Diverse Image-to-Image Translation via Disentangled Representations

Hsin-Ying Lee*[1], Hung-Yu Tseng*[1], Jia-Bin Huang[2], Maneesh Singh[3], Ming-Hsuan Yang[1,4]

[1]University of California, Merced [2]Virginia Tech [3]Verisk Analytics [4]Google Cloud

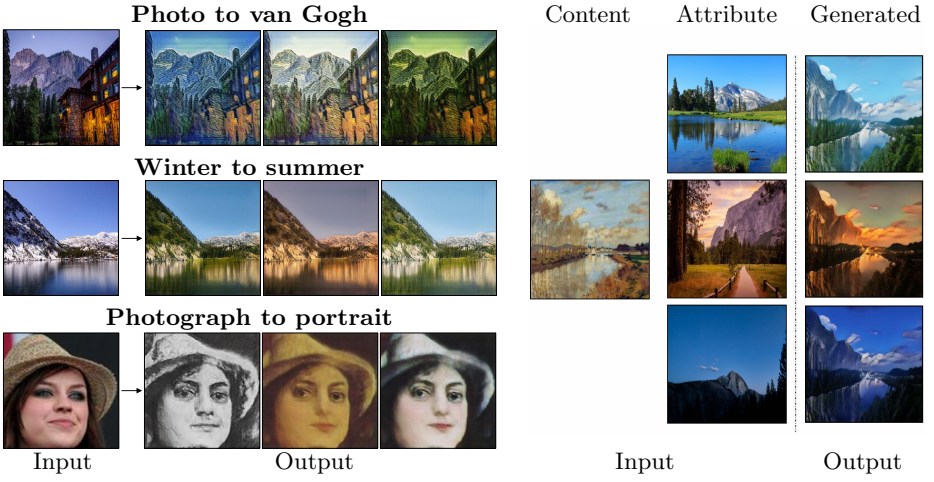
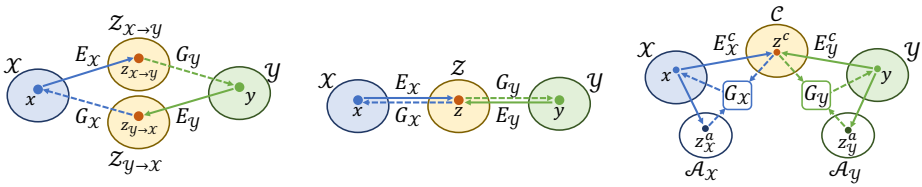


Fig. 1: **Unpaired diverse image-to-image translation.** (*Left*) Our model learns to perform diverse translation between two collections of images without aligned training pairs. (*Right*) Example-guided translation.

Abstract. Image-to-image translation aims to learn the mapping between two visual domains. There are two main challenges for many applications: 1) the 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 producing diverse outputs without paired training images. To achieve diversity, 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 the attribute vectors sampled from the attribute space to produce diverse outputs at test time. To handle unpaired training data, we introduce a novel 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 comparisons, we measure realism with user study and diversity with a perceptual distance metric. We apply the proposed model to domain adaptation and show competitive performance when compared to the state-of-the-art on the MNIST-M and the LineMod datasets.

* equal contribution



(a) CycleGAN [48] (b) UNIT [27] (c) Ours

Fig. 2: **Comparisons of unsupervised I2I translation methods.** Denote x and y as images in domain $\mathcal{X}$ and $\mathcal{Y}$: (a) CycleGAN [48] maps x and y onto *separated* latent spaces. (b) UNIT [27] 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.

1 Introduction

Image-to-Image (I2I) translation aims to learn the mapping between different visual domains. Many vision and graphics problems can be formulated as I2I translation problems, such as colorization [23,46] (grayscale → color), super-resolution [25,22,26] (low-resolution → high-resolution), and photorealistic image synthesis [6,42] (label → image). Furthermore, I2I translation has recently shown promising results in facilitating domain adaptation [3,36,16,32].

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 a random noise vector to the generator for modeling the 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. Pix2pix [18] applies conditional generative adversarial network to I2I translation problems. Nevertheless, the training process requires paired data. A number of recent work [48,27,44,38,9] relaxes the dependency on paired training data for learning I2I translation. These methods, however, produce a single output conditioned on the given input image. As shown in [18,49], simply incorporating noise vectors as additional inputs to the generator does not lead the increased variations of the generated outputs due to the mode collapsing issue. The generators in these methods are inclined to overlook the added noise vectors. Very recently, BicycleGAN [49] tackles the problem of generating diverse outputs in I2I problems 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. Specifically, we propose to embed images onto two spaces: 1) a domain-invariant content space and 2) a domain-specific attribute space as shown in Figure 2. Our generator learns to perform I2I translation conditioned on content features and a latent at-

Table 1: **Feature-by-feature comparison of image-to-image translation networks.** Our model achieves multimodal translation without using aligned training image pairs.

