International Journal of Computer Vision
https://doi.org/10.1007/s11263-019-01284-z

DRIT++: Diverse Image-to-Image Translation via Disentangled Representations

Hsin-Ying Lee[1] · Hung-Yu Tseng[1] · Qi Mao[2] · Jia-Bin Huang[3] · Yu-Ding Lu[1] · Maneesh Singh[4] · Ming-Hsuan Yang[1]

Received: 26 April 2019 / Accepted: 15 December 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Image-to-image translation aims to learn the mapping between two visual domains. There are two main challenges for this task: (1) lack of aligned training pairs and (2) multiple possible outputs from a single input image. In this work, we present an approach based on disentangled representation for generating diverse outputs without paired training images. To synthesize diverse outputs, we propose to embed images onto two spaces: a domain-invariant content space capturing shared information across domains and a domain-specific attribute space. Our model takes the encoded content features extracted from a given input and attribute vectors sampled from the attribute space to synthesize diverse outputs at test time. To handle unpaired training data, we introduce a cross-cycle consistency loss based on disentangled representations. Qualitative results show that our model can generate diverse and realistic images on a wide range of tasks without paired training data. For quantitative evaluations, we measure realism with user study and Fréchet inception distance, and measure diversity with the perceptual distance metric, Jensen–Shannon divergence, and number of statistically-different bins.

1 Introduction

Image-to-Image (I2I) translation aims to learn the mapping between different visual domains. Numerous vision and graphics problems can be formulated as I2I translation problems, such as colorization (Larsson et al. 2016; Zhang et al. 2016) (grayscale → color), super-resolution (Lai et al. 2017; Ledig et al. 2017; Li et al. 2016, 2019) (low-resolution → high-resolution), and photorealistic image synthesis (Chen and Koltun 2017; Park et al. 2019; Wang et al. 2018) (label → image). In addition, I2I translation can be applied to synthesize images for domain adaptation (Bousmalis et al. 2017; Chen et al. 2019; Hoffman et al. 2018; Murez et al. 2018; Shrivastava et al. 2017).

Learning the mapping between two visual domains is challenging for two main reasons. First, aligned training image pairs are either difficult to collect (e.g., day scene ↔ night scene) or do not exist (e.g., artwork ↔ real photo). Second, many such mappings are inherently multimodal—a single input may correspond to multiple possible outputs. To handle multimodal translation, one possible approach is to inject

Communicated by Jun-Yan Zhu, Hongsheng Li, Eli Shechtman, Ming-Yu Liu, Jan Kautz, Antonio Torralba.

Hsin-Ying Lee, Hung-Yu Tseng and Qi Mao have contributed equally to this work.

✉ Ming-Hsuan Yang
mhyang@ucmerced.edu

Hsin-Ying Lee
hlee246@ucmerced.edu

Hung-Yu Tseng
htseng6@ucmerced.edu

Qi Mao
qimao@pku.edu.cn

Jia-Bin Huang
jbhuang@vt.edu

Yu-Ding Lu
ylu52@ucmerced.edu

Maneesh Singh
maneesh.singh@verisk.com

[1] Electrical Engineering and Computer Science, University of California at Merced, Merced, CA 95343, USA

[2] Electrical Engineering and Computer Science, Peking University, Beijing, China

[3] Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060, USA

[4] Verisk Analytics, Jersey City, NJ 07310, USA

Published online: 03 February 2020

a random noise vector to the generator for modeling the multimodal data distribution in the target domain. However, mode collapse may still occur easily since the generator often ignores the additional noise vectors.

Several recent efforts have been made to address these issues. The Pix2pix (Isola et al. 2017) method applies conditional generative adversarial network to I2I translation problems. Nevertheless, the training process requires paired data. A number of recent approaches (Choi et al. 2018; Liu et al. 2017; Taigman et al. 2017; Yi et al. 2017; Zhu et al. 2017a) relax the dependency on paired training data for learning I2I translation. These methods, however, generate a single output conditioned on the given input image. As shown in Isola et al. (2017) and Zhu et al. (2017b), the strategy of incorporating noise vectors as additional inputs to the generator does not increase variations of generated outputs due to the mode collapse issue. The generators in these methods are likely to overlook the added noise vectors. Most recently, the BicycleGAN (Zhu et al. 2017b) algorithm tackles the problem of generating diverse outputs in I2I translation by encouraging the one-to-one relationship between the output and the latent vector. Nevertheless, the training process of BicycleGAN requires paired images.

In this paper, we propose a disentangled representation framework for learning to generate *diverse* outputs with *unpaired* training data. We propose to embed images onto two spaces: (1) a domain-invariant content space and (2) a domain-specific attribute space as shown in Fig. 2. Our generator learns to perform I2I translation conditioned on content features and a latent attribute vector. The domain-specific attribute space aims to model variations within a domain given the same content, while the domain-invariant content space captures information across domains. We disentangle the representations by applying a content adversarial loss to encourage the content features *not* to carry domain-specific cues, and a latent regression loss to encourage the invertible mapping between the latent attribute vectors and the corresponding outputs. To handle unpaired datasets, we propose a *cross-cycle consistency loss* using the proposed disentangled representations. Given a pair of unaligned images, we first perform a cross-domain mapping to obtain intermediate results by swapping the attribute vectors from both images. We can then reconstruct the original input image pair by applying the cross-domain mapping one more time and use the proposed cross-cycle consistency loss to enforce the consistency between the original and the reconstructed images. Furthermore, we apply the mode seeking regularization (Mao et al. 2019) to further improve the diversity of generated images. At test time, we can use either 1) randomly sampled vectors from the attribute space to generate diverse outputs or 2) the transferred attribute vectors extracted from existing images for example-guided translation. Figure 1 shows examples of diverse outputs produced by our model (Fig. 2).

We evaluate the proposed model with extensive qualitative and quantitative experiments. For various I2I tasks, we show diverse translation results with randomly sampled attribute vectors and example-guided translation with transferred attribute vectors from existing images. In addition to the common dual-domain image-to-image translation, we extend our proposed framework to the more general multi-domain image-to-image translation and demonstrate diverse translation among domains. We measure realism of our results with a user study and the Fréchet inception distance (FID) (Heusel et al. 2017), and evaluate diversity using perceptual distance metrics (Zhang et al. 2018b). However, the diversity metric alone does not effectively measure similarity between the distribution of generated images and the distribution of real data. Therefore, we use the Jensen-Shannon Divergence (JSD) distance which measures the similarity between distributions, and the Number of Statistically-Different Bins (NDB) (Richardson and Weiss 2018) metric which determines the relative proportions of samples within clusters predetermined by real data.

We make the following contributions in this work:

(1) We introduce a disentangled representation framework for image-to-image translation. We apply a content discriminator to facilitate the factorization of domain-invariant content space and domain-specific attribute space, and a cross-cycle consistency loss that allows us to train the model with unpaired data.
(2) Extensive qualitative and quantitative experiments show that our model performs favorably against existing I2I models. Images generated by our model are both diverse and realistic.
(3) The proposed disentangled representation and cross-cycle consistency can be applied to multi-domain image-to-image translation for generating diverse images.

2 Related Work

Generative adversarial networks. The recent years have witnessed rapid advances of generative adversarial networks (GANs) (Arjovsky et al. 2017; Goodfellow et al. 2014; Radford et al. 2016) for image generation. The core idea of GANs lies in the adversarial loss that enforces the distribution of generated images to match that of the target domain. The generators in GANs can map from noise vectors to realistic images. Several recent efforts exploit *conditional* GAN in various contexts including conditioned on text (Reed et al. 2016), audio (Lee et al. 2019), low-resolution images (Ledig et al. 2017), human pose (Ma et al. 2017; AlBahar and Huang 2019), video frames (Vondrick et al. 2016), and image (Isola et al. 2017). Our work focuses on using GAN conditioned

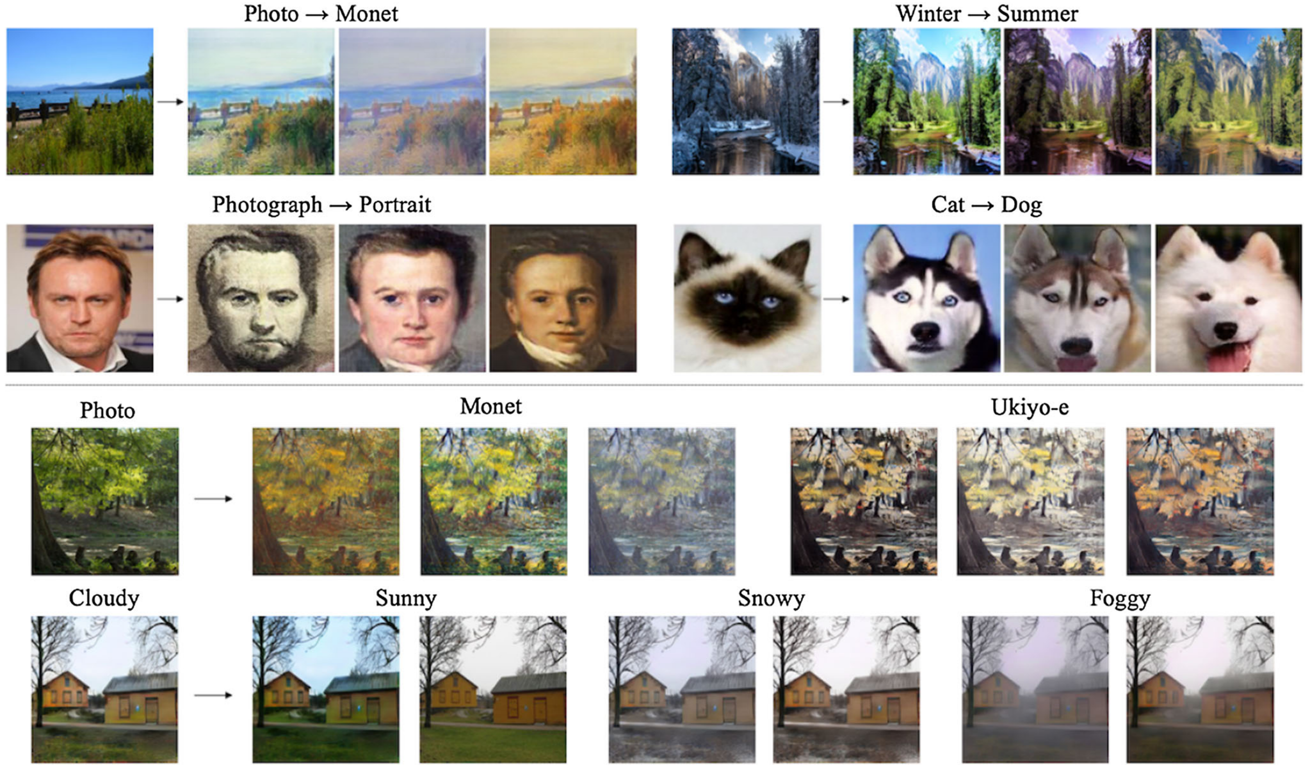
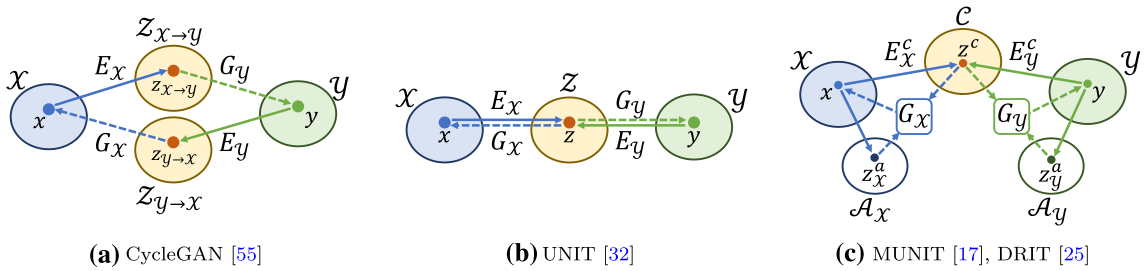


