SMIT: Stochastic Multi-Label Image-to-Image Translation

Andrés Romero
BCV Lab
Universidad de Los Andes
rv.andres10@uniandes.edu.co

Pablo Arbeláez
BCV Lab
Universidad de Los Andes
pa.arbelaez@uniandes.edu.co

Luc Van Gool
ETH Zürich
KU Leuven
vangool@ethz.ch

Radu Timofte
CV Lab
ETH Zürich
timofter@ethz.ch

Abstract

*Cross-domain mapping has been a very active topic in recent years. Given one image, its main purpose is to translate it to the desired target domain, or multiple domains in the case of multiple labels. This problem is highly challenging due to three main reasons: (i) unpaired datasets, (ii) multiple attributes, and (iii) the multimodality (*e.g. *style) associated with the translation. Most of the existing state-of-the-art has focused only on two reasons* i.e.*, either on (i) and (ii), or (i) and (iii). In this work, we propose a joint framework (i, ii, iii) of diversity and multi-mapping image-to-image translations, using a single generator to conditionally produce countless and unique fake images that hold the underlying characteristics of the source image. Our system does not use style regularization, instead, it uses an embedding representation that we call domain embedding for both domain and style. Extensive experiments over different datasets demonstrate the effectiveness of our proposed approach in comparison with the state-of-the-art in both multi-label and multimodal problems. Additionally, our method is able to generalize under different scenarios: continuous style interpolation, continuous label interpolation, and fine-grained mapping.*

1. Introduction

The ability of humans to easily imagine how a black haired person would look like if they were blond, or with a different type of eyeglasses, or to imagine a winter scene as summer is formulated as the image-to-image (I2I) translation problem in the computer vision community. Since the recent introduction of Generative Adversarial Networks (GANs) [19], a plethora of problems such as video analysis [51, 7], super resolution [33, 9], semantic synthesis [26, 10], photo enhancement [24, 25], photo editing [49, 14], and most recently domain adaptation [21, 43] have been addressed as I2I translation problems.

Initially, translating from one domain into another required paired datasets that exactly matched both domains [26] *e.g.*, edges↔shoes or edges↔handbags datasets. However, this approach is unpractical because the full representation of the cross-domain mapping is, in most cases, intractable. Existing techniques try to perform deterministic I2I translation with unpaired images to map from one domain into another (one-to-one) [55, 4, 37, 25], or into multiple domains (one-to-many) [12, 46, 20]. Nevertheless, many problems are fundamentally stochastic as there are countless mappings from one domain to another *e.g.*, a day↔night or cat↔dog translation.

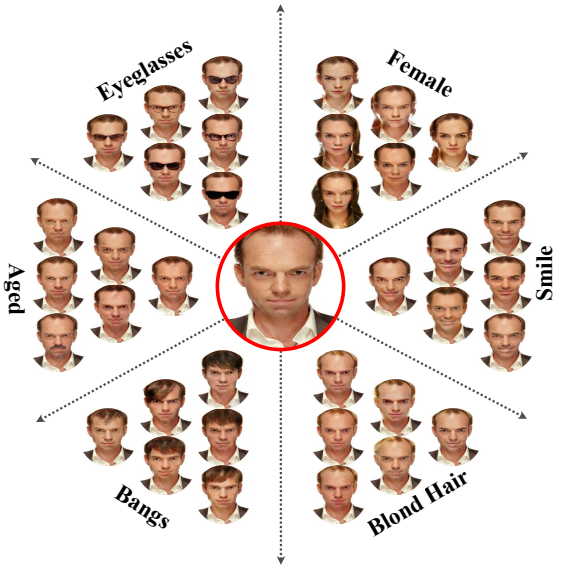

Figure 1. **Stochastic Multi-Label Image-to-Image Translation (SMIT)**. Our model learns a full diverse representation for multiple attributes using a single generator.

Recent techniques [34, 23, 39] have successfully addressed the multimodal representation for one-to-one domain translation. These methods are based on the idea developed on traditional I2I approaches [55, 56], in which the generator tends to overlook a noise injection. As a consequence, these techniques studied the problem of disentangling representation as style transfer, including a shared

