content latent representation of GMM-UNIT and the generalisation properties associated to Gaussian mixture modeling.

2 Related work

Our work is best placed in the literature of image-to-image translation, where the challenge is to translate one image from a visual domain (e.g. summer) to another one (e.g. winter). This problem is inherently ill-posed, as there could be many mappings between two images. Thus, researchers have tried to tackle the problem from many different perspectives. The most impressive results on this task are undoubtedly related to GANs, which aim to synthesize new images as similar as possible to the real data through an adversarial approach between a Discriminator and a Generator. The former continuously learns to recognize real and fake images, while the latter tries to generate new images that are indistinguishable from the real data, and thus to fool the Discriminator. These networks can be effectively conditioned and thus generate new samples from a specific class (Chen et al. (2016)) and a latent vector extracted from the images. For example, Isola et al. (2017) and Wang et al. (2018) trained a conditional GAN to encode the latent features that are shared between images of the same domain and thus decode the features to images of the target domain in a one-to-one mapping. However, this approach is limited to supervised settings, where pairs of corresponding images in different domains are available (e.g. a photos-sketch image pair). In many cases, it is too expensive and unrealistic to collect a large amount of paired data.

Unsupervised Domain Translation. Translating images from one domain to another without a paired supervision is particularly difficult, as the model has to learn how to represent both the content and the domain. Thus, constraints are needed to narrow down the space of feasible mappings between images. Taigman et al. (2017) proposed to minimize the feature-level distance between the generated and input images. Liu et al. (2017) created a shared latent space between the domains, which encourages different images to be mapped in the same latent space. Zhu et al. (2017a) proposed CycleGAN, which uses a cycle consistency loss that requires a generated image to be translated back to the original domain. Similarly, Kim et al. (2017) used a reconstruction loss applying the same approach to both the target and input domains. Mo et al. (2019) later expanded the previous approach to the problem of translating multiple instances of objects in the same image. All these methods, however, are limited to a one-to-one domain mapping, thus requiring training multiple models for cross-domain translation. Recently, Choi et al. (2018) proposed StarGAN, a unified framework to translate images in a **multi-domain** setting through a single GAN model. To do so, they used a conditional label and a domain classifier ensuring network consistency when translating between domains. However, StarGAN is limited to a deterministic mapping between domains.

Style transfer. A related problem is style transfer, which aims to transform the style of an image but not its content (e.g. from a photo to a Monet painting) to another image (Gatys et al., 2015; Huang & Belongie, 2017; Tenenbaum & Freeman, 1997; Donahue et al., 2018). Differently from domain translation, usually the style is extracted from a single reference image. We willshow that our model could be applied to style transfer as well.

Multi-modal Domain Translation. Most existing image-to-image translation methods are deterministic, thus limiting the diversity of the translated outputs. However, even in a one-to-one domain translation such as when we want to translate people's hair from blonde to black, there could be multiple hair styles that are not modeled in a deterministic mapping. The straightforward solution would be to inject noise in the model, but it turned out to be worthless as GANs tend to ignore this injected noise (Mathieu et al., 2015; Isola et al., 2017; Zhu et al., 2017b). To address this prob-

lem, Zhu et al. (2017b) proposed BicycleGAN, which encourages the multi-modality in a paired setting through GANs and Variational Auto-Encoders (VAEs). Almahairi et al. (2018) have instead augmented CycleGAN with two latent variables for the input and target domains and showed that it is possible to increase diversity by marginalizing over these latent spaces. Huang et al. (2018) proposed MUNIT, which assumes that domains share a common content space but different style spaces. Then, they showed that by sampling from the style space and using Adaptive Instance Normalization (AdaIN) (Huang & Belongie, 2017), it is possible to have diverse and multimodal outputs. In a similar vein, Ma et al. (2019) focused on the semantic consistency during the translation, and applied AdaIN to the feature-level space. Recently, Mao et al. (2019) proposed a mode seeking loss to encourage GANs to better explore the modes and help the network avoiding the mode collapse.

Altogether, the models in the literature are either multi-modal or multi-domain. Thus, one has to choose between generating diverse results and training one single model for multiple domains. Here, we propose a unified model to overcome this limitation. Concurrent to our work, DRIT++ (Lee et al. (2019)) also proposed a multi-modal and multi-domain model using a discrete domain encoding and assuming, however, a zero-mean unit-variance Gaussian shared space for multiple modes. We instead propose a content-attribute disentangled representation, where the attribute space fits a GMM distribution. A variational loss forces the latent representation to follow this GMM, where each component is associated to a domain. This is the key to provide for both multi-modal and multi-domain translation. In addition, GMM-UNIT is the first method proposing a continuous encoding of the domains, as opposed to the discrete encoding used in the literature. This is important because it allows for domain interpolation and extrapolation with very few or no data (few/zero-shot generation). The main properties of GMM-UNIT compared to the literature are shown in Table 1.

Table 1: A comparison of the state of the art methods for domain to domain translation.

Method	**Unpaired**	**Multi-domain (MDom)**	**Multi-modal (MMod)**	**Domain encoding**
CycleGAN (Zhu et al., 2017a)	✓			None
BicycleGAN (Zhu et al., 2017b)			✓	None
MUNIT (Huang et al., 2018)	✓		✓	None
StarGAN (Choi et al., 2018)	✓	✓		Discrete
DRIT++ (Lee et al., 2019)	✓	✓	✓	Discrete
GMM-UNIT (Proposed)	✓	✓	✓	Continuous

3 GMM-UNIT

GMM-UNIT is an image-to-image translation model that maps an image to multiple domains in a stochastic fashion. Following recent seminal works (Huang et al., 2018; Lee et al., 2018), our model assumes that each image can be decomposed in a domain-invariant content space and a domain-specific attribute space. In this paper, we model the attribute latent space through Gaussian Mixture Models (GMMs), formally with a K-component Z-dimensional GMM:

$$p(\mathbf{z}) = \sum_{k=1}^{K} \phi_k \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^Z$ denotes a random attribute vector sample, ϕ_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote respectively the weight, mean vector and covariance matrix of the k-th GMM component ($\phi_k \geq 0$, $\sum_{k=1}^{K} \phi_k = 1$, $\boldsymbol{\mu}_k \in \mathbb{R}^Z$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{Z \times Z}$ is symmetric and positive definite). $p(\mathbf{z})$ denotes the probability density of this GMM at $\mathbf{z}$. In the proposed representation, the domains are Gaussian components in a mixture. This simple yet effective model has two prominent advantages. Differently from previous works where each domain is a category and the one-hot vector representation grows linearly with the number of domains, we can encode many more domains than the dimension of the latent attribute space Z. Moreover, the continuous encoding of the domains we are introducing in this paper allows us to navigate in the attribute latent space, thus generating images corresponding to domains that have never (or very little) been observed and allowing to interpolate between two domains.