Fig. 1 Unpaired diverse image-to-image translation. (Top) Our model learns to perform diverse translation between two collections of images without aligned training pairs. (Bottom) Multi-domain image-to-image translation

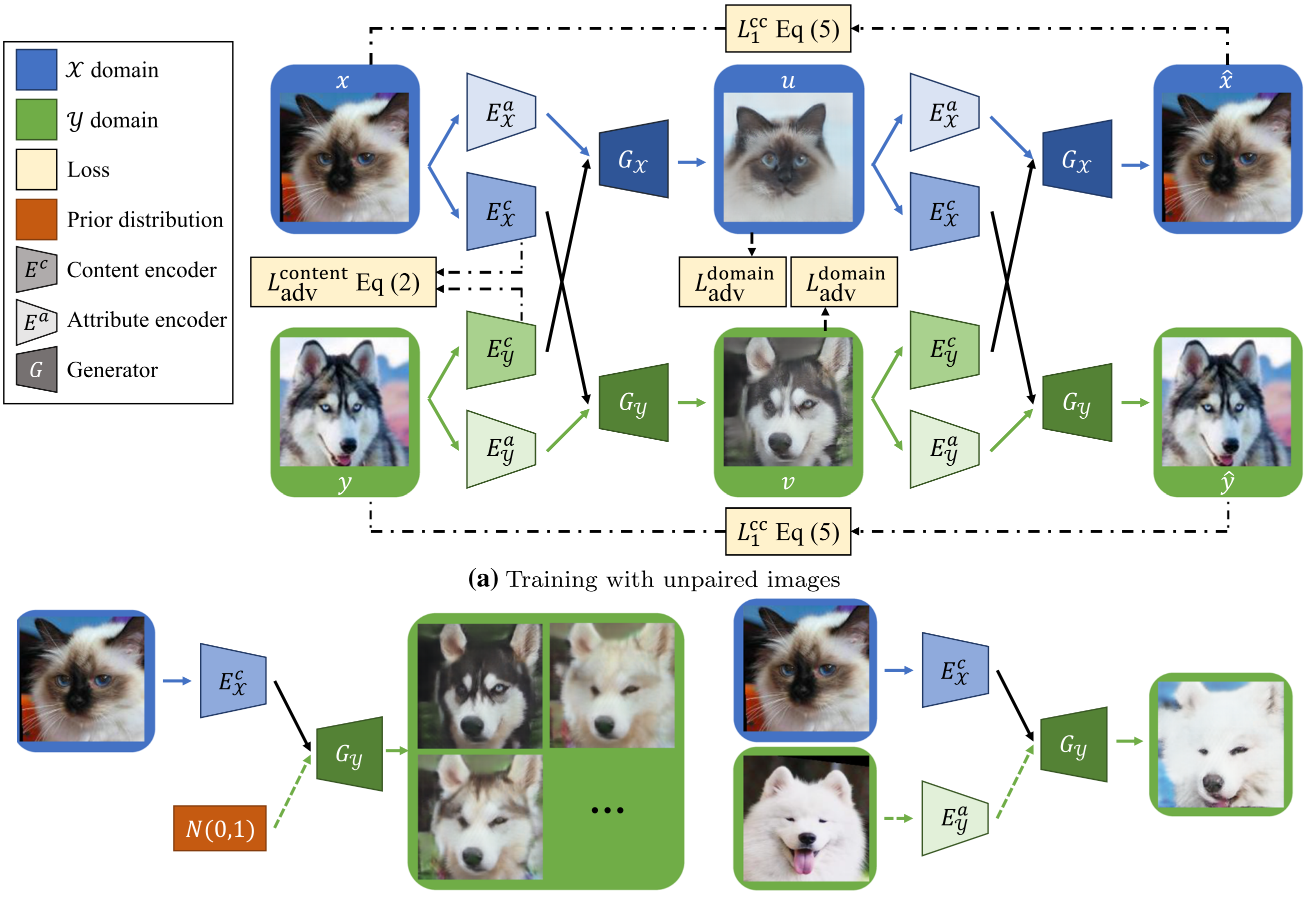


(a) CycleGAN [55] **(b)** UNIT [32] **(c)** MUNIT [17], DRIT [25]

Fig. 2 Comparisons of unsupervised I2I translation methods. Denote x and y as images in domain $\mathcal{X}$ and $\mathcal{Y}$: **a** CycleGAN (Zhu et al. 2017a) maps x and y onto *separated* latent spaces. **b** UNIT (Liu et al. 2017) assumes x and y can be mapped onto a *shared* latent space. **c** Our approach disentangles the latent spaces of x and y into a shared content space $\mathcal{C}$ and an attribute space $\mathcal{A}$ of each domain

on an input image. In contrast to several existing conditional GAN frameworks that require paired training data, our model generates diverse outputs without paired data. As such, our method has wider applicability to problems where paired training datasets are scarce or not available.

Image-to-image translation. I2I translation aims to learn the mapping from a source image domain to a target image domain. The Pix2pix (Isola et al. 2017) method applies a conditional GAN to model the mapping function. Although high-quality results have been shown, the model training requires paired training data. To train with unpaired data, the CycleGAN (Zhu et al. 2017a), DiscoGAN (Kim et al. 2017), and UNIT (Liu et al. 2017) schemes leverage cycle consistency to regularize the training. However, these methods perform generation conditioned solely on an input image and thus produce one single output. Simply injecting a noise vector to a generator is usually not an effective solution to achieve multimodal generation due to the lack of regularization between the noise vectors and the target domain. On the other hand, the BicycleGAN (Zhu et al. 2017b) algorithm enforces the bijection mapping between the latent and target space to tackle the mode collapse problem. Nevertheless, the method is only applicable to problems with paired training



(a) Training with unpaired images

(b) Testing with random attributes

(c) Testing with a given attribute

Fig. 3 Method overview. **a** With the proposed content adversarial loss $L_{adv}^{content}$ (Sect. 3.1) and the cross-cycle consistency loss L_1^{cc} (Sect. 3.2), we are able to learn the multimodal mapping between the domain $\mathcal{X}$ and $\mathcal{Y}$ with unpaired data. Thanks to the proposed disentangled representation, we can generate output images conditioned on either **b** random attributes or **c** a given attribute at test time

data. Unlike existing work, our method enables I2I translation with diverse outputs in the absence of paired training data.

We note several concurrent methods (Almahairi et al. 2018; Cao et al. 2018; Huang et al. 2018; Lin et al. 2018a, b; Ma et al. 2018) (all independently developed) also adopt disentangled representations similar to our work for learning diverse I2I translation from unpaired training data. Furthermore, several approaches (Choi et al. 2018; Liu et al. 2018) extend the conventional dual-domain I2I to general multi-domain settings. However, these methods can only achieve one-to-one mapping among domains.

Disentangled representations. The task of learning disentangled representation aims at modeling the factors of data variations. Previous work makes use of labeled data to factorize representations into class-related and class-independent components (Cheung et al. 2015; Kingma et al. 2014; Makhzani et al. 2016; Mathieu et al. 2016). Recently, numerous unsupervised methods have been developed (Chen et al. 2016; Denton and Birodkar 2017) to learn disentangled representations. The InfoGAN (Chen et al. 2016) algorithm achieves disentanglement by maximizing the mutual information between latent variables and data variation. Similar to DrNet (Denton and Birodkar 2017) that separates time-independent and time-varying components with an adversarial loss, we apply a content adversarial loss to disentangle an image into domain-invariant and domain-specific representations to facilitate learning diverse cross-domain mappings.

3 Disentangled Representation for I2I Translation

Our goal is to learn a multimodal mapping between two visual domains $\mathcal{X} \subset \mathbb{R}^{H\times W\times 3}$ and $\mathcal{Y} \subset \mathbb{R}^{H\times W\times 3}$ without paired training data. As illustrated in Fig. 3, our framework consists of content encoders $\{E^c_{\mathcal{X}}, E^c_{\mathcal{Y}}\}$, attribute encoders $\{E^a_{\mathcal{X}}, E^a_{\mathcal{Y}}\}$, generators $\{G_{\mathcal{X}}, G_{\mathcal{Y}}\}$, and domain discriminators $\{D_{\mathcal{X}}, D_{\mathcal{Y}}\}$

for both domains, and a content discriminators $D^{\mathrm{c}}_{\mathrm{adv}}$. Taking domain $\mathcal{X}$ as an example, the content encoder $E^c_{\mathcal{X}}$ maps images onto a shared, domain-invariant content space ($E^c_{\mathcal{X}} : \mathcal{X} \rightarrow \mathcal{C}$) and the attribute encoder $E^a_{\mathcal{X}}$ maps images onto a domain-specific attribute space ($E^a_{\mathcal{X}} : \mathcal{X} \rightarrow \mathcal{A}_{\mathcal{X}}$). The generator $G_{\mathcal{X}}$ synthesizes images conditioned on both content and attribute vectors ($G_{\mathcal{X}} : \{\mathcal{C}, \mathcal{A}_{\mathcal{X}}\} \rightarrow \mathcal{X}$). The discriminator $D_{\mathcal{X}}$ aims to discriminate between real images and translated images in the domain $\mathcal{X}$. In addition, the content discriminator D^c is trained to distinguish the extracted content representations between two domains. To synthesize multimodal outputs at test time, we regularize the attribute vectors so that they can be drawn from a prior Gaussian distribution $N(0, 1)$.