Method	Pix2Pix [18]	CycleGAN [48]	UNIT [27]	BicycleGAN [49]	Ours
Unpaired	-	✓	✓	-	✓
Multimodal	-	-	-	✓	✓

tribute vector. The domain-specific attribute space aims to model the variations within a domain given the same content, while the domain-invariant content space captures information across domains. We achieve this representation disentanglement 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 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. 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 the two testing modes.

We evaluate the proposed model through extensive qualitative and quantitative evaluation. In a wide variety of I2I tasks, we show diverse translation results with randomly sampled attribute vectors and example-guided translation with transferred attribute vectors from existing images. We evaluate the realism of our results with a user study and the diversity using perceptual distance metrics [47]. Furthermore, we demonstrate the potential application of unsupervised domain adaptation. On the tasks of adapting domains from MNIST [24] to MNIST-M [12] and Synthetic Cropped LineMod to Cropped LineMod [15,43], we show competitive performance against state-of-the-art domain adaptation methods.

We make the following contributions:

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 compares favorably against existing I2I models. Images generated by our model are both diverse and realistic.

3) We demonstrate the application of our model on unsupervised domain adaptation. We achieve competitive results on both the MNIST-M and the Cropped LineMod datasets.

Our code, data and more results are available at https://github.com/HsinYingLee/DRIT/.

2 Related Work

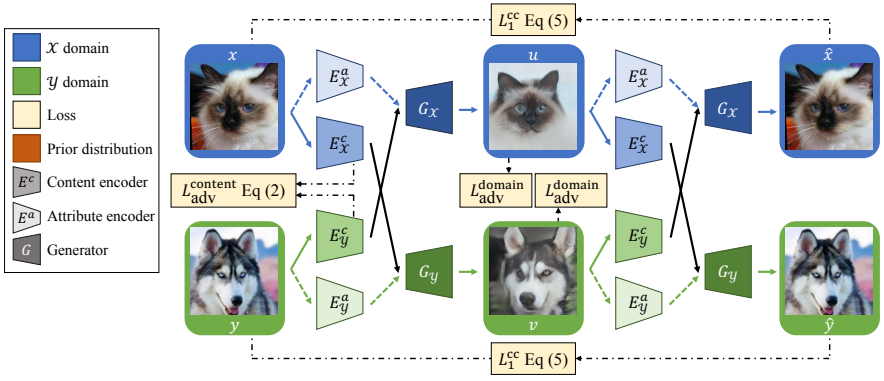
Generative adversarial networks. Recent years have witnessed rapid progress on generative adversarial networks (GANs) [14,34,2] for image generation. The core idea of GANs lies in the adversarial loss that enforces the distribution of generated images to match that of the target domain. The generators in GANs can map from noise vectors to realistic images. Several recent efforts explore *conditional* GAN in various contexts including conditioned on text [35], low-resolution images [25], video frames [41], and image [18]. Our work focuses on using GAN conditioned on an input image. In contrast to several existing conditional GAN frameworks that require paired training data, our model produces diverse outputs without paired data. This suggests that 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. Pix2pix [18] 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, CycleGAN [48], DiscoGAN [19], and UNIT [27] 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, BicycleGAN [49] 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 data. Table 1 shows a feature-by-feature comparison among various I2I models. Unlike existing work, our method enables I2I translation with diverse outputs in the absence of paired training data.

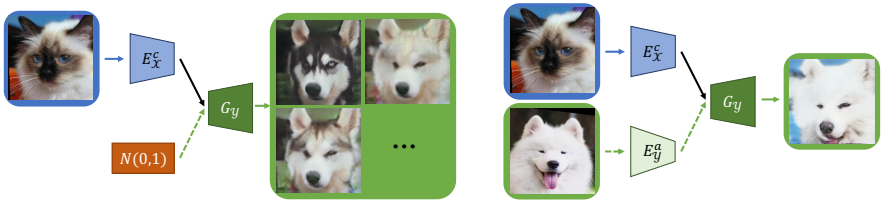
Very recently, several concurrent works [1,17,5,29] (all independently developed) also adopt a disentangled representation similar to our work for learning diverse I2I translation from unpaired training data. We encourage the readers to review these works for a complete picture.

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 [8,21,30,31]. Recently, the unsupervised setting has been explored [7,10]. InfoGAN [7] achieves disentanglement by maximizing the mutual information between latent variables and data variation. Similar to DrNet [10] 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.