We note that the state of the art models can be traced back as a particular case of GMMs. Existing multi-domain models such as Choi et al. (2018) or Pumarola et al. (2018) can be modelled with $K = |\text{domain in the training data}|$ and $\forall k\ \boldsymbol{\Sigma}_k = 0$, thus only allowing the generation of a single

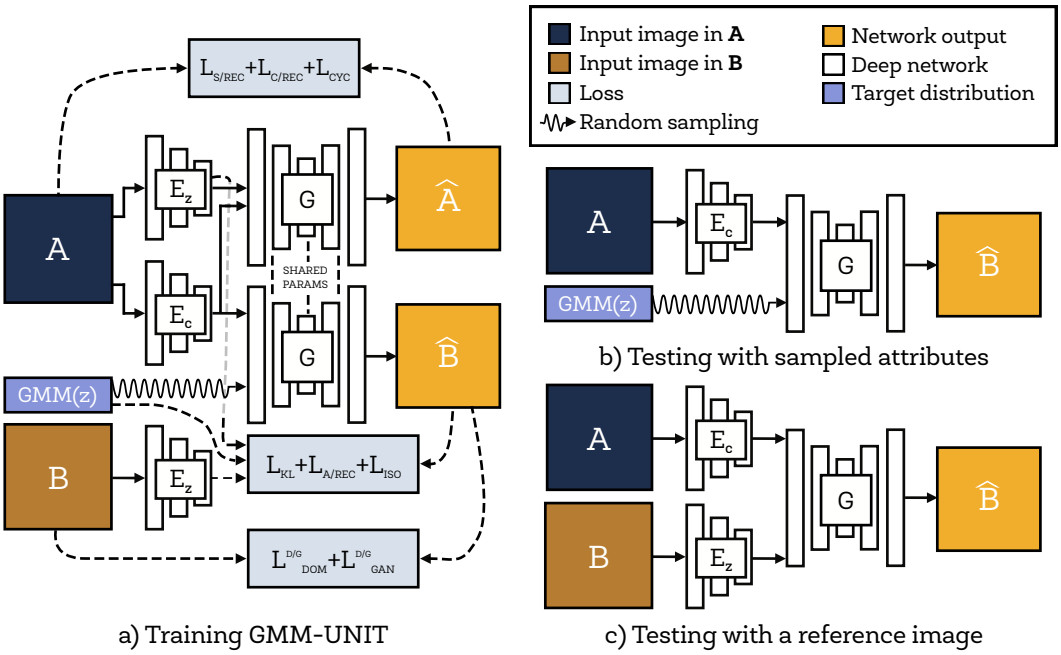

Figure 2: Overview of the GMM-UNIT framework: a) Training phase to translate an image from domain $\boldsymbol{A}$ to $\boldsymbol{B}$. The generator uses the content of the input image (extracted by E_c) and the attribute of the target image (extracted by E_z) to train the network to fit the GMM. b) Testing with target attributes sampled from the GMM distribution of the attributes of domain $\boldsymbol{B}$; c) Testing with an attribute extracted from an image belonging to the target domain $\boldsymbol{B}$. The style of this image is inspired from Zhu et al. (2017b).

result per domain translation. Then, when $K = 1$, $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\Sigma} = \mathbf{I}$ it is possible to model the state of the art approaches in multi-modal translation (Huang et al., 2018; Zhu et al., 2017b), which share a unique latent space where every domain is overlapped, and it is thus necessary to train $N(N-1)$ models to achieve the multi-domain translation. Finally, we can obtain the approach of Lee et al. (2019) by separating the latent space from the domain code. The former is a GMM with $K = 1$, $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\Sigma} = \mathbf{I}$, while the latter is another GMM with $K = |\text{domain in the training data}|$ and $\forall k\ \Sigma_k = 0$. Thus, our GMM-UNIT is a generalization of the existing state of the art. the In the next sections, we formalize our model and show that the use of GMMs for the latent space allows learning multi-modal and multi-domain mappings, and also few/zero-shot image generation.

3.1 The content-attribute disentangling approach

GMM-UNIT follows the generative-discriminative philosophy. The generator inputs a *content* latent code $\mathbf{c} \in \boldsymbol{\mathcal{C}} = \mathbb{R}^C$ and an *attribute* latent code $\mathbf{z} \in \boldsymbol{\mathcal{Z}} = \mathbb{R}^Z$, and outputs a generated image $G(\mathbf{c}, \mathbf{z})$. This image is then fed to a discriminator that must discern between "real" or "fake" images ($D_{\text{r/f}}$), and must also recognize the domain of the generated image (D_{dom}). For an image $\mathbf{x}^n$ from domain $\boldsymbol{\mathcal{X}}^n$ (i.e. $\mathbf{x}^n \sim p_{\boldsymbol{\mathcal{X}}^n}$), its latent attribute $\mathbf{z}^n$ is assumed to follow the n-th Gaussian component of the GMM, $\mathbf{z}^n \sim \mathcal{N}(\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n)$, thus $K = N$. The attribute and content latent representations need to be learned, and they will be modeled by two architectures, namely a *content extractor* E_c and an *attribute extractor* E_z. See Figure 2 for a graphical representation of GMM-UNIT.

In addition to tackling the problem of multi-domain and multi-modal translation, we would like these two extractors, content and attribute, to be *disentangled*. This would constrain the learning and hopefully yield better domain translation, since the content would be as independent as possible from the attributes. Formally, the following two properties must hold:

Sampled attribute translation

$$G(E_c(\mathbf{x}^m), \mathbf{z}^n) \sim p_{\boldsymbol{\mathcal{X}}^n} \quad \forall\, \mathbf{z}^n \sim \mathcal{N}(\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n),\ \mathbf{x}^m \sim p_{\boldsymbol{\mathcal{X}}^m},\ n, m \in \{1, \dots, N\}. \tag{2}$$

Extracted attribute translation

$$G(E_c(\mathbf{x}^m), E_z(\mathbf{x}^n)) \sim p_{\boldsymbol{\mathcal{X}}^n} \quad \forall\, \mathbf{x}^n \sim p_{\boldsymbol{\mathcal{X}}^n},\ \mathbf{x}^m \sim p_{\boldsymbol{\mathcal{X}}^m},\ n, m \in \{1, \dots, N\}. \tag{3}$$

3.2 TRAINING THE GMM-UNIT

The encoders E_c and E_z, and the generator G need to be learned to satisfy three main properties. **Consistency**: When traveling through the network, the generated/extracted codes and images must be consistent with the original samples. **Fit**: The distribution of the attribute latent space must follow a GMM. **Realism**: The generated images must be indistinguishable of real images. In the following we discuss different losses used to force the overall pipeline to satisfy these properties.

In the textbfconsistency term, we include image, attribute and content reconstruction, as well as cycle consistency. More formally, we use the following losses:

Self-reconstruction of any input image from its extracted content and attribute vectors:

$$\mathcal{L}_{\text{s/rec}} = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}} \left[\| G(E_c(\mathbf{x}), E_z(\mathbf{x})) - \mathbf{x} \|_1 \right] \tag{4}$$

Content reconstruction from an image, translated into any domain:

$$\mathcal{L}_{\text{c/rec}} = \sum_{n,m=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}, \mathbf{z}\sim\mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)} \left[\| E_c(G(E_c(\mathbf{x}), \mathbf{z})) - E_c(\mathbf{x}) \|_1 \right] \tag{5}$$

Attribute reconstruction from an image translated with any content:

$$\mathcal{L}_{\text{a/rec}} = \sum_{n,m=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}, \mathbf{z}\sim\mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)} \left[\| E_z(G(E_c(\mathbf{x}), \mathbf{z})) - \mathbf{z} \|_1 \right] \tag{6}$$

In practice, this loss needs to be complemented with an isometry loss:

$$\mathcal{L}_{\text{iso}} = \sum_{n,m=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}, \mathbf{z}, \mathbf{z}'\sim\mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)} \left[\left| \| E_z(G(E_c(\mathbf{x}), \mathbf{z})) - E_z(G(E_c(\mathbf{x}), \mathbf{z}')) \|_1 - \| \mathbf{z} - \mathbf{z}' \|_1 \right| \right] \tag{7}$$

Cycle consistency when translating an image back to the original domain:

$$\mathcal{L}_{\text{cyc}} = \sum_{n=1}^{N} \sum_{m\neq n} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}, \mathbf{z}\sim\mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)} \left[\| G(E_c(G(E_c(\mathbf{x}), \mathbf{z})), E_z(\mathbf{x})) - \mathbf{x} \|_1 \right] \tag{8}$$

In the **fit** term we encourage both the attribute latent variable to follow the Gaussian mixture distribution and the generated images to follow the domain's distribution. We set two loss functions.

