SDIT: Scalable and Diverse Cross-domain Image Translation

Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, Luis Herranz
{yaxing,agonzalez,joost,lherranz}@cvc.uab.es
Computer Vision Center Universitat Autònoma de Barcelona, Spain

ABSTRACT

Recently, image-to-image translation research has witnessed remarkable progress. Although current approaches successfully generate diverse outputs or perform scalable image transfer, these properties have not been combined into a single method. To address this limitation, we propose SDIT: Scalable and Diverse image-to-image translation. These properties are combined into a single generator. The diversity is determined by a latent variable which is randomly sampled from a normal distribution. The scalability is obtained by conditioning the network on the domain attributes. Additionally, we also exploit an attention mechanism that permits the generator to focus on the domain-specific attribute. We empirically demonstrate the performance of the proposed method on face mapping and other datasets beyond faces.

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; *Machine learning algorithms.*

KEYWORDS

Generative adversarial networks; Image generation; Image translation

ACM Reference Format:
Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, Luis Herranz . 2019. SDIT: Scalable and Diverse Cross-domain Image Translation. In *MM '19: ACM International Conference on Multimedia, October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Image-to-image translation aims to build a model to map images from one domain to another. Many computer vision tasks can be interpreted as image-to-image translation, e.g. style transfer [10], image dehazing [52], colorization [56], surface normal estimation [8], and semantic segmentation [26]. Face translation has always been of great interest in the context of image translation, and several methods [5, 35, 36] have shown outstanding performance. Image-to-image translation can be formulated in a supervised manner when corresponding image pairs from both domains are provided, and unsupervised otherwise. In this paper, we focus on unsupervised image-to-image translation with the two-fold goal of learning a model that has both scalability and diversity (see Figure 1(a)).

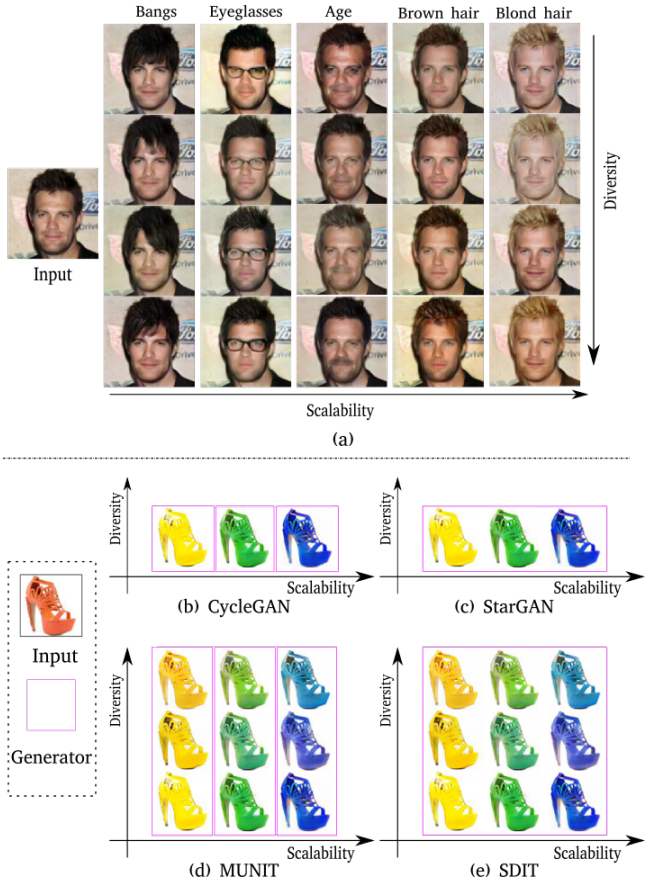

Figure 1: (a) Example of diverse image translations for various attributes of our method generated by a single model. (b-e) Comparison to current unpaired image-to-image translation methods. Given four color subsets (*orange, yellow, green, blue*), the task is to translate images between the domains. (b) CycleGAN requires three independent generators (indicated by pink lines) which produce deterministic results. (c) StarGAN only requires a single generator but produces deterministic results. (d) MUNIT requires separate generators but is able to produce diverse results. (e) SDIT produces diverse results from a single generator.

Recently, Isola *et al.* [15] consider a conditional generative adversarial network to perform image mapping from input to output with paired training samples. One of the drawbacks, however, is that this method produces a deterministic output for a given input image. BicycleGAN [59] extended image-to-image translation to one-to-many mappings between images by training the model to reconstruct the noise used in the latent space, effectively forcing it to use it in the translations. To address the same concern,

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Gonzalez-Garcia *et al.* [12] explicitly exploit the feature representation, disentangling the latent feature into shared and exclusive representations, the latter being aligned with the input noise.

The above methods, however, need paired images during the training process. For many image-to-image translation cases, obtaining abundant annotated data remains very expensive or, in some cases, even impossible. To relax the requirement of paired training images, recent approaches have made efforts to address this issue. The cyclic consistency constraint [19, 49, 58] was initially proposed for unpaired image-to-image translation. Liu *et al.* [24] assumes a shared joint latent distribution between the encoder and the decoder, then learns the unsupervised translation.

Nonetheless, previous methods perform a deterministic one-to-one translation and lack diversity on its outputs, as shown in Figure 1(b). For example, given the task from orange (domain A) to yellow (domain B) the generator taking the orange shoes as input only synthesizes shows with a single shade of yellow. Recently, the idea of non-deterministic outputs was extended to unpaired methods [14, 22] by disentangling the latent feature space into content and style and aligning the style code with a known distribution (typically Gaussian or uniform). During inference, the model is able to generate diverse outputs by sampling different style codes from the distribution. The main drawback of these methods is that they lack scalability. As shown in Figure 1(d) the orange shoes can be translated into many possible green shoes with varying green shades. As the number of colors increases, however, the number of required domain-specific encoder-decoder pairs rises quadratically.

IcGAN [35] initially performs face editing by combining cGAN [30] with an attribute-independent encoder, and at the inference stage conducts face mapping for given face attributes. Recently, Yunjey *et al.* [5] proposed StarGAN, a domain-independent encoder-decoder architecture for face translation that concatenates the domain label to the input image. Unlike the aforementioned non-scalable approaches [14, 22], StarGAN is able to perform scalable image-to-image translation between multi-domains (Figure 1(b)). StarGAN, however, fails to synthesize diverse translation outputs.