Domain adaptation. Domain adaptation techniques focus on addressing the domain-shift problem between a source and a target domain. Domain Adversarial Neural Network (DANN) [11,13] and its variants [40,4,39] tackle domain



(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 $L_{\text{adv}}^{\text{content}}$ (Section 3.1) and the cross-cycle consistency loss L_1^{cc} (Section 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.

adaptation through learning domain-invariant features. Sun et al. [37] aims to map features in the source domain to those in the target domain. I2I translation has been recently applied to produce simulated images in the target domain by translating images from the source domain [11,16]. Different from the aforementioned I2I based domain adaptation algorithms, our method does not utilize source domain annotations for I2I translation.

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 Figure 3, our framework consists of content encoders $\{E_{\mathcal{X}}^c, E_{\mathcal{Y}}^c\}$, attribute encoders $\{E_{\mathcal{X}}^a, E_{\mathcal{Y}}^a\}$, generators $\{G_{\mathcal{X}}, G_{\mathcal{Y}}\}$, and domain discriminators $\{D_{\mathcal{X}}, D_{\mathcal{Y}}\}$ for both domains, and a content discriminators D_{adv}^c. Take domain $\mathcal{X}$ as an example, the content encoder $E_{\mathcal{X}}^c$ maps images onto a shared, domain-invariant content space ($E_{\mathcal{X}}^c : \mathcal{X} \to \mathcal{C}$) and the attribute encoder $E_{\mathcal{X}}^a$ maps images onto a domain-specific attribute space ($E_{\mathcal{X}}^a : \mathcal{X} \to \mathcal{A}_{\mathcal{X}}$). The generator $G_{\mathcal{X}}$ generates images conditioned on both content and attribute vectors ($G_{\mathcal{X}} : \{\mathcal{C}, \mathcal{A}_{\mathcal{X}}\} \to \mathcal{X}$). The

discriminator $D_{\mathcal{X}}$ aims to discriminate between real images and translated images in the domain $\mathcal{X}$. Content discriminator D^c is trained to distinguish the extracted content representations between two domains. To enable multimodal generation at test time, we regularize the attribute vectors so that they can be drawn from a prior Gaussian distribution $N(0, 1)$.

In this section, we first discuss the strategies used to disentangle the content and attribute representations in Section 3.1 and then introduce the proposed cross-cycle consistency loss that enables the training on unpaired data in Section 3.2. Finally, we detail the loss functions in Section 3.3.

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 $\mathcal{A}_{\mathcal{Y}}$.

$$\begin{aligned} \{z_x^c, z_x^a\} &= \{E_{\mathcal{X}}^c(x), E_{\mathcal{X}}^a(x)\} \qquad z_x^c \in \mathcal{C}, z_x^a \in \mathcal{A}_{\mathcal{X}} \\ \{z_y^c, z_y^a\} &= \{E_{\mathcal{Y}}^c(y), E_{\mathcal{Y}}^a(y)\} \qquad z_y^c \in \mathcal{C}, z_y^a \in \mathcal{A}_{\mathcal{Y}} \end{aligned} \tag{1}$$

To achieve representation disentanglement, we apply two strategies: weight-sharing and a content discriminator. First, similar to [27], based on the assumption that two domains share a common latent space, we share the weight between the last layer of $E_{\mathcal{X}}^c$ and $E_{\mathcal{Y}}^c$ and the first layer of $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$. Through weight sharing, we force the content representation to be mapped onto the same space. However, sharing the same high-level mapping functions cannot guarantee the same content representations encode the same information for both domains. Therefore, we propose a content discriminator D^c which aims to distinguish the domain membership of the encoded content features z_x^c and z_y^c. 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:

$$\begin{aligned} L_{\text{adv}}^{\text{content}}(E_{\mathcal{X}}^c, E_{\mathcal{Y}}^c, D^c) = \mathbb{E}_x[\frac{1}{2}\log D^c(E_{\mathcal{X}}^c(x)) + \frac{1}{2}\log\left(1 - D^c(E_{\mathcal{X}}^c(x))\right)] \\ + \mathbb{E}_y[\frac{1}{2}\log D^c(E_{\mathcal{Y}}^c(y)) + \frac{1}{2}\log\left(1 - D^c(E_{\mathcal{Y}}^c(y))\right)] \end{aligned} \tag{2}$$

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 [48] (i.e., $\mathcal{X} \to \mathcal{Y} \to \mathcal{X}$) 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.

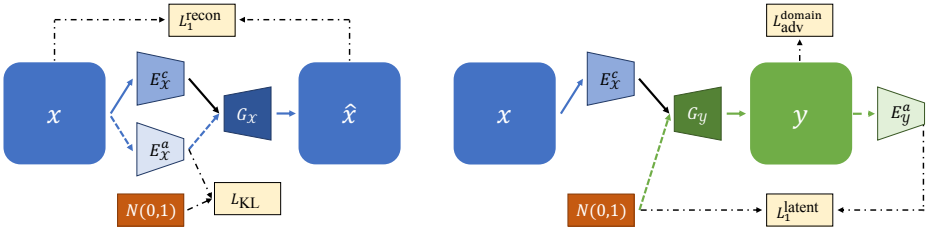


Fig. 4: **Loss functions.** In addition to the cross-cycle reconstruction loss L_1^{cc} and the content adversarial loss $L_{\text{adv}}^{\text{content}}$ described in Figure 3, we apply several additional loss functions in our training process. The self-reconstruction loss L_1^{recon} facilitates training 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. More details can be found in Section 3.3.