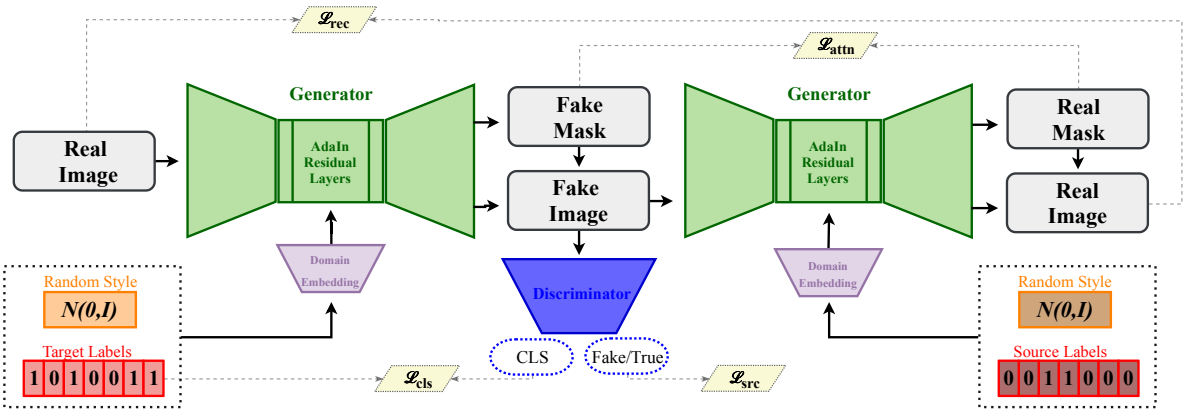

Figure 2. **Overview of SMIT.** We translate an image by jointly taking as input a random style and target attributes into the generator. The Domain Embedding is a map projection that uses random and fixed parameters for the embedding. The discriminator aims at classifying only the source and the attributes, *i.e*. no style regularization. We use the original source attributes and a different style to recover the real image.

content space representation and a style encoder network.

In this paper, we propose Stochastic Multi-Label Image-to-Image Translation (SMIT), a novel and robust framework that includes multiple labels and diversity, and does not require either style or content regularization. Moreover, we build our entire approach using a single generator that does not ignore the noise perturbation, *i.e*. for different level of noise our method produces different styles with the underlying characteristics and structure of the target domain[1]. As illustrated in Figure 1, SMIT learns a full distribution for each attribute, so it can perform diverse translation for different fine-grained or broader attributes. It is important to remark that in contrast to [12, 46, 30] the trainable parameters in the SMIT generator are not label-dependent, that is there is a negligible difference either on computational time or on memory consumption when learning as many as 40 attributes instead of just 2 labels. Figure 2 presents an overview of our model. We radically depart from mainstream approaches [12, 46, 30], where the target domain is inserted through the spatial concatenation, instead we indirectly inject the style and the target labels through Adaptive Instance Normalization (AdaIN) [22] layers in the generator, and the discriminator aims at recovering only the labels, *i.e*. we remark the importance of no style regularization.

We perform a comprehensive quantitative evaluation of SMIT either for disentanglement or multiple domain I2I problems, demonstrating the advantages of our method in comparison with existing state-of-the-art models. We also show qualitative results on several datasets that validate the effectiveness of our approach under varied and challenging settings.

More precisely, our main contribution is to propose a single and end-to-end system with an agnostic-domain generator capable of performing style transformation, multi-label mapping, style interpolation, and continuous label interpolation with no need of style regularization. For reproducibility, we plan to release our source code and trained models.

[1]Hereafter, we refer to domains as the number of labels per dataset, and style as the diversity induced by noise.

2. Related Work

Generative Adversarial Networks (GANs) [19] have proven to be a powerful approach to learn statistical data distributions. GANs rely on game theory where there are two networks (discriminator and generator) optimizing a Minimax function, a training scheme also known as adversarial training. The discriminator learns to distinguish real images from fake ones produced by the generator, and the generator learns to fool the discriminator by producing realistic fake images. Since their introduction, GANs have provided remarkable results in several computer vision problems, such as image generation [47, 11, 29], image translation [26, 55, 3, 37], video translation [51, 7] and resolution enhancement [6, 33, 2]. As our approach lies in the domain of image-to-image translation, it is the focus of our related work review.

Conditional GANs (cGANs) In vanilla GANs [19], the information regarding the domain is unknown. Conversely, on conditional GANs (cGANs) [44], the discriminator not only distinguishes between real and fake, but it also trains an auxiliary classifier for the conditional data distribution. cGANs have been applied in image-to-image translation problems for semantic layouts [26, 10], super resolution [33], photo editing [49], and for multi-target domains [12, 30, 46]. While traditional cGANs exploit the underlying conditional distribution of the data, they are constrained to produce deterministic outputs, *i.e*. given an input

	CycleGAN [55]	BiCycleGAN [56]	StarGAN [12]	MUNIT&alike [23, 3, 39]	DRIT [34]	GANimation [46]	**SMIT (ours)**
Unpaired Training	✓		✓	✓	✓	✓	✓
Multimodal Generation		✓		✓	✓		✓
Multiple Attributes			✓			✓	✓
One Single Generator			✓			✓	✓
Fine-grained Transformation			✓			✓	✓
Continuous Label Interpolation						✓	✓
Style Transformation				✓	✓		✓
Style Interpolation				✓	✓		✓
Attention Mechanism						✓	✓

Table 1. **Feature comparison with state-of-the-art approaches in I2I translation.** SMIT uses a single generator trained with unpaired data to produce disentangled representations of a multi-targeted domain.

and a target label, the output is always the same. In comparison, our approach introduces a style randomness in the generation process.