3.1 Disentangle Content and Attribute Representations

Our approach embeds input images onto a shared content space $\mathcal{C}$, and domain-specific attribute spaces, $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{\mathcal{Y}}$. Intuitively, the content encoders should encode the common information that is *shared* between domains onto $\mathcal{C}$, while the attribute encoders should map the remaining domain-specific information onto $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{\mathcal{Y}}$.

$$\begin{aligned} \{z^c_x, z^a_x\} &= \{E^c_{\mathcal{X}}(x), E^a_{\mathcal{X}}(x)\} \qquad z^c_x \in \mathcal{C}, z^a_x \in \mathcal{A}_{\mathcal{X}}, \\ \{z^c_y, z^a_y\} &= \{E^c_{\mathcal{Y}}(y), E^a_{\mathcal{Y}}(y)\} \qquad z^c_y \in \mathcal{C}, z^a_y \in \mathcal{A}_{\mathcal{Y}}. \end{aligned} \tag{1}$$

To achieve representation disentanglement, we apply two strategies: weight-sharing and a content discriminator. First, similar to Liu et al. (2017), based on the assumption that two domains share a common latent space, we share the weight between the last layer of $E^c_{\mathcal{X}}$ and $E^c_{\mathcal{Y}}$ and the first layer of $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$. Through weight sharing, we enforce the content representation to be mapped onto the same space. However, sharing the same high-level mapping functions does not guarantee the same content representations encode the same information for both domains. Thus, we propose a content discriminator D^c which aims to distinguish the domain membership of the encoded content features z^c_x and z^c_y. On the other hand, content encoders learn to produce encoded content representations whose domain membership cannot be distinguished by the content discriminator D^c. We express this content adversarial loss as:

$$\begin{aligned} &L^{\mathrm{content}}_{\mathrm{adv}}(E^c_{\mathcal{X}}, E^c_{\mathcal{Y}}, D^c) \\ &\quad = \mathbb{E}_x\left[\frac{1}{2}\log D^c(E^c_{\mathcal{X}}(x)) + \frac{1}{2}\log\left(1 - D^c(E^c_{\mathcal{X}}(x))\right)\right] \\ &\qquad + \mathbb{E}_y\left[\frac{1}{2}\log D^c(E^c_{\mathcal{Y}}(y)) + \frac{1}{2}\log\left(1 - D^c(E^c_{\mathcal{Y}}(y))\right)\right]. \end{aligned} \tag{2}$$

3.2 Cross-Cycle Consistency Loss

With the disentangled representation where the content space is shared among domains and the attribute space encodes intra-domain variations, we can perform I2I translation by combining a content representation from an arbitrary image and an attribute representation from an image of the target domain. We leverage this property and propose a *cross-cycle consistency*. In contrast to cycle consistency constraint in Zhu et al. (2017a) (i.e., $\mathcal{X} \rightarrow \mathcal{Y} \rightarrow \mathcal{X}$) which assumes one-to-one mapping between the two domains, the proposed cross-cycle constraint exploit the disentangled content and attribute representations for cyclic reconstruction.

Our cross-cycle constraint consists of two stages of I2I translation.

Forward translation. Given a non-corresponding pair of images x and y, we encode them into $\{z^c_x, z^a_x\}$ and $\{z^c_y, z^a_y\}$. We then perform the first translation by swapping the attribute representation (i.e., z^a_x and z^a_y) to generate $\{u, v\}$, where $u \in \mathcal{X}, v \in \mathcal{Y}$.

$$u = G_{\mathcal{X}}(z^c_y, z^a_x) \quad v = G_{\mathcal{Y}}(z^c_x, z^a_y). \tag{3}$$

Backward translation. After encoding u and v into $\{z^c_u, z^a_u\}$ and $\{z^c_v, z^a_v\}$, we perform the second translation by once again swapping the attribute representation (i.e., z^a_u and z^a_v).

$$\hat{x} = G_{\mathcal{X}}(z^c_v, z^a_u) \quad \hat{y} = G_{\mathcal{Y}}(z^c_u, z^a_v). \tag{4}$$

Here, after two I2I translation stages, the translation should reconstruct the original images x and y (as illustrated in Fig. 3). To enforce this constraint, we formulate the *cross-cycle consistency loss* as:

$$\begin{aligned} &L^{\mathrm{cc}}_1(G_{\mathcal{X}}, G_{\mathcal{Y}}, E^c_{\mathcal{X}}, E^c_{\mathcal{Y}}, E^a_{\mathcal{X}}, E^a_{\mathcal{Y}}) \\ &\quad = \mathbb{E}_{x,y}[\|G_{\mathcal{X}}(E^c_{\mathcal{Y}}(v), E^a_{\mathcal{X}}(u)) - x\|_1 \\ &\qquad + \|G_{\mathcal{Y}}(E^c_{\mathcal{X}}(u), E^a_{\mathcal{Y}}(v)) - y\|_1], \end{aligned} \tag{5}$$

where $u = G_{\mathcal{X}}(E^c_{\mathcal{Y}}(y), E^a_{\mathcal{X}}(x))$ and $v = G_{\mathcal{Y}}(E^c_{\mathcal{X}}(x), E^a_{\mathcal{Y}}(y))$, respectively.

3.3 Other Loss Functions

In addition to the proposed content adversarial loss and cross-cycle consistency loss, we also use several other loss functions to facilitate network training. We illustrate these additional losses in Fig. 4. Starting from the top-right, in the counter-clockwise order:

Domain adversarial loss. We impose an adversarial loss $L^{\mathrm{domain}}_{\mathrm{adv}}$ where $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ attempt to discriminate between

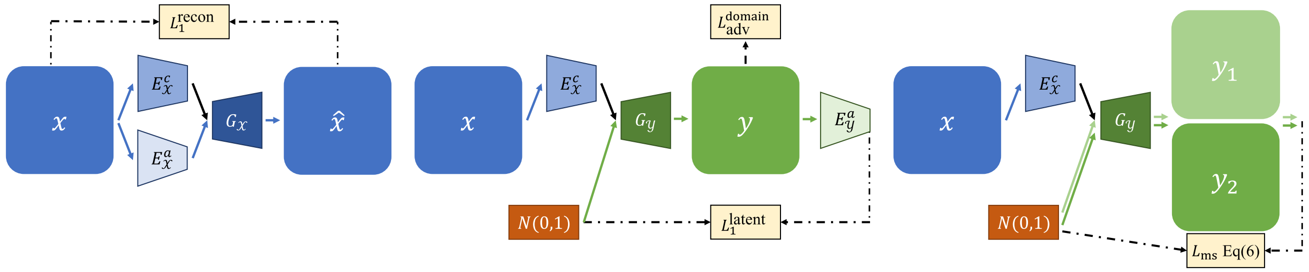


Fig. 4 Additional loss functions. In addition to the cross-cycle reconstruction loss L_1^{cc} and the content adversarial loss $L_{\text{adv}}^{\text{content}}$ described in Fig. 3, we apply several additional loss functions in our training process. The self-reconstruction loss L_1^{recon} facilitates training with self-reconstruction; the KL loss L_{KL} aims to align the attribute representation with a prior Gaussian distribution; the adversarial loss $L_{\text{adv}}^{\text{domain}}$ encourages G to generate realistic images in each domain; and the latent regression loss L_1^{latent} enforces the reconstruction on the latent attribute vector. Finally, the mode seeking regularization L_{ms} further improves the diversity. More details can be found in Sects. 3.3 and 3.4

real images and generated images in each domain, while $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ attempt to generate realistic images.

Self-reconstruction loss. In addition to the cross-cycle reconstruction, we apply a self-reconstruction loss L_1^{rec} to facilitate the training process. With encoded content and attribute features $\{z_x^c, z_x^a\}$ and $\{z_y^c, z_y^a\}$, the decoders $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ should decode them back to original input x and y. That is, $\hat{x} = G_{\mathcal{X}}(E_{\mathcal{X}}^c(x), E_{\mathcal{X}}^a(x))$ and $\hat{y} = G_{\mathcal{Y}}(E_{\mathcal{Y}}^c(y), E_{\mathcal{Y}}^a(y))$.

Latent regression loss. To encourage invertible mapping between the image and the latent space, we apply a latent regression loss L_1^{latent} similar to Zhu et al. (2017b). We draw a latent vector z from the prior Gaussian distribution as the attribute representation and attempt to reconstruct it with $\hat{z} = E_{\mathcal{X}}^a(G_{\mathcal{X}}(E_{\mathcal{X}}^c(x), z))$ and $\hat{z} = E_{\mathcal{Y}}^a(G_{\mathcal{Y}}(E_{\mathcal{Y}}^c(y), z))$.

The full objective function of our network is:

$$L_{D,D^c} = \lambda_{\text{adv}}^{\text{content}} L_{\text{adv}}^{\text{c}} + \lambda_{\text{adv}}^{\text{domain}} L_{\text{adv}}^{\text{domain}}, \tag{6}$$

$$\begin{aligned} L_{G,E^c,E^a} = &- L_{D,D^c} + \lambda_1^{\text{cc}} L_1^{\text{cc}} + \lambda_1^{\text{recon}} L_1^{\text{recon}} \\ &+ \lambda_1^{\text{latent}} L_1^{\text{latent}}, \end{aligned} \tag{7}$$

where the hyper-parameters λs control the importance of each term.

3.4 Mode Seeking Regularization

We incorporate the mode seeking regularization (Mao et al. 2019) method to alleviate the mode-collapse problem in conditional generation tasks. Given a conditional image $\mathbf{I}$, latent vectors $\mathbf{z}_1$ and $\mathbf{z}_2$, and a conditional generator G, we use the mode seeking regularization term to maximize the ratio of the distance between $G(\mathbf{I}, \mathbf{z}_1)$ and $G(\mathbf{I}, \mathbf{z}_2)$ with respect to the distance between $\mathbf{z}_1$ and $\mathbf{z}_2$,

$$\mathcal{L}_{\text{ms}} = \max_G \left(\frac{d_{\mathbf{I}}(G(\mathbf{I}, \mathbf{z}_1), G(\mathbf{I}, \mathbf{z}_2))}{d_{\mathbf{z}}(\mathbf{z}_1, \mathbf{z}_2)} \right), \tag{8}$$

where $d_*(\cdot)$ denotes the distance metric.