In this paper, we propose a compact and general architecture that allows for diversity and scalability in a single model, as shown in Figure 1(e). Our motivation is that scalability and diversity are orthogonal properties that can be independently controlled. Scalability is obtained by using the domain label to train a single multi-domain image translator, preventing the need to train a encoder-decoder for each domain. Inspired by [7], we employ Conditional Instance Normalization (CIN) layers in the generator to introduce the latent code and generate diverse outputs. We explore the reasons behind CIN's success (Fig. 6) and discover the following limitation: CIN affects the entirety of the latent features and could possibly modify areas that do not correspond to the specific target domain. To prevent this from happening, we include an attention mechanism that helps the model focus on domain-specific areas of the input image.

Our contributions are as follows:

- We propose a compact and effective framework that combines both scalability and diversity in a single model. Note that current models only possess one of these desirable properties, whereas our model achieves both simultaneously.
- We empirically demonstrate the effectiveness of the attention technique for multi-domain image-to-image translation.
- We conduct extensive qualitative and quantitative experiments. The results show that our method is able to synthesize diverse outputs while being scalable to multiple domains.

2 RELATED WORK

Generative adversarial networks. Typical GANs [13] are composed of two modules: a generator and a discriminator. The aim of the generator is to synthesize images to fool the discriminator, while the discriminator distinguishes between fake images and real images. There have been many variants of GANs [13] and they show remarkable performance on a wide variety of image-to-image translation tasks [14, 15, 22, 36, 49, 58], super-resolution [21], image compression [38], and conditional image generation such as text to image[27, 53, 54], segmentation to image[18, 43] and domain adaptation [9, 11, 39, 42, 48, 55, 60].

Conditional GANs. Exploiting conditional image generation is an active topic in GAN research. Early methods considered incorporating into the model category information [5, 30–32] or text description [17, 37, 54] for image synthesis. More recently, a wide variety of ideas have been proposed and used in several tasks such as image super-resolution [21], video prediction [28], and photo editing [41]. Similarly, we consider image-to-image translation conditioned on an input image and the label of the target domain.

Image-to image-translation. The goal of image-to-image translation is to learn a mapping between images of the source domain and images of the target domain. Given pairs of data samples, pix2pix [15] initially performed this mapping by using conditional GANs and relying on the real images. This model, however, fails to conduct one-to-many mappings, namely, it cannot generate diverse outputs from a single input. BicycleGAN [59] explicitly modeled the mapping between output and latent space, and aligned the latent distribution with a known distribution. Finally, the diverse outputs are performed by sampling from the latent distribution. Gonzalez-Garcia *et al.* [12] disentangle the latent space into disjoint elements, which allows them to successfully perform cross-domain retrieval as well as one-to-many translation. Although these methods allow to synthesize diverse results, the requirement of paired data limits their application. Recently, the cycle consistency loss [19, 49, 58] is enforced into models to explicitly reconstruct the source sample, which is translated into the target domain and back, thus enabling translation using unpaired data. In addition, UNIT [24] aligns the latent space in two domains by assuming the similar domains share the same content. Although this approach shows remarkable results without paired data, they fail to perform diverse outputs. More recently, several image-to-image translation methods [1, 6, 36, 45] enable diverse results with the usage of noise or labels.

Diversity of image-to-image translation. Most recently, several approaches [3, 12, 14, 16, 22, 23, 51] consider to disentangle factors in feature space by enforcing a latent structure or regulating the structure distribution. Exploiting this disentangled representation enables the generator to synthesize diverse outputs by controlling style distribution. The key difference with the proposed method is that our method additionally performs *scalable* image-to-image translation while still having diversity.

Scalability of image-to-image translation. The scalability aim is to conduct image-to-image translation across multiple domains by a single generator. MMNet [45] uses a shared encoder and a domain-independent decoder, not only allowing to perform style learning but zero-pair image-to-image translation. Anoosheh *et al.* [2] additionally consider encoder-decoder pairs for each domain as well as the used techniques in CycleGAN [58]. IcGAN [35] and StarGAN [5] condition the domain label on the latent space and input, respectively. Our approach also works by imposing domain labels in a single generator, but simultaneously enabling the model to synthesize diverse outputs.

Attention learning. Attention mechanisms have been successfully employed for image-to-image translation. Current approaches [4, 29] learn an attention mask to enforce the translation to focus only on the objects of interest and preserve the background area. GANimation [36] uses *action units* to choose regions from the input images that are relevant for facial animation. These methods exploit attention mechanisms at the image level. Our method, on the other hand, learns feature-wise attention maps, which enables us to control which features are modified during translation. Therefore, our attention maps are highly effective at restricting the translation to change only domain-specific areas (e.g. forehead region when modifying the 'bangs' attribute).

3 SCALABLE AND DIVERSE IMAGE TRANSLATION

Our method must be able to perform multi-domain image-to-image translation. We aim to learn a model with both scalability and diversity. By scalability we refer to the property that a single model can be used to perform translations between multiple domains. By diversity we refer to the property that given a single input image, we can obtain multiple plausible output translations by sampling from a random variable.

3.1 Method Overview

Here we consider two domains: source domain $\mathcal{X} \subset \mathbb{R}^{H\times W\times 3}$ and target domain $\mathcal{Y} \subset \mathbb{R}^{H\times W\times 3}$ (it can trivially be extended to multiple domains). As illustrated in Figure 2, our framework is composed of four neural networks: encoder E, generator G, multilayer perceptron M, and discriminator D. Let $x \in \mathcal{X}$ be the input source image and $y \in \mathcal{Y}$ the target output, with corresponding labels $l_{sc} \in \{1,\dots,C\}$ for the source and $l_{tg} \in \{1,\dots,C\}$ for the target. In addition, let $z \in \mathbb{R}^Z$ be the latent code, which is sampled from a Gaussian distribution.

An overview of our method is provided in Figure 2. To address the problem of scalability we introduce the target domain as a conditioning label to the encoder, $E(x, l_{tg})$. The diversity is introduced by the latent variable z, which is mapped to the input parameters of a Conditional Instance Normalization (CIN) layer [7] by means of the multilayer perceptron $M(z)$. The CIN learns an additive (β) and a multiplicative term (γ) for each feature layer. Both the output of the encoder E and the multilayer perceptron M are used as input to the generator $G(E(x, l_{tg}), M(z))$. The generator G outputs a sample y of the target domain. Sampling different z results into different output results y. The unpaired domain translation is enforced by a cycle consistency [19, 49, 58]: taking as input the output y and the source category l_{sc}, we reconstruct the input image x as $G(E(G(E(x, l_{tg}), M(z)), l_{sc}), M(z))$. The encoder E, the multilayer perceptron M, and the generator G are all shared.