Kullback-Leibler divergence between the extracted latent code and the model. Since the KL divergence between two GMMs is not analytically tractable, we resort on the fact that we know from which domain are we sampling and define:

$$\mathcal{L}_{\text{KL}} = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}} \left[\mathcal{D}_{KL}(E_z(\mathbf{x}) \| \mathcal{N}(\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n)) \right] \tag{9}$$

where $\mathcal{D}_{KL}(p\|q) = -\int p(t) \log \frac{p(t)}{q(t)} dt$ is the Kullback-Leibler divergence.

Domain classification of generated and original images. For any given input image $\mathbf{x}$, we would like the method to classify it as its original domain, and to be able to generate from its content an image in any domain. Therefore, we need two different losses, one directly applied to the original images, and a second one applied to the generated images:

$$\mathcal{L}^D_{\text{dom}} = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}, d_{\boldsymbol{\mathcal{X}}^n}} \left[-\log D_{\text{dom}}(d_{\boldsymbol{\mathcal{X}}^n} | \mathbf{x}) \right], \tag{10}$$

$$\mathcal{L}^G_{\text{dom}} = \sum_{n,m=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}, \mathbf{z}\sim\mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m), d_{\boldsymbol{\mathcal{X}}^m}} \left[-\log D_{\text{dom}}(d_{\boldsymbol{\mathcal{X}}^m} | G(E_c(\mathbf{x}), \mathbf{z})) \right], \tag{11}$$

where $d_{\boldsymbol{\mathcal{X}}^n}$ is the label of domain n. Importantly, while the generator is trained using the second loss only, the discriminator D_{dom} is trained using both.

The **realism** term tries to making the generated images indistinguishable from real images; we adopt the adversarial loss to optimize both the real/fake discriminator $D_{\text{r/f}}$ and the generator G:

$$\mathcal{L}^D_{\text{GAN}} = \sum_{n,m=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^n}}[-\log D_{\text{r/f}}(\mathbf{x})] + \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^m}, \mathbf{z}\sim\mathcal{N}(\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n)}[-\log(1 - D_{\text{r/f}}(G(E_c(\mathbf{x}), \mathbf{z})))] \tag{12}$$

$$\mathcal{L}^G_{\text{GAN}} = \sum_{n,m=1}^{N} \mathbb{E}_{\mathbf{x}\sim p_{\boldsymbol{\mathcal{X}}^m}, \mathbf{z}\sim\mathcal{N}(\boldsymbol{\mu}^n, \boldsymbol{\Sigma}^n)}[-\log(D_{\text{r/f}}(G(E_c(\mathbf{x}), \mathbf{z})))] \tag{13}$$

The full objective function of our network is:

$$\mathcal{L}_D = \lambda_{\text{GAN}}\mathcal{L}^D_{\text{GAN}} + \lambda_{\text{dom}}\mathcal{L}^D_{\text{dom}} \tag{14}$$

$$\begin{aligned}\mathcal{L}_G =& \lambda_{\text{GAN}}\mathcal{L}^G_{\text{GAN}} + \lambda_{\text{s/rec}}\mathcal{L}_{\text{s/rec}} + \lambda_{\text{c/rec}}\mathcal{L}_{\text{c/rec}} + \lambda_{\text{a/rec}}\mathcal{L}_{\text{a/rec}} + \\ & \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{iso}}\mathcal{L}_{\text{iso}} + \lambda_{\text{dom}}\mathcal{L}^G_{\text{dom}}\end{aligned} \tag{15}$$

where $\{\lambda_{\text{GAN}}, \lambda_{\text{s/rec}}, \lambda_{\text{c/rec}}, \lambda_{\text{a/rec}}, \lambda_{\text{cyc}}, \lambda_{\text{KL}}, \lambda_{\text{iso}}, \lambda_{\text{dom}}\}$ are hyper-parameters of weights for corresponding loss terms. The value of most of these parameters come from the literature. We refer to Appendix A for the details.

4 EXPERIMENTS

We perform extensive quantitative and qualitative analysis in three real-world tasks, namely: edges-shoes, digits and faces. First, we test GMM-UNIT on a simple task such as a one-to-one domain translation. Then, we move to the problem of multi-domain translation where each domain is independent from each other. Finally, we test our model on multi-domain translation where each domain is built upon different combinations of lower level attributes. Specifically, for this task, we test GMM-UNIT in a dataset containing over 40 labels related to facial attributes such as hair color, gender, and age. Each domain is then composed by combinations of these attributes, which might be mutually exclusive (e.g. either male or female) or mutually inclusive (e.g. blonde and black hair).

Additionally, we show how the learned GMM latent space can be used to interpolate attributes and generate images in previously unseen domains, thus showing the first example of few- or zero-shot generation in image-to-image translation. Finally, GMM-UNIT will be applied to the Style transfer task.

We compare our model to the state of the art of both multi-modal and multi-domain image translation problems. In the former, we select BicycleGAN (Zhu et al., 2017b), MUNIT (Zhu et al., 2017a) and MSGAN (Mao et al., 2019). In the latter, we compare with StarGAN (Choi et al., 2018) and DRIT++ (Lee et al., 2019), which is the only multi-modal and multi-domain method in the literature. However, since StarGAN is not multi-modal we additionally test a simple modification of the model where we inject noise in the network. We call this version StarGAN*. More details are in Appendix A.

4.1 METRICS

We quantitatively evaluate the performance of our method through image quality and diversity of generated images. The former is evaluated through the Fréchet Inception Distance (FID) Heusel et al. (2017) and the Inception Score (Salimans et al., 2016). We evaluate the latter through the LPIPS Distance (Zhang et al., 2018), NDB and JSD (Richardson & Weiss, 2018) metrics. In addition, we also show the overall number of parameters used for all domains (Params).

FID We use FID to measure the distance between the generated and real distributions. Lower FID values indicate better quality of the generated images. We estimate the FID using 100 input images and 100 samples per input v.s. randomly selected 10000 images from the target domain.

IS To estimate the IS, we use Inception-v3 (Szegedy et al., 2016) fine-tuned on our specific datasets as classifier for 100 input images and 100 samples per input image. Higher IS means higher generated image quality.

LPIPS The LPIPS distance is defined as the $\mathcal{L}_2$ distance between the features extracted by a deep learning model of two images. This distance has been demonstrated to match well the human perceptual similarity (Zhang et al., 2018). Thus, following Zhu et al. (2017b); Huang et al. (2018); Lee et al. (2018), we randomly select 100 input images and translate them to different domains. For each domain translation, we generate 10 images for each input image and evaluate the average LPIPS distance between the 10 generated images. Finally, we get the average of all distances. Higher LPIPS distance indicates better diversity among the generated images.