The regularization term can be easily incorporated into the proposed framework:

$$\mathcal{L}_{\text{new}} = \mathcal{L}_{\text{ori}} + \lambda_{\text{ms}} \mathcal{L}_{\text{ms}}, \tag{9}$$

where $\mathcal{L}_{\text{ori}}$ denote the full objective.

3.5 Multi-Domain Image-to-Image Translation

In addition to the translation between two domains, we apply the proposed disentangle representation to the multi-domain setting. Different from typical I2I designed for two domains, multi-domain I2I aims to perform translation among multiple domains with a single generator G.

We illustrate the framework for multi-domain I2I in Fig. 5. Given k domains $\{N_i\}_{i=1\sim k}$, two images (x, y) and their one-hot domain codes (z_x^d, z_y^d) are randomly sampled $(x \in N_n, y \in N_m, Z^d \subset \mathbb{R}^k)$. We encode the images onto a shared content space $\mathcal{C}$, and domain-specific attribute spaces $\{\mathcal{A}_i\}_{i=1\sim k}$:

$$\begin{aligned} \{z_x^c, z_x^a\} &= \{E^c(x), E^a(x, z_x^d)\} \quad z_x^c \in \mathcal{C}, z_x^a \in \mathcal{A}_n, \\ \{z_y^c, z_y^a\} &= \{E^c(y), E^a(y, z_y^d)\} \quad z_y^c \in \mathcal{C}, z_y^a \in \mathcal{A}_m. \end{aligned} \tag{10}$$

We then perform the forward and backward translation similar to the dual-domain translation.

$$\begin{aligned} u &= G(z_y^c, z_x^a, z_x^d) \quad v = G(z_x^c, z_y^a, z_y^d), \\ \hat{x} &= G(z_v^c, z_u^a, z_u^d) \quad \hat{y} = G(z_u^c, z_v^a, z_v^d). \end{aligned} \tag{11}$$

In addition to the loss functions used in the dual-domain translation, we leverage the discriminator D as an auxiliary domain classifier. That is, the discriminator D not only aims to discriminate between real images and translated images

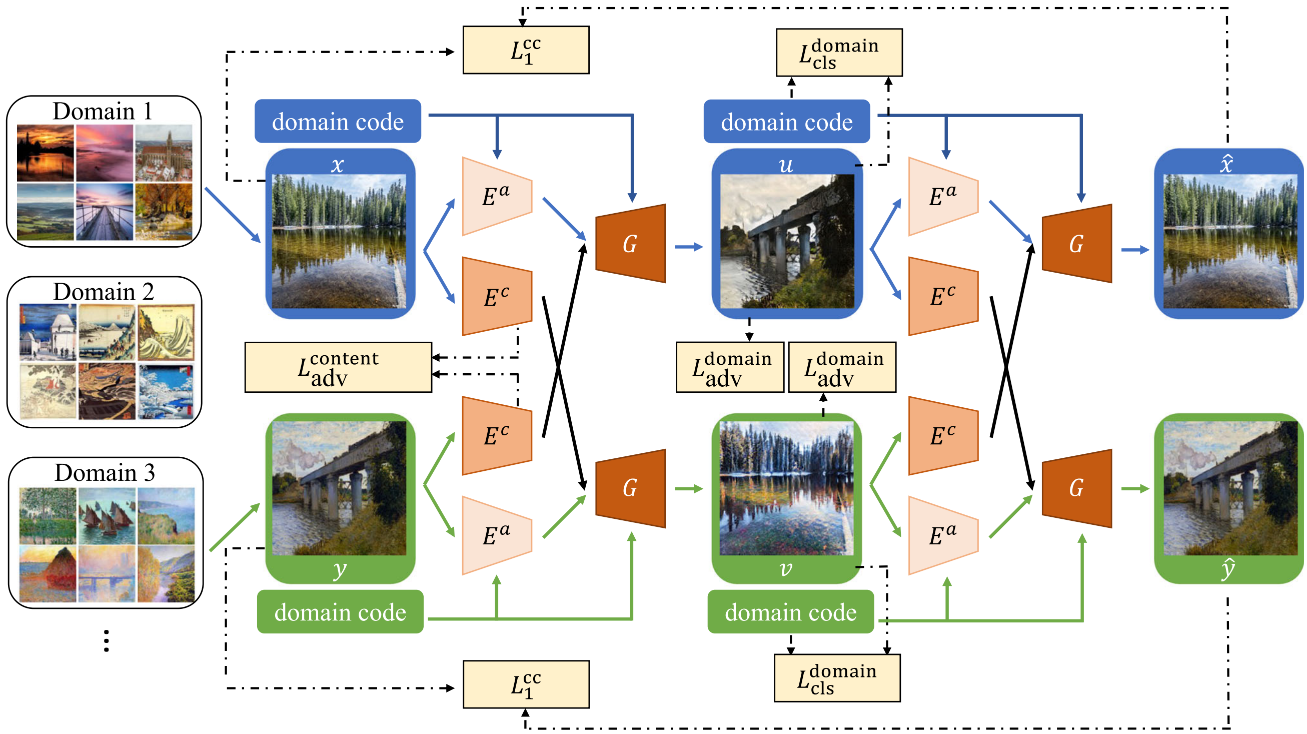


Fig. 5 Multi-domains I2I framework. We further extend the proposed disentangle representation framework to a more general multi-domain setting. Different from the class-specific encoders, generators, and discriminators used in dual-domain I2I, all networks in multi-domain are shared among all domains. Furthermore, one-hot domain codes are used as inputs and the discriminator will perform domain classification in addition to discrimination

(D_{dis}), but also performs domain classification ($D_{\text{cls}} : N_i \rightarrow Z^d$).

$$\begin{aligned}\mathcal{L}_{\text{cls}}^{\text{domain}} &= \mathbb{E}_{x,z_x^d}[-\log D_{\text{cls}}(z_x^d|x)] \\ &\quad + \mathbb{E}_{x,y,z_y^d}[-\log D_{\text{cls}}(z_y^d|G(z_x^c, z_y^a, z_y^d)].\end{aligned} \tag{12}$$

Thus, our new objective function is:

$$\begin{aligned}L_{D,D^c} &= \lambda_{\text{adv}}^{\text{content}} L_{\text{adv}}^{\text{c}} + \lambda_{\text{adv}}^{\text{domain}} L_{\text{adv}}^{\text{domain}} \\ &\quad + \lambda_{\text{cls}}^{\text{domain}} \mathcal{L}_{\text{cls}}^{\text{domain}},\end{aligned} \tag{13}$$

$$\begin{aligned}L_{G,E^c,E^a} &= -L_{D,D^c} + \lambda_1^{\text{cc}} L_1^{\text{cc}} + \lambda_1^{\text{recon}} L_1^{\text{recon}} \\ &\quad + \lambda_1^{\text{latent}} L_1^{\text{latent}} + \lambda_{\text{KL}} L_{\text{KL}} + \lambda_{\text{cls}}^{\text{domain}} \mathcal{L}_{\text{cls}}^{\text{domain}}.\end{aligned} \tag{14}$$

4 Experimental Results

Implementation details. We implement the proposed model with PyTorch (Paszke et al. 2017). We use the input image size of 216 × 216 for all of our experiments. For the content encoder E^c, we use an architecture consisting of three convolution layers followed by four residual blocks. For the attribute encoder E^a, we use a CNN architecture with four convolution layers followed by fully-connected layers. We set the size of the attribute vector to $z^a \in R^8$ for all experiments. For the generator G, we use an architecture consisting of four residual blocks followed by three fractionally strided convolution layers.

For training, we use the Adam optimizer (Kinga and Adam 2015) with a batch size of 1, a learning rate of 0.0001, and exponential decay rates $(\beta_1, \beta_2) = (0.5, 0.999)$. In all experiments, we set the hyper-parameters as follows: $\lambda_{\text{adv}}^{\text{content}} = 1$, $\lambda_{\text{cc}} = 10$, $\lambda_{\text{adv}}^{\text{domain}} = 1$, $\lambda_1^{\text{rec}} = 10$, $\lambda_{\text{ms}} = 1$, $\lambda_1^{\text{latent}} = 10$, and $\lambda_{\text{KL}} = 0.01$. We also apply an L1 weight regularization on the content representation with a weight of 0.01. We follow the procedure in DCGAN (Radford et al. 2016) for training the model with adversarial loss. More results can be found at http://vllab.ucmerced.edu/hylee/DRIT_pp/. The source code and trained models will be made available to the public (Table 1).

Datasets. We evaluate the proposed model on several datasets include Yosemite (Zhu et al. 2017a) (summer and winter scenes), pets (cat and dog) cropped from Google images, artworks (Zhu et al. 2017a) (Monet), and photo-to-portrait cropped from subsets of the WikiArt dataset[1] and the CelebA dataset (Liu et al. 2015).

[1] https://www.wikiart.org/

Table 1 Summary of the components used in each method

Method	Mode-seeking	Multi-domain	High-resolution
DRIT	–	–	–
DRIT++ (two-domain)	✓	–	–
DRIT++ (multi-domain)	✓	✓	–
DRIT++ (high-resolution)	✓	–	✓