The function of the discriminator D is threefold. It produces three outputs: $x \rightarrow \{D_{src}(x), D_{cls}(x), F_{rec}(x)\}$. Both $D_{src}(x)$ and $D_{cls}(x)$ represent probability distributions, while $F_{rec}(x)$ is a regressed code. The goal of $D_{src}(x)$ is to distinguish between real samples and generated images in the target domain. The auxiliary classifier $D_{cls}(x)$ predicts the target label and allows the generator to perform domain-specific output conditioned on it. This was found to improve the quality of the conditional GAN [32]. Similarly to previous methods [3, 14] we reconstruct the latent input code in the output $F_{rec}(x)$. This was found to lead to improved diversity. Note that F_{rec} is just used for generated samples, as F_{rec} aims to reconstruct the latent code, which is not defined for real images.

We shortly summarize here the differences of our method with respect to the most similar approaches. StarGAN [5] can also generate outputs on multiple domains, but: (1) it learns a scalable but deterministic model, while our method additionally obtains diversity via the latent code; (2) we explicitly exploit an attention mechanism to focus the generator on the object of interest. Comparing against both MUNIT [14] and DRIT [22], which perform diverse image-to-image translation but without being scalable, our method: (1) employs the domain label to control the target domain, allowing to conduct image-to-image translation among multiple domains with a single generator; (2) avoids the need for domain-specific style encoders, effectively saving computational resources; (3) considers attention to avoid undesirable changes in the translation; and (4) experimentally proves that the bias of CIN is the key factor to make the generator achieve the diversity, whereas the multiplicative term was only found to play a minor role.

3.2 Training Losses

The full loss function consists of several losses: the *adversarial loss* that discriminates the distribution of synthesized data and the real distribution in target domain, *domain classification loss* which contributes to the model $\{E, G\}$ to learn the specific attribute for a given target label, the *latent code reconstruction loss* regularizes the latent code to improve diversity and avoids the problem of partial mode collapse, and the *image reconstruction loss* that guarantees that the translated image keeps the structure of the input images.

Adversarial loss. We employ GANs [13] to distinguish the generated images from the real images

$$\begin{aligned}\mathcal{L}_{GAN} &= \mathbb{E}_{x\sim\mathcal{X}}\left[\log D_{src}(x)\right] \\ &+ \mathbb{E}_{x\sim\mathcal{X}, z\sim p(z)}\left[\log(1 - D_{src}(G(E(x, l_{tg}), M(z))))\right],\end{aligned} \tag{1}$$

where the discriminator tries to differentiate between generated images from the generator and real images, while G tries to fool the discriminator taking the output of M and the output of E as input. The final loss function is optimized by the minimax game

$$\left\{E^*, G^*, D^*\right\} = \arg\min_{E,G}\max_{D} \mathcal{L}_{GAN}. \tag{2}$$

Domain classification loss. In this paper, we consider Auxiliary Classifier GANs (AC-GAN) [32] to control domains. The discriminator aims to output a probability distribution over given input images

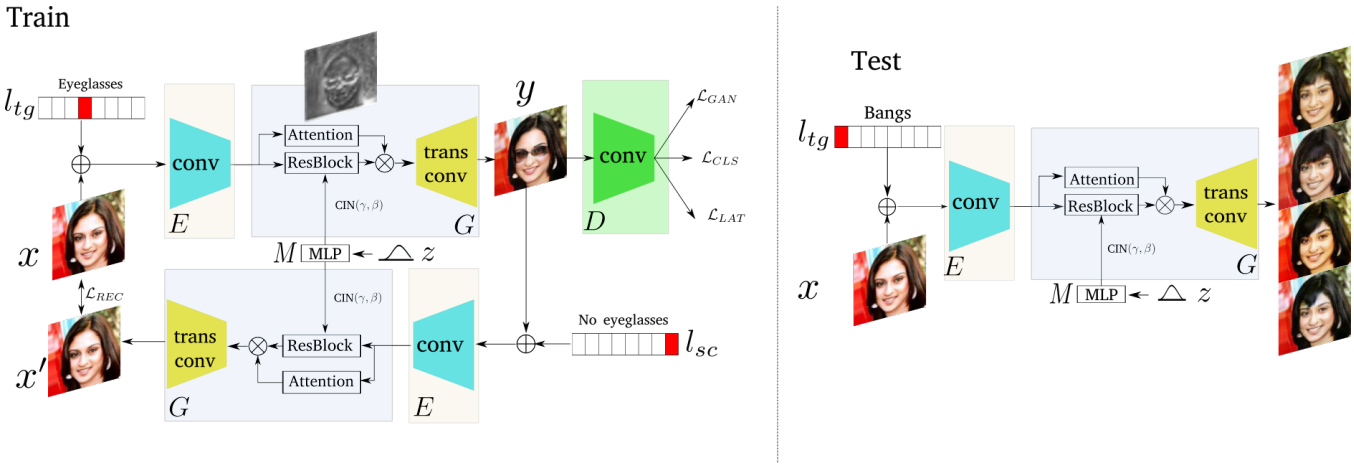

Figure 2: Model architecture. (*Left*) The proposed approach is composed of two main parts: a discriminator D to distinguish the generated images and the real images; and the set of the encoder E, multilayer perceptron M and the generator G, containing the attention block, residual blocks with CIN, and the transposed convolutional layers. (*Right*) At test time, we can generate multiple plausible translations in the desired domain using a single model.

y and domain label, in consequence E and G synthesize the domain-specific images. We share the discriminator model except for the last layer and optimize the triplet $\{E, G, D\}$ by the cross-entropy loss. The final domain classification loss for generated samples, real samples, and total are

$$\mathcal{L}_{FAKE}(E,G) = -\mathbb{E}_{x\sim\mathcal{X}, z\sim p(z)}\left[\log\left(D_{cls}\left(l_{tg}|G(E(x,l_{tg}),M(z))\right)\right)\right], \quad (3)$$

$$\mathcal{L}_{REAL}(D) = -\mathbb{E}_{x\sim\mathcal{X}}\left[\log\left(D_{cls}\left(l_{sc}|x\right)\right)\right], \quad (4)$$

$$\mathcal{L}_{CLS} = \mathcal{L}_{REAL} + \mathcal{L}_{FAKE}, \quad (5)$$

respectively. Given domain labels l_{sc} and l_{tg} these objectives are able to minimize the classification loss so that the model explicitly generates domain-specific outputs.