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 $\{z_x^c, z_x^a\}$ and $\{z_y^c, z_y^a\}$. We then perform the first translation by swapping the attribute representation (i.e., z_x^a and z_y^a) to generate $\{u, v\}$, where $u \in \mathcal{X}, v \in \mathcal{Y}$.

$$u = G_{\mathcal{X}}(z_y^c, z_x^a) \quad v = G_{\mathcal{Y}}(z_x^c, z_y^a) \tag{3}$$

Backward translation. After encoding u and v into $\{z_u^c, z_u^a\}$ and $\{z_v^c, z_v^a\}$, we perform the second translation by once again swapping the attribute representation (i.e., z_u^a and z_v^a).

$$\hat{x} = G_{\mathcal{X}}(z_v^c, z_u^a) \quad \hat{y} = G_{\mathcal{Y}}(z_u^c, z_v^a) \tag{4}$$

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

$$\begin{aligned} L_1^{\text{cc}}(G_{\mathcal{X}}, G_{\mathcal{Y}}, E_{\mathcal{X}}^c, E_{\mathcal{Y}}^c, E_{\mathcal{X}}^a, E_{\mathcal{Y}}^a) = \mathbb{E}_{x,y}[\|G_{\mathcal{X}}(E_{\mathcal{Y}}^c(v), E_{\mathcal{X}}^a(u)) - x\|_1 \\ + \|G_{\mathcal{Y}}(E_{\mathcal{X}}^c(u), E_{\mathcal{Y}}^a(v)) - y\|_1], \end{aligned} \tag{5}$$

where $u = G_{\mathcal{X}}(E_{\mathcal{Y}}^c(y)), E_{\mathcal{X}}^a(x))$ and $v = G_{\mathcal{Y}}(E_{\mathcal{X}}^c(x)), E_{\mathcal{Y}}^a(y))$.

3.3 Other Loss Functions

Other than 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 Figure 4. Starting from the top-right, in the counter-clockwise order:
Domain adversarial loss. We impose adversarial loss $L_{\text{adv}}^{\text{domain}}$ where $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ attempt to discriminate between real images and generated images in each domain, while $G_{\mathcal{X}}$ and $G_{\mathcal{Y}}$ attempt to generate realistic images.

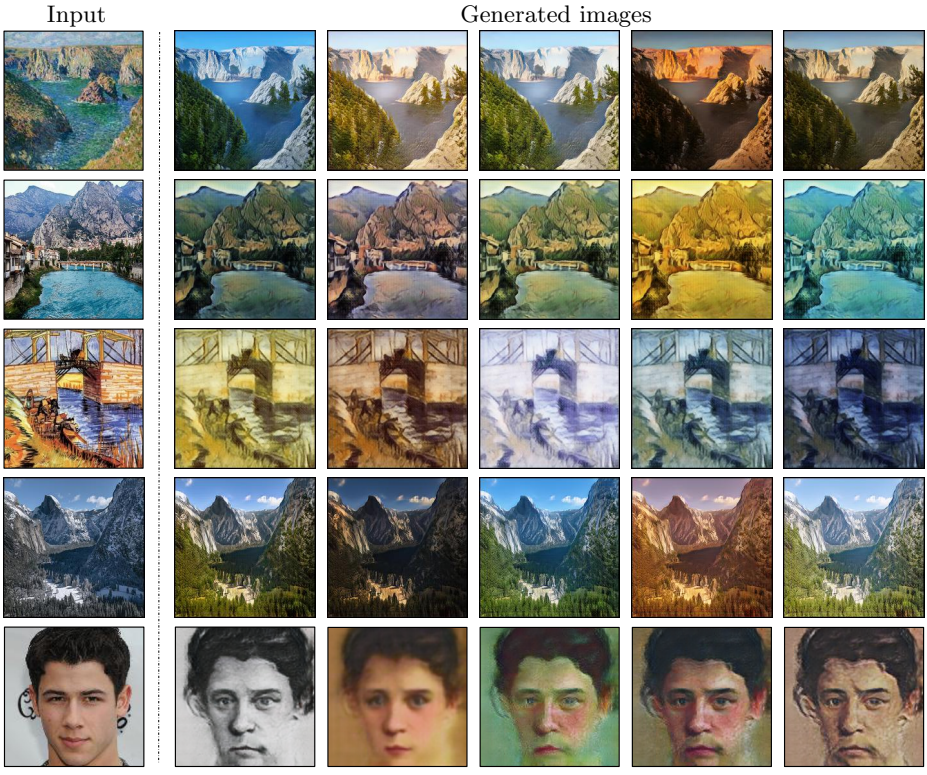


Fig. 5: **Sample results.** We show example results produced by our model. The left column shows the input images in the source domain. The other five columns show the output images generated by sampling random vectors in the attribute space. The mappings from top to bottom are: Monet → photo, photo → van Gogh, van Gogh → Monet, winter → summer, and photograph → portrait.