We desciribe the differences among DRIT, DRIT++, and variants

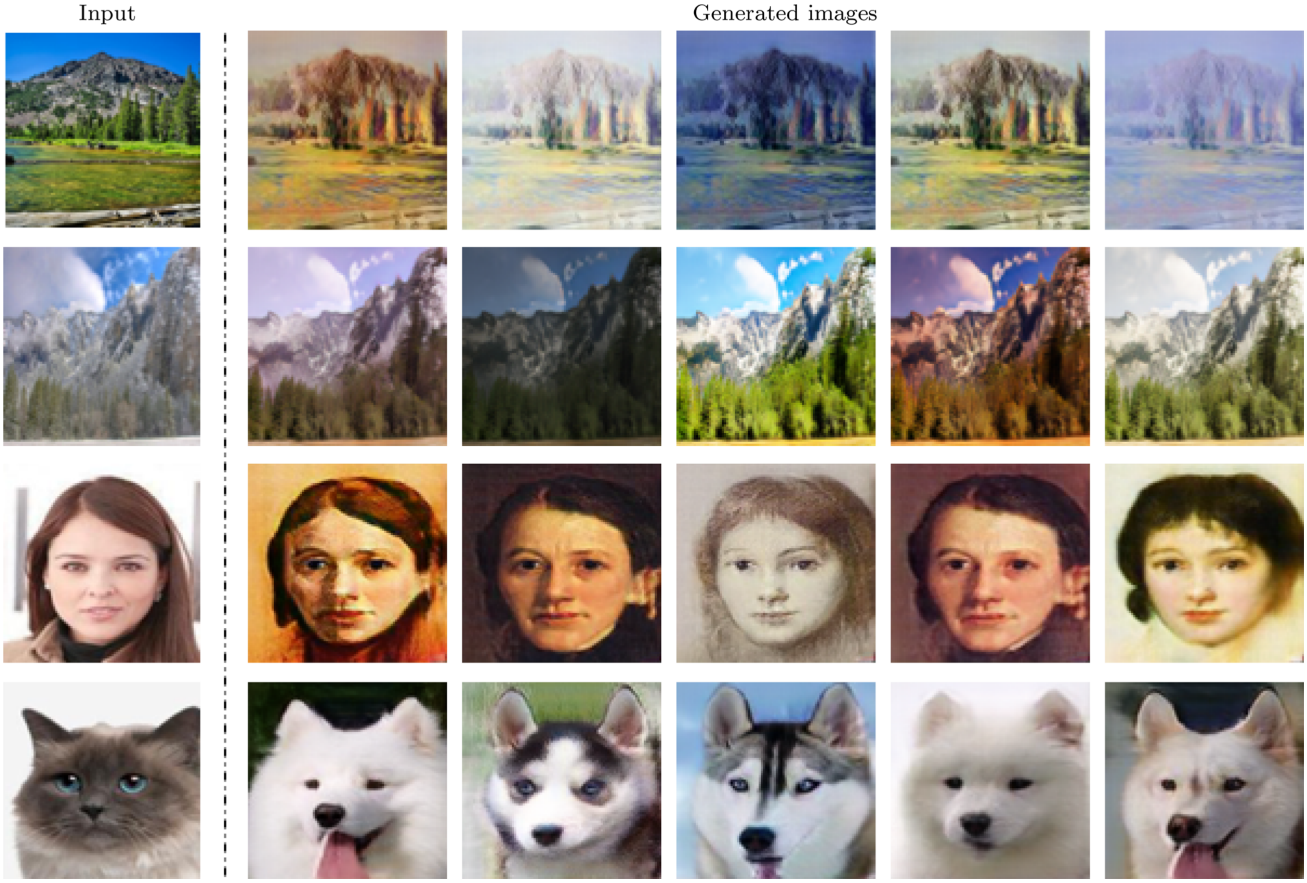


Fig. 6 Sample results. We show example results produced by our model. The left column shows the input images in the source domain. The other five columns show the output images generated by sampling random vectors in the attribute space. The mappings from top to bottom are: Photo → Monet, winter → summer, photograph → portrait, and cat → dog

Evaluated methods. We perform the evaluation on the following algorithms:

- **DRIT++:** The proposed model.
- **DRIT** (Lee et al. 2018), and **MUNIT** (Huang et al. 2018): Multimodal generation frameworks trained with unpaired data.
- **DRIT w/o** D^c: DRIT model without the content discriminator.
- **Cycle/Bicycle:** We construct a baseline using a combination of CylceGAN and BicycleGAN. Here, we first train CycleGAN on unpaired data to generate corresponding images as *pseudo* image pairs. We then use this pseudo paired data to train BicycleGAN.
- **CycleGAN** (Zhu et al. 2017a), and **BicycleGAN** (Zhu et al. 2017b)

The proposed DRIT++ method extends the original DRIT method by (1) incorporating mode-seeking regularization for improving sample diversity and (2) generalizing the two-domain model to handle multi-domain image-to-image translation problems. The DRIT++ (multi-domain) algorithm is *backward compatible* with the DRIT++ (two-domain) and DRIT methodss with comparable performance (as shown in Sect. 4.2). Thus, the DRIT++ (two-domain) method can be viewed as a special case of the DRIT++ (multi-domain) algorithm. The DRIT++ (two-domain) algorithm can improve the visual quality slightly over the DRIT++ (multi-domain)

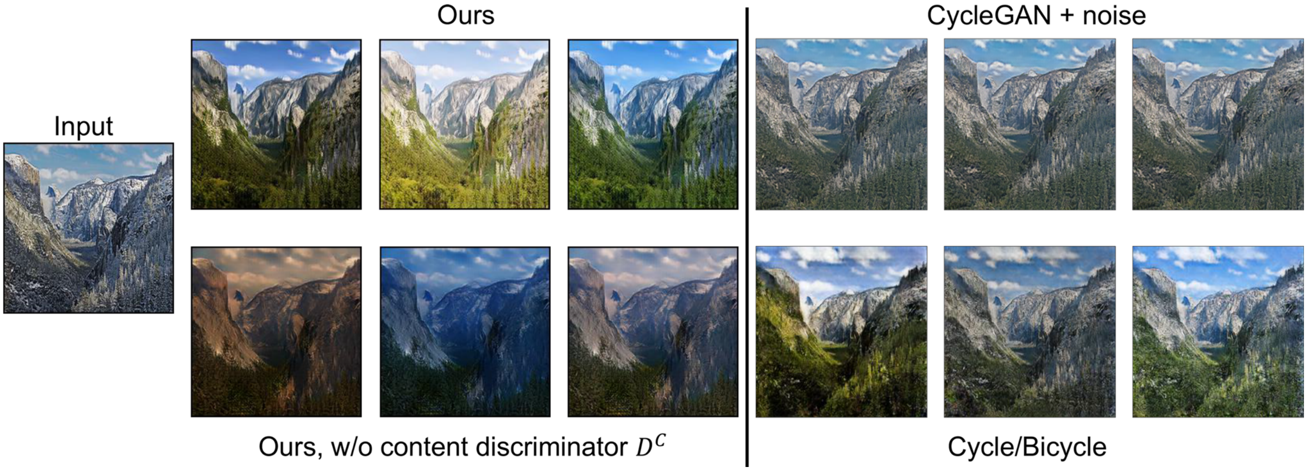


Fig. 7 Baseline artifacts. On the winter → summer translation task, our model produces more diverse and realistic samples over baselines

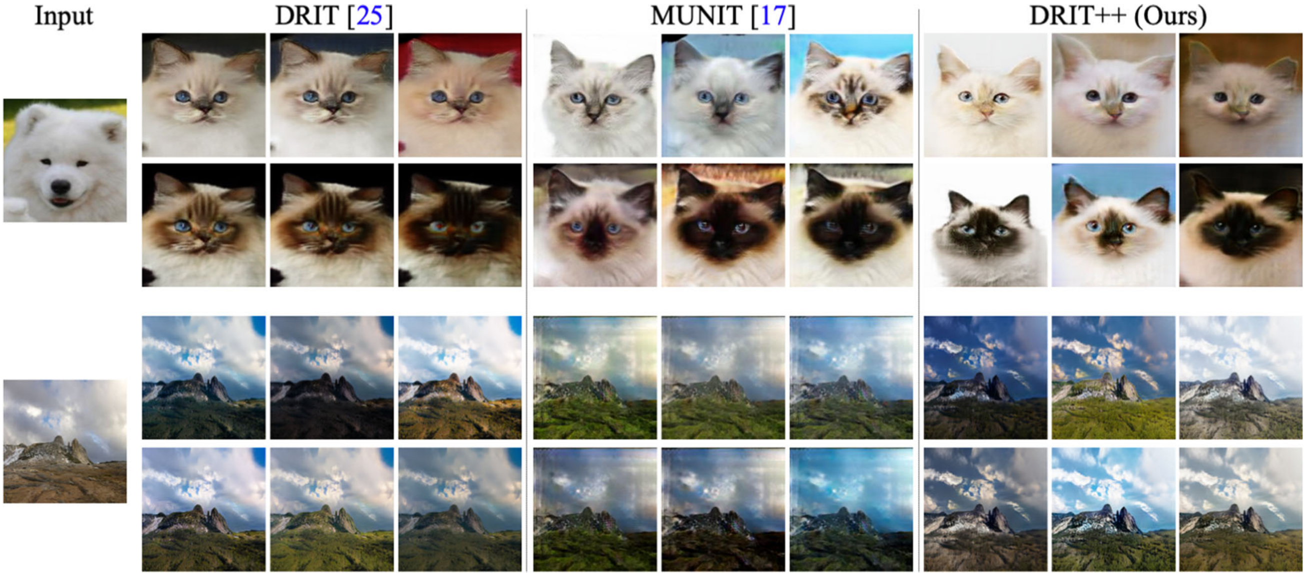


Fig. 8 Effectiveness of mode seeking regularization. Mode seeking regularization helps improve the diversity of translated images while maintaining the visual quality

scheme with a category-specific generator and discriminator under the two-domain setting.

4.1 Qualitative Evaluation

Diversity. We first compare the proposed model with other methods in Fig. 6. In Fig. 7, demonstrate the visual artifacts of images generated by baseline methods. Both our model without D^c and Cycle/Bicycle can generate diverse results. However, the results contain clearly visible artifacts. Without the content discriminator, our model fails to capture domain-specific details (e.g., the color of tree and sky). Therefore, the variations of synthesized images lie in global color differences. As the Cycle/Bicycle methods are trained on pseudo paired data generated by CycleGAN, the quality of the pseudo paired data is not high. As a result, the generated images contain limited diversity.

To better analyze the learned domain-specific attribute space, we perform linear interpolation between two given attributes and generate the corresponding images as shown in Fig. 9. The interpolation results validate the continuity in the attribute space and show that our model can generalize in the distribution, rather than simply retain visual information.

Mode seeking regularization. We demonstrate the effectiveness of the mode seeking regularization term in Fig. 8. The mode seeking regularization term substantially alleviates the mode collapse issue in DRIT (Lee et al. 2018), particularly in the challenging shape-variation translation (i.e., dog-to-cat translation) (Fig. 9).