Image-to-Image Translation (I2I) Isola *et al*. [26] introduced a framework in which they trained cGANs using paired datasets. This work led to a new set of previously unexplored I2I problems. Based on these findings, Zhu *et al*. [55] extended the framework by introducing the cycle-consistency loss, which allowed to perform cross-domain mapping using unpaired datasets. Although CycleGAN [55] is currently one of the most common backbones for I2I models and frameworks, it is constrained to one-to-one domain translation, hence it needs one generator per domain. In contrast, our method uses a single generator regardless of the number of domains.

Other works [12, 46] extended the cycle-consistency insight in order to cope with multiple domains, by using a single generator. These methods take the label as independent features to the first layer of the generator, hence constraining the generator weights to restricted applications. Similarly, additional methods [30, 20] tackled the multilabel mapping problem from a VAE-GAN [32] perspective. Our approach neither uses a variational autoencoder representation nor does it depend on label weights, since the generator has always the same number of parameters regardless of the application.

Disentangled Representations A recurrent limitation in traditional I2I methods is their deterministic output. In image generation problems [47, 11, 28], disentangled representations are achieved by injecting random noise in the generator. Nevertheless, this idea cannot be used on the seminal CycleGAN, as this framework learns to ignore the noise vector due to the lack of regularization [55].

Recently, there have been efforts [10, 56, 8] to produce diverse representations from a single input. For instance, BiCycleGAN [56] bypassed the regularization issues of CycleGAN and it included a random noise vector in the training scheme, thus generating images of higher quality than CycleGAN. However, this approach requires paired data to train, which makes it unfeasible to scale in real-world scenarios.

Furthermore, generating multimodal images can also be studied as a problem of style transfer [17, 18] between two images. Inspired by the work of Gatys *et al*. [17], recent approaches [23, 39, 34] split the generator encoder into a two-stream content and style encoder, where the content stream extracts the underlying structure, shape and main information to be preserved on the image, and the style one draws the rendering attributes it aims at transferring. These disentangled representations are similar in spirit with the CycleGAN cycle-consistency adversarial loss since they perform a cross-domain mapping for the style and content space. Consequently, it is difficult to perform fine-grained translations. In comparison, our proposed approach does not suffer in this regard, since we neither constrain the content nor the style distributions. Moreover, as the experiments will show, SMIT is suitable for both coarser translations and subtle local appearances *e.g*., art in-painting or facial expressions, respectively.

Continuous Interpolation On the one hand, Pumarola *et al*. [46] introduced a cGAN framework that takes as input continuous rather than discrete labels. This approach enables the generation of examples with continuous labels at inference time, however, it does not handle diversity for the same input. On the other hand, for binary problems, Lee *et al*. [34] and Huang *et al*. [23] performed continuous interpolation between two styles in order to produce a pseudo-animated style transferring with images that belong to the same domain. Our work uses both target and style continuous interpolation.

Table 1 summarizes our main differences with respect to the literature for either multi-label or multimodal translation. SMIT has richer capabilities that those of existing methods as we perform fine-grained local transformation, style transformation, continuous style interpolation, continuous label interpolation, and multi-label transferring using one single generator.

3. Stochastic Multi-Label Image-to-Image Translation (SMIT)

Our final goal is to generate multi-attribute images with different styles using a single generator. As illustrated in Figure 2, our method is an ensemble of three different networks: a generator, a discriminator, and a domain embedding (DE). The generator takes the source image as input and translates it. The discriminator does not only differentiate between real and fake samples, but it also approximates the output distribution of the real target by means of an auxiliary classifier. Finally, SMIT uses the DE to merge both target style and target labels into the generator.

3.1. Problem Formulation

Let $\mathcal{X}_r \in \mathbb{R}^{H\times W\times 3}$ be the real image. $\mathcal{X}_r$ is encoded by a set of N discrete or continuous labels $y_r \in \mathbb{R}^N$. Additionally, for each possible $\mathcal{X}_r$, there is an unknown style distribution $s_r \in \mathbb{R}^S$. Given a target label y_f, and a target style s_f, we want to learn a mapping function $\mathbb{G}$ to produce a fake image $\mathcal{X}_f$, without having access to the joint distribution $p(\mathcal{X}_r, \mathcal{X}_f)$:

$$\mathbb{G}(\mathcal{X}_r, y_f, s_f) \rightarrow \mathcal{X}_f \in \mathbb{R}^{H\times W\times 3} \tag{1}$$

As it is common in cGANs [12, 46, 11, 47], we have a discriminator $\mathbb{D}$ that outputs the source domain probability, *i.e*. true or fake, and a classification/regression estimator, namely, $\mathbb{D}(\mathcal{X}_f) \rightarrow \{0, y_f\}$ and $\mathbb{D}(\mathcal{X}_r) \rightarrow \{1, y_r\}$.