Latent code reconstruction loss. The lack of constraints on the latent code results in the generated images suffering from partial mode collapse as the latent code is ignored. We use the discriminator to predict the latent code, which forces the network to use it for generation:

$$\mathcal{L}_{LAT}(E,G,D) = \mathbb{E}_{x\sim\mathcal{X}, z\sim p(z)}\left[\|F_{rec}(x) - z\|_1\right] \quad (6)$$

Image reconstruction loss. Both adversarial loss and classification loss fail to keep the structure of the input. To avoid this, we formulate the image reconstruction loss as

$$\begin{aligned} y &= G\left(E\left(x, l_{tg}\right), M(z)\right), \\ x' &= G\left(E\left(y, l_{sc}\right), M(z)\right), \\ \mathcal{L}_{REC} &= \mathbb{E}_{x\sim\mathcal{X}, x'\sim\mathcal{X}'}\left[\left\|x - x'\right\|_1\right]. \end{aligned} \quad (7)$$

Full Objective. The full objective function of our model is:

$$\begin{aligned} &\min_{E,G}\max_{D} \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_{FAKE}\mathcal{L}_{FAKE} \\ &+ \lambda_{REAL}\mathcal{L}_{REAL} + \lambda_{LAT}\mathcal{L}_{LAT} + \lambda_{REC}\mathcal{L}_{REC} \end{aligned} \quad (8)$$

where λ_{GAN}, λ_{FAKE}, λ_{REAL}, λ_{LAT}, λ_{REC} are hyper-parameters that balance the importance of each iterm.

3.3 Attention-guided generator

The attention mechanism encourages the generator to locate the domain-specific area relevant to the target domain label. Let $e = E\left(x, l_{tg}\right)$ be the output of the encoder. We propose to localize the CIN operation by introducing an attention mechanism. Only part of the encoder output e should be changed to obtain the desired diversity. We separate the signal e into two parallel residual blocks T^c and T^a. The CIN is applied to the residual block according to $f = T^c\left(e, M\left(z\right)\right)$. We estimate the attention with a separate residual block according to $a = T^a\left(e\right)$. We then combine the original encoder output and the CIN output using attention:

$$h = (1 - a)\cdot e + a\cdot f. \quad (9)$$

In [36], an *attention loss* regularizes the attention maps, since they quickly saturate to 1. In contrast, we employ the attention in the bottleneck features, and experimentally prove that the attention masks can be easily learned. This makes the task easier due to lower resolution in the bottleneck, and avoids the need to tune the attention hyperparameter. Finally, our attention mechanism does not add any new terms to the overall optimization loss in (8).

4 EXPERIMENTAL SETUP

Training setting. Our model is composed of four sub-networks: encoder E, multilayer perceptron M, generator G, and discriminator D. The encoder contains 3 convolutional layers and 6 blocks. Each convolutional layer uses 4×4 filters with stride 2, except for the first one which uses 7×7 with stride 1, and each block contains two convolutional layers with 3×3 filters and stride of 1. M consists of two fully connected layers with 256 and 4096 units. The generator G comprises ResBlock layers, attention layers and two fractionally strided convolutional layers. The ResBlock consists of 6 residual blocks, as in the encoder E, but including CIN layers. The CIN layers take the output of E and the ouput of the M as input. Except for six blocks like the CIN layers, the attention layers also use additional convolutional layers with sigmoid activations on top. For the discriminator D, we use six convolutional layers with 4×4 and stride 2, followed by three parallel sub-networks, each of them

containing one convolutional layer with 3×3 filters and stride 1, except for the branch to output F_{rec} which uses an additional fully connected layer from 32 units to 8. Note how M adds around 1M parameters to the architecture.

All models are implemented in PyTorch [34] and released[1]. We randomly initialize the weights following a Gaussian distribution, and optimize the model using Adam [20] with batch size 16 and 4 for face and non-face datasets, respectively. The learning rate is 0.0001, followed the exponential decay rates $(\beta_1, \beta_2) = (0.5, 0.999)$. In all experiments, we use the following hyper-parameters: $\lambda_{GAN} = 10$, $\lambda_{FAKE} = 1$, $\lambda_{REAL} = 1$, $\lambda_{LAT} = 10$ and $\lambda_{REC} = 800$. We use Gaussian noise to the latent code with zero mean and a standard deviation of 1.

4.1 Datasets

We consider several datasets to evaluate our models. In order to verify the generality of our method, the datasets were chosen to cover a variety of cases, including faces (CelebA), object (Color), and scenes (Artworks).

CelebA [25]. The Celeb Faces Attributes is a face dataset of celebrities with 202,599 images and 40 attribute labels per face. To explicitly preserve the face ratio, we crop the face size of 178×218 and resize it to 128×128. We leave out 2000 random images for test and train with the rest.

Color dataset [50]. We use the dataset collected by Yu *et.al* [50], which consists of 11 color labels, each category containing 1000 images. In order to easily compare to the non-scalable baselines which need train one independent model for each domain pair, we use only four colors (*green*, *yellow*, *blue*, *orange*). We resize all images to 128×128. We collected 3200 images for the train set and 800 images for the test set.

Artworks [58]. We also illustrate SDIT in an artwork setting [58]. This includes real images (*photo*) and three artistic styles (*Monet*, *Ukiyo-e*, and *Cezanne*). The training set contains 3000 (photo), 700 (Ukiyo-e), 500 (Cezanne) and 1000 (Monet) images, while the test set are: 300 (photo), 100 (Ukiyo-e), 100 (Cezanne) and 200 (Monet) images. All image are resized to 256×256.

4.2 Evaluation Metrics

To validate our approach, we consider the three following metrics.

LPIPS. In this paper, LPIPS [57] is used to compute the similarity of pairs of images from the same attribute. LPIPS takes larger values if the generator has more diversity. In our setting, we generate 10 samples given an input image via different random codes.

ID distance. The key point of face mapping is to preserve the *identity* of the input, since an identity change is unacceptable for this task. To measure whether two images depict the same identity, we consider *ID distance* [44], which represents the difference in identity between pairs of input and translated faces. More concretely, given a pair of input and output faces, we extract the identity features represented by the VGGFace [33] network, and compute the distance between these features. VGGFace is trained on a large face dataset and is robust to appearance changes (e.g. illumination, age, expression, etc.). Therefore, two images of the same person should have a very small value. We only use this evaluation metric for CelebA. We use all 2000 test images as input and generate 10 output images, which in total amounts to 20,000 pairs.

