Unsupervised Multi-Domain Multimodal Image-to-Image Translation with Explicit Domain-Constrained Disentanglement

Weihao Xia[a], Yujiu Yang[b,*] and Jing-Hao Xue[c]

[a]Tsinghua University, China
[b]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[c]Department of Statistical Science, University College London, UK

ARTICLE INFO

Keywords:
Deep neural networks
Generative adversarial network
Image-to-image translation

ABSTRACT

Image-to-image translation has drawn great attention during the past few years. It aims to translate an image in one domain to a target image in another domain. However, three big challenges remain in image-to-image translation: 1) the lack of large amounts of aligned training pairs for various tasks; 2) the ambiguity of multiple possible outputs from a single input image; and 3) the lack of simultaneous training for multi-domain translation with a single network. Therefore in this paper, we propose a unified framework for learning to generate diverse outputs using unpaired training data and allow for simultaneous multi-domain translation via a single model. Moreover, we also observed from experiments that the implicit disentanglement of content and style could lead to undesirable results. Thus we investigate how to extract domain-level signal as explicit supervision so as to achieve better image-to-image translation. Extensive experiments show that the proposed method outperforms or is comparable with the state-of-the-art methods for various applications.

1. Introduction

Image-to-image translation aims to learn a mapping that can transfer an image from a source domain to a target domain, while maintaining the representative content of the input image. It has received significant attention since many problems in computer vision can be formulated as cross-domain image-to-image translation tasks (Isola et al., 2017; Zhu et al., 2017a,b), including super-resolution (Ledig et al., 2017), image inpainting (Yu et al., 2018a,b; Nazeri et al., 2019) and style transfer (Gatys et al., 2016), for example.

Despite of the great success, learning the mapping between two visually different domains is still challenging in three crucial aspects. First, exquisite large-scale datasets with thousands of aligned training pairs for various tasks are often unavailable. Second, in many scenarios, such mappings of interest are inherently multi-modal (i.e., a single input may correspond to multiple possible outputs). Third, for multi-domain image translation tasks, most existing methods learn individual two-domain mappings separately, and thus with n domains, they needs to train $\binom{n}{2} = O\left(n^2\right)$ models. They are incapable of jointly learning the mapping between all available domains in different datasets. To address these issues, several recent efforts have been made.

To tackle the limitation of paired training data, many studies propose unsupervised learning frameworks for image-to-image translation. Most methods are inspired by the intuition that the unpaired images from two domains should be consistent with their reconstructions in a cyclic mapping (Zhu et al., 2017a) or primal-dual relation (Yi et al., 2017). Superiority of this cycle consistency loss has been demonstrated on several tasks where paired training data hardly exist. However, these methods fail to produce multi-modal outputs conditioned on the given input image.

To capture the full distribution of possible outputs, simply incorporating noise vectors as additional inputs often leads to the mode collapsing issue and thus does not increase the variation of the generated images. Zhu et al. (2017b) try to encourage the one-to-one relationship between the output and the latent vector to generate diverse outputs. However, the training process of Zhu et al. (2017b) requires paired images to supervise. Very recently, Lee et al. (2018) and Huang et al. (2018) propose the disentangled representation framework to generate diverse outputs with unpaired training data. These two multi-modal unsupervised image-to-image translation methods assume that the latent space (Zhang et al., 2017, 2020) of images can be decomposed into a content latent space and a style latent space, and the images in different domains vary in the style but share a common content. Thus multi-modality can be achieved by recombining the content vector of an image from the source domain with a random style vector in the target style latent space.

To simultaneously train multi-domain translation with a single model, Choi et al. (2018) use a label (e.g., binary or one-hot vector) to represent domain information. They input both images and the corresponding domain information to the model, and learn to flexibly translate the images from the source domain to the target domain. By controlling domain labels, an image can be translated into any desired domain. Instead of using domain labels to represent domain characteristics as in Choi et al. (2018), Lin et al. (2019) use domain information as explicit supervision. They pre-train a classification network to classify images into domains. The classification features, together with the latent content features of the image in the source domain, are used to generate an image in the target domain. Such features extracted from the pre-trained network is used to represent domain information, thus they can be called domain features and the training with domain features can be called domain supervision. However, both Choi et al. (2018) and Lin et al. (2019)

*Corresponding author
xiawh3@outlook.com (W. Xia); yang.yujiu@sz.tsinghua.edu.cn (Y. Yang); jinghao.xue@ucl.ac.uk (J. Xue)
ORCID(s): 0000-0003-0087-3525 (W. Xia); 0000-0002-6427-1024 (Y. Yang); 0000-0003-1174-610X (J. Xue)

Table 1
Feature-by-feature comparison of image-to-image translation methods. Our model achieves unsupervised multi-domain multi-modal image-to-image translation with explicit domain-constrained disentanglement.

	Pix2pix	CycleGAN	BicycleGAN	StarGAN	DosGAN	MUNIT	DRIT	Ours
Unsupervised learning	-	✓	-	✓	✓	✓	✓	✓
Multi-modal	-	-	✓	-	-	✓	✓	✓
Multi-domain	-	-	-	✓	✓	-	-	✓
Disentangled representation	-	-	-	-	✓	✓	✓	✓
Domain supervision	-	-	-	-	✓	-	-	✓

produce a single output and are lack of output diversity.

Several recent methods (Lee et al., 2018; Huang et al., 2018; Liu et al., 2018) adopt disentangled representations for unsupervised image-to-image translation, but we observed in experiments that implicit disentanglement learning can confuse content with style in some cases. As shown in Figure 3, if the framework of Lee et al. (2018) is adapted for image de-blurring tasks, the de-blurred images may have different face contour from the original one, which means that the attribute extractor of Lee et al. (2018) has not only learned the blur distortion pattern but also misrecognized some content representations like face contour as style. This can be attributed to the ambiguous and implicit disentanglement of content and style.

What's more, domain information is currently underexploited in the area of image-to-image translation. For photo-to-art translation, we can distinguish that the generated image is either in the style of Pablo Picasso or in the style of Isaac Levitan. Similarly, different weather conditions, such as sunny, foggy, rainy, snowy and cloudy, should contain specific modalities, and the same is true for seasons. That is, the style itself can be learned from the collected data of a unique domain (e.g., artist and weather) and then exploited for image-to-image translation.

Therefore, in this paper we propose a new approach termed unsupervised **M**ulti-domain **M**ultimodal **I**mage-to-image **T**ranslation with explicit **D**omain-**C**onstrained disentanglement (DCMIT).To the best of our knowledge, DCMIT is the first approach to tackling all the aforementioned challenges and issues in image-to-image translation. DCMIT is a unified framework for learning to generate diverse outputs with unpaired training data and allow for simultaneous multi-domain translation with a single model. Furthermore, DCMIT utilizes domain information and explicitly constrains the disentanglement for desired unsupervised image-to-image translation.

To sum up, our key contributions are:

- We introduce the first unsupervised image-to-image translation method that achieves diverse outputs and simultaneous training of multi-domain translation with a single model.
- We propose explicit disentanglement learning constraints with domain supervision. We investigate how to extract domain supervision information so as to learn explicit disentangled representations to avoid the confusion of content and style.
- Extensive qualitative and quantitative experiments are conducted on multiple datasets, and they show that the proposed method outperforms or is comparable with the state-of-the-art methods for various applications.

2. Related work

We initially provide an overview of the recent advances with Generative Adversarial Networks (GANs), then introduce some existing image-to-image translation methods and disentangled representations. We also give a brief introduction to style transfer and domain adaptation, two tasks closely related with image-to-image translation.

2.1. Generative adversarial network

The GAN framework (Goodfellow et al., 2014; Schmidhuber, 2020) has achieved excellent results in many tasks such as image super-resolution (Ledig et al., 2017) and image inpainting (Yu et al., 2018a,b; Nazeri et al., 2019). GANs usually consist of a generator G and a discriminator D. The training procedure for GANs is a minimax game between G and D, where D is trained to distinguish whether the input image is real or fake, and G is trained to fool D with the generated samples. The ideal solution is the Nash equilibrium where G and D could not improve their cost unilaterally (Heusel et al., 2017).

Various improvements have been proposed to handle challenges in GANs including model generalization and training stability. Arjovsky et al. (2017) and Gulrajani et al. (2017) propose to minimize the Wasserstein distance between the model and data distributions. Berthelot et al. (2017) optimize a lower bound of the Wasserstein distances between auto-encoder loss distributions on real and fake data. Mao et al. (2017) propose a least square loss for the discriminator, which implicitly minimizes the Pearson $\mathcal{X}^2$ divergence, leading to stable training, high image quality and considerable diversity.