3.2. Model

Generator ($\mathbb{G}$) We build upon the CycleGAN generator [55]. It is inspired in an encoder-decoder architecture, which consists of down-sampling layers, residual blocks, and up-sampling layers. Importantly, we use Instance Normalization (IN) [15, 52], Adaptive Instance Normalization (AdaIN) [22], and Layer Normalization (LN) [5] for the three stages, respectively. The main reason we only use IN during the first stage and not in the up-sampling is because they introduce undesirable properties to the global mean and variance that are modified by AdaIN in the residual Layers.

Domain Embedding (DE) We indirectly input the target attribute and the style randomness through AdaIN [22] weights. AdaIN normalization is computed from Equation 2, where x is the input and z are the adaptive parameters.

$$AdaIN(x, z) = z_w \frac{x - \mu(x)}{\sigma(x)} + z_b \tag{2}$$

$$z = DE(y, s) \tag{3}$$

As the AdaIN parameters depend entirely on the number of feature maps of the input x, they are agnostic to both style and label domains, which makes the generator entirely label and style independent. This key property makes SMIT highly suitable for transfer learning, addressing a drawback of cGANs in real-world scenarios.

It is important to mention that since the style and label dimensions may differ from the z dimensions, we use a projection embedding representation to encode style and label inputs to a fixed size suitable for AdaIN (Equation 3).

We remark that the DE does not require any training scheme, instead it is inspired by Language Modeling methods [40, 13, 36, 41, 45] that uses random initialization to map the input to a space embedding distribution. Particularly, we use a simple random embedding, *i.e*. a fully connected layer to map from style and labels concatenation to the AdaIN parameters. Our rationale is as follows: By always ensuring different z, we guarantee different normalization parameters, which means different fake images. We study the DE behaviour in more detail in Section 5.1.

Discriminator ($\mathbb{D}$) As previously stated, the discriminator has two outputs: source domain (*src*) and auxiliary classifier (*cls*). First, we use the idea of patch-GAN [26], to tell whether the source is fake or true based on a patch rather than a single number ($\mathbb{D}_{src}$). Second, we have a binary cross entropy loss function for the conditional labels ($\mathbb{D}_{cls}$). If continuous labels are used, then a regression objective loss should be applied. However, as we will discuss Section 5.2, our approach is capable of generating continuous labels even if it was trained with discrete ones.

3.2.1 Training Framework

In order to approximate function $\mathbb{G}$ in Equation 1, we split our general loss function for clarity.

Adversarial Loss We use the recently introduced averaged Relativistic Adversarial Loss (RGAN) [27] and the hinge version [42] loss to train the adversarial loss. RGAN relies on the idea that the discriminator not only estimates whether images are real or fake, but it also estimates the probability that the given real images are more realistic than the fake ones.

$$\begin{aligned} L_D &= \mathbb{D}_{src}(\mathcal{X}_r) - ||\mathbb{D}_{src}(\mathcal{X}_f)||_1 \\ L_G &= \mathbb{D}_{src}(\mathcal{X}_f) - ||\mathbb{D}_{src}(\mathcal{X}_r)||_1 \\ \mathcal{L}_{adv} &= L_D + L_G \end{aligned} \tag{4}$$

Conditional Loss The adversarial loss does not include any regularization for the conditional labels, yet the generator must be able to produce both realistic and conditioned images. To solve this issue, we define the conditional loss as:

$$\mathcal{L}_{cls} = \mathbb{D}_{cls}(\mathcal{X}) \log(y) + (1 - \mathbb{D}_{cls}(\mathcal{X})) \log(1 - y) \tag{5}$$

Recovery Loss In order to produce $\mathcal{X}_f$, we jointly input the target label and the target style. Therefore, the cycle consistency loss employed to recover the original image can be naively defined as:

$$\mathcal{X}_r \approx \mathcal{X}_{rec} = \mathbb{G}(\mathbb{G}(\mathcal{X}_r, y_f, s_f), y_r, s_r)$$

Note that the original style (s_r) is an unknown parameter. Nonetheless, we assume that s_r is drawn from a known normal distribution, and therefore reformulate the reconstruction loss by adding a different random style s'_f. We assume random styles during the whole training process. Thus, we compute the reconstruction or cycle consistency loss as:

$$\begin{aligned} \mathcal{X}_{rec} &= \mathbb{G}(\mathbb{G}(\mathcal{X}_r, y_f, s_f), y_r, s'_f) \\ \mathcal{L}_{rec} &= ||\mathcal{X}_r - \mathcal{X}_{rec}||_1 \end{aligned} \tag{6}$$

Attention Loss Until this point, there is no guarantee that the output of our generator will preserve background details *e.g.*, the underlying structure, or the identity of a person. To solve this particular issue, we regularize our model with the unsupervised attention mechanism proposed by Pumarola *et al*. [46]. We add a new and parallel layer to the generator output ($\mathcal{X}_f$) that works as the attention mask ($\mathcal{M}$).