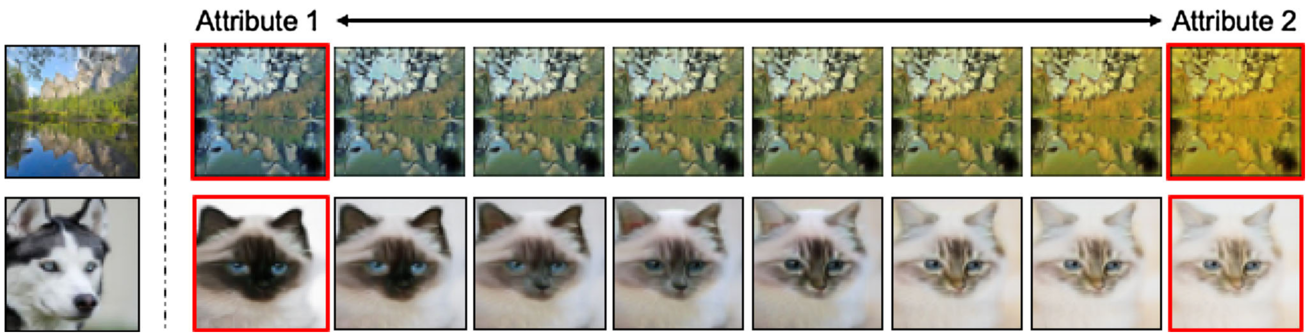
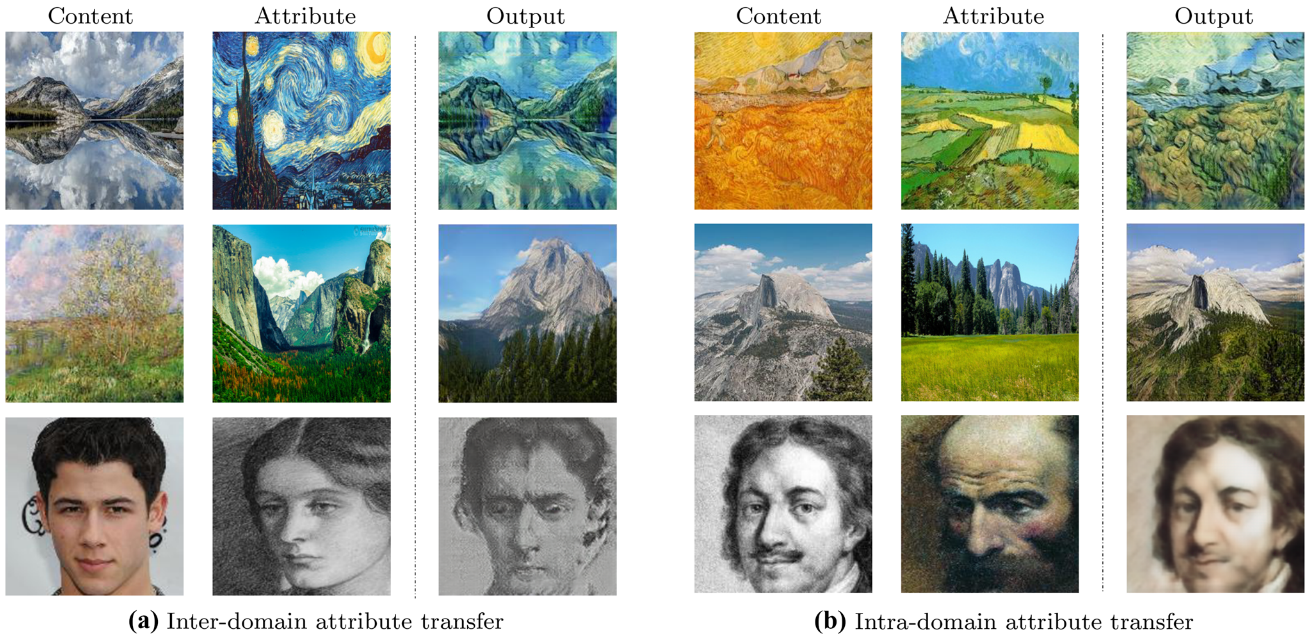


Fig. 9 Linear interpolation between two attribute vectors. Translation results with linear-interpolated attribute vectors between two attributes (highlighted in red) (Color figure online)



(a) Inter-domain attribute transfer

(b) Intra-domain attribute transfer

Fig. 10 Attribute transfer. At test time, in addition to random sampling from the attribute space, we can also perform translation with the query images with the desired attributes. Since the content space is shared across the two domains, we not only can achieve **a** inter-domain, but also **b** intra-domain attribute transfer. Note that we do not explicitly involve intra-domain attribute transfer during training

Attribute transfer. We demonstrate the results of the attribute transfer in Fig. 10. By disentangling content and attribute representations, we are able to perform attribute transfer from images of desired attributes, as illustrated in Fig. 3c. Furthermore, since the content space is shared between two domains, we can generate images conditioned on content features encoded from either domain. Thus our model can achieve not only inter-domain but also intra-domain attribute transfer. Note that intra-domain attribute transfer is not explicitly involved in the training process.

Multi-domain I2I. Figure 11 shows the results of applying the proposed method on the multi-domain I2I. We perform translation among three domains (real images and two artistic styles) and four domains (different weather conditions). Using one single generator, the proposed model is able to perform diverse translation among multiple domains.

4.2 Quantitative Evaluation

Metrics We conduct quantitative evaluations using the following metrics:

- **FID.** To evaluate the quality of the generated images, we use the FID (Heusel et al. 2017) metric to measure the distance between the generated distribution and the real one through features extracted by Inception Network (Szegedy et al. 2015). Lower FID values indicate better quality of the generated images.

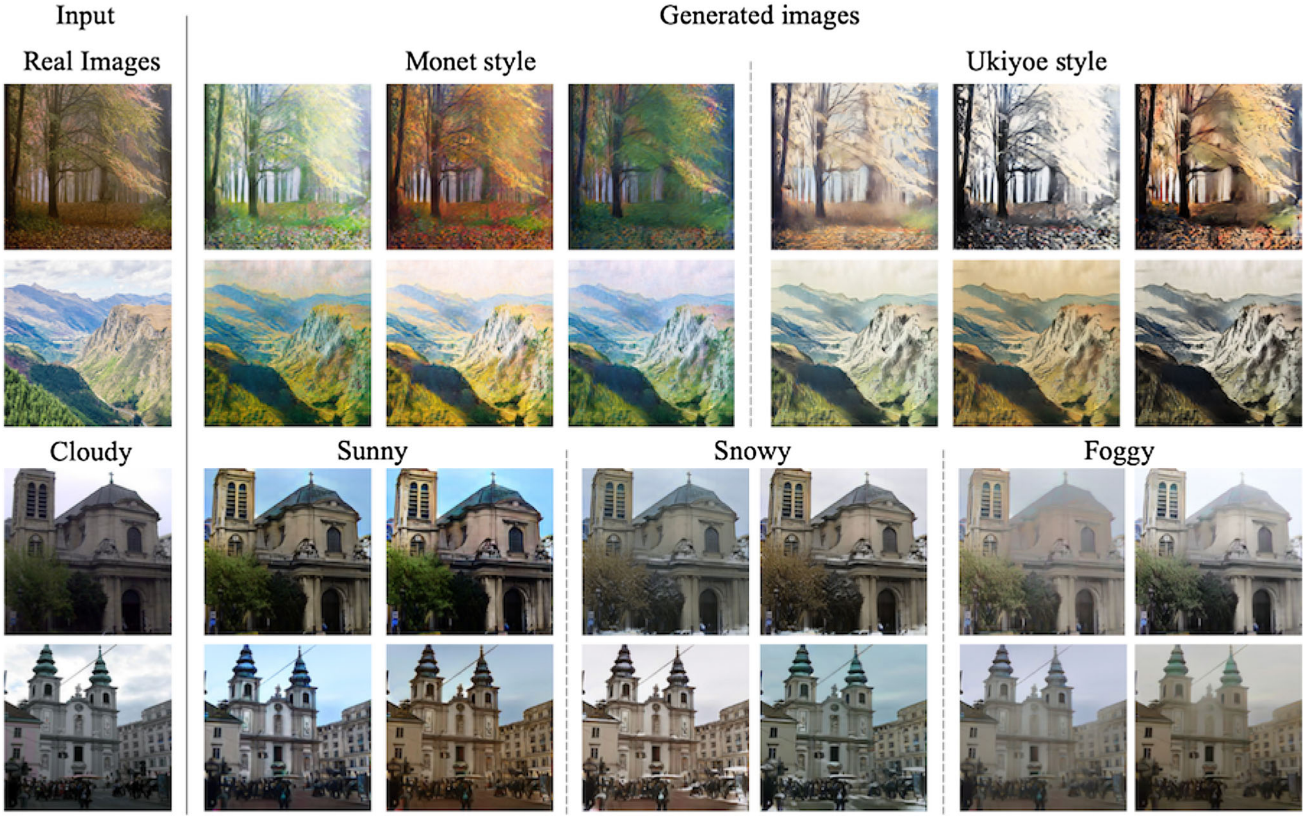


Fig. 11 Multi-domain I2I. We show example results of our model on the multi-domain I2I task. We demonstrate the translation among real images and two artistic styles (Monet and Ukiyoe), and the translation among different weather conditions (sunny, cloudy, snowy, and foggy)

Table 2 Quantitative results of the Yosemite (Summer ⇌ Winter) and the Cat ⇌ Dog dataset

	Cycle/bicycle	DRIT	MUNIT	DRIT++
Datasets	Winter → Summer			
FID ↓	67.04 ± 0.60	41.34 ± 0.20	57.09 ± 0.37	**41.02 ± 0.24**
NDB↓	9.36 ± 0.69	9.38 ± 0.74	9.53 ± 0.64	**9.22 ± 0.97**
JSD↓	0.290 ± 0.086	0.304 ± 0.075	0.293 ± 0.062	**0.222 ± 0.070**
LPIPS↑	0.0974 ± 0.0003	0.0965 ± 0.0004	0.1136 ± 0.0008	**0.1183 ± 0.0007**
Datasets	Cat → Dog			
FID↓	54.008 ± 1.590	24.306 ± 0.329	22.127 ± 0.712	**17.253 ± 0.648**
NDB↓	9.23 ± 0.84	8.16 ± 1.60	8.21 ± 1.17	**7.57 ± 1.25**
JSD↓	0.262 ± 0.072	0.075 ± 0.046	0.132 ± 0.066	**0.041 ± 0.014**
LPIPS↑	0.147 ± 0.001	0.245 ± 0.002	0.244 ± 0.002	**0.280 ± 0.002**

Bold values indicate the best performance in the comparisons

- **LPIPS.** To evaluate diversity, we employ LPIPS (Zhang et al. 2018b) metric to measure the average feature distances between generated samples. Higher LPIPS scores indicate better diversity among the generated images.
- **JSD and NDB.** To measure the similarity between the distribution between real images and generated one, we adopt two bin-based metrics, JSD and NDB (Richardson and Weiss 2018). These metrics evaluate the extent of mode missing of generative models. Similar to Richardson and Weiss (2018), we first cluster the training samples using K-means into different bins. These bins can be viewed as modes of the real data distribution. We then assign each generated sample to the bin of its nearest neighbor. We compute the bin-proportions of the training samples and the synthesized samples to evaluate the difference between the generated distribution and the real data distribution. The NDB and JSD metrics of the bin-proportion are then computed to measure the level of mode collapse. Lower NDB and JSD scores mean the generated data distribution approaches the real data dis-