Self-reconstruction loss. In addition to the cross-cycle reconstruction, we apply a self-reconstruction loss L_1^{rec} to facilitate the training. With encoded content/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 $\hat{y} = G_{\mathcal{Y}}(E_{\mathcal{Y}}^c(y), E_{\mathcal{Y}}^a(y))$.

KL loss. In order to perform stochastic sampling at test time, we encourage the attribute representation to be as close to a prior Gaussian distribution. We thus apply the loss $L_{\text{KL}} = \mathbb{E}[D_{\text{KL}}((z_a)\|N(0,1))]$, where $D_{\text{KL}}(p\|q) = -\int p(z) \log \frac{p(z)}{q(z)} \mathrm{d}z$.

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 [49]. 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:

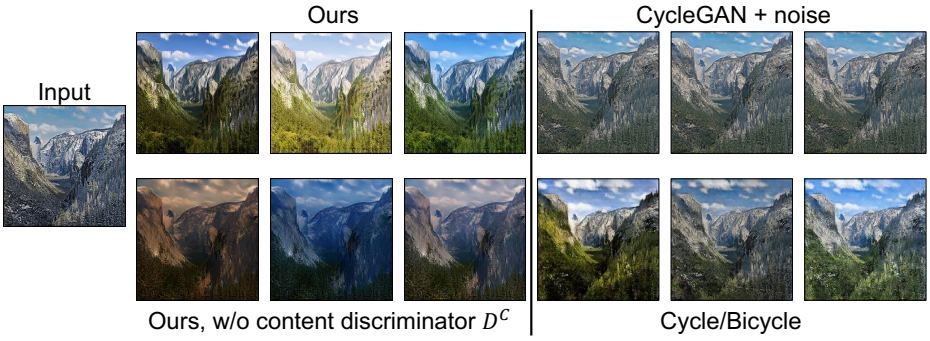


Fig. 6: **Diversity comparison.** On the winter → summer translation task, our model produces more diverse and realistic samples over baselines.

Fig. 7: **Linear interpolation between two attribute vectors.** Translation results with linear-interpolated attribute vectors between two attributes (highlighted in red).

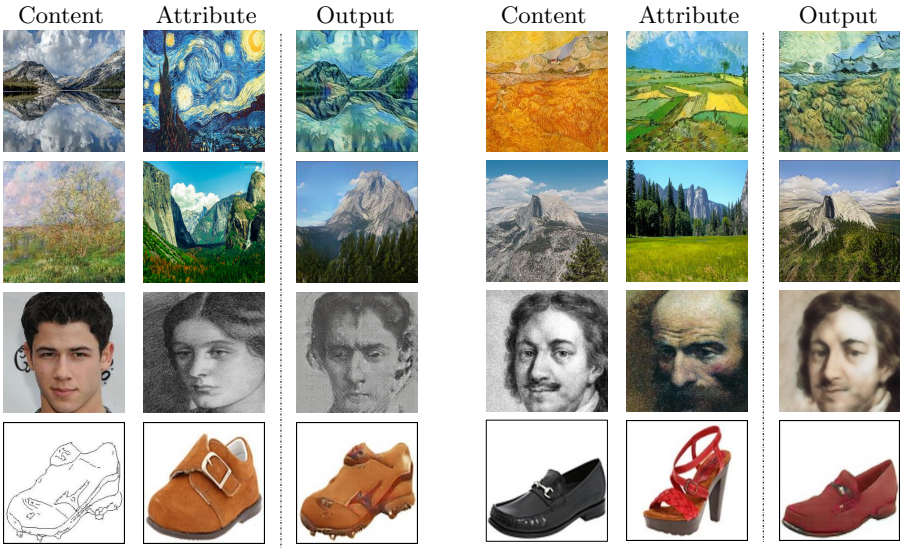
$$\min_{G,E^c,E^a} \max_{D,D^c} \lambda_{\text{adv}}^{\text{content}} L_{\text{adv}}^{\text{c}} + \lambda_1^{\text{cc}} L_1^{\text{cc}} + \lambda_{\text{adv}}^{\text{domain}} L_{\text{adv}}^{\text{domain}} + \lambda_1^{\text{recon}} L_1^{\text{recon}} + \lambda_1^{\text{latent}} L_1^{\text{latent}} + \lambda_{\text{KL}} L_{\text{KL}} \tag{6}$$

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

4 Experimental Results

Implementation details. We implement our model with PyTorch [33]. We use the input image size of 216×216 for all of our experiments except domain adaptation. 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 containing four residual blocks followed by three fractionally strided convolution layers. For more details of architecture design, please refer to the supplementary material.