2.2. Image-to-image translation

Isola et al. (2017) propose the first general image-to-image translation method (pix2pix) based on conditional GANs. Wang et al. (2018a) propose an HD version of pix2pix by utilizing a coarse-to-fine generator, several multi-scale discriminators, and a feature matching loss, which increase the

resolution to 2048×1024. Since it is usually time-consuming and expensive to collect such an exquisite large-scale dataset with thousands of image pairs, many studies have also attempted to tackle the paired training data limitation. Zhu et al. (2017a), Kim et al. (2017), Yi et al. (2017) and Liu et al. (2017) leverage cycle consistency to regularize the unsupervised training process Song et al. (2020). Many methods aim to produce diverse outputs, including Zhu et al. (2017b), Lee et al. (2018) and Huang et al. (2018). Some other methods such as Choi et al. (2018), Lin et al. (2019), Liu et al. (2018) and Anoosheh et al. (2018) are proposed to improve the scalability of unsupervised image-translation methods. Table 1 shows a feature-by-feature comparison among some existing image-to-image translation methods.

2.3. Disentangled representations

There are many recent studies on disentangled representation learning. For example, Lu et al. (2019) disentangle content from blur; Denton et al. (2017) separate time-independent and time-varying parts; Johnson et al. (2016) iteratively optimize the image by minimizing a content loss and a style loss, which can also be regarded as an implicit disentanglement of content and style; Zhu et al. (2017b) combine cLR-GAN and cVAE-GAN to model the distribution of possible outputs; and Chen et al. (2016a) decompose representation by maximizing the mutual information between the latent factors and the synthesized images without utilizing paired training data. Some other studies (Xiao et al., 2018; Liu et al., 2018; Lee et al., 2018; Huang et al., 2018) focus on disentanglement of content and style or attribute. It is difficult to explicitly define content or style and different methods adopt different definitions due to their specific tasks. In our setting, we refer to content as domain-invariant visual elements that can be shared across domains and style as domain-specific visual elements. We disentangle an image into domain-invariant and domain-specific representations to facilitate learning diverse cross-domain mappings.

3. Methodology

Our goal is to achieve unsupervised multi-domain multimodal image-to-image translation via disentangled representations with a single model. The pipeline of our method is shown in Figure 1. For multi-domain translation, we design an intra-domain and inter-domain supervision mechanism, which is able to represent the essence of different domains and translate images across different domains with only one single model. For multi-modal generation between two domains, we regularize the style codes in the training phase so that they can be represented by a Gaussian distribution. By controlling the parameters of style codes, multi-modalities of generated images are possible. The model architecture and loss functions are also coherently designed for diverse and realistic image-to-image translation.

3.1. Problem formulation

Assume in a dataset there are n different domains $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_n\}$. Our goal is to achieve unsupervised multi-domain multimodal image-to-image translation with explicit domain-constrained disentanglement by using a single model. For each image $x_i \in \mathcal{D}_i$, the unique disentangled representations of content latent code $c \in \mathcal{C}$ and the style latent code $s_i \in \mathcal{S}_i$ can be extracted from content encoder $E^c_{\mathcal{D}_i}$ and style encoder $E^s_{\mathcal{D}_i}$. The generator G_i can produce an image of certain style if given specific style latent code and content code. For example, let $x_1 \in \mathcal{D}_1$ and $x_2 \in \mathcal{D}_2$ be images from two different domains, the content encoders E^c_1 and E^c_2 map images onto a domain-invariant content space ($E^c_i : \mathcal{D}_i \to \mathcal{C}$) and the style encoders E^s_1 and E^s_2 map images onto the domain-specific style spaces ($E^s_i : \mathcal{D}_i \to \mathcal{S}_i$). The generator G_i generates images conditioned on the given content codes and style codes ($G_i : \{\mathcal{C}, \mathcal{S}_i\} \to \mathcal{D}_i$). We postulate that only the content latent part can be shared across domains and that the style part is domain-specific.

3.2. Intra-domain and inter-domain supervisions

To exploit domain information and explicitly constrain the disentanglement of content and style, we propose explicit domain-constrained disentanglement by first introducing intra-domain and inter-domain supervisions.

Let $x_{1\to2}$ be a sample produced by translating image x_1 in domain $\mathcal{D}_1$ to its counterpart x_2 in domain $\mathcal{D}_2$ (similarly for $x_{2\to1}$). Then for a pair images $\left(x_1, x_2\right)$, we have

$$\begin{aligned} x_1 &= G_1(c, s_1), \quad & x_2 &= G_2(c, s_2), \\ x_{1\to2} &= G_2(c, s_2), \quad & x_{2\to1} &= G_1(c, s_1). \end{aligned} \tag{1}$$

Since s_1 and s_2 are domain-specific style codes extracted from single images x_1 and x_2, respectively, from different domains, we call this translation inter-domain translation.

The style code extracted from a single image contains more information than the generalized style of a collection of images. In the training phase, the model may extract some content features as style features incorrectly as illustrated in Figure 3. To alleviate this situation, we design an intra-domain supervision to constrain the disentangled representation learning and represent the essence of different domains. The main idea to achieve this is relatively simple: "Two images from the same domain exchange their style codes, the generated images should be consistent with themselves." Different from the style codes extracted from a single image $s_i \in \mathcal{S}_i$, the style codes extracted at the domain level should be domain-specific and represent generalized domain style representations. For n domains $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_n\}$, we have n domain style representations $\{\mathcal{S}_{\mathcal{D}_1}, \mathcal{S}_{\mathcal{D}_2}, \cdots, \mathcal{S}_{\mathcal{D}_n}\}$. We can call this translation as intra-domain translation.

As shown in Figure 2, inter-domain and intra-domain translation can be represented as

$$\begin{aligned} x_1 &= G_1(c, s_1), \quad & x_2 &= G_2(c, s_2), \\ x_{1\to2} &= G_2(c, s_2), \quad & x_{2\to1} &= G_1(c, s_1), \\ x_{1\to1'} &= G_1(c, \mathcal{S}_{\mathcal{D}_1}), \quad & x_{2'\to2} &= G_2(c, \mathcal{S}_{\mathcal{D}_2}). \end{aligned} \tag{2}$$

The intra-domain translation aims to learn the essence style of a domain, which means that the learned style representations of images from the same domain do not vary to

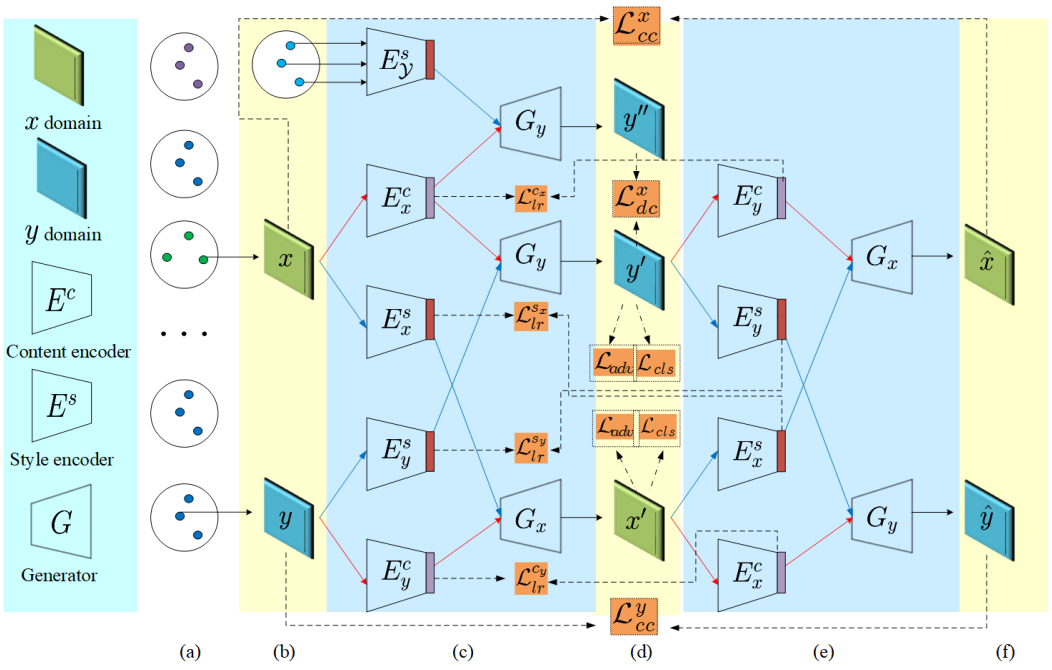

Figure 1: The pipeline of our method: (a) n domains; (b) two batches of images $x \in \mathcal{D}_x$, $y \in \mathcal{D}_y$ with labels, randomly selected from two different domains; (c) the first translation; (d) style-swapped images; (e) the second translation; and (f) cycle-reconstructed images. To achieve image translation between domains, we first randomly select two domains, then load two batches of images $x \in \mathcal{D}_x$, $y \in \mathcal{D}_y$ with labels. Images from different domains are encoded as domain-invariant content representations c and domain-specific style representations s. The two translations are achieved by swapping the style codes and using generator G to produce the translated output images. The first translation constrains the translated images x' and y' with the proposed disentanglement constrained loss. The second translation constrains the image reconstruction with the cycle consistency loss. Due to the disentangled representations, the style representations are constrained to match the prior Gaussian distribution, so that we can generate several possible outputs by random sampling from this prior. The domain style representations are extracted by the pre-trained feature extractor E^s_y from the collections of a certain style and used to constrain the disentanglement of content and style (similarly for y, which is omitted for simplicity of the diagram). The multi-domain simultaneous training is implemented by adding specific discriminative labels for the domains.