NDB and JSD These are measuring the similarity between the distributions of real and generated images. We use the same testing data as for FID. Lower NDB and JSD mean the generated data distribution approaches better the real data distribution.

4.2 EDGES ↔ SHOES: TWO-DOMAINS+ TRANSLATION

We first evaluate our model on a simpler task than multi-domain translation: two-domain translation (e.g. edges to shoes). We use the dataset provided by Isola et al. (2017); Zhu et al. (2017a) containing images of shoes and their edge maps generated by HED (Xie & Tu, 2015). We train a single model for edges ↔ shoes without using paired information. Figure 3 displays examples of shoes generated from the same sketch by GMM-UNIT. Table 2 shows the quantitative evaluation and comparison with the state-of-the-art. Our model generates images with high diversity and quality using half the parameters of the state of the art. We refer to Appendix B.1 for additional results on this task.

Table 2: Benchmark on the Edges → Shoes dataset. MMod/MDom stand for multi-modal/domain.

Method	MMod	MDom	IS↑	FID↓	NDB↓	JSD↓	LPIPS↑	Params[§]↓
MUNIT	✓		2.874	**54.52**	34.67±3.68	.098±.006	.227±0.107	23.52M×2
MSGAN			3.125	111.19	35.67±2.62	.121±.001	.220±.118	65.03M×2
StarGAN*		✓	**3.479**	140.41	62.33±1.25	.192±.002	.002±.007	53.23M×1
DRIT++	✓	✓	3.038	123.87	41.33±4.50	.148±.006	**.236±.113**	54.06M×1
GMM-UNIT	✓	✓	3.245	80.78	**31.33±2.62**	**.098±.001**	.209±.110	**23.52M×1**

[§] Number of all parameters for the models trained in all the domains.

Figure 3: Examples of edges → shoes translation with the proposed GMM-UNIT.

4.3 DIGITS: SINGLE-ATTRIBUTE MULTI-DOMAIN TRANSLATION

We then evaluate our model in a multi-domain translation problem where each domain is composed by digits collected in different scenes. We use the Digits-Five dataset introduced in Xu et al. (2018), from which we select three different domains, namely MNIST (LeCun et al., 1998), MNIST-M (Ganin & Lempitsky, 2014), a colorized version of MNIST for domain adaptation, and Street View House Numbers (SVHN) (Netzer et al., 2011). We compare our model with the state-of-the-art on multi-domain translation, and we show in Figure 4 and Table 3 the qualitative and quantitative results respectively.

From these results we conclude that StarGAN* fails at generating diversity, thus confirming the findings of previous studies that adding noise does not increase diversity (Mathieu et al., 2015; Isola et al., 2017; Zhu et al., 2017b). GMM-UNIT instead generates images with higher quality and diversity than all the state-of-the-art models. We note, however, that StarGAN* achieves a higher IS, probably due to the fact that it solves a simpler task. Additional experiments carried out implementing a StarGAN*-like GMM-UNIT (i.e. setting $\boldsymbol{\sigma}_k = 0,\ \forall k$) indeed produced similar results. Specifically, the StarGAN*-like GMM-UNIT tends to generate for each input image one single (deterministic) output and thus the corresponding LPIPS scores are around zero. We refer to Appendix B.2 for additional results on this task.

Table 3: Benchmark on the Digits dataset. [1]FID fails in some domains because of low diversity.

Method	MMod	MDom	IS↑	FID↓	NDB↓	JSD↓	LPIPS↑	Params ↓
StarGAN*		✓	**3.35**	88.89	62.55±3.59	.14±.03	.005±.016	**11.18×1**
DRIT++	✓	✓	3.06	–[1]	77.2±3.49	.27±.01	.055±.046	24.49M×1
GMM-UNIT	✓	✓	3.23	**64.96**	**58.33±3.96**	**.12±.00**	**.107±.076**	16.37M×1

[1] FID fails in some of the domains because of the low diversity in the results.

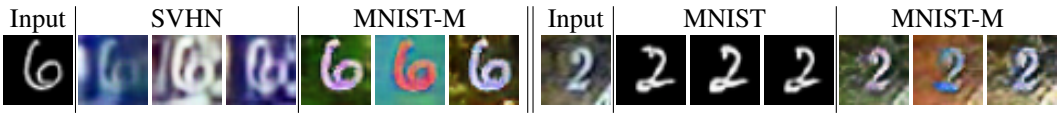

Figure 4: Samples of domain translation of GMM-UNIT trained on Digits.

4.4 Faces: Multi-attribute Multi-domain Translation

We also evaluate GMM-UNIT in the complex setting of multi-domain translation in a dataset of facial attributes. We use the CelebFaces Attributes (CelebA) dataset (Liu et al., 2015), which contains 202,599 face images of celebrities where each face is annotated with 40 binary attributes. We resize the initial 178×218 size images to 128×128. We randomly select 2,000 images for testing and use all remaining images for training. This dataset is composed of some attributes that are mutually exclusive (e.g. either male or female) and those that are mutually inclusive (e.g. people could have both blonde and black hair). Thus, we model each attribute as a different GMM component. For this reason, we can generate new images for all the combinations of attributes by sampling from the GMM. As aforementioned, this is not possible for state-of-the-art models such as StarGAN and DRIT++, as they use one-hot domain codes to represent the domains. For the purpose of this experiment we show five binary attributes: hair color (*black, blond, brown*), gender (*male/female*), and age (*young/old*). These five attributes allow GMM-UNIT to generate 32 domains.

Figure 5 shows some generated results of our model. We can see that GMM-UNIT learns to translate images to simple attributes such as blonde hair, but also to translate images with combinations of them (e.g. blonde hair and male). Moreover, we can see that the rows show different realizations of the model thus demonstrating the stochastic approach of GMM-UNIT. These results are corroborated by Table 4 that shows that our model is superior to StarGAN* in both quality and diversity of generated images. We also note in this experiment that the IS is higher in StarGAN*. Additional results are on Appendix B.3.

Table 4: Benchmark on the CelebA dataset. [1]FID fails in all domains because of low diversity.

Method	MMod	MDom	IS↑	FID↓	NDB↓	JSD↓	LPIPS↑	Params ↓
StarGAN*		✓	**3.345**	–[1]	66.33±2.58	.185±.012	.003±.004	53.24×1
GMM-UNIT	✓	✓	2.67	**72.04**	**56.11±1.65**	**.143±.007**	**.062±.041**	**23.52M×1**

[1] FID fails in all the domains because of the low diversity in the results.

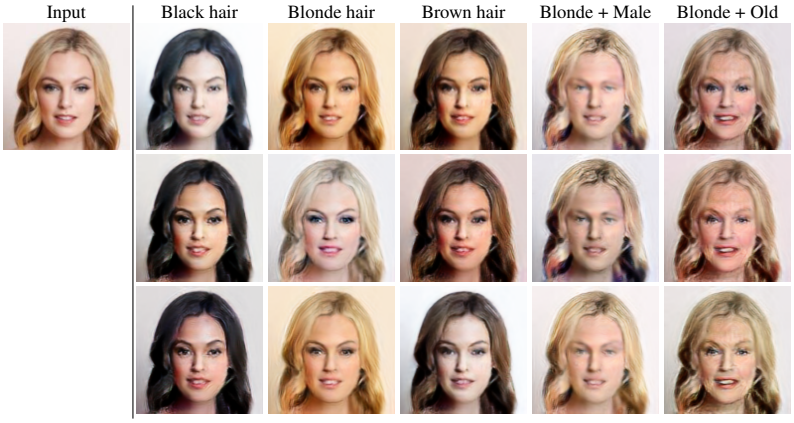

Figure 5: Facial expression synthesis results on the CelebA dataset with different attribute combinations. Each row represents a different output sampled from the model.