For training, we use the Adam optimizer [20] 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: $\lambda_{\text{adv}}^{\text{content}} = 1$, $\lambda_{\text{cc}} = 10$, $\lambda_{\text{adv}}^{\text{domain}} = 1$, $\lambda_1^{\text{rec}} = 10$, $\lambda_1^{\text{latent}} = 10$, 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 [34] for training the model with adversarial loss.



(a) Inter-domain attribute transfer (b) Intra-domain attribute transfer

Fig. 8: **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.

Datasets. We evaluate our model on several datasets include Yosemite [48] (summer and winter scenes), artworks [48] (Monet and van Gogh), edge-to-shoes [45] and photo-to-portrait cropped from subsets of the WikiArt dataset [1] and the CelebA dataset [28]. We also perform domain adaptation on the classification task with MNIST [24] to MNIST-M [12], and on the classification and pose estimation tasks with Synthetic Cropped LineMod to Cropped LineMod [15,43].

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

- **DRIT:** We refer to our proposed model, Disentangled Representation for Image-to-Image Translation, as DRIT.
- **DRIT w/o D^c:** Our proposed model without the content discriminator.
- **CycleGAN** [48], **UNIT** [27], **BicycleGAN** [49]
- **Cycle/Bicycle:** As there is no previous work addressing the problem of multimodal generation from unpaired training data, 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.

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

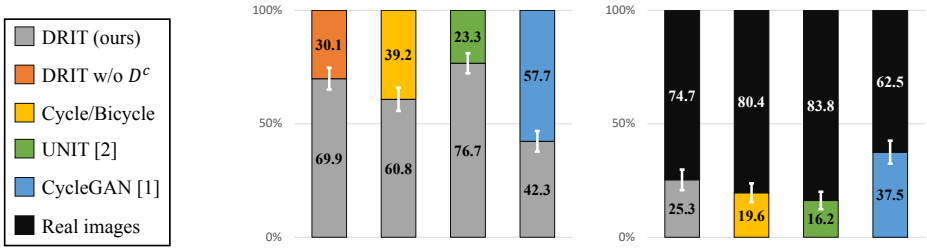


Fig. 9: **Realism preference results.** We conduct a user study to ask subjects to select results that are *more realistic* through pairwise comparisons. The number indicates the percentage of preference on that comparison pair. We use the winter → summer translation on the Yosemite dataset for this experiment.

Table 2: **Diversity.** We use the LPIPS metric [47] to measure the diversity of generated images on the Yosemite dataset.

Method	Diversity
real images	.448 ± .012
DRIT	**.424** ± .010
DRIT w/o D^c	.410 ± .016
UNIT [27]	.406 ± .022
CycleGAN [48]	.413 ± .008
Cycle/Bicycle	.399 ± .009

Table 3: **Reconstruct error**. We use the edge-to-shoes dataset to measure the quality of our attribute encoding. The reconstruction error is $\|y - G_{\mathcal{Y}}(E^c_{\mathcal{X}}(x), E^a_{\mathcal{Y}}(y))\|_1$. * BicycleGAN uses *paired* data for training.

Method	Reconstruct error
BicycleGAN [49]*	**0.0945**
DRIT	0.1347
DRIT, w/o D^c	0.2076

4.1 Qualitative Evaluation

Diversity. We first demonstrate the diversity of the generated images on several different tasks in Figure 5. In Figure 6, we compare the proposed model with other 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-related details (e.g., the color of tree and sky). Therefore, the variations take place in global color difference. Cycle/Bicycle is trained on pseudo paired data generated by CycleGAN. The quality of the pseudo paired data is not uniformly ideal. As a result, the generated images are of ill-quality.

To have a better understanding of the learned domain-specific attribute space, we perform linear interpolation between two given attributes and generate the corresponding images as shown in Figure 7. The interpolation results verify the continuity in the attribute space and show that our model can generalize in the distribution, rather than memorize trivial visual information.

Attribute transfer. We demonstrate the results of the attribute transfer in Figure 8. Thanks to the representation disentanglement of content and attribute, we are able to perform attribute transfer from images of desired attributes, as

illustrated in Figure 3(c). Moreover, 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.