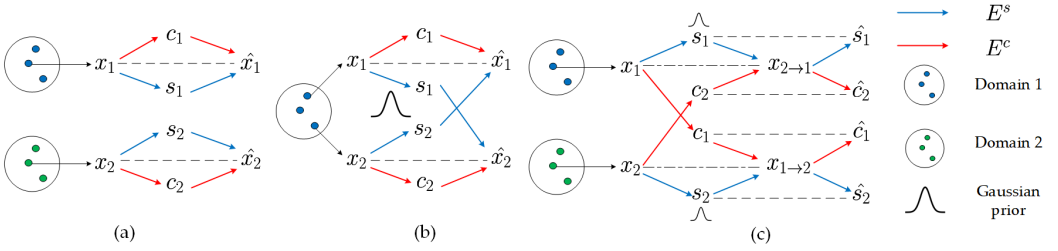

Figure 2: Illustration of (a) self translation, (b) intra-domain translation and (c) inter-domain translation. For better comparison, we follow the representations as with MUNIT (Huang et al., 2018), and to avoid unnecessary confusion, we change the descriptions. Our model consists of two types of auto-encoders (denoted by red and blue arrows, respectively). Similarly to MUNIT (Huang et al., 2018) and DRIT (Lee et al., 2018), the latent code of each auto-encoder is composed of a content code c and a style code s. The model is trained with adversarial objectives (dashed dotted lines) that ensure the translated images to be indistinguishable from real images in the target domain, as well as with bidirectional reconstruction objectives (dashed lines) that reconstruct both images and latent codes.

an unreasonable degree. Specifically, all images converge to the "mean" style of the domain. After the training on carefully selected images, this constraint helps the content and style encoders learn explicit disentangled representations for accurate inter-domain translation. We can readily control its influence by changing the weight parameters.

3.3. Pre-training of domain style representation extractor

Differently from many previous works regarding multiple domains as different sources of images, we treat each domain as explicit supervision. Similarly to Lin et al. (2019), we pre-train a domain style representation extractor for each domain

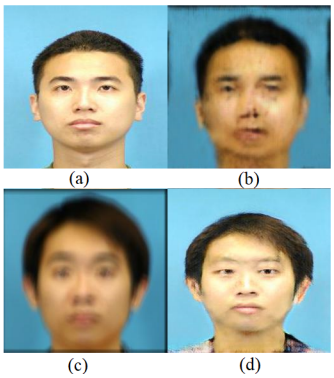

Figure 3: Translation results of DRIT (Lee et al., 2018) for image de-blurring task: (a) real image; (b) blurred version of (a); (c) real blurred image; and (d) de-blurred version of (c) by DRIT. We can see that when DRIT is adopted for image de-blurring, the de-blurred image has different face contours from the original ones, which means that the attribute extractor has not only learned blur distortion pattern but also misrecognized some content representations such as face contour as attribute. It might be attributed to the ambiguous and implicit disentanglement of content and style.

as explicit domain supervision.

For domain supervision, Lin et al. (2019) train a classifier that tries to correctly distinguish images of different domains. Then they regard the output of second-to-last layer of the classifier as the domain style. Different from this ambiguous and implicit definition, we try to learn the domain style representations by intra-domain translation.

Given images from n different domains, we train a CNN by switching style codes of images from the same domain. The goal of this CNN, which we call domain representation extractor $E^s_{\mathcal{D}_i}$, is to learn domain-specific style representations $\mathcal{S}_{\mathcal{D}_i}$ for domain $\mathcal{D}_i$ and to correctly classify the domain of an image. Then this pre-trained model $E^s_{\mathcal{D}_i}$ is used as explicit domain supervision for inter-domain translation.

3.4. Model

As aforementioned, the pipeline of our model is shown in Figure 1. Similar to other unsupervised image-to-image translation via disentangled representations (Liu et al., 2018; Lee et al., 2018; Huang et al., 2018), our model consists of content encoder E^c_i , style encoder E^s_i, decoder G and discriminator D_i for each domain $\mathcal{D}_i$, $i = 1, 2, \ldots, n$. Moreover, we have the domain classifier D_{cls} pre-trained together with the domain style representation extractor $E^s_{\mathcal{D}_i}$.

As shown, to achieve image translation between two domains $\{\mathcal{D}_1, \mathcal{D}_2\}$, images x_1, x_2 from different domains are encoded as domain-invariant content representations $c_1 = E^c_1(x_1)$, $c_2 = E^c_2(x_2)$, and domain-specific style representations $s_1 = E^s_1(x_1)$, $s_2 = E^s_2(x_2)$. Then swap the style codes and use G_2 to produce the translated output image $x_{1\to 2} = G_2(c_1, s_2)$.

3.5. Network architecture

Figure 4 shows the network architecture of our model. It consists of a content encoder, a style encoder and a decoder.

Content encoder. The content encoder consists of several convolutional layers to down-sample the input images to get high-dimensional features and several basic blocks for further processing. There are many choices for basic block such as residual block (He et al., 2016), residual dense block (Zhang et al., 2018b), and residual in residual dense block (Wang et al., 2018b). Here we use the traditional residual block for simplicity and replace Batch Normalization (BN) (Ioffe and Szegedy, 2015) with Instance Normalization (IN) (Ulyanov et al., 2016). For diversity, we add noise in the last two basic blocks as with Lee et al. (2018).

Style encoder. The style encoder includes several strided convolutional layers, followed by an adaptive average pooling layer and a convolutional layer. We do not use IN layers in the style encoder, as IN removes the original feature mean and variance which contain important style information.

Decoder. The decoder generates images from their content codes and style codes. For multi-domain translation, we also add the domain class as input. Specifically, the domain class and style codes are concatenated by channel and then fed into a multi-layer perceptron (MLP). The content codes and outputs generated by the MLP are further processed via several concatenation blocks. We equip the residual blocks with Adaptive Instance Normalization (AdaIN) layers (Huang and Belongie, 2017), whose parameters are dynamically generated by the MLP from the style codes y:

$$\mathrm{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y), \tag{3}$$

in which we simply scale the normalized content input x with the variance $\sigma(y)$ and shift it with the mean $\mu(y)$.

Discriminator and domain classifier. The architecture of discriminator is similar with Choi et al. (2018). The domain classifier is built on top of the discriminator, as shown in Figure 5. It consists of six convolution layers with kernel size 4×4 and stride 2, followed by two separated convolutional branches that are implemented for discriminative output and domain class.

Domain style representation extractor. The domain style representation extractor shares the same architecture with the style encoder. Specifically, it consists of one convolution layer with kernel size 4×4 and stride 1; six convolution layers with kernel size 4×4, stride 2 and ReLU followed by an adaptive average pooling layer and a convolutional layer with kernel size 1×1, stride 1.

3.6. Loss functions

Our loss functions are designed for unsupervised multi-domain multi-modal image-to-image translation. For unsupervised training, we adopt the image reconstruction loss