The attention loss encourages fake images to change only certain regions with respect to the real input, and it is decomposed by the following terms:

$$\begin{aligned} [\mathcal{X}_f \in \mathbb{R}^{H\times W\times 3},\ \mathcal{M} \in \mathbb{R}^{H\times W}] &= \mathbb{G}(\mathcal{X}_r, y_f, s_f) \\ \mathcal{X}_f &= \mathcal{M}\cdot\mathcal{X}_r + (1-\mathcal{M})\cdot\mathcal{X}_f \\ \mathcal{L}_{attn} &= ||\mathcal{M}||_1 \end{aligned} \tag{7}$$

Identity Loss To further stabilize the training framework, we regularize our model with the identity loss that is defined as follows:

$$\mathcal{L}_{idt} = ||\mathcal{X}_r - (\mathbb{G}(\mathcal{X}_r, y_r, s''_f))||_1 \tag{8}$$

Overall Loss We define our full objective function in Equation 9, as the weighed sum of the previous losses:

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{attn}\mathcal{L}_{attn} + \lambda_{idt}\mathcal{L}_{idt} \tag{9}$$

Remarkably, our method does not require style regularization [23, 34] since we use a training framework that can easily bypass it.

4. Experimental Setup

We validate our method over several and very different datasets and tasks, such as instance facial synthesis [38], emotion recognition [31], Yosemite summer↔winter [26], and edges-to-object generation [26].

In the supplementary material, we extend our qualitative results to painters [4], Alps seasons [4], RafD [31], BP4D [54], EmotionNet [16], and full CelebA [38] with 40 attributes.

4.1. Evaluation Metrics

Diverse Translation The LPIPS metric [53] allows us to quantify the similarity between two different images. LPIPS computes the L2 distance between pairs of deep features (*e.g.*, AlexNet, VGG, etc) images.

Multi-label Translation Besides the LPIPS score, we also compute the Inception Score (IS) [48] that is a popular score for I2I problems. The IS employs an Inception Network [50] to classify fake images and thus rank them according to their scores with respect to the prior distribution. Additionally, we report the Conditional Inception Score (CIS) [23] that quantifies both high quality and diverse mapping.

4.2. Evaluation Framework

Given the unique nature of our approach, we unfold the quantitative evaluation into two different schemes: multimodal evaluation, and multi-label evaluation.

Multimodal Evaluation We directly use MUNIT [23] and DRIT [34] to compare our method in GAN-based disentangled representations. For fair comparison under this setting, we work within the same datasets Edges [26] and Yosemite [55]. To this end, we train MUNIT and DRIT and report the corresponding LPIPS over the whole test set.

We use the LPIPS score to measure the diversity of the generated images. As there is no standard evaluation framework for the diversity in GAN-based problems, we use a set of two metrics. First, as in MUNIT, we compute the diversity one-vs-all across the entire dataset (D), using the diversity in the real data as a reference. Then, we use one single fixed style to produce the cross-mapping in order to compute the diversity along the entire fake dataset. Second, as in DRIT, given a single image, we measure the partial diversity (PD) across different modalities (20 different styles) and report the average and standard deviation over each image, over the whole set.

Multi-label Evaluation Additionally, for purely multi-label I2I methods, we train an Inception network [50] on a RafD train set (90%) and report the IS and CIS over the remaining test set (10%). We retrain StarGAN and GANimation [46] under exactly the same settings in order to make a fair comparison.

4.3. Implementation Details

We use an ensemble of three different convolutional networks: Generator, Discriminator, and a Domain Embedding (DE).

Similar to previous methods [23, 34], we assume the style to be drawn from a prior Gaussian distribution with

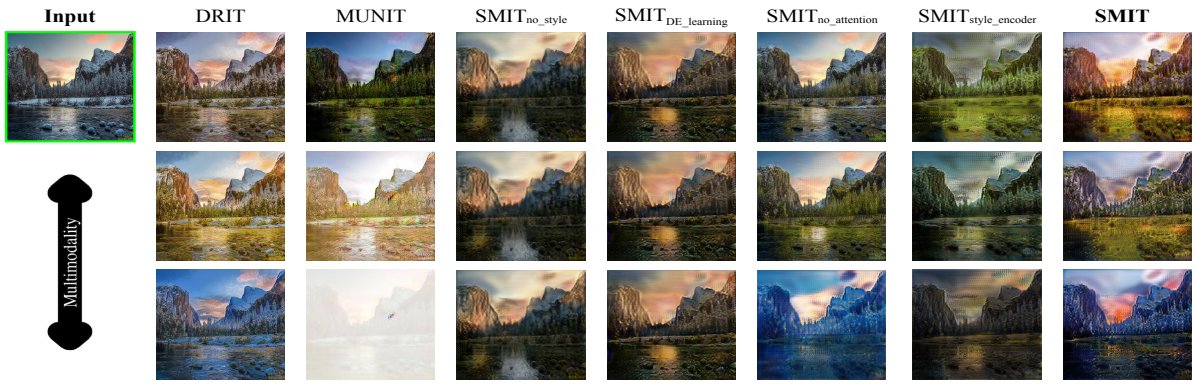

Figure 3. **Ablation experiments.** Qualitative comparisons over the Yosemite dataset [26]. Given the same input, we report the output for the related work [34, 23] and for each ablation experiment. Each row depicts different styles.