4.5 STYLE TRANSFER

We evaluate our model on style transfer, which is a specific task where the style is usually extracted from a single reference image. Thus, we randomly select two input images and synthesize new images where, instead of sampling from the GMM distribution, we extract the style (through E_z) from some reference images. Figure 6 shows that the generated images are sharp and realistic, showing that our method can also be effectively applied to Style transfer.

Figure 6: Examples of GMM-UNIT applied on the Style transfer task. The style is here extracted from a single reference images provided by the user.

4.6 DOMAIN INTERPOLATION AND EXTRAPOLATION

In addition, we evaluate the ability of GMM-UNIT to synthesize new images with attributes that are extremely scarce or non present in the training dataset. To do so, we select three combinations of attributes consisting of less than two images in the CelebA dataset: *Black hair+Blonde hair+Male+Young*, *Black hair+Blonde hair+Female+Young* and *Black hair+Blonde hair+Brown+Young*.

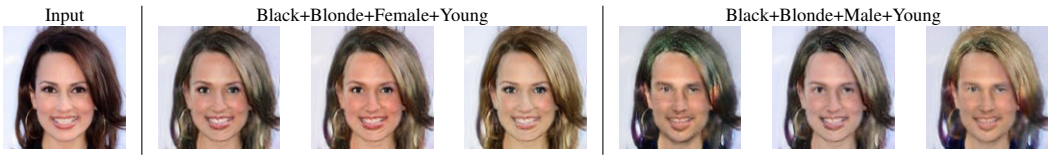

Figure 7: Generated images in previously unseen combinations of attributes.

Figure 7 shows that learning the continuous and multi-modal latent distribution of attributes allow to effectively generate images as zero- or few-shot generation. At the best of our knowledge, we are the first ones being able to translate images in previously unseen domains. This can be extremely important in tasks that are extremely imbalanced.

Finally, we show that by learning the full latent distribution of the attributes we can do attribute interpolation both intra- and inter-domains. In contrast, state of the art methods such as Lee et al. (2019) can only do intra-domain interpolations due to their discrete domain encoding. Figure 8 shows some generated images through a linear interpolation between two given attributes, while in Appendix B.3 we show that we can also do intra-domain interpolations.

4.7 ABLATION STUDY

We compare GMM-UNIT with three variants of the model that ablate $\mathcal{L}_{\text{cyc}}$, $\mathcal{L}_{\text{d/rec}}$ and $\mathcal{L}_{\text{iso}}$ in the Digits dataset. Table 5 shows the results of the ablation. As expected, $\mathcal{L}_{\text{cyc}}$ is needed to have higher image quality. When $\mathcal{L}_{\text{d/rec}}$ is removed image quality decreases, but $\mathcal{L}_{\text{iso}}$ still helps to learn the attributes space. Finally, without $\mathcal{L}_{\text{iso}}$ we observe that both diversity and quality decrease, thus confirming the need of all these losses. We refer to Appendix B.4 for the additional ablation results broken down by domain.

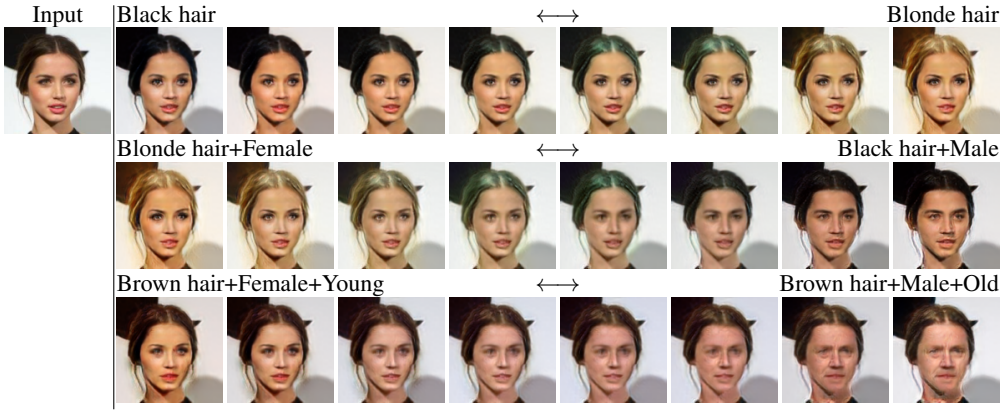

Figure 8: Examples of domain interpolation given an input image.

Table 5: Ablation study performance on the Digits dataset.

Model	**IS**↑	**FID**↓	**NDB**↓	**JSD**↓	**LPIPS**↑
GMM-UNIT w/o $\mathcal{L}_{\text{cyc}}$	3.03	85.78	63.44±2.48	0.174±0.004	0.138±.081
GMM-UNIT w/o $\mathcal{L}_{\text{d/rec}}$	3.26	67.95	61.33±2.35	0.144±0.005	0.121±.008
GMM-UNIT w/o $\mathcal{L}_{\text{iso}}$	3.28	68.56	59.44±4.25	0.139±0.004	0.117±.078
GMM-UNIT	3.23	64.96	58.33±3.96	0.119±0.003	0.124±.078

5 Conclusion

In this paper, we present a novel image-to-image translation model that maps images to multiple domains and provides a stochastic translation. GMM-UNIT disentangles the content of an image from its attributes and represents the attribute space with a GMM, which allows us to have a continuous encoding of domains. This has two main advantages: first, it avoids the linear growth of the dimension of the attribute space with the number of domains. Second, GMM-UNIT allows for interpolation across-domains and the translation of images into previously unseen domains.

We conduct extensive experiments in three different tasks, namely two-domain translation, multi-domain translation and multi-attribute multi-domain translation. We show that GMM-UNIT achieves quality and diversity superior to state of the art, most of the times with fewer parameters. Future work includes the possibility to thoroughly learn the mean vectors of the GMM from the data and extending the experiments to a higher number of GMM components per domain.

References

Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, pp. 3722–3731, 2017.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, pp. 2172–2180, 2016.

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*, pp. 8789–8797, 2018.

Chris Donahue, Akshay Balsubramani, Julian McAuley, and Zachary C. Lipton. Semantically decomposing the latent spaces of generative adversarial networks. In *ICLR*, 2018.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pp. 184–199. Springer, 2014.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2014.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pp. 2414–2423, 2016.

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 *NIPS*, 2017.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pp. 1501–1510, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pp. 172–189, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pp. 1125–1134, 2017.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, pp. 1857–1865. JMLR. org, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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*, pp. 35–51, 2018.

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. *arXiv preprint arXiv:1905.01270*, 2019.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, pp. 700–708, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pp. 3730–3738, 2015.

Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *ICLR*, 2019.

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.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

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

Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *CVPR*, pp. 4500–4509, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pp. 2536–2544, 2016.

Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, pp. 818–833, 2018.

Eitan Richardson and Yair Weiss. On gans and gmms. In *NIPS*, pp. 5847–5858, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pp. 2234–2242, 2016.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pp. 2818–2826, 2016.