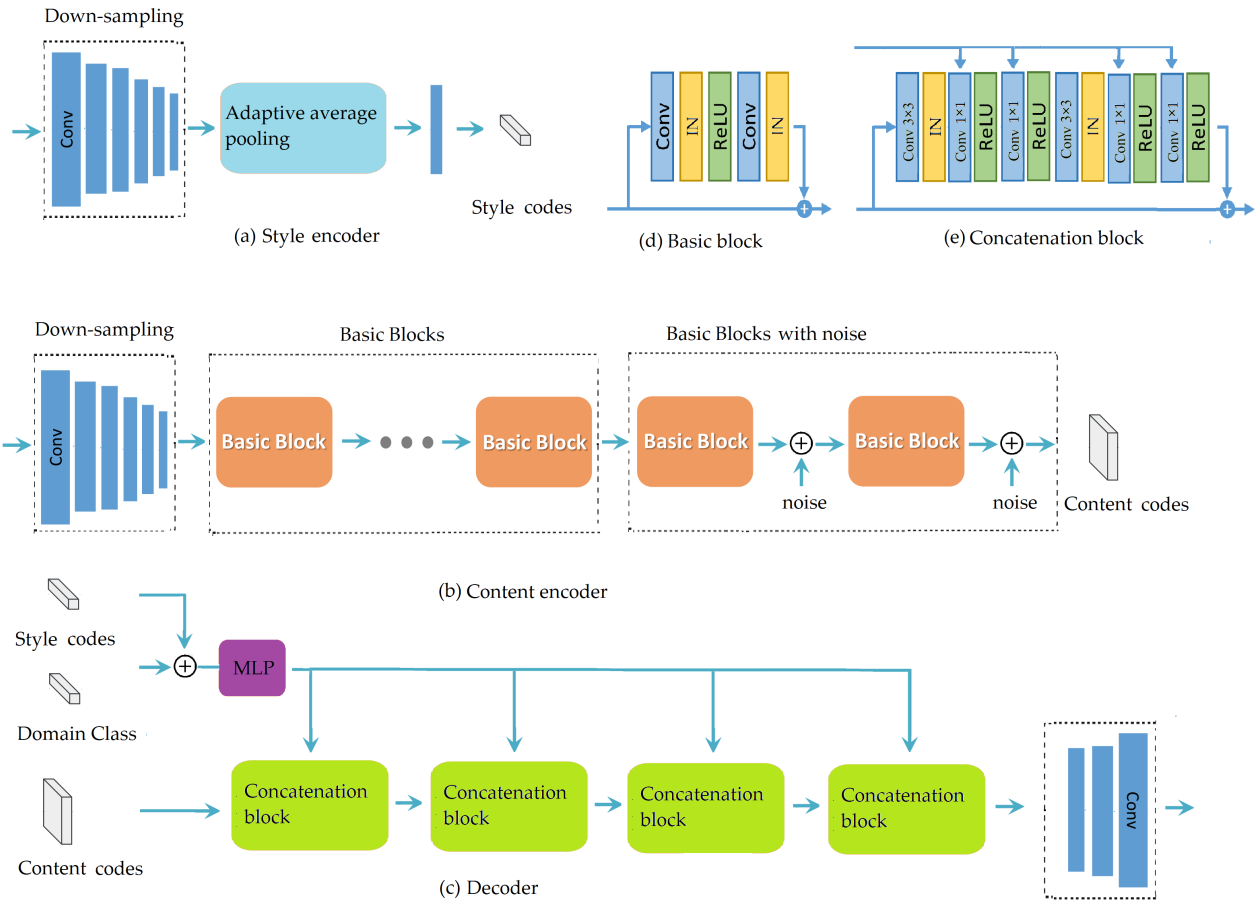

Figure 4: Network architecture. For more details, refer to Section 3.5.

and the latent reconstruction loss based on the cycle consistent loss. We also add constraints to improve the representations of content and style codes by the self-reconstruction loss. For multi-modality, we introduce a distribution matching loss to make the style codes extracted by the content encoder close to a prior Gaussian distribution. By doing this, we are able to sample style codes from the prior Gaussian distribution at the test phase. Since the sampled style codes are stochastic, the decoder can produce diverse samples sharing the same content. For simultaneous training of multiple different domains, we use the domain classification loss.

Self-reconstruction loss. Given an image from a certain domain, we should be able to reconstruct itself after encoding and decoding. Thus, for example for x_1 from domain $\mathcal{D}_1$, the self-reconstruction loss $\mathcal{L}_{\text{sr}}$ can be written as

$$\mathcal{L}_{\text{sr}}^{x_1} = \mathbb{E}_{x_1 \sim p(x_1)} \left[\left\| G_1 \left(E_1^c \left(x_1 \right), E_1^s \left(x_1 \right) \right) - x_1 \right\|_1 \right]. \quad (4)$$

Image reconstruction loss. Given an image sampled from the data distribution, we should be able to reconstruct it after encoding and decoding. The image reconstruction loss $\mathcal{L}_{\text{cc}}$ is adopted in two stages. In the pre-training of domain representations, we use the image reconstruction loss to obtain a domain-specific style representation extractor $E_{\mathcal{D}_i}^s$ during image reconstruction; e.g. when $i = 1$:

$$\mathcal{L}_{\text{cc}}^{x} = \mathbb{E}_{x \sim p(x)} \left[\left\| G_1 \left(E_1^c \left(x \right), E_{\mathcal{D}_1}^s \left(x' \right) \right) - x \right\|_1 \right], \quad (5)$$

where x and x' are from the same domain $\mathcal{D}_1$.

In inter-domain translation, the image reconstruction loss $\mathcal{L}_{\text{cc}}$ is used on the style from a single image. The image reconstruction loss can be represented as

$$\begin{aligned} \mathcal{L}_{\text{cc}}^{x} &= \mathbb{E}_{x,y} \left[\left\| G_1 \left(E_2^c(y'), E_1^s(x') \right) - x \right\|_1 \right], \\ \mathcal{L}_{\text{cc}}^{y} &= \mathbb{E}_{x,y} \left[\left\| G_2 \left(E_1^c(x'), E_2^s(y') \right) - y \right\|_1 \right], \end{aligned} \quad (6)$$

where

$$x' = G_1 \left(E_2^c(y), E_1^s(x) \right), \; y' = G_2 \left(E_1^c(x), E_2^s(y) \right). \quad (7)$$

Disentanglement constrained loss. To utilize domain information and explicitly constrain the disentanglement, we

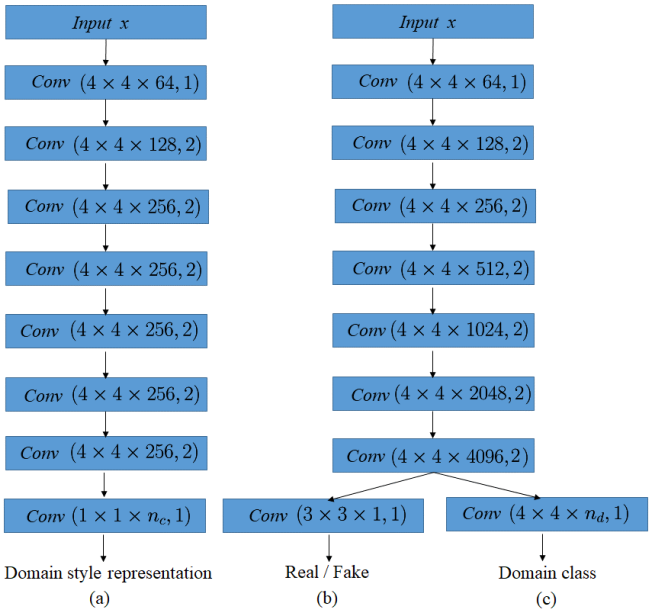

Figure 5: Discriminator and domain classifier.

propose the disentanglement loss. For a style extracted in the domain style representation, the disentanglement constrained loss $\mathcal{L}_{\text{dc}}$ can be expressed as

$$\mathcal{L}_{\text{dc}}^{x} = \mathbb{E}_{x,y}\left[\left\|y' - y''\right\|_1\right], \tag{8}$$

where $y'' = G_2\left(E_1^c(x), S_y\right)$, and S_y is the extracted domain style.

Latent reconstruction loss. Given latent (style and content) codes sampled from the latent distribution at the translation time, we should be able to reconstruct them after decoding and encoding. Take the path to reconstruct x' for example:

$$\begin{aligned}\mathcal{L}_{\text{lr}}^{c_1} &= \mathbb{E}_{c_1\sim p(c_1),s_2\sim q(s_2)}\left[\left\|E_2^c\left(G_2\left(c_1,s_2\right)\right) - c_1\right\|_1\right],\\ \mathcal{L}_{\text{lr}}^{s_2} &= \mathbb{E}_{c_1\sim p(c_1),s_2\sim q(s_2)}\left[\left\|E_2^s\left(G_2\left(c_1,s_2\right)\right) - s_2\right\|_1\right].\end{aligned} \tag{9}$$

$\mathcal{L}_{\text{lr}}^{c_2}$ and $\mathcal{L}_{\text{lr}}^{s_1}$ are similarly defined at the path to reconstruct y', as shown in Figure 1.

Distribution matching loss. We adopt a distribution matching loss to make the style codes close to a prior Gaussian distribution. At the test phase, we are able to sample randomly from the prior Gaussian distribution and regard it as a style code. The measure of distance between two distributions can be covariance, Maximum Mean Discrepancy (MMD) or KL divergence. Instead of implementing the KL divergence as in MUNIT (Huang et al., 2018) and DRIT (Lee et al., 2018), here we choose MMD. We will illustrate the reasons for this choice in Section 4.7.

The distribution matching loss $\mathcal{L}_{\text{dm}}$ described by MMD can be written as

$$\mathcal{L}_{\text{dm}} = \mathbb{E}\left[D_{\text{MMD}}\left(q(z)|N(0,1)\right)\right], \tag{10}$$

where

$$\begin{aligned}D_{\text{MMD}}(q|p) = \mathbb{E}_{p(z),p(z')}\left[k\left(z,z'\right)\right] - 2\mathbb{E}_{q(z),p(z')}\left[k\left(z,z'\right)\right]\\ + \mathbb{E}_{q(z),q(z')}\left[k\left(z,z'\right)\right],\end{aligned} \tag{11}$$

$q(z)$ is attribute representation and $k(\cdot,\cdot)$ can be any positive definite kernel, such as Gaussian kernel $k\left(z,z'\right) = e^{-\frac{\left\|z-z'\right\|^2}{2\sigma^2}}$.