0 mean and identity variance, namely $\mathcal{N}(0, I)$. Therefore, the DE takes this 20-dimensional style vector and the N-dimensional target domain (one hot encoded) as inputs to produce the corresponding AdaIN number of parameters.

We provide a more detailed description of the architecture of our networks and training details in the supplementary material.

5. Results

We quantitatively and qualitatively demonstrate the effectiveness of SMIT in several settings. First, we perform ablation experiments, then we show qualitative results over different datasets, and finally we perform an extensive quantitative evaluation and compare our results against the state-of-the-art.

5.1. Ablation Study

We establish different baselines that define the main components of our framework: DE learning, removing the style randomness, adding style regularization, and removing the attention mechanism. We perform a qualitative and quantitative comparison for each of them, and we report our findings in Figure 3 and Table 2, respectively.

DE learning Studying DE parameters is one of our main interests as it is the only controller between the style and labels, and the mapped image. We observe that the generator can easily fall in mode collapse if the DE weights are learned, thus producing almost the same images for different styles. In order to overcome this problem, we analyze the DE contribution to the general system either with learned or fixed random parameters. As we can see in Figure 3, SMIT$_{DE_learning}$, learning the DE parameters leads to full mode collapse, since the style has a negligible impact on the AdaIN generator parameters. This behaviour is due to the fact that the gradients that come from the auxiliary classifier force the domain embedding to produce stable outputs, and therefore the same output thanks to the lack of specialized and per domain style regularization. Conversely, by establishing fixed weights on the DE, we guarantee diversity, *i.e.*, from Equation 2 we observe that for different *scale* and *bias*, we ensure different behaviour on the normalization, hence different outputs.

	Yosemite [26]	
	D	PD
SMIT$_{no_style}$	0.412±0.046	-
SMIT$_{DE_learning}$	0.413±0.044	0.004±0.003
SMIT$_{no_atention}$	0.406±0.041	0.105±0.071
SMIT$_{style_encoder}$	0.418±0.043	0.133±0.063
SMIT	**0.419±0.048**	**0.145±0.072**

Table 2. **Ablation quantitative evaluation.** We report the diversity (D) and the partial diversity (PD) for every ablation study in our method.

Among all the datasets [38, 26, 16, 4, 55, 31] in which we validate our system, we observed that for datasets that contain a small number of samples with only two different domains (*e.g.*, Yosemite [55], ∼1k images per domain), there is a decline in the quality of the fake images when the DE has fixed random parameters. More precisely, even though the auxiliary classifier is highly confident after a few iterations, and the generator learns to fool the discriminator, the generator produces pixelation in the images. Nevertheless, given the simplicity of the dataset, pixelated images fulfill the conditions to fool the discriminator, *i.e.* fake images are realistic enough and they fall into the statistical representation of the labels. We further study this behaviour by combining two different settings: DE training (no pixelation and deterministic) and DE fixed (pixelation and stochastic). We split the AdaIN parameters into different small networks with different behaviours (learned or fixed weights), which share the input (target domain and style). We found that learning either small or big parts of the AdaIN layers induces mode collapse to the whole system. Nonetheless, with enough training iterations the pixelation issue is nuanced and not too evident. Surprisingly, we

Figure 4. **Qualitative results for facial analysis mapping [38].** Example results for an image in the wild. For each attribute (column), we show the corresponding translation for four different modalities (rows).

observed that the partial diversity metric (PD) is higher for highly pixelated images than smooth yet diverse ones. This finding indicates that the partial diversity is not related to both quality and diversity, but only to diversity at any cost (*e.g.*, change in color, pixelation, etc).

This form of coupling style and domain information is in line with [35, 17] as to use global statistics is better suited for the purpose of style transferring, rather than spatially connected features (*e.g.*, concatenating the image and the labels) as other methods usually employ [34, 12, 46].

No Style By removing the style, our network behaves on a fully deterministic way since fixed labels always impose the same statistics over the generator (Figure 3 and Table 2, $\text{SMIT}_{\text{no\_style}}$).