[1]The codes are available at https://github.com/yaxingwang/SDIT

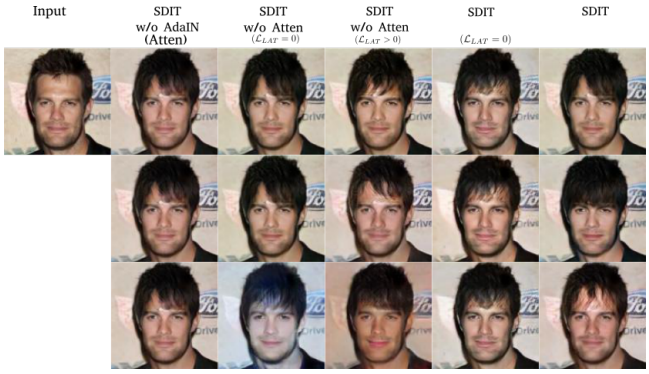

Figure 3: Ablation study of different variants of our method. We show results for the face task of adding 'bangs'. We display three random outputs for each variant of the method.

Method	Atten	CIN	$\mathcal{L}_{LAT}$	ID Distance	LPIPS
SDIT w/o CIN (Atten)	Y	N	N	0.061	0.408
SDIT w/o Atten ($\mathcal{L}_{LAT}=0$)	N	Y	N	0.063	0.409
SDIT w/o Atten ($\mathcal{L}_{LAT}>0$)	N	Y	Y	0.070	**0.432**
SDIT ($\mathcal{L}_{LAT}=0$)	Y	Y	N	0.063	0.412
SDIT	Y	Y	Y	**0.060**	0.424

Table 1: ID distance (lower, better) / LPIPS (higher, better) for different variants of our method. Atten: attention, Y: yes, N: no.

Reverse classification. One of the methods to evaluate conditional image-to-image translation is to train a reference classifier on real images and test it on generated images [46, 47]. The reference classifier, however, fails to evaluate diversity, since it may still report a high accuracy even when the generator encounters mode-collapse for a specific domain, as shown on the third column of Figure 3. Following [40, 47], we use the *reverse classifier* which is trained using translated images for each target domain and evaluated on real images for which we know the label. Lower classification errors indicate more realistic and diverse translated images.

5 EXPERIMENTAL RESULTS

In Section 5.1 we introduce several baselines against which we compare our model, as well as multiple variants of our model. Next, we evaluate the model on faces in Section 5.2. Finally, in Section 5.3 and Section 5.4, we analyze the generality of the model to color translation and scene translation.

5.1 Baselines and variants

We compare our method with the following baselines. For all baselines, we use the authors' original implementations and recommended hyperparameters. We also consider different configurations of our proposed SDIT approach. In particular, we study variants with and without CIN, attention, and latent code reconstruction.

CycleGAN [58]. CycleGAN is composed of two pairs of domain-specific encoders and decoders. The full objective is optimized with an adversarial loss and a cycle consistency loss.

MUNIT [14]. MUNIT disentangles the latent distribution into the content space which is shared between two domains, and the style space which is domain-specific and aligned with a Gaussian

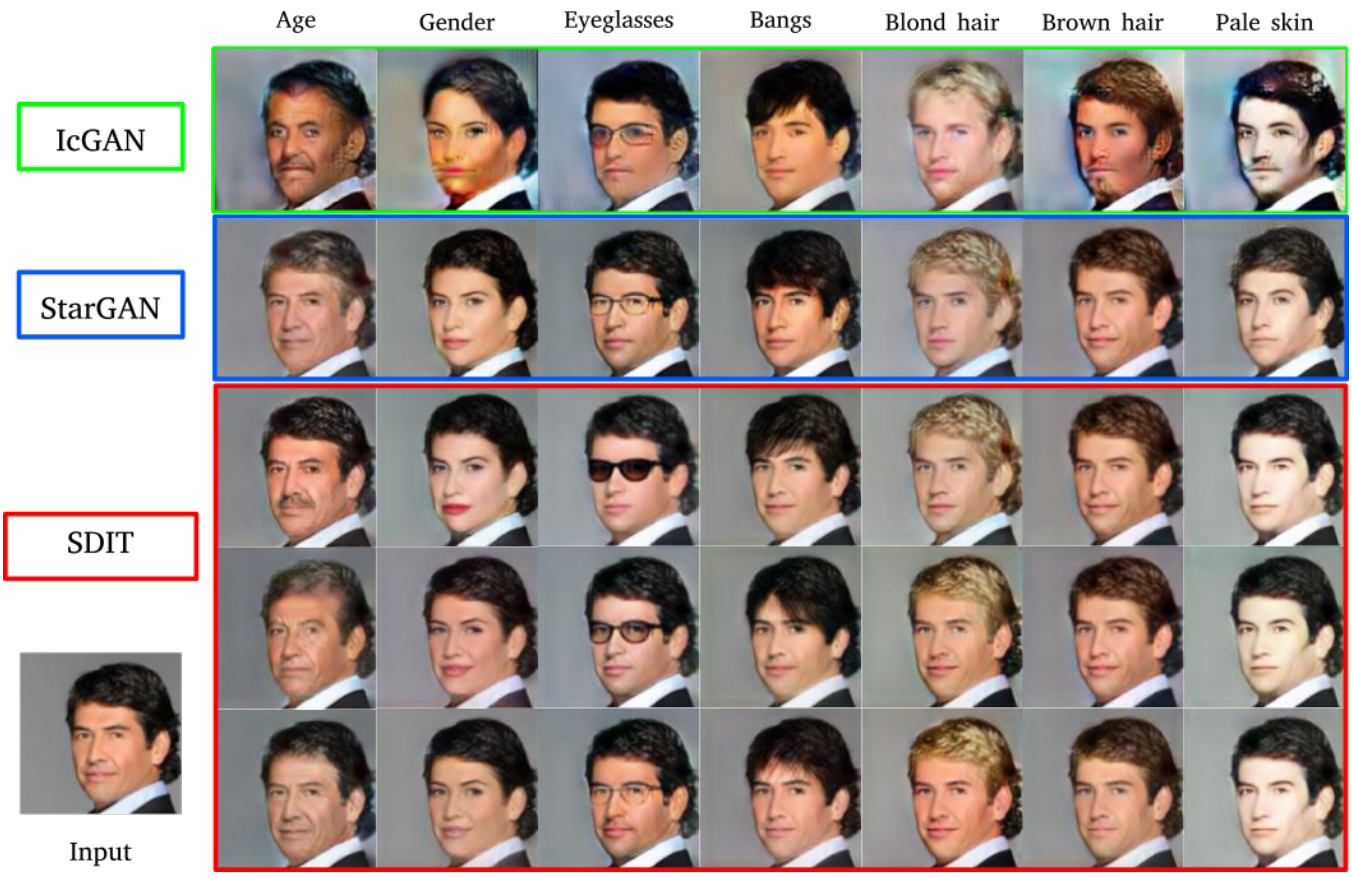

Figure 4: Qualitative comparison to the baselines. The input face image is at the left bottom and the remaining columns show the attribute-specific mapped images. The first two lines show the translated results of the IcGAN [35] and StarGAN [5], respectively, while the remaining rows are from the proposed method.