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

Joshua B Tenenbaum and William T Freeman. Separating style and content. In *NIPS*, pp. 662–668, 1997.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, pp. 6924–6932, 2017.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pp. 8798–8807, 2018.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pp. 1395–1403, 2015.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *CVPR*, pp. 3964–3973, 2018.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE ICCV*, pp. 2223–2232, 2017a.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, pp. 465–476, 2017b.

A IMPLEMENTATION DETAILS

Our deep neural models are built upon the state-of-the-art methods MUNIT (Huang et al., 2018), BicycleGAN (Zhu et al., 2017b) and StarGAN (Choi et al., 2018), as shown in Table 6 with details. We apply Instance Normalization (IN) (Ulyanov et al., 2017) to the content encoder E_c and Adaptive Instance Normalization (AdaIN) (Huang & Belongie, 2017) and Layer Normalization (LN) (Ba et al., 2016) for the decoder G. For the discriminator network, we use Leaky ReLU (Xu et al., 2015) with a negative slope of 0.2. We use the following notations: $\mathcal{D}$: the number of domains, N: the number of output channels, K: kernel size, S: stride size, P: padding size, CONV: a convolutional layer, GAP: a global average pooling layer, UPCONV: a $2\times$bilinear upsampling layer followed by a convolutional layer. Note that we reduce the number of layers of the discriminator on the Digits dataset.

We use the Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and an initial learning rate of 0.0001. The learning rate is decreased by half every 100,000 iterations. In all experiments, we use a batch size of 1 for Edges2shoes and Faces and batch size of 32 for Digits. And we set the loss weights to $\lambda_{\text{GAN}} = 1$, $\lambda_{\text{s/rec}} = 10$, $\lambda_{\text{c/rec}} = 1$, $\lambda_{\text{a/rec}} = 1$, $\lambda_{\text{cyc}} = 10$, $\lambda_{\text{KL}} = 0.1$, $\lambda_{\text{iso}} = 0.1$ and $\lambda_{\text{dom}} = 1$. We use the domain-invariant perceptual loss with weight 0.1 in all experiments. Random mirroring is applied during training.

Table 6: Network architecture.

Part	Input → Output Shape	Layer Information
E_c	$(h, w, 3) \rightarrow (h, w, 64)$	CONV-(N64, K7x7, S1, P3), IN, ReLU
	$(h, w, 64) \rightarrow (\frac{h}{2}, \frac{w}{2}, 128)$	CONV-(N128, K4x4, S2, P1), IN, ReLU
	$(\frac{h}{2}, \frac{w}{2}, 128) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	CONV-(N256, K4x4, S2, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), IN, ReLU
E_z	$(h, w, 3) \rightarrow (h, w, 64)$	CONV-(N64, K7x7, S1, P3), ReLU
	$(h, w, 64) \rightarrow (\frac{h}{2}, \frac{w}{2}, 128)$	CONV-(N128, K4x4, S2, P1), ReLU
	$(\frac{h}{2}, \frac{w}{2}, 128) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	CONV-(N258, K4x4, S2, P1), ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (1, 1, 256)$	GAP
G	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), AdaIN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), AdaIN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), AdaIN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	Residual Block: CONV-(N256, K3x3, S1, P1), AdaIN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{2}, \frac{w}{2}, 128)$	UPCONV-(N128, K5x5, S1, P2), LN, ReLU
	$(\frac{h}{2}, \frac{w}{2}, 128) \rightarrow (h, w, 64)$	UPCONV-(N64, K5x5, S1, P2), LN, ReLU
	$(h, w, 64) \rightarrow (h, w, 3)$	CONV-(N3, K7x7, S1, P3), Tanh
D	$(h, w, 3) \rightarrow (\frac{h}{2}, \frac{w}{2}, 64)$	CONV-(N64, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{2}, \frac{w}{2}, 64) \rightarrow (\frac{h}{4}, \frac{w}{4}, 128)$	CONV-(N128, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{4}, \frac{w}{4}, 128) \rightarrow (\frac{h}{8}, \frac{w}{8}, 256)$	CONV-(N256, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{8}, \frac{w}{8}, 256) \rightarrow (\frac{h}{16}, \frac{w}{16}, 512)$	CONV-(N512, K4x4, S2, P1), Leaky ReLU
	$(\frac{h}{16}, \frac{w}{16}, 512) \rightarrow (\frac{h}{16}, \frac{w}{16}, 1)$	CONV-(N1, K1x1, S1, P0)
	$(\frac{h}{16}, \frac{w}{16}, 512) \rightarrow (1, 1, n)$	CONV-(Nn, K$\frac{h}{16}$x$\frac{w}{16}$, S1, P0)

A.1 GMM

In our experiments we use a simplified version of the GMM, which satisfies the following properties:

- The mean vectors are placed on the vertices of $(N-1)$-dimensional regular simplex, so that the mean vectors are equidistant.
- The covariance matrices are diagonal, with the same on all the components. In other words, each Gaussian component is *spherical*, formally: $\mathbf{\Sigma}_k = \boldsymbol{\sigma}_k \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

B Additional results

B.1 Edges ↔ shoes: Two-domain translation

In this section, we present the additional results for the one-to-one domain translation. As shown in Figure 9, we qualitatively compare GMM-UNIT with the state-of-the-art. We observe that while all the methods (multi-domain and not) achieve acceptable diversity, it seems that DRIT++ suffers from problems of realism. As expected, StarGAN* does not generate diverse results. Appendix B.1 shows instead the quantitative results of paired image translation. Even though BicycleGAN solves the much simpler task of supervised learning, GMM-UNIT surprisingly shows similar results.

B.2 Digits: single-attribute multi-domain translation

Figure 10 shows the qualitative comparison with the state of the art, while Table 8 show the breakdown, per domain, of the quantitative results. We observe, as expected, that StarGAN* fails at generating diverse results.

Figure 9: Visual comparisons of state of the art methods on Edge ↔ Shoes dataset. We note that BicycleGAN, MUNIT and MSGAN are one-to-one domain translation models, while StarGAN* is a multi-domain (deterministic) model. Finally DRIT++ and GMM-UNIT are multi-modal and multi-domain methods.

B.3 FACES: MULTI-ATTRIBUTE MULTI-DOMAIN TRANSLATION

In Table 9 we show the quantitative results on the CelebA datset, broken down per domain. In Figure 11 we show some generated images in comparison with StarGAN*. Figure 12 shows the possibility to do attribute interpolation inside a domain.

Table 7: Comparison between GMM-UNIT and the reference of Paired, multi-modal, generation (BicycleGAN) on the Edges → Shoes dataset.

Metric	BicycleGAN†	GMM-UNIT
IS↑	2.956	3.245
FID↓	47.43	80.78
NDB↓	27.33±1.70	31.33±2.62
JSD↓	.089±.005	.098±.001
LPIPS↑	.196±0.091	.209±.110
Params§↓	64.30M×2	23.52M×1

§ Number of all parameters for the models trained in all the domains.
† Trained on paired data.

Table 8: Quantitative comparison on the Digits dataset.