Style Encoder MUNIT [23] and DRIT [34] share a common practice by using a style encoder, where they regularize the style or noise previously injected in the generator. We also evaluate the necessity of such a mechanism. To this end, we deploy a separate network for style encoding ($\mathbb{S}$), whose purpose is to extract the style that is injected to fake images, *i.e.* computing $s'_f \approx \mathbb{S}(\mathcal{X}_f)$. As we depict in Figure 3 ($\text{SMIT}_{\text{style\_encoder}}$) and Table 2, there are no qualitative or quantitative differences by using this regularizer. However, the style encoder is a different network as big as the discriminator, so it increases the training time and memory consumption. Moreover, we argue that having a fixed random embedding as DE is enough to produce diversity because we force the generator to always produce different images regardless of the lack of regularization in the style. Therefore, the style encoder is not performing a critical role within our system. It is worth noting that the style encoder in conjunction with the DE-training has no effect on the diversity.

Due to the nature of multi-label problems, the style regularization is unhelpful in its simple form because of the high label entanglement. Thus, for any style encoding, it would require different styles for different labels using as many domain embeddings as domains, and perform cycle-consistency in a way that styles are tied to labels, which is difficult in practice.

Attention Mask We observe that the attention mechanism plays a critical role for the entire training scheme for those fine-grained datasets *e.g.*, CelebA, EmotionNet, BP4D. Without this loss, our framework takes the easiest way in the translation process, *i.e.* uniformly changing the color of the input (Figure 3, MUNIT). We argue that with enough iterations, this undesirable property leads to higher partial diversity due to diversity in color.

Furthermore, our Domain Embedding differs from the Multi-Layer Perceptron (MLP) proposed by MUNIT [23] as they use domain-specific yet trainable networks in order to transform from the style vector representation to the AdaIN number of parameters, which prevents the mode collapse problem. Note that we only use a single Domain Embedding regardless the multi-domain nature.

5.2. Qualitative Results

We now proceed to highlight the SMIT capabilities over the CelebA dataset. In Figure 4, we demonstrate the effectiveness of our method for 10 different attributes, switching one attribute at a time (columns) for different styles (rows). From these transformations, we observe that our model is indeed learning a fully continuous representation for the attributes, as it generalizes across different modalities either for subtle or broader transformations such as eyeglasses or smiling, or gender or hair colors, respectively. Similarly,

	Edges2Shoes [26]		Edges2Handbags [26]		Yosemite [26]		# Parameters (Generator)
	D	PD	D	PD	D	PD	
CycleGAN [55]	0.272±0.048	-	0.293±0.081	-	0.272±0.048	-	2x11.4M
DRIT [34]	0.237±0.149	0.028±0.030	0.296±0.181	0.056±0.060	0.398±0.038	0.126±0.019	2x21.3M
MUNIT [23]	0.295±0.051	0.077±0.057	0.365±0.052	0.123±0.067	0.335±0.045	0.208±0.034	2x15.0M
SMIT (ours)	0.303±0.058	0.072±0.056	0.367±0.048	0.096±0.072	**0.437±0.041**	0.145±0.072	**8.4M**
Real Data	0.313±0.052	-	0.374±0.051	-	0.447±0.049	-	-

Table 3. **Multimodal quantitative evaluation.** We report the LPIPS score to compare the diversity (D) and partial diversity (PD) with respect to the multimodal approaches. Better results are boldfaced according to their significant values.

	RafD [31]			
	CIS	IS	D	PD
StarGAN [12]	1.00±0.00	1.66±0.38	0.15±0.01	-
GANimation[46]	1.00±0.00	1.51±0.33	0.16±0.01	-
SMIT (ours)	**1.25±0.06**	**2.51±0.70**	**0.17±0.01**	**0.004±0.001**
Real Data	-	1.18±0.18	0.16±0.01	-

Table 4. **Multi-label quantitative evaluation.** We report the results for Inception Score (IS), Conditioned Inception Score (CIS), and LPIPS diversity metric (D and PD), for multi-label frameworks.

in the supplementary material we depict different emotion translations and compare against state-of-the-art methods.

Moreover, in the supplementary material, we show that, given fixed style and fixed labels, our model is able to generate always the same attributes for different people, *i.e*. the same eyeglasses, bangs, etc. We also report the attention mask visualizations. Additionally, we show translations for painters [4], Alps [4], RafD [31], edges2objects [26], BP4D [54], EmotionNet [16], and full CelebA [38] datasets. We also depict qualitative differences with StarGAN, GANimation, and FaceApp [1] over the CelebA dataset.

Interpolations Following common practice within Multimodal Image-to-Image translation methods [34, 23], where we assume that each style is randomly sampled from a normal probability distribution, our method also benefits from style interpolation going from one style to another by performing a spherical interpolation.

Even though the labels are binary attributes at train time, the DE transforms them into a higher dimensional representation given by the number of channels in the AdaIN layers. Inserting the labels in the form of continuous labels into the generator is of importance as we can easily perform continuous inference before the DE. The absence or presence of any label is correlated with different representations in the AdaIN parameters.

In the supplementary material, we show visualizations for style and label interpolation. Note that we do not explicitly train with continuous labels.