Domain classification loss. To achieve simultaneous training of multiple domains with a single model, we assign a unique class label for each domain as with Choi et al. (2018). While translating input image x_1 with domain class c_1 to image x_2 with class c_2, the auxiliary domain classifier tries to distinguish images from different domains. The corresponding domain classification loss can be defined as

$$\begin{aligned}\mathcal{L}_{\text{cls}}^{real} &= \mathbb{E}_{x,c'}\left[-\log D_{\text{cls}}\left(c'|x\right)\right],\\ \mathcal{L}_{\text{cls}}^{fake} &= \mathbb{E}_{x,c}\left[-\log D_{\text{cls}}(c|G(x,c))\right],\end{aligned} \tag{12}$$

where $D_{\text{cls}}\left(c'|x\right)$ represents the probability of a domain label calculated by D. The goal of this term is that D can correctly classify a real image x to its original domain c' and G tries to generate images that can be recognized as from target domain c by D.

This auxiliary domain classifier is built on top of discriminator D. In the training phase, the domain classification loss of real images is used to optimize parameters of discriminator D and the domain classification loss of fake images is used to optimize G.

In our experiment, the domain classifier D_{cls} is pretrained together with the domain style representation extractor $E_{D_i}^s$.

Adversarial loss. For high image quality, stable training and considerable diversity, we use the least-squares GAN proposed by Mao et al. (2017). Thus $\mathcal{L}_{\text{adv}}$ can be formulated as

$$\begin{aligned}\min_{D_1}\mathcal{L}_{\text{adv}}\left(D_1\right) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\left(D_1(\boldsymbol{x}) - b\right)^2\right] +\\ & \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\left(D_1\left(G_1(\boldsymbol{z})\right) - a\right)^2\right],\\ \min_{G_1}\mathcal{L}_{\text{adv}}\left(G_1\right) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\left(D_1\left(G_1(\boldsymbol{z})\right) - c\right)^2\right].\end{aligned} \tag{13}$$

Overall training loss. The total training loss functions of the encoder E, decoder G and discriminator D are defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{EoG}}^{\text{total}} &= \mathcal{L}_{\text{adv}} + \lambda_{\text{sr}}\mathcal{L}_{\text{sr}} + \lambda_{\text{cc}}\mathcal{L}_{\text{cc}} + \lambda_{\text{dc}}\mathcal{L}_{\text{dc}}\\ &+ \lambda_{\text{dm}}\mathcal{L}_{\text{dm}} + \lambda_{\text{lr}}\mathcal{L}_{\text{lr}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}^{fake},\end{aligned} \tag{14}$$

$$\mathcal{L}_{D}^{\text{total}} = \mathcal{L}_{\text{adv}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}^{real}, \tag{15}$$

where hyper-parameters λ_{sr}, λ_{cc}, λ_{dc}, λ_{dm}, λ_{lr} and λ_{cls} are weights to control the importance of each term.

Figure 6: Samples from datasets. We mainly use three multi-domain datasets for experiments: Art, Season and Weather. Each contains four domains.

The overall process. The overall process is summarized in Algorithm 1. The training process consists of two phases: the intra-domain style representation extractor training and the inter-domain translation training. Both phases share almost the same network architecture and loss functions except the following differences. Since we want to learn the domain style representation from each domain and adopt it to inter-domain translation as domain supervision, we select images from the same domain and swap their style codes. Ideally, the style-exchanged images should be consistent with the original ones. Only one-step translation is required to get the domain style representation. So the loss functions of the domain style representation extractor training can be defined as

$$\begin{aligned}\mathcal{L}_{\text{E}\circ\text{G}}^{\text{total}} &= \mathcal{L}_{\text{adv}} + \lambda_{\text{sr}}\mathcal{L}_{\text{sr}} + \lambda_{\text{dm}}\mathcal{L}_{\text{dm}} + \lambda_{\text{cls}}L_{\text{cls}}^{fake} \\ &+ \lambda_{\text{cc}}L_{\text{cc}}, \qquad (16)\\ \mathcal{L}_{D}^{\text{total}} &= \mathcal{L}_{\text{adv}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}^{real}, \qquad (17)\end{aligned}$$

where hyper-parameters λ_{sr}, λ_{dm}, λ_{cls} and λ_{cc} are weights to control the importance of each term.

Thus we get the domain style representation extractor $E_{D_i}^{s}$. It is mainly used in image reconstruction loss to constrain feature disentanglements.

4. Experiments

4.1. Experiment settings

For training, we adopt the Adam optimizer with a batch size 8, a learning rate of 0.0001 with exponential decay rates $\beta_1 = 0.5$, $\beta_2 = 0.999$. We resize all input images into 216×216 in experiments. The hyper-parameters are set as $\lambda_{\text{sr}} = 10$, $\lambda_{\text{cc}} = 10$, $\lambda_{\text{dm}} = 0.01$, $\lambda_{\text{lr}} = 10$ and $\lambda_{\text{dc}} = 0.15$. Parameter λ_{cls} of G is 5.0 and λ_{cls} of D is 1.0. We do not implement domain supervision if the training data are paired.

4.2. Datasets

We use three multi-domain datasets for experiments: Art, Weather, Season. Notice that all images in these datasets are not paired.

Art: This dataset contains four domains: real images, Monet, Ukiyo-e and Van Gogh. These art images can be downloaded from Wikiart[1] and the real photos are from Flickr with tags *landscape* and *landscapephotography*. We use the datasets of monet2photo, vangogh2photo, ukiyoe2photo and cezanne2photo collected by Zhu et al. (2017a).

Weather: This dataset contains four domains: sunny, cloudy, snowy and foggy, which is randomly selected from the Image2Weather (Chu et al., 2017).

Season: This dataset consists of approximately 6,000 images of the Alps mountain range scraped from Flickr. The original dataset collected by Anoosheh et al. (2018) categorizes photos individually into four seasons based on the provided timestamp of when it was taken. But this lead to many misclassifications. We revise each category by deleting ambiguous images or misclassified images to the right category to make them more distinguishable.

Since Zhu et al. (2017b) need paired data for training, we evaluate multi-modality on **edges → shoes** and **edges → handbags**. The edges → shoes dataset contains 50k training images from the UT Zappos50K dataset (Yu and Grauman, 2014). The edges → handbags dataset contains 137K Amazon Handbag images from Zhu et al. (2016). Edges are

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

Figure 7: Results of StarGAN, DosGAN, ComboGAN and ours on the Season dataset. Images in the first column are input images randomly selected from the four seasons. Following are results generated by ours, StarGAN, ComboGAN and DosGAN. For each method, the four columns are arranged successively as spring, summer, autumn and winter. Better look by zooming in.

computed by the HED edge detector (Xie and Tu, 2015) and post-processing. Both datasets can be downloaded at CycleGAN (Zhu et al., 2017a) website[2].

Samples from these three datasets are visually demonstrated in Figure 6 to illustrate their styles. Table 2 lists domain information and corresponding number of training data.

[2]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

Figure 8: Multi-domain multi-modal image translation results on Art. The Art dataset contains four domains: real image, Monet, Van Gogh and Ukiyo-e. Better look by zooming in.

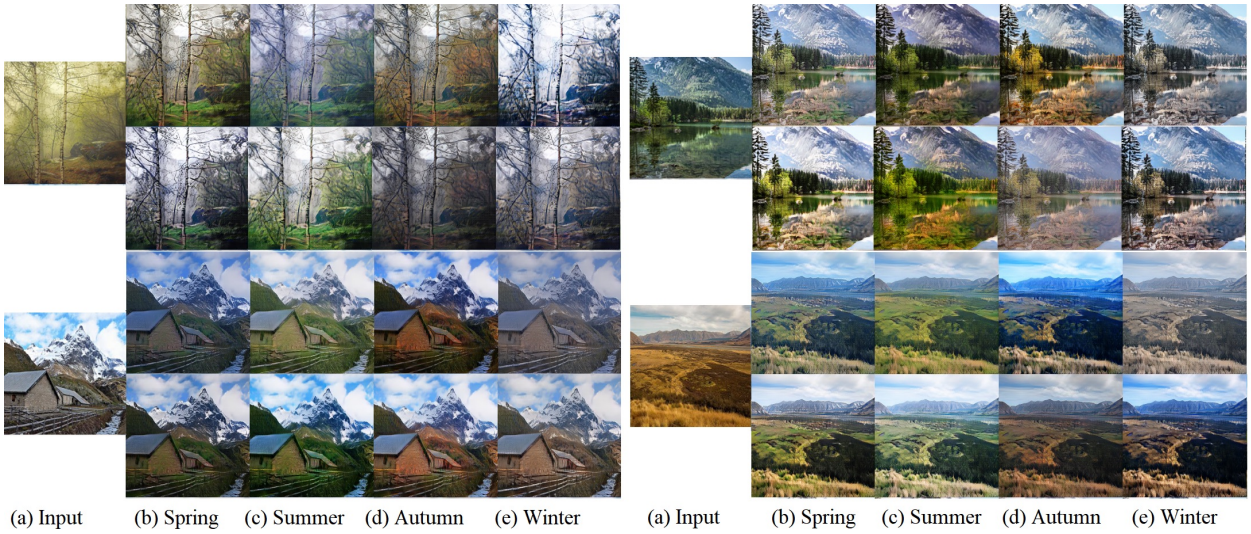

Figure 9: Multi-domain multi-modal translation results on Season. The Season dataset contains four domains: spring, summer, autumn and winter. Notice that all these image are generated via one training process. Better look by zooming in.