Method	Bangs	Age	Gender	Smiling	Wearing hat	Pale skin	Brown hair	Blond hair	Eyeglasses	Mouth open	Mean
StarGAN [5]	**0.067**/0.427	**0.065**/0.428	**0.068**/0.428	**0.061**/0.427	**0.075**/0.427	**0.064**/0.421	0.060/0.418	0.067/0.426	**0.066**/0.435	0.059/0.429	**0.065**/0.427
IcGAN [35]	0.118/0.430	0.097/0.431	0.094/0.430	0.121/0.430	0.102/0.429	0.10/0.430	0.127/0.424	0.113/0.421	0.097/0.425	0.116/0.438	0.108/0.432
SDIT	0.068/**0.456**	**0.065**/**0.447**	0.069/**0.444**	**0.061**/**0.449**	0.076/**0.458**	0.065/**0.439**	**0.058**/**0.443**	**0.067**/**0.442**	**0.066**/**0.458**	**0.058**/**0.457**	**0.065**/**0.451**
Real data	-/0.486	-/0.483	-/0.484	-/0.480	-/0.489	-/0.479	-/0.492	-/0.490	-/0.492	-/0.489	-/0.486

Table 2: ID distance (lower, better) / LPIPS (higher, better) on CelebA dataset.

distribution. At test time, MUNIT takes as input the source image and different style codes to achieve diverse outputs.

IcGAN [35]. IcGAN explicitly maps the input face into a latent feature, followed by a decoder which is conditioned on the latent feature and a target face attribute. In addition, the face attribute can be explicitly reconstructed by an inverse encoder.

StarGAN [5]. StarGAN shares the encoders and decoders for all domains. The full model is trained by optimizing the adversarial loss, the reconstruction loss and the cross-entropy loss, which controls that the input image is translated into a target image.

5.2 Face translation

We firstly conduct an experiment on the CelebA [25] dataset to compare against ablations of our full model. Next, we compare SDIT to the baselines. For this case, we consider IcGAN and StarGAN, both of which show outstanding results for face synthesis.

Ablation study. We performed an ablation study comparing several variants of SDIT in terms of model diversity. We consider five attributes, namely *bangs*, *blond hair*, *brown hair*, *young*, and *male*. Figure 3 shows the translated images obtained with different variants of our method. As expected, SDIT with only *attention* (second column of Figure 3) fails to synthesize diverse outputs, since the model lacks the additional factors (e.g. noise) to control this. Both the third and fourth columns show that adding CIN to our method without attention generates diverse images. Their quality, however, is unsatisfactory and the model suffers from partial mode collapse, since CIN operates on the entire image, rather than being localized by the attention mechanism to the desired area (e.g. the bangs). Combining both CIN and attention but without the latent code reconstruction ($\mathcal{L}_{LAT} = 0$) leads to little diversity, as shown in the fifth column. Finally, our full model (last column) achieves the best results in terms of quality and diversity.

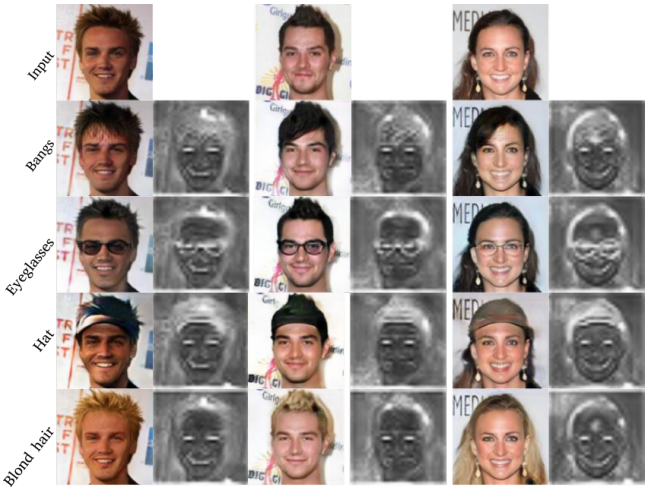

Figure 5: Generated images and learned attention maps for three input images. For each of them we present multi-domain outputs and attribute-specific attention.

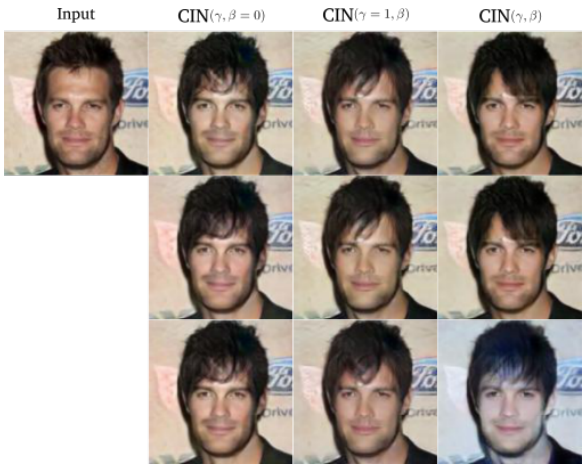

Figure 6: Ablation study on CIN. We compare three cases: (γ, β = 0), where γ is learnable; (γ = 1, β), where β is learnable; and (γ, β), where both γ and β are learnable.

For quantitative evaluation, we report the results in terms of the ID distance and LPIPS. As shown in Table 1, the SDIT models without CIN or $\mathcal{L}_{LAT}$ generate less diverse outputs according to LPIPS scores. Using $\mathcal{L}_{LAT}$ without attention contributes to improve the diversity. It has a higher LPIPS, but this could be because it is adding unwanted diversity (e.g. the red lips in the fourth column of Figure 3). This may explain its higher ID distance. Combining both attention and $\mathcal{L}_{LAT} > 0$ (i.e. the full SDIT model) encourages the results to have better targeted diversity, as reported in the last row of Table 1. The preservation of identity is crucial for the facial attribute transfer task, and thus we keep both attention and the reconstruction loss in the following sections.