5.3. Quantitative Results

Next, we quantitatively compare SMIT with respect to the literature. We separate our experiments into two strategies due to the lack of both multi-label and multimodal translation methods.

Multimodal Evaluation As we depict in Table 3, we compare directly with DRIT and MUNIT over *edges2shoes*, *edges2handbags* and *Yosemite* datasets.

Our method produces higher LPIPS for the entire test set (D), and competitive results across partial diversity (PD) with respect to the state-of-the-art since there is no significant differences with MUNIT. We hypothesize that MUNIT's [23] good performance in the PD score is because this method is focused on color transformation and rendering rather than texture or content (Figure 3, MUNIT column). MUNIT constrains the content latent space, producing thus highly diverse mappings across a batch, and low general diversity if the style is fixed. As we retrain DRIT and MUNIT, it is worth to mention that DRIT's poor performance on edges2shoes and edges2handbags is due to the lack of diversity for object→edge mapping.

Remarkably, due to the reduced number of parameters (Table 3, number of parameters), SMIT takes less computational resources than baseline approaches to training, that is SMIT fits four times the batch size used in DRIT [34] and MUNIT [23], using one Titan X GPU.

We provide more quantitative results for each domain independently in the supplementary material.

Multi-label Evaluation Table 4 shows our results for StarGAN, GANimation, and SMIT. For each image, we perform 7 different translations (ignoring the ground truth translation). As we expected, StarGAN and GANimation obtain a constant CIS (1.0) and high IS scores, which indicates their lack of diversity but good qualitative translations. SMIT significantly overcomes related methods in diversity and image quality. Note that SMIT also outperforms the IS and D for the real images, demonstrating thus the effectiveness in both quality and diversity beyond the original dataset. In the supplementary material, we discriminate CIS and IS over each label independently.

Even though StarGAN and GANimation use a single generator and share a similar number of parameters, it is important to remark that they reshape the label vector into the input image size. This issue arises in high-resolution image to image translation as neither the number of parameters nor the computational time are negligible. By contrast, SMIT is suitable either for low or high resolution as it is label-agnostic dependent.

6. Conclusions

In this paper, we presented a novel, robust yet simple method for automatically performing stochastic image-to-image translation for multiple domains using a single generator. We demonstrated the capability of our approach with respect to the state-of-the-art in both disentangled and multi-label scenarios by achieving jointly high quality and diversity representations for both coarse or fine-grained translations. Moreover, SMIT is directly suitable for either multimodal interpolation or continuous interpolation in style and label intensity domains, respectively.

References

[1] Faceapp. http://www.faceapp.com. 2018.

[2] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool. Extreme learned image compression with gans. In *CVPR Workshops*, 2018.

[3] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*, 2018.

[4] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool. Combogan: Unrestrained scalability for image domain translation. In *CVPR Workshops*, 2018.

[5] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[6] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem. Finding tiny faces in the wild with generative adversarial network. In *CVPR*, 2018.

[7] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh. Recycle-gan: Unsupervised video retargeting. In *ECCV*, 2018.

[8] A. Bansal, Y. Sheikh, and D. Ramanan. Pixelnn: Example-based image synthesis. *arXiv preprint arXiv:1708.05349*, 2017.

[9] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. 2018 pirm challenge on perceptual image super-resolution. *arXiv preprint arXiv: 1809.07517*, 2018.

[10] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017.

[11] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[12] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] B. Dolhansky and C. C. Ferrer. Eye in-painting with exemplar generative adversarial networks. In *CVPR*, 2018.

[15] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. In *ICLR*, 2017.

[16] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, 2016.

[17] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[18] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *CVPR*, 2017.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[20] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Arbitrary facial attribute editing: Only change what you want. *arXiv preprint arXiv:1711.10678*, 2017.

[21] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

[22] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[23] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[24] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017.

[25] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. Wespe: Weakly supervised photo enhancer for digital cameras. *arXiv preprint arXiv:1709.01118*, 2017.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

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

[28] T. Kaneko, K. Hiramatsu, and K. Kashino. Generative attribute controller with conditional filtered generative adversarial networks. In *CVPR*, 2017.

[29] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICML*, 2018.

[30] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *NIPS*, 2017.

[31] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[32] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[33] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[34] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.

[35] Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[36] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017.

[37] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[38] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[39] L. Ma, X. Jia, S. Georgoulis, T. Tuytelaars, and L. Van Gool. Exemplar guided unsupervised image-to-image translation. *arXiv preprint arXiv:1805.11145*, 2018.

[40] E. Margffoy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. *ECCV*, 2018.

[41] B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. In *NIPS*, 2017.

[42] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

[43] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. Image to image translation for domain adaptation. In *CVPR*, 2018.

[44] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.

[45] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[46] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, 2018.

[47] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[48] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016.

[49] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[51] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.

[52] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, 2017.

[53] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[54] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[56] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017.