Target Domain	Metric	Method		
		StarGAN*	DRIT++	GMM-UNIT
MNIST	IS↑	2.628	1.926	2.120
	FID↓	103.76	×[1]	78.19
	NDB↓	62.67±3.68	82.67±0.47	57.33±4.78
	JSD↓	.132±.014	.418±.012	.101±.003
	LPIPS↑	.002±.006	.001±.004	.067±.058
SVHN	IS↑	3.576	3.463	3.293
	FID↓	77.86	89.81	46.22
	NDB↓	65.67±2.49	78.33±5.44	65.33±1.70
	JSD↓	.179±.006	.252±.005	.178±.002
	LPIPS↑	.008±.023	.048±.040	.115±.078
MNIST-M	IS↑	3.853	3.795	4.274
	FID↓	85.07	114.51	70.48
	NDB↓	59.33±3.30	70.67±2.62	52.33±4.64
	JSD↓	.132±.000	.132±.001	.078±.004
	LPIPS↑	.006±.014	.116±.070	.191±.095
	Params§↓	11.18M×1	24.49M×1	16.37M×1

[1] FID fails because of the low diversity in the results

B.4 Ablation study

In Table 10 we show additional, per domain, ablation results on the Digits dataset.

C Visualization of the Attribute Latent space

Figure 13 shows that the attributes sampled from the distribution and those extracted by the encoder E_z are mapped and well projected in the latent space of the attributes.

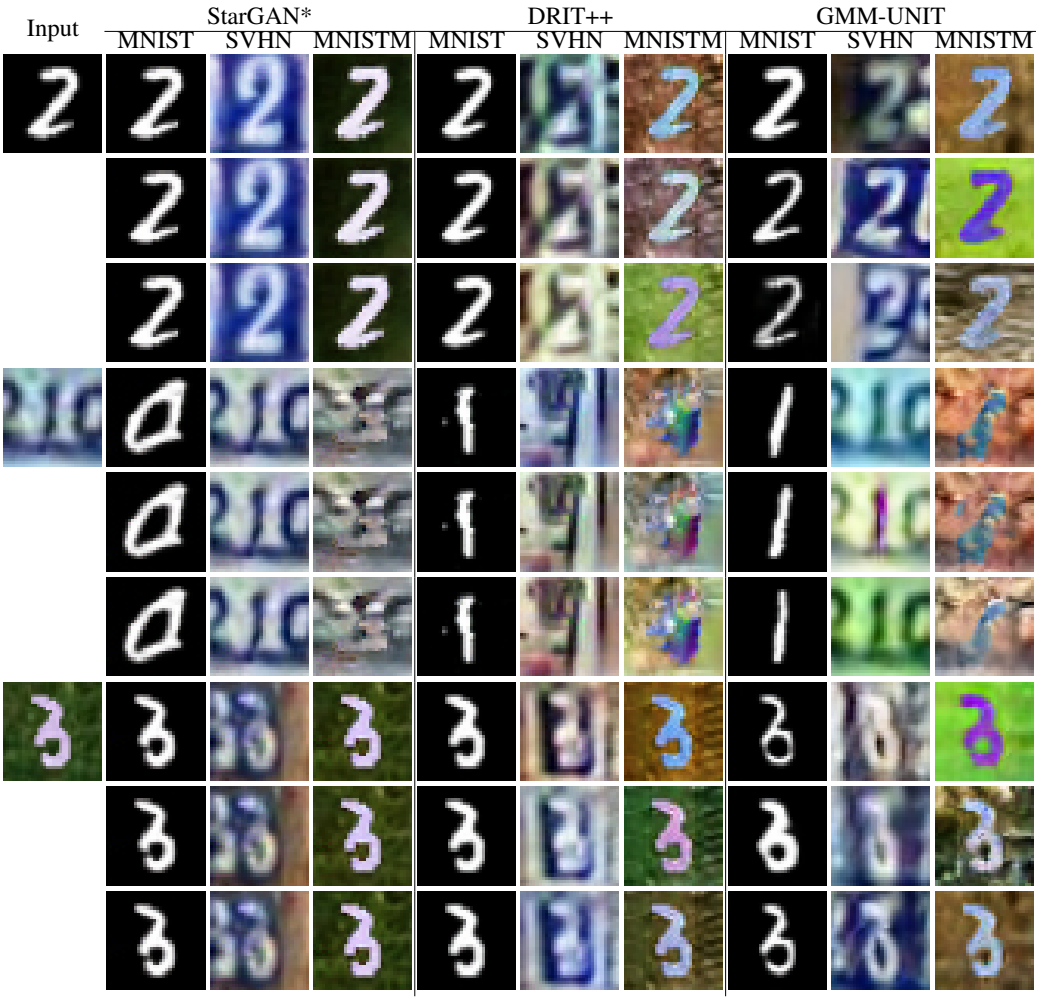

Figure 10: Visual comparisons of state of the art methods on the digits dataset. We note that StarGAN* is a multi-domain (deterministic) model, while DRIT++ and GMM-UNIT are multi-modal and multi-domain methods.