Attention. Figure 5 shows the attention maps for several translations from the face dataset. We note that our method explicitly learns the attribute-specific attention for a given face image (e.g. *eyeglasses*), and generates the corresponding outputs. In this way, attention enables to modify only attribute-specific areas of the input image. This is a key factor to restrict the effect of the CIN, which otherwise would globally process the entire feature representation.

CIN learning. We explain here how CIN contributes to the diversity of the generator. In this experiment, we only consider CIN without attention nor latent code reconstruction. The operation performed by CIN on a feature e is given by:

$$\text{CIN}(e; z) = \gamma(z)\left(\frac{e - \mu(e)}{\delta(e)}\right) + \beta(z) \tag{10}$$

where e and z are the output of encoder E and latent code z, respectively; γ, β are affine parameters learned from M and $\mu(e)$, $\delta(e)$ are the mean and standard deviation. As shown in the second column of Figure 6, only learning γ fails to output diverse images, while only learning β already generates diverse results (third column of Figure 6), clearly indicating that β is the key factor to diversity. Updating the two parameters obtains a similar performance in this task. However, β could be ignored by the network. Therefore we introduced the latent code reconstruction loss, Eq. 6, which helps to avoid this.

Comparison against baselines. Figure 4 shows the comparison to the baselines on test data. We consider ten attributes: *bangs*, *blond hair*, *brown hair*, *young*, *male*, *mouth slightly open*, *smiling*, *pale skin*, *wearing hat*, and *eyeglasses*. Although both IcGAN and StarGAN are able to perform image-to-image translation to each domain, they fail to synthesize diverse outputs. Moreover, the performance of IcGAN is unsatisfactory and it fails to keep the personal identity. Our method not only enables the generation of realistic and diverse outputs, but also allows scalable image-to-image translation. Note that both StarGAN and our method use a single model. The visualization shows that scalability and diversity can be successfully integrated in a single model without conflict. Taking adding *bangs* as an example translation; the generated bangs with different directions do not impact the classification performance or the adversarial learning, in fact possibly contribute to the adversarial loss, since the CIN layer slightly reduces the compactness of the network, which increases the freedom of the generator.

As we can see in Table 2, our method obtains the best scores in both LPIPS and ID distance. In the case of LPIPS, the mean value of our method is 0.451, while IcGAN and StarGAN achieve 0.432 and 0.427 respectively. This clearly indicates that SDIT can successfully generate multimodal outputs using a single model. Moreover, the low ID distance indicates that SDIT effectively preserves the identity, achieving a competitive performance with StarGAN. Note that here we do not compare to CycleGAN and MUNIT because these methods require a single generator to be trained for each pair of domains. This is unfeasible for this task, because each attribute combination would require a different generator.

5.3 Object translation

The experiments in the previous section were conducted on a face dataset, in which all images have a relatively similar content and structure (a face on a background). Here we consider the color object dataset to show that SDIT can be applied to datasets that lack a common structure. This dataset contains a wide range of

Figure 7: Examples of scalable and diverse inference of multi-domain translations on (a) color dataset and (b) artworks dataset. In both cases, the first column is the input, the next three show results for CycleGAN [58], IcGAN [35], and StarGAN [5], respectively, followed by three samples from MUNIT [14] in next three columns and three samples from SDIT in the last three. Each row indicates a different domain.

Method	Yellow	Blue	Green	Orange	Mean	Num E/G
CycleGAN	93.4/0.599	95.1/0.601	93.4/0.584	92.3/0.587	93.5/0.592	6/6
IcGAN	92.2/0.581	93.5/0.592	92.8/0.579	92.1/0.589	92.6/0.585	1/1
StarGAN	95.9/0.591	95.3/0.602	96.0/0.590	94.2/0.584	95.3/0.591	1/1
MUNIT	97.3/0.607	**97.1**/0.603	**97.2**/0.599	96.8/0.621	**97.2**/0.608	6/6
SDIT	**97.6**/**0.610**	96.6/**0.607**	**97.3**/**0.604**	97.1/**0.627**	**97.1**/**0.612**	1/1
Real image	98.5/0.652	98.6/0.652	97.8/0.653	98.8/0.652	98.4/0.652	-/-

Table 3: Reverse classification accuracy (%) and LPIPS on the color dataset. For both metrics, the higher the better.

Method	Photo	Cezanne	Ukiyoe	Monet	Mean	Num E/G
CycleGAN	52.8/0.684	57.4/0.654	56.1/0.674	60.9/0.648	56.8/0.665	6/6
IcGAN	50.9/0.697	56.8/0.663	55.1/0.677	59.7/0.651	55.6/0.671	1/1
StarGAN	60.1/0.694	61.5/0.667	61.3/0.689	62.7/0.663	61.3/0.678	1/1
MUNIT	**66.2**/0.763	**67.9**/0.784	**67.2**/0.791	63.9/0.778	**66.3**/0.779	6/6
SDIT	65.6/**0.816**	63.4/**0.806**	65.3/**0.829**	**66.4**/**0.802**	65.1/**0.828**	1/1
Real image	70.2/0.856	72.4/0.874	69.9/0.884	71.7/0.864	71.1/0.869	-/-

Table 4: Reverse classification accuracy (%) and LPIPS on the artworks dataset. For both metrics, the higher the better.

different objects which greatly vary in shape, scale, and complexity. This makes the translation task more challenging.

Qualitative results. Figure 7(a) compares image-to-image translations obtained with CycleGAN [58], IcGAN [35], StarGAN [5], MUNIT [14] and the proposed method. We can see how SDIT clearly generates highly realistic and attribute-specific bags with different color shades, which is comparable to the results of MUNIT. Other baselines, however, only generate one color shade. The main advantage of SDIT is the *scalability*, as SDIT explicitly synthesizes the target color image (*yellow*, *green*, or *blue*) using a single generator.

Quantitative results. The qualitative observations above are validated here by quantitative evaluations. Table 3 compares the results of SDIT to the baseline methods. Our method outperforms both baseline methods on LPIPS despite only using a single model. For the classification accuracy, CycleGAN, IcGAN and StarGAN produce a lower score, since it is not able to generate diverse outputs for a given test samples. Both MUNIT and SDIT have a similar performance. However, for both CycleGAN and MUNIT training all pairwise translation would in case of N domains require $N \times (N-1)/2$ generators. Since we consider $N = 3$ here, we have trained a total of 6 generators for CycleGAN and MUNIT. The advantage of SDIT with respect to this non-scalable models would be even more evident for an increased number of domains.