4.2 Quantitative Evaluation

Realism vs. diversity. Here we have the quantitative evaluation on the realism and diversity of the generated images. We conduct the experiment using winter → summer translation with the Yosemite dataset. For realism, we conduct a user study using pairwise comparison. Given a pair of images sampled from real images and translated images generated from various methods, users need to answer the question "Which image is more realistic?" For diversity, similar to [49], we use the LPIPS metric [47] to measure the similarity among images. We compute the distance between 1000 pairs of randomly sampled images translated from 100 real images.

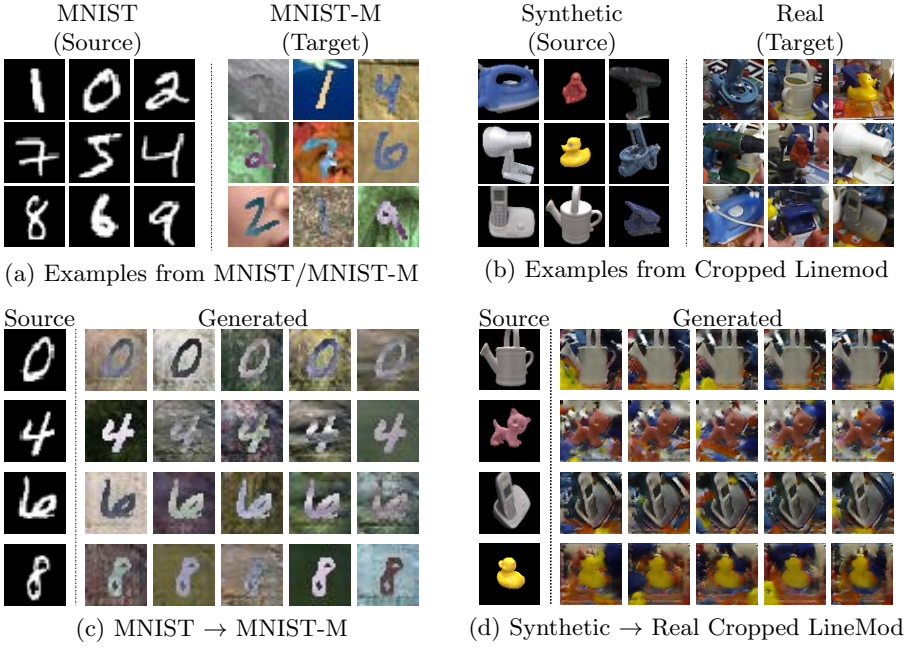
Figure 9 and Table 2 show the results of realism and diversity, respectively. UNIT obtains low realism score, suggesting that their assumption might not be generally applicable. CycleGAN achieves the highest scores in realism, yet the diversity is limited. The diversity and the visual quality of Cycle/Bicycle are constrained by the data CycleGAN can generate. Our results also demonstrate the need for the content discriminator.

Reconstruction ability. In addition to diversity evaluation, we conduct an experiment on the edge-to-shoes dataset to measure the quality of the disentangled encoding. Our model was trained using unpaired data. At test time, given a paired data $\{x, y\}$, we can evaluate the quality of content-attribute disentanglement by measuring the reconstruction errors of y with $\hat{y} = G_{\mathcal{Y}}(E^c_{\mathcal{X}}(x), E^a_{\mathcal{Y}}(y))$.

We compare our model with BicycleGAN, which requires paired data during training. Table 3 shows our model performs comparably with BicycleGAN despite training without paired data. Moreover, the result suggests that the content discriminator contributes greatly to the quality of disentangled representation.

4.3 Domain Adaptation

We demonstrate that the proposed image-to-image translation scheme can benefit unsupervised domain adaptation. Following PixelDA [3], we conduct experiments on the classification and pose estimation tasks using MNIST [24] to MNIST-M [12], and Synthetic Cropped LineMod to Cropped LineMod [15,43]. Several example images in these datasets are shown in Figure 10 (a) and (b). To evaluate our method, we first translate the labeled source images to the target domain. We then treat the generated labeled images as training data and train the classifiers of each task in the target domain. For a fair comparison, we use the classifiers with the same architecture as PixelDA. We compare the proposed method with CycleGAN, which generates the most realistic images in the target



(a) Examples from MNIST/MNIST-M

(b) Examples from Cropped Linemod

(c) MNIST → MNIST-M

(d) Synthetic → Real Cropped LineMod

Fig. 10: **Domain adaptation experiments.** We conduct the experiment on (a) MNIST to MNIST-M, and (b) Synthetic to Realistic Cropped LineMod. (c)(d) Our method can generate diverse images that benefit the domain adaptation.

domain according to our previous experiment, and three state-of-the-art domain adaptation algorithms: PixelDA, DANN [13] and DSN [4].