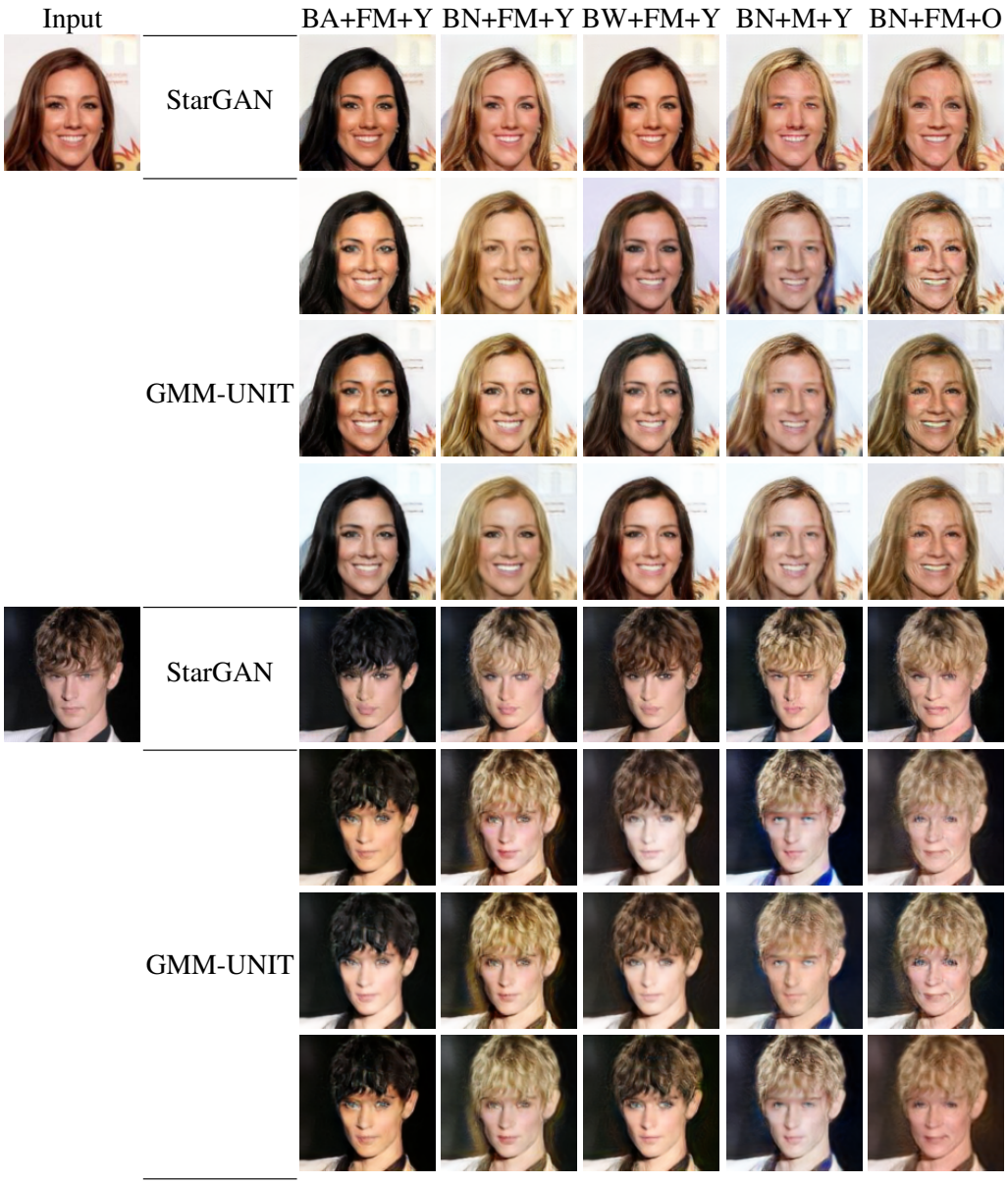

Figure 11: Comparisons on CelebA dataset. BA: Black hair, BN: blondehair, BW: Brown hair, M: Male, FM: Female, Y: Young, O: Old.

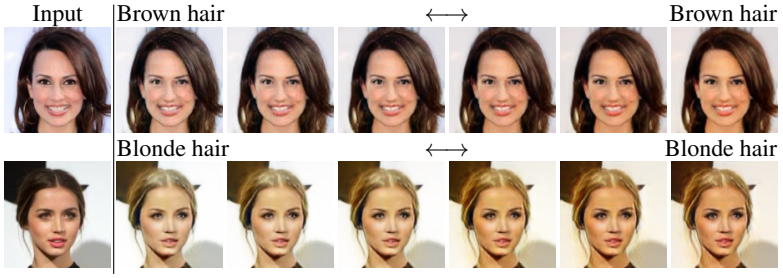

Figure 12: An example of attribute intra-domain interpolation.

Table 9: Quantitative comparison on the CelebA dataset.

Target Domain	Metric	Method	
		StarGAN*	GMM-UNIT
Black hair + Female + Young	IS↑	3.432	2.878
	FID↓	×[1]	68.32
	NDB↓	63.33±3.09	51.00±1.63
	JSD↓	.159±.011	.101±.011
	LPIPS↑	0.002±0.003	.060±.039
Blonde hair + Female + Young	IS↑	3.325	2.529
	FID↓	×[1]	81.08
	NDB↓	72.33±2.05	64.33±1.89
	JSD↓	.211±.014	.188±.003
	LPIPS↑	0.004±0.006	.072±.046
Brown hair + Female + Young	IS↑	3.277	2.608
	FID↓	×[1]	66.73
	NDB↓	63.33±2.49	53.00±1.41
	JSD↓	.184±.012	.142±.006
	LPIPS↑	0.003±0.005	.056±.036

[1] FID fails because of the low diversity in the results.

Table 10: Ablation study performance on the Digits dataset.

Target Domain	Model	IS↑	FID↓	NDB↓	JSD↓	LPIPS↑
MNIST	GMM-UNIT w/o $\mathcal{L}_{cyc}$	2.088	68.69	51.00±2.16	.073±.004	.067±.055
	GMM-UNIT w/o $\mathcal{L}_{d/rec}$	2.046	79.49	61.66±2.87	.119±.007	.063±.057
	GMM-UNIT w/o $\mathcal{L}_{iso}$	2.050	81.16	61.00±3.27	.123±.006	.072±.062
	GMM-UNIT	2.120	78.19	57.33±4.78	.101±.003	.067±.058
SVHN	GMM-UNIT w/o $\mathcal{L}_{cyc}$	2.705	69.94	72.33±3.34	.247±.005	.166±.083
	GMM-UNIT w/o $\mathcal{L}_{d/rec}$	3.290	52.05	71.67±2.36	.212±.004	.110±.078
	GMM-UNIT w/o $\mathcal{L}_{iso}$	3.357	49.40	68.00±2.16	.219±.001	.107±.077
	GMM-UNIT	3.293	46.22	65.33±1.70	.178±.002	.115±.078
MNISTM	GMM-UNIT w/o $\mathcal{L}_{cyc}$	4.303	118.73	67.00±1.63	.201±.002	.182±.101
	GMM-UNIT w/o $\mathcal{L}_{d/rec}$	4.457	72.32	50.67±1.67	.102±.002	.192±.101
	GMM-UNIT w/o $\mathcal{L}_{iso}$	4.430	75.12	49.33±6.24	.074±.004	.172±.091
	GMM-UNIT	4.274	70.48	52.33±4.64	.078±.004	.191±.095

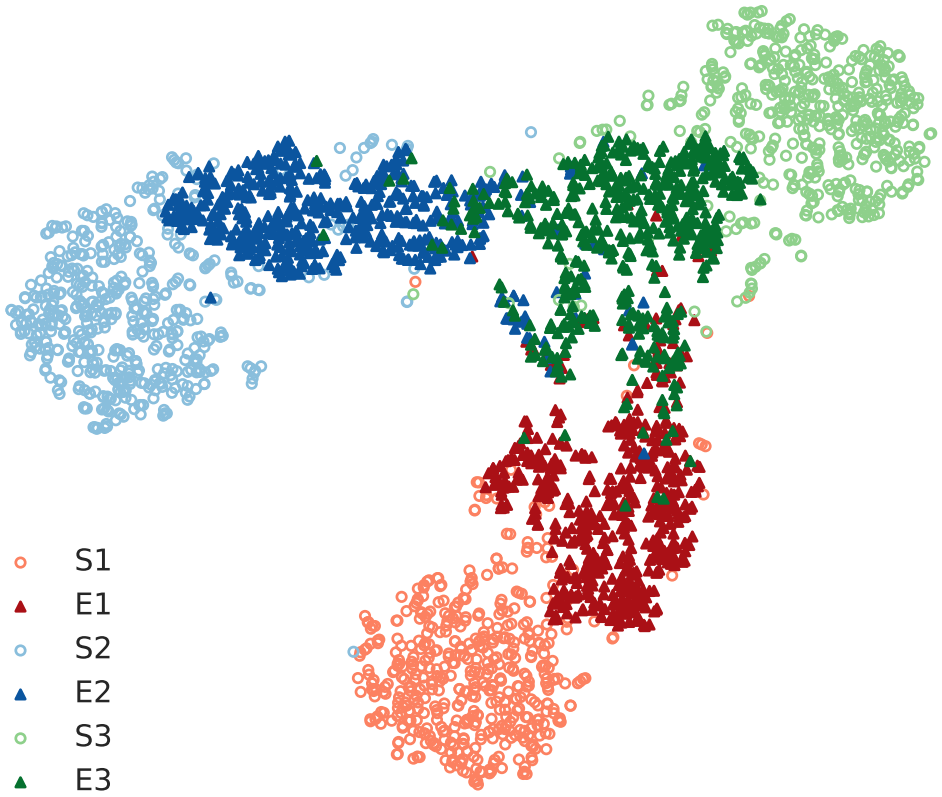

Figure 13: Visualization of the attribute vectors in a 2D space via t-SNE method. "S" refers to randomly sampling from GMM components (1: black hair, 2: blondehair, 3: brown hair) and "E" refers to extracting attribute vectors by the encoder E_z from the real data.