5.4 Scene translation

Finally, we train our model on the photo and artworks dataset [58]. Differently from the model used for faces and color objects, here we consider the variant of our model without attention. This difference is due to the fact that previous datasets had a foreground that needed to be changed (object) and a fixed background, whereas in the scene case we need the generator to learn a global image translation instead of a local one, and thus background must also be changed.

Figure 7(b) shows several representative examples of the different methods. The conclusions are similar to previous experiments: SDIT maps the input (*photo*) to other domains with diversity while using a single model. Table 4 also confirms this, showing how the proposed method achieves excellent scores with only one scalable model.

6 CONCLUSION

We have introduced SDIT to perform image-to-image translation with scalability and diversity using a simple and compact network. The key challenge lies in controlling the two functions separately without conflict. We achieve scalability by conditioning the encoder with the target domain label, and diversity by applying conditional instance normalization in the bottleneck. In addition, the use of attention on the latent represent further improves the performance of image translation, allowing the model to mainly focus on domain-specific areas instead of the unrelated ones. The model has limited applicability for domains with large variations (for example, faces and paintings in a single model) and works better when the domains have characteristics in common.

Acknowledgements. Y. Wang acknowledges the Chinese Scholarship Council (CSC) grant No.201507040048. L. Herranz acknowledges the European Union research and innovation program under the Marie Skłodowska-Curie grant agreement No. 6655919. This work was supported by TIN2016-79717-R, and the CHISTERA project M2CR (PCIN-2015-251) of the Spanish Ministry, the CERCA Program of the *Generalitat de Catalunya*. We also acknowledge the generous GPU support from NVIDIA.

REFERENCES

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

[2] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. 2018. ComboGAN: Unrestrained Scalability for Image Domain Translation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Jun 2018). https://doi.org/10.1109/cvprw.2018.00122

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2172–2180.

[4] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. 2018. Attention-GAN for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 164–180.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[7] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. [n. d.]. A learned representation for artistic style. ([n. d.]).

[8] David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the International Conference on Computer Vision*. 2650–2658.

[9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. 1180–1189.

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

[11] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2066–2073.

[12] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. 2018. Image-to-image translation for cross-domain disentanglement. In *Advances in Neural Information Processing Systems*. 1294–1305.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[14] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*. 172–189.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).

[16] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. 2018. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*. 4020–4031.

[17] Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1219–1228.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

[19] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. *International Conference on Machine Learning* (2017).

[20] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations* (2014).

[21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4681–4690.

[22] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. 2018. Diverse Image-to-Image Translation via Disentangled Representations. In *Proceedings of the European Conference on Computer Vision*.

[23] Jerry Li. 2018. Twin-GAN–Unpaired Cross-Domain Image Translation with Weight-Sharing GANs. *arXiv preprint arXiv:1809.00946* (2018).

[24] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. *Advances in Neural Information Processing Systems* (2017).

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[27] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. 2018. DA-GAN: Instance-level image translation by deep attention generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5657–5666.

[28] Michael Mathieu, Camille Couprie, and Yann LeCun. 2016. Deep multi-scale video prediction beyond mean square error. *International Conference on Learning Representations* (2016).

[29] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. 2018. Unsupervised Attention-guided Image-to-Image Translation. In *Advances in Neural Information Processing Systems*. 3697–3707.

[30] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[31] Augustus Odena. 2016. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583* (2016).

[32] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*. JMLR. org, 2642–2651.

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference*.

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[35] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. 2016. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355* (2016).

[36] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision*. 818–833.

[37] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. *International Conference on Machine Learning* (2016).

[38] Oren Rippel and Lubomir Bourdev. 2047. Real-time adaptive image compression. In *International Conference on Machine Learning*. JMLR. org, 2922–2930.

[39] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. *International Conference on Machine Learning* (2017).

[40] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. 2018. How good is my GAN?. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 213–229.

[41] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. 2017. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5541–5550.

[42] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. 2018. Learning to adapt structured output space for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018).

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8798–8807.

[44] Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, and Luis Herranz. 2019. Controlling biases and diversity in diverse image-to-image translation. *arXiv preprint arXiv:1907.09754* (2019).

[45] Yaxing Wang, Joost van de Weijer, and Luis Herranz. 2018. Mix and match networks: encoder-decoder alignment for zero-pair image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5467–5476.

[46] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. 2018. Transferring GANs: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 218–234.

[47] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. 2018. Memory Replay GANs: Learning to Generate New Categories without Forgetting. In *Advances in Neural Information Processing Systems*. 5966–5976.

[48] Zuxuan Wu, Xintong Han, Yen-Liang Lin, Mustafa Gkhan Uzunbas, Tom Goldstein, Ser Nam Lim, and Larry S Davis. 2018. DCAN: Dual Channel-wise Alignment Networks for Unsupervised Scene Adaptation. In *Proceedings of the European Conference on Computer Vision*.

[49] Zili Yi, Hao Zhang, Ping Tan Gong, et al. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *Proceedings of the International Conference on Computer Vision*.

[50] Lu Yu, Yongmei Cheng, and Joost van de Weijer. 2018. Weakly Supervised Domain-Specific Color Naming Based on Attention. *arXiv preprint arXiv:1805.04385* (2018).

[51] Xiaoming Yu, Xing Cai, Zhenqiang Ying, Thomas Li, and Ge Li. 2018. SingleGAN: Image-to-Image Translation by a Single-Generator Network using Multiple Generative Adversarial Learning. In *Proceedings of the Asian Conference on Computer Vision*.

[52] He Zhang and Vishal M Patel. 2018. Densely Connected Pyramid Dehazing Network. In *CVPR*.

[53] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. 2017. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[54] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the*

International Conference on Computer Vision. 5907–5915.
[55] Lichao Zhang, Abel Gonzalez-Garcia, Joost van de Weijer, Martin Danelljan, and Fahad Shahbaz Khan. 2019. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing* 28, 4 (2019), 1837–1850.
[56] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*. Springer, 649–666.
[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.
[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision*.
[59] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*. 465–476.
[60] Yang Zou, Zhiding Yu, B.V.K. Vijaya Kumar, and Jinsong Wang. 2018. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In *Proceedings of the European Conference on Computer Vision*.