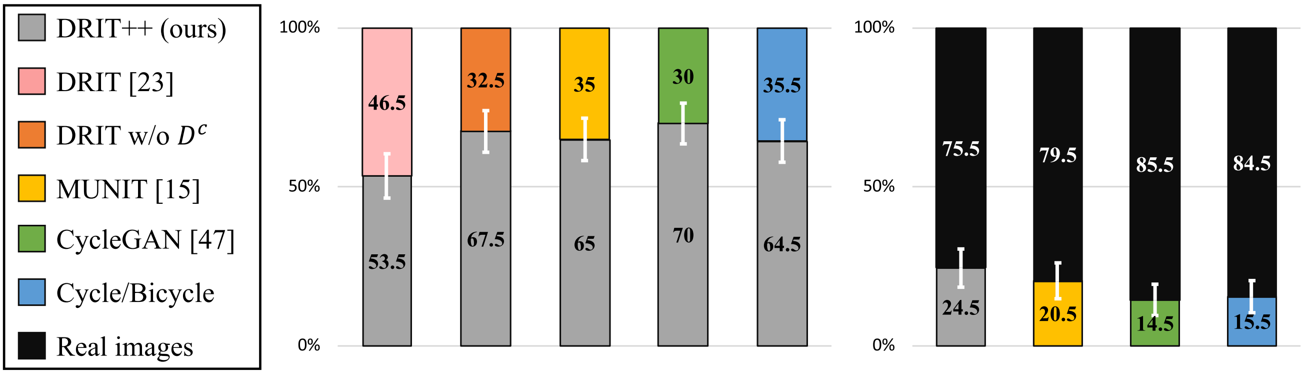


Fig. 12 Realism of synthesized images. We conduct a user study to ask subjects to select results that are *more realistic* through pairwise comparisons. The number indicates the percentage of preference for that comparison pair. We use the winter → summer and the cat → dog translation for this experiment

Table 3 Quantitative results of DRIT++ (multi-domain) on the Yosemite (Summer ⇌ Winter) dataset

	DRIT++ (two-domain)	DRIT++ (multi-domain)
FID ↓	**41.02 ± 0.24**	44.86 ± 0.33
NDB ↓	**9.22 ± 0.97**	9.20 ± 0.88
JSD ↓	**0.222 ± 0.070**	0.254 ± 0.051
LPIPS↑	0.1183 ± 0.0007	**0.1204 ± 0.0004**

Bold values indicate the best performance in the comparisons

Table 4 Average distance between latent content representations of two domains

Dataset	DRIT++	DRIT++ w/o content discriminator D_c
cat2dog	**10.45**	55.45
Yosemite	**31.56**	58.69

Bold values indicate the best performance in the comparisons

tribution better by fitting more modes. More discussions on these metrics can be found in Richardson and Weiss (2018).

- **User preference.** For evaluating realism of synthesized images, we conduct a user study using pairwise comparison. Given a pair of images sampled from real images and translated images generated from various methods, each subject needs to answer the question "Which image is more realistic?"

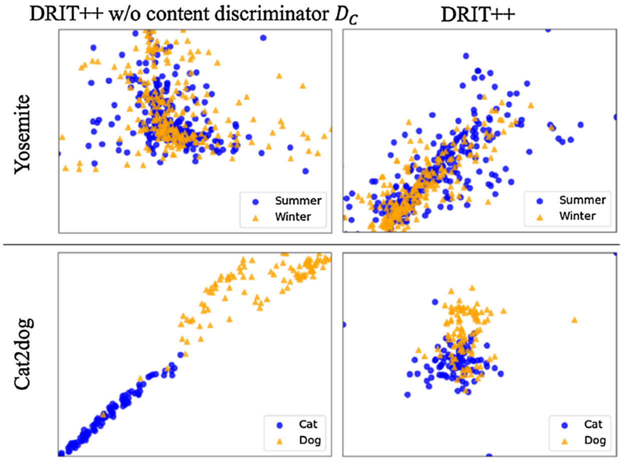


Fig. 13 Visualization of the latent content representations of two domains using t-SNE. Each data point is a content representation encoded from an image of that domain

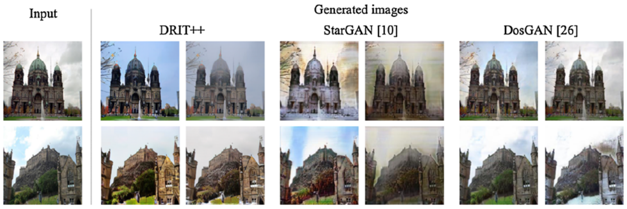


Fig. 14 Comparisons of different multi-domain translation model on the weather dataset

Realism vs. diversity. We conduct the experiment using winter → summer and cat → dog translation with the Yosemite and pets datasets, respectively. Tables 2, 5, and Fig. 12 present the quantitative comparisons with other methods as well as baseline methods. In Table 2, the DRIT++ method performs well on all metrics. The DRIT++ method generates images that are not only realistic, but also diverse and close to the original data distribution. Table 5 validates the effectiveness of the content discriminator, latent regression loss, and mode-seeking regularization in the proposed algorithm. Figure 12 shows the results of user study. The DRIT++ algorithm performs favorably against the state-of-the-art approaches as well as baseline methods.

Multi-domain translation We compare the performance of DRIT++, StarGAN (Choi et al. 2018), and DosGAN (Lin et al. 2018a) in terms of realism on the weather dataset. For each trial, We translate 1000 testing images to one of four domains and measure the visual quality (in terms of FID) and

Table 5 Ablation study

	DRIT w/o D^c	DRIT w/o KL	DRIT w/o L_1^{latent}	DRIT	DRIT++
FID ↓	46.92 ± 0.35	$\mathbf{40.08 \pm 0.33}$	53.12 ± 0.16	41.34 ± 0.20	41.02 ± 0.24
NDB↓	9.36 ± 0.72	9.47 ± 0.70	9.97 ± 0.17	9.38 ± 0.74	$\mathbf{9.22 \pm 0.97}$
JSD↓	0.277 ± 0.077	0.289 ± 0.066	0.494 ± 0.045	0.304 ± 0.075	$\mathbf{0.222 \pm 0.070}$
LPIPS↑	0.0954 ± 0.0006	0.0957 ± 0.0007	0.0158 ± 0.0003	0.0965 ± 0.0004	$\mathbf{0.1183 \pm 0.0007}$

Bold values indicate the best performance in the comparisons

We demonstrate the effect of content discriminator, latent regression loss, and mode-seeking regularization in the proposed algorithm

Table 6 Multi-domain translation comparison

	DRIT++	StarGAN	DosGAN
FID↓	$\mathbf{61.51 \pm 3.11}$	82.38 ± 3.91	67.98 ± 2.38
LPIPS↑	0.676 ± 0.008	$\mathbf{0.692 \pm 0.010}$	0.650 ± 0.005

Bold values indicate the best performance in the comparisons

We compare the visual quality and diversity of DRIT++ (multi-domian) with two multi-domain translation model on the weather dataset. The results are averaged after 5 trials. StarGAN gets highest score on LPIPS due to its lower visual quality

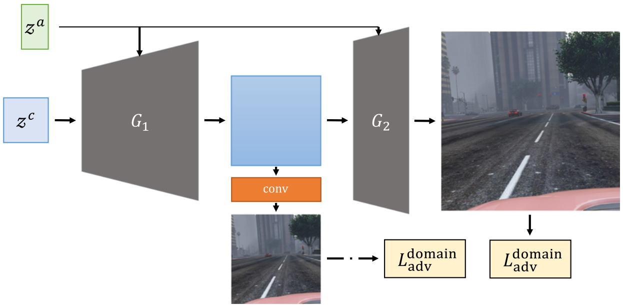
diversity (using the LPIPS metric). We report the averaged results of 5 trials. Table 6 shows that the disentangled representations by our method not only enable diverse translation, but also improve the quality of generated images. Figure 14 presents qualitative results by the evaluated methods.

Multi-domain model on two-domain translation Two-domain translation is a special case of multi-domain translation problems. We conduct an experiment under the same settings described in Table 2 and 3. As shown in Table 3, our multi-domain model performs well in all metrics against the two-domain translation model that consists of the domain-specific generator and discriminator.

Ablation study on the content discriminator In practice, the content discriminator helps align distributions of the latent content representations of two domains. We conduct experiments on both cat2dog and the Yosemite datasets to illustrate this. The distance between the means of the content representations from two domains is measured by:

$$D = \left\| \frac{1}{N_A^{\text{test}}} \sum_{i=1}^{N_A^{\text{test}}} f_A^{\text{content}} - \frac{1}{N_B^{\text{test}}} \sum_{i=1}^{N_B^{\text{test}}} f_B^{\text{content}} \right\|_1^1 . \qquad (15)$$

Table 4 shows the quantitative results. Furthermore, Fig. 13 visualizes the distributions of the latent content representations from two domains using t-SNE. The distance (15) between the content representations of the two domains is much smaller with the help of the content discriminator (Fig. 13, Table 4).



z^a
z^c
G_1
G_2
conv
$L_{\text{adv}}^{\text{domain}}$
$L_{\text{adv}}^{\text{domain}}$

Fig. 15 Multi-scale generator-discriminator. To enhance the quality of generated high-resolution images, we adopt a multi-scale generator-discriminator architecture. We generate low-resolution images from the intermediate features of the generator. An additional adversarial domain loss is applied on the low-resolution images

Table 7 Ablation study on multi-scale generator-discriminator architecture

	DRIT++ w/ 2 more layers	DRIT++ (high-resolution)
FID ↓	37.19 ± 0.21	$\mathbf{28.62 \pm 0.38}$
LPIPS ↑	0.616 ± 0.038	$\mathbf{0.621 \pm 0.004}$

Bold values indicate the best performance in the comparisons

We improvement using two more layers in the multi-scale architecture

4.3 High Resolution I2I

We demonstrate that the proposed scheme can be applied to the translation tasks with high-resolution images. We perform image translation on the street scene [GTA (Richter et al. 2016) ↔ Cityscape (Cordts et al. 2016)] dataset. The size of the input image is 720 × 360 pixels. During the training, we randomly crop the image to the size of 340 × 340 for memory efficiency consideration. To enhance the quality of the generated high-resolution images, we adopt a multi-scale generator-discriminator structure similar to the StackGAN (Zhang et al. 2018a) scheme. As shown in Fig. 15, we extract the intermediate feature of the generator and pass through a convolutional layer to generate low-resolution images. We utilize an additional discriminator which takes low-resolution images as input. This discriminator enforces the first few layers of the generator to capture the distribution of low-level variations such as colors and image