4.3. Baselines

We perform the evaluation on the following baseline approaches:

BicycleGAN. BicycleGAN (Zhu et al., 2017b) is the first image-to-image translation model that aims to generate continuous and multi-modal output images. However, it needs paired images for training.

DRIT and **MUNIT.** DRIT (Lee et al., 2018) and MUNIT (Huang et al., 2018) propose to simultaneously generate diverse outputs given the same input image without the requirement of paired supervision via disentangled representations. It is designed for translation between two domains.

StarGAN. StarGAN (Choi et al., 2018) aims to handle scalability of unsupervised image-to-image translation problems. It uses one generator and discriminator in common for all domains by adding domain labels. The generator requires images and the desired domain label specifying the target domain as inputs, and the discriminator is trained to classify the domain labels of generated images and judge whether it is real or fake. By doing this, it is able to take any number of domains as input. However, the model was just applied to the task of face attribute translation in its original paper.

Algorithm 1: TRAINING PROCESS.

Input: n different domains $\mathcal{D}_k$ $\forall k \in [n]$, batch size K, learning rate η;
Stage 1: domain style representation extractor training
Randomly initialize the parameters Θ_E of domain representation extractor $E_{\mathcal{D}}^s$;
Randomly select one domain $\mathcal{D}_k$, $k \in [n]$. Get a mini-batch of data $\mathcal{D}_k$ satisfying $D_k \subset \mathcal{D}_k$ and $\left|D_k\right| = K$;
Update the network as follows:
$\Theta_{E\circ G} \leftarrow \Theta_{E\circ G} - \eta\nabla_{\Theta}\Theta_{E\circ G}\mathcal{L}_{\text{E}\circ\text{G}}^{\text{total}}\left(D_S\right)$
$\Theta_D \leftarrow \Theta_D - \eta\nabla_{\Theta_D}\mathcal{L}_D^{\text{total}}\left(D_S\right)$
where $\mathcal{L}_{\text{E}\circ\text{G}}^{\text{total}}\left(D_S\right)$ and $\mathcal{L}_D^{\text{total}}\left(D_S\right)$ are defined in Eq.(16) and Eq.(17), respectively.
Repeat from step 3 until convergence.
Stage 2: cross-domain translation training
Randomly initialize the parameters $\Theta_{E\circ G}$ of content encoder E^c, style encoder E^s, decoder G and parameters θ_G of discriminator D;
Randomly select two different domains $\mathcal{D}_A, \mathcal{D}_B, A, B \in [n]$. For each selected domain $\mathcal{D}_l$ where $l \in \{A, B\}$, get a mini-batch of data D_l satisfying $D_l \subset \mathcal{D}_l$ and $\left|D_l\right| = K$.
if *Training* **then**
  Update the parameters as follows:
  $\Theta_{E\circ G} \leftarrow \Theta_{E\circ G} - \eta\nabla_{\Theta}\Theta_{E\circ G}\mathcal{L}_{\text{E}\circ\text{G}}^{\text{total}}\left(D_A\right)$
  $\Theta_D \leftarrow \Theta_D - \eta\nabla_{\Theta_D}\mathcal{L}_D^{\text{total}}\left(D_A\right)$
where $\mathcal{L}_{\text{E}\circ\text{G}}^{\text{total}}\left(D_A\right)$ and $\mathcal{L}_D^{\text{total}}\left(D_A\right)$ are defined in Eq.(14) and Eq.(15), respectively.
Repeat from step 8 until convergence.

Table 2
Features of each dataset.

Art	Num.	Weather	Num.	Season	Num.
Photos	2853	Sunny	70601	Spring	1382
Monet	1074	Cloudy	45662	Summer	1512
Van Gogh	401	Foggy	357	Autumn	1606
Ukiyo-e	1433	Snowy	1252	Winter	993

It did not validate on datasets with various categories. Furthermore, it did not pay attention to the problem of multi-modality.

DosGAN. DosGAN (Lin et al., 2019) shares the similar idea of StarGAN (Choi et al., 2018) to achieve simultaneous training for multi-domains. It further introduces domain supervision, which uses domain-level information as supervision and pre-trains a classifier to predict which domain an image is from. The authors believe that the classifier should carry rich domain signal. Therefore, the output of the second-to-last layer of this classifier can be leveraged to extract the domain features of an image. Still, it has the same drawback in diversity as with Choi et al. (2018).

ComboGAN. Unlike StarGAN (Choi et al., 2018) and DosGAN (Lin et al., 2019), ComboGAN (Anoosheh et al., 2018) does not use domain labels to achieve simultaneous training for multi-domains. Instead, it uses n generators and discriminators for translations among n domains. Specifically, it divides each generator network in half, labeling each one as an encoder and decoder, respectively, and then assigns an encoder and decoder to each domain.

Since that these methods are designed for different purposes, we conduct comparisons in two scenarios. For simultaneous training, we compare our approach with StarGAN (Choi et al., 2018), DosGAN (Lin et al., 2019) and ComboGAN (Anoosheh et al., 2018). For multi-modality, we compare our method with BicycleGAN (Zhu et al., 2017b), DRIT (Lee et al., 2018) and MUNIT (Huang et al., 2018).

4.4. Evaluation metrics

We use the Fréchet inception distance and the LPIPS distance to evaluate the quality and diversity of the generated images.

LPIPS distance. Similarly to Zhu et al. (2017b), we use the Learned Perceptual Image Patch Similarity (LPIPS) metric (Zhang et al., 2018a) to measure translation diversity. The LPIPS distance is calculated by a weighted $\mathcal{L}_2$ distance between deep features of randomly-sampled translation results from the same input. It has been shown to correlate well with human perceptual similarity.

FID score. The Fréchet inception distance (FID) (Heusel et al., 2017) is a measure of similarity between two datasets of images. It was shown to correlate well with human judgement of visual quality and is most often used to evaluate the quality of samples of Generative Adversarial Networks. FID is calculated by computing the Fréchet inception distance between two Gaussians fitted to feature representations of the Inception network.

4.5. Qualitative evaluation

Qualitative comparisons of our approach with baselines for simultaneous multi-domain translation on the Season dataset are illustrated in Figure 7. The results produced by StarGAN (Choi et al., 2018) have clear artifacts. DosGAN (Lin et al., 2019) generates fewer artifacts than StarGAN; however, its results are still unpleasing and lack diversity for different seasons: in most cases, the translated spring and summer images are almost indistinguishable, and all four translated season images are nearly the same in the last row of DosGAN. ComboGAN (Anoosheh et al., 2018) generates better results in terms of both realism and diversity than DosGAN and StarGAN. However, it needs 8 generators and 4 discriminators to achieve conversion of four seasons between any two. The green boxes on the panel of ComboGAN in Figure 7 indicate the input images of ComboGAN: as ComboGAN translates an input image into the other $n-1$ domains, we use green boxes to highlight the positions (and thus the corresponding domains) of the input images, as well as to distinguish them from the images generated by ComboGAN.

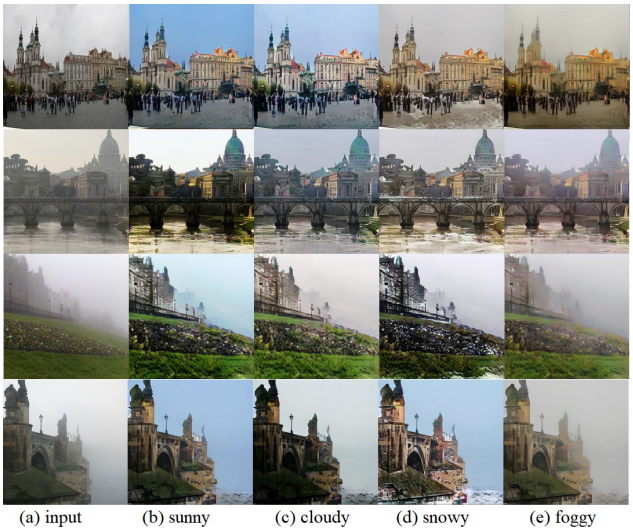

Figure 10: Results of our method on Weather. The images in the first row demonstrate that our method can handle images with complex and elaborate structures; the rest images show its potential for image defogging. Better look by zooming in.

Figure 11: Results of edges → shoes and edges → handbags translations. The first column shows the input and ground truth image. Each following column shows three random possible outputs from a method. Better look by zooming in.

Compared with the baseline methods, our approach generates high-quality images which are more photo-realistic and diverse. In terms of realism, the real input images are not better than the four images generated by our method; while in terms of diversity, the four generated images can be easily classified into corresponding seasons.

More results of our method on art, season and weather translations can be found in Figure 8, Figure 9 and Figure 10, respectively. For example, Figure 10 shows the results of our methods conducted on the Weather dataset. The images in the first row demonstrate that our method can handle images with complex and elaborate structures. The rest images show its potential to achieve well for image defogging tasks.

Qualitative comparisons of our approach with baselines for multi-modality translation are illustrated in Figure 11 on the edges → shoes and edges → handbags datasets.