We present the quantitative comparisons in Table 4 and visual results from our method in Figure 10(c)(d). Since our model can generate diverse output, we generate one time, three times, and five times (denoted as $\times 1, \times 3, \times 5$) of target images using the same amount of source images. Our results validate that the proposed method can simulate diverse images in the target domain and improve the performance in target tasks. While our method does not outperform PixelDA, we note that unlike PixelDA, we do not leverage label information during training. Compared to CycleGAN, our method performs favorably even with the same amount of generated images (i.e., $\times 1$). We observe that CycleGAN suffers from the mode collapse problem and generates images with similar appearances, which degrade the performance of the adapted classifiers.

4.4 Limitations

Our method has the following limitations. 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 Figure 11(a). Second, it remains difficult when the domain characteristics differ significantly. For example, Figure 11(b) shows a failure case on the human figure due to the lack of human-related portraits in Monet collections.

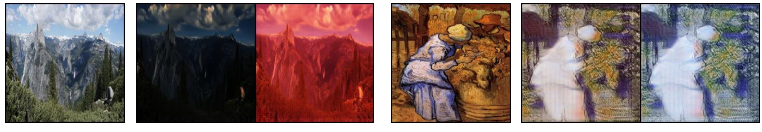
Table 4: **Domain adaptation results**. We report the classification accuracy and the pose estimation error on MNIST to MNIST-M and Synthetic Cropped LineMod to Cropped LineMod. The entries "Source-only" and "Target-only" represent that the training uses either image only from the source and target domain. Numbers in parenthesis are reported by PixelDA, which are slightly different from what we obtain.

(a) MNIST-M

Model	Classification Accuracy (%)
Source-only	56.6
CycleGAN [48]	74.5
Ours, ×1	86.93
Ours, ×3	90.21
Ours, ×5	**91.54**
DANN [13]	77.4
DSN [4]	83.2
PixelDA [3]	**95.9**
Target-only	96.5

(b) Cropped LineMod

Model	Classification Accuracy (%)	Mean Angle Error (°)
Source-only	42.9 (47.33)	73.7 (89.2)
CycleGAN [48]	68.18	47.45
Ours, ×1	95.91	42.06
Ours, ×3	97.04	37.35
Ours, ×5	**98.12**	**34.4**
DANN [13]	99.9	56.58
DSN [4]	**100**	53.27
PixelDA [3]	99.98	**23.5**
Target-only	100	12.3 (6.47)



(a) Summer → Winter (b) van Gogh → Monet

Fig. 11: **Failure Cases.** Typical cases: (a) Attribute space not fully exploited. (b) Distribution characteristic difference.

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. We also apply the proposed method to domain adaptation and achieve competitive performance compared to the state-of-the-art methods.

Acknowledgements

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

References

1. Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., Courville, A.: Augmented cyclegan: Learning many-to-many mappings from unpaired data. arXiv preprint arXiv:1802.10151 (2018)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. In: ICML (2017)
3. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: CVPR (2017)
4. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: NIPS (2016)
5. Cao, J., Katzir, O., Jiang, P., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y.: Dida: Disentangled synthesis for domain adaptation. arXiv preprint arXiv:1805.08019 (2018)
6. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: ICCV (2017)
7. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: NIPS (2016)
8. Cheung, B., Livezey, J.A., Bansal, A.K., Olshausen, B.A.: Discovering hidden factors of variation in deep networks. In: ICLR workshop (2015)
9. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR. vol. 1711 (2018)
10. Denton, E.L., Birodkar, V.: Unsupervised learning of disentangled representations from video. In: NIPS (2017)
11. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015)
12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR (2016)
13. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR (2016)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
15. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: ACCV (2012)
16. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: ICML (2018)
17. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
19. Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML (2017)
20. Kinga, D., Adam, J.B.: A method for stochastic optimization. In: ICLR (2015)

21. Kingma, D.P., Rezende, D., Mohamed, S.J., Welling, M.: Semi-supervised learning with deep generative models. In: NIPS (2014)
22. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate superresolution. In: CVPR (2017)
23. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: ECCV (2016)
24. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE (1998)
25. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
26. Li, Y., Huang, J.B., Ahuja, N., Yang, M.H.: Deep joint image filtering. In: ECCV (2016)
27. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS (2017)
28. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
29. Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Van Gool, L.: Exemplar guided unsupervised image-to-image translation. arXiv preprint arXiv:1805.11145 (2018)
30. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. In: ICLR workshop (2016)
31. Mathieu, M., Zhao, J., Sprechmann, P., Ramesh, A., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: NIPS (2016)
32. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: CVPR (2018)
33. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS workshop (2017)
34. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)
35. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
36. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
37. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: AAAI (2016)
38. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In: ICLR (2017)
39. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018)
40. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
41. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NIPS (2016)
42. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)

43. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: CVPR (2015)
44. Yi, Z., Zhang, H.R., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV (2017)
45. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: CVPR (2014)
46. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep networks as a perceptual metric. In: CVPR (2018)
48. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
49. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: NIPS (2017)