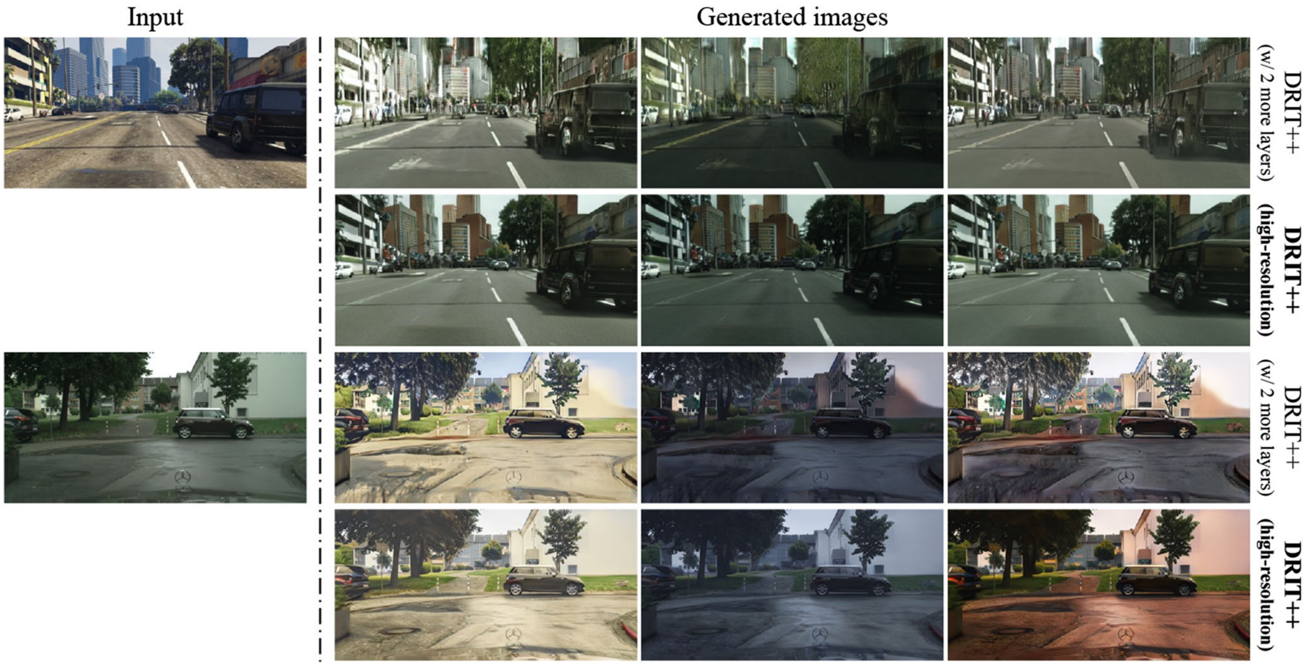


Fig. 16 High-resolution translations. We show sample results produced by our model with multi-scale generator-discriminator architecture. The mappings from top to bottom are: GTA → Cityscape, Cityscape → GTA

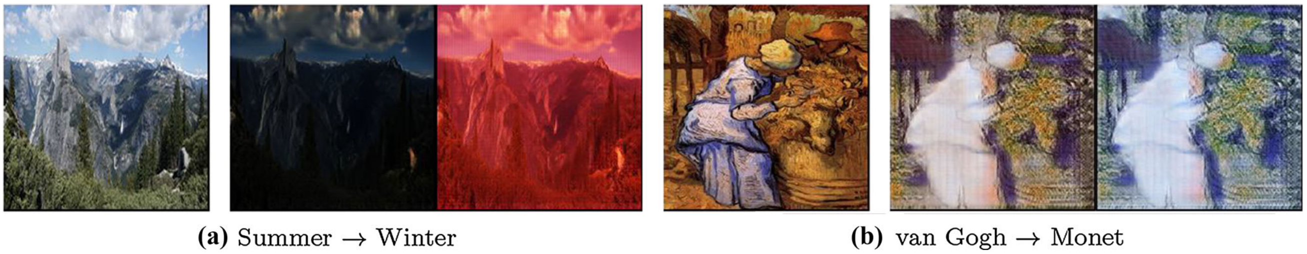


Fig. 17 Failure examples. Typical cases: **a** attribute space not fully exploited. **b** Distribution characteristic difference

structures. We find such multi-scale generator-discriminator structure facilitate the training and yields more realistic images on high-resolution translation task. To validate the effectiveness of the multi-scale architecture, we show the comparison between (1) adding two more layers to generators and (2) using the multi-scale generator-discriminator architecture in Table 7 and Fig. 16. We report the FID and LPIPS scores of the generated images by the two methods on the GTA5 → Cityscape translation task. As shown in Table 7, using the multi-scale architecture we can generate more photo-realistic images on the translation task with high-resolution images.

4.4 Limitations

The performance of the proposed algorithm is limited in several aspects. First, due to the limited amount of training data, the attribute space is not fully exploited. Our I2I translation fails when the sampled attribute vectors locate in under-sampled space, see Fig. 17a. Second, it remains difficult when the domain characteristics differ significantly. For example, Fig. 17b shows a failure case on the human figure due to the lack of human-related portraits in Monet collections. Third, we use multiple encoders and decoders for the cross-cycle consistency during training, which requires large memory usage. The memory usage limits the application on high-resolution image-to-image translation.

5 Conclusions

In this paper, we present a novel disentangled representation framework for diverse image-to-image translation with unpaired data. we propose to disentangle the latent space to a content space that encodes common information between domains, and a domain-specific attribute space that can model the diverse variations given the same content. We apply a content discriminator to facilitate the representation disentanglement. We propose a cross-cycle consistency loss for cyclic reconstruction to train in the absence of paired data. Qualitative and quantitative results show that the proposed model produces realistic and diverse images.

Acknowledgements This work is supported in part by the NSF CAREER Grant #1149783, the NSF Grant #1755785, and gifts from Verisk, Adobe and Google.

References

AlBahar, B., & Huang, J. B. (2019). Guided image-to-image translation with bi-directional feature transformation. In *ICCV*.

Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., & Courville, A. (2018). Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. In *ICML*.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*.

Cao, J., Katzir, O., Jiang, P., Lischinski, D., Cohen-Or, D., Tu, C., et al. (2018). Dida: Disentangled synthesis for domain adaptation. arXiv preprint arXiv:1805.08019.

Chen, Q., & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. In *ICCV*.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*.

Chen, Y. C., Lin, Y. Y., Yang, M. H., & Huang, J. B. (2019). Crdoco: Pixel-level domain transfer with cross-domain consistency. In *CVPR*.

Cheung, B., Livezey, J. A., Bansal, A. K., & Olshausen, B. A. (2015). Discovering hidden factors of variation in deep networks. In *ICLR workshop*.

Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR* (Vol. 1711).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *CVPR*.

Denton, E. L., & Birodkar, V. (2017). Unsupervised learning of disentangled representations from video. In *NIPS*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *NIPS*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*.

Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., et al. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *ECCV*.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *CVPR*.

Kim, T., Cha, M., Kim, H., Lee, J., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *ICML*.

Kinga, D., & Adam, J. B. (2015). A method for stochastic optimization. In *ICLR*.

Kingma, D. P., Rezende, D., Mohamed, S. J., & Welling, M. (2014). Semi-supervised learning with deep generative models. In *NIPS*.

Lai, W. S., Huang, J. B., Ahuja, N., & Yang, M. H. (2017). Deep laplacian pyramid networks for fast and accurate superresolution. In *CVPR*.

Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *ECCV*.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

Lee, H. Y., Tseng, H. Y., Huang, J. B., Singh, M. K., & Yang, M. H. (2018) Diverse image-to-image translation via disentangled representations. In *ECCV*.

Lee, H. Y., Yang, X., Liu, M. Y., Wang, T. C., Lu, Y. D., Yang, M. H., et al. (2019). Dancing to music. In *NeurIPS*.

Li, Y., Huang, J. B., Ahuja, N., & Yang, M. H. (2016). Deep joint image filtering. In *ECCV*.

Li, Y., Huang, J. B., Ahuja, N., & Yang, M. H. (2019). Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(8), 1909–1923.

Lin, J., Xia, Y., Liu, S., Qin, T., Chen, Z., & Luo, J. (2018a). Exploring explicit domain supervision for latentspace disentanglement in unpaired image-to-image translation. arXiv preprint arXiv:1902.03782.

Lin, J., Xia, Y., Qin, T., Chen, Z., & Liu, T. Y. (2018b). Conditional image-to-image translation. In *CVPR*.

Liu, A., Liu, Y. C., & Wang, F. Y. C. (2018). A unified feature disentangler for multi-domain image translation and manipulation. In *NIPS*.

Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. In *NIPS*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *ICCV*.

Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., & Van Gool, L. (2018). Exemplar guided unsupervised image-to-image translation. In *ICLR*.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L. (2017). Pose guided person image generation. In *NIPS*.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2016). Adversarial autoencoders. In *ICLR workshop*.

Mao, Q., Lee, H. Y., Tseng, H. Y., Ma, S., & Yang, M. H. (2019). Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*.

Mathieu, M., Zhao, J., Sprechmann, P., Ramesh, A., & LeCun, Y. (2016). Disentangling factors of variation in deep representation using adversarial training. In *NIPS*.

Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., & Kim, K. (2018). Image to image translation for domain adaptation. In *CVPR*.

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in pytorch. In *NIPS workshop*.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In *ICML*.

Richardson, E., & Weiss, Y. (2018). On GANs and GMMs. In *NIPS*.

Richter, S. R., Vineet, V., Roth, S., & Koltun, V. (2016). Playing for data: Ground truth from computer games. In *ECCV*.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *CVPR*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *CVPR*.

Taigman, Y., Polyak, A., & Wolf, L. (2017). Unsupervised cross-domain image generation. In *ICLR*.

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. In *NIPS*.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*.

Yi, Z., Zhang, H. R., Tan, P., & Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., et al. (2018a). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. In *TPAMI*.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *ECCV*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018b). The unreasonable effectiveness of deep networks as a perceptual metric. In *CVPR*.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.

Zhu, J. Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., et al. (2017b). Toward multimodal image-to-image translation. In *NIPS*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.