4.6. Quantitative evaluation

We conduct the quantitative evaluation of the methods in terms of the realism and diversity of season cross-domain translation (Anoosheh et al., 2018).

Table 3
Performance in terms of the fooling rate and the season classification accuracy on the Season dataset. We conduct the user study to select results that are more realistic through pairwise comparisons and distinguish which season of an image is. The **best** and second best results are highlighted. For details refer to Section 4.6.

Method	Fooling Rate	Accuracy
Real photos	-	48.9%
Choi et al. (2018)	5.3%	41.3%
Lin et al. (2019)	27.2%	54.2%
Anoosheh et al. (2018)	**47.33%**	55.6%
Ours	37.8%	**65.8%**

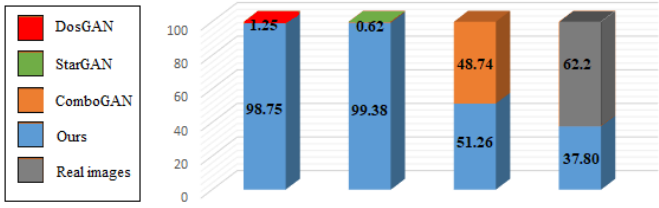

Figure 12: Realism preference results. We conduct a user study to ask people to select a more realistic one between our output and those of DosGAN (Lin et al., 2019), StarGAN (Choi et al., 2018) and ComboGAN (Anoosheh et al., 2018), as well as between ours and real image. The number indicates the percentage of preference on a pairwise comparison. We use the season translation for this experiment.

For realism, we conduct a user study using pairwise comparisons. Given a pair of images sampled from real images and translation outputs, users need to answer two questions: "Which image in this pair is more realistic?" and "Which season is this image?" They are given unlimited time to select their preferences. For each comparison, we randomly generate 100 questions and each question is answered by 30 different persons. Table 3 show the results of fooling rate and season classification accuracy. ComboGAN (Anoosheh et al., 2018) gets the highest fooling rate of 47.33% and ours rank the second highest. Notice that ComboGAN use several encoders and decoders to achieve this and our method only use one model. For the season classification accuracy, since many images in the Season dataset are too ambiguity to classify it into a certain season, the classification accuracy of the real images is like random guess, 48.9%. But the image-to-image translation methods tend to learn the general properties, the generated images are endowed with more distinguishable properties of certain season. DosGAN (Lin et al., 2019) and ComboGAN (Anoosheh et al., 2018) get higher classification accuracy than using real images, i.e., 54.2% and 55.6%; and ours achieve the highest accuracy of 65.8%, which means the domain-specific styles are better captured by our proposed method.

Table 4
Performance as the LPIPS and FID scores on the Season dataset. The **best** and second best results are highlighted. For details refer to Section 4.6.

Method	LPIPS	FID
Choi et al. (2018)	0.4273	221.7
Lin et al. (2019)	0.2503	145.3
Anoosheh et al. (2018)	0.4349	109.99
Ours	**0.4810**	**73.47**

Table 5
Diversity. We use the LPIPS and FID metrics to measure the diversity of generated images on the edges → shoes and edges → handbags translations. The **best** and second best results are highlighted.

Method	edges → shoes		edges → handbags	
	LPIPS	FID	LPIPS	FID
Zhu et al. (2017b)	0.2443	115.87	0.3180	184.56
Lee et al. (2018)	0.2631	**62.67**	0.3760	90.89
Huang et al. (2018)	**0.2652**	65.87	**0.3820**	91.43
Ours	0.2639	64.46	0.3759	**89.19**

We conduct another user study to ask people to select a more realistic one between our output images and those generated by StarGAN (Choi et al., 2018), DosGAN (Lin et al., 2019), ComboGAN (Anoosheh et al., 2018) or real images. Figure 12 plots the results of this pairwise realism preference study. It shows that ours are significantly more preferred as being realistic than DosGAN and StarGAN, as well as slightly more preferred than ComboGAN.

For diversity, similarly to BicycleGAN (Zhu et al., 2017b), we use the LPIPS metric to measure the similarity among images. Additionally, we implement the FID to acquire perceptual scores. We compute the distance between 1000 pairs of randomly sampled images translated from 100 real images. As shown in Table 4, our method achieves the lowest FID scores, which implies the best results in both high-level similarity and perceptual judgement, and the highest LPIPS scores, which means the most diverse results.

As BicycleGAN (Zhu et al., 2017b) need paired data for training, we evaluate the multi-modality performance on edges → shoes and edges → handbags translations. We use the LPIPS and FID metrics to compare our method with the existing state-of-the-art methods, i.e., BicycleGAN (Zhu et al., 2017b), DRIT (Lee et al., 2018), and MUNIT (Huang et al., 2018). As shown in Table 5, our method outperforms the supervised method BicycleGAN (Zhu et al., 2017b) and produce comparable results with other unsupervised methods DRIT and MUNIT (Lee et al., 2018; Huang et al., 2018).

4.7. Ablation study

The effect of domain supervision. As illustrated in Figure 3, the de-blurred images of DRIT (Lee et al., 2018) have different face contours from the original ones, which means that their attribute extractor has not only learned blur distortion pattern but also misrecognized some content representations such as face contour as attribute. This can be caused by the ambiguous and implicit disentanglement of content and style. Thus we introduce explicit domain constraint for disentanglement of content and style aiming to better utilize domain information by explicitly constraining the disentanglement learning.

Figure 13 shows the style-swapped reconstruction results

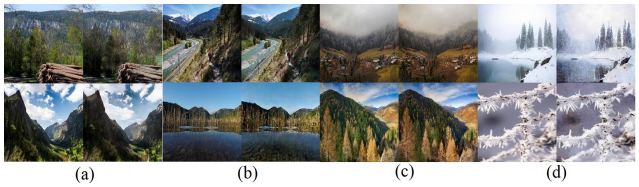

Figure 13: Results of intra-domain translation on Season dataset: (a) spring; (b) summer; (c) autumn; and (d) winter. The two columns in each panel represent the original input and the intra-domain translation result, respectively. Better look by zooming in.

Figure 14: Results of adapting our method for image deblurring. The images in the first row are blurred image generated by using the same method as in (Yu et al., 2018c) on the CUFS dataset (Wang and Tang, 2009); the second row is for de-blurred results of our method; The third row is ground truth. Better look by zooming in.

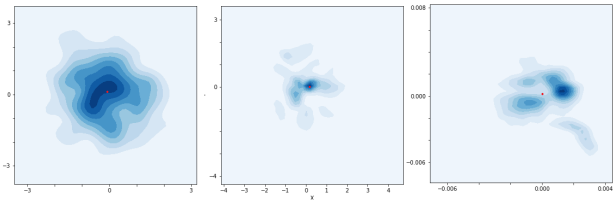

Figure 15: Comparing the prior Gaussian distribution $p(z)$ (left), the distribution $q_\phi(z)$ estimated by using MMD (middle), and that by KL (right). The red dots represent $(0,0)$. It clearly demonstrates that with the $q_\phi(z)$ from KL matches the prior Gaussian distribution $p(z)$ poorly, while $q_\phi(z)$ from MMD matches $p(z)$ significantly better.

of intra-domain translation. It shows that the pre-trained model could reconstruct style-swapped images from the same domain. To further validate that the domain supervision can help the explicit disentanglement learning of content and style, and thus the effectiveness of domain supervision, we adapt our method for image de-blurring tasks, and we can compare the results with those of DRIT (Lee et al., 2018) in Figure 3. The blurred images are generated using the same method as in Yu et al. (2018c) on the CUFS dataset (Wang and Tang, 2009). The results of image de-blurring after adding the proposed disentanglement constrained loss are shown in Figure 14. Compared with the results of DRIT in Figure 3, our generated images are more consistent with the original images, and the blur distortion are removed.

Furthermore, we also found that the perceptual loss (Johnson et al., 2016) can achieve similar disentangled constraint. The perceptual loss is based on perceptual similarity, which is often computed as the distance of two activated features in a pre-trained deep neural network between the output and the reference image:

$$\mathcal{L}_{\text{percep}} = \mathbb{E}\left[\sum_i \frac{1}{N_i}\left\|\phi_i\left(\mathrm{I}_{\text{gt}}\right) - \phi_i(\mathrm{I}_{\text{pred}})\right\|_1\right], \tag{18}$$

where ϕ_i donates the feature maps of the pre-trained VGG-19 network. However, the perceptual loss and disentanglement restrained loss constrain the learning of disentangled representations in different aspects. The former, which is at image level, forces the generated images sharing the same content with the input ones. The latter, which is at the collection level, restrains the style encoder from learning any content of images.

The measure of distribution distance. In the training phase, the style representations are constrained to match the prior Gaussian distribution, so that later we can generate several possible outputs by random sampling from this prior. Many measures can be used to estimate the distance between probability distributions. DRIT (Lee et al., 2018) and MUNIT (Huang et al., 2018) adopt the Kullback-Leibler (KL) divergence as the measure of distribution distance, which can be expressed in the distribution matching loss as

$$L_{\text{dm}} = \mathbb{E}\left[D_{\text{KL}}\left(p(z)|N(0,1)\right)\right], \tag{19}$$

where $D_{\text{KL}}(p|q) = -\int p(z)\log\frac{p(z)}{q(z)}\mathrm{d}z$.

Since the style representations are randomly sampled from the prior Gaussian distribution for multimodal outputs in the test phase, it is important to match the distribution of the style latent codes with the prior Gaussian distribution. However, researchers have noticed that the KL divergence can be too restrictive (Bowman et al., 2015; Sønderby et al., 2016; Chen et al., 2016b; Bińkowski et al., 2018). Sometimes it failed to learn any meaningful latent representation. Several methods (Bowman et al., 2015; Sønderby et al., 2016; Chen et al., 2016b) try to alleviate this problem, but do not completely solve the issue. Borgwardt et al. (2006) propose the MMD as a relevant criterion for comparing distributions based on the Reproducing Kernel Hilbert Space (RKHS). It is a framework to quantify the distance of two distributions by calculating all of their moments. It can be efficiently implemented by using the kernel trick:

$$\begin{aligned} D_{\text{MMD}}(q|p) = \mathbb{E}_{p(z),p(z')}\left[k\left(z,z'\right)\right] - 2\mathbb{E}_{q(z),p(z')}\left[k\left(z,z'\right)\right] \\ + \mathbb{E}_{q(z),q(z')}\left[k\left(z,z'\right)\right], \end{aligned} \tag{20}$$

where $k(\cdot,\cdot)$ can be any positive definite kernel, such as Gaussian kernel $k\left(z,z'\right) = e^{-\frac{\left\|z-z'\right\|^2}{2\sigma^2}}$. Therefore, the distance between distributions of two samples can be well-estimated by the distance between the means of the two samples mapped into an RKHS.

In this ablation study, we do not compare KL and MMD in the distribution matching loss $\mathcal{L}_{\text{dm}}$ (Eq.10), as it is not so reasonable to use the FID and LPIPS scores to measure the distribution from the perspective of diversity. Instead, to illustrate the advantage of using MMD over KL in a simple and intuitive way, we conduct experiments on the MNIST dataset (LeCun et al., 1998) and make the latent code have two dimensions for convenient visualization. It can be seen from Figure 15 that with KL, the obtained distribution matches the prior Gaussian distribution $p(z)$ poorly, while with MMD, it matches the prior significantly better. The results in Table 6 also demonstrate that MMD provides better reconstruction than KL.

For quantitative evaluation, Table 6 shows the reconstruction error and log likelihood of adopting KL and MMD in experiments, respectively. The reconstruction error indicates the quality of style encoding and the log likelihood represents.

Table 6
MMD versus KL on the MNIST dataset. MMD provides better reconstruction than KL.

Method	Reconstruction error
KL	0.04367
MMD	**0.03605**

5. Conclusion and future work

In this paper, we propose a unified framework for learning to generate diverse outputs with unpaired training data and allow simultaneous multi-domain translation through a single network. Furthermore, we also investigate how to extract domain information so as to utilize domain supervision and explicitly constrain the disentanglement of content and style. Qualitative and quantitative experiments on different datasets show that the proposed method outperforms or is comparable to the state-of-the-art methods. We also show the potential of our method for image de-blurring and image de-hazing. In the future, we will explore the feasibility

of extending our method to a more challenging task, multi-degradation image restoration.

6. acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No. 2018YFB1601102), and Shenzhen special fund for the strategic development of emerging industries (No. JCYJ20170412170118573). In addition, we would like to thank the anonymous reviewers for their constructive suggestions.

References

Anoosheh, A., Agustsson, E., Timofte, R., Van Gool, L., 2018. ComboGAN: Unrestrained scalability for image domain translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 783–790.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: International Conference on Machine Learning, pp. 214–223.

Berthelot, D., Schumm, T., Metz, L., 2017. BEGAN: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 .

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying MMD GANs. arXiv preprint arXiv:1801.01401 .

Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J., 2006. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22, e49–e57.

Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S., 2015. Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349 .

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016a. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems 29, 2016, pp. 2172–2180.

Chen, X., Kingma, D.P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., Abbeel, P., 2016b. Variational lossy autoencoder. arXiv preprint arXiv:1611.02731 .

Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., Choo, J., 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pp. 8789–8797.

Chu, W.T., Zheng, X.Y., Ding, D.S., 2017. Camera as weather sensor: Estimating weather information from single images. Journal of Visual Communication and Image Representation 46, 233–249.

Denton, E.L., et al., 2017. Unsupervised learning of disentangled representations from video, in: Advances in neural information processing systems, pp. 4414–4423.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: International Conference on Neural Information Processing Systems, pp. 2672–2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of Wasserstein GANs, in: Advances in Neural Information Processing Systems, 2017, pp. 5769–5779.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Computer Vision and Pattern Recognition, pp. 770–778.

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: Advances in Neural Information Processing Systems, pp. 6626–6637.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510.

Huang, X., Liu, M., Belongie, S.J., Kautz, J., 2018. Multimodal unsupervised image-to-image translation, in: Computer Vision - ECCV 2018 - 15th European Conference, Proceedings, Part III, pp. 179–196.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 5967–5976.

Johnson, J., Alahi, A., Li, F.F., 2016. Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, pp. 694–711.

Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J., 2017. Learning to discover cross-domain relations with generative adversarial networks, in: Proceedings of the 34th International Conference on Machine Learning, pp. 1857–1865.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 105–114.

Lee, H., Tseng, H., Huang, J., Singh, M., Yang, M., 2018. Diverse image-to-image translation via disentangled representations, in: Computer Vision - ECCV 2018 - 15th European Conference, Proceedings, Part I, pp. 36–52.

Lin, J., Liu, S., Xia, Y., Zhao, S., Qin, T., Chen, Z., 2019. Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. IEEE Trans. Pattern Analysis and Machine Intelligence .

Liu, A.H., Liu, Y., Yeh, Y., Wang, Y.F., 2018. A unified feature disentangler for multi-domain image translation and manipulation, in: Advances in Neural Information Processing Systems 31, NeurIPS 2018, pp. 2595–2604.

Liu, M., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems, pp. 700–708.

Lu, B., Chen, J.C., Chellappa, R., 2019. Unsupervised domain-specific deblurring via disentangled representations. arXiv preprint arXiv:1903.01594 .

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M., 2019. EdgeConnect: Generative image inpainting with adversarial edge learning, in: ICCV Workshop.

Schmidhuber, J., 2020. Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991). Neural Networks 127, 58 – 66.

Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O., 2016. Ladder variational autoencoders, in: Advances in neural information processing systems, pp. 3738–3746.

Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X., 2020. Unsupervised domain adaptive re-identification: Theory and practice. Pattern Recognit. 102, 107173.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 .

Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B., 2018a. High-resolution image synthesis and semantic manipulation with conditional GANs, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pp. 8798–8807.

Wang, X., Tang, X., 2009. Face photo-sketch synthesis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31, 1955–1967.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C., 2018b. ESRGAN: enhanced super-resolution generative adversarial networks, in: Computer Vision - ECCV 2018 Workshops, Proceedings, Part V, pp. 63–79.

Xiao, T., Hong, J., Ma, J., 2018. ELEGANT: Exchanging latent encodings with gan for transferring multiple face attributes, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–187.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection, in: Proceedings of IEEE International Conference on Computer Vision, pp. 1395–1403.

Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. DualGAN: Unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2849–2857.

Yu, A., Grauman, K., 2014. Fine-grained visual comparisons with local learning, in: Proceedings of IEEE International Conference on Computer Vision, pp. 192–199.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018a. Free-form image inpainting with gated convolution. arXiv preprint arXiv:1806.03589 .

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018b. Generative image inpainting with contextual attention. arXiv preprint arXiv:1801.07892 .

Yu, K., Dong, C., Lin, L., Loy, C.C., 2018c. Crafting a toolchain for image restoration by deep reinforcement learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2443–2452.

Zhang, C., Fu, H., Hu, Q., Cao, X., Xie, Y., Tao, D., Xu, D., 2020. Generalized latent multi-view subspace clustering. IEEE Trans. Pattern Anal. Mach. Intell. 42, 86–99.

Zhang, C., Fu, H., Hu, Q., Zhu, P., Cao, X., 2017. Flexible multi-view dimensionality co-reduction. IEEE Trans. Image Processing 26, 648–659.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018a. The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y., 2018b. Residual dense network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481.

Zhu, J., Krähenbühl, P., Shechtman, E., Efros, A.A., 2016. Generative visual manipulation on the natural image manifold, in: Computer Vision - ECCV 2016, Proceedings, Part V, pp. 597–613.

Zhu, J., Park, T., Isola, P., Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, ICCV 2017, pp. 2242–2251.

Zhu, J., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E., 2017b. Toward multimodal image-to-image translation, in: Advances in Neural Information Processing Systems 30, 2017, pp. 465–476.

Appendix

In this appendix, we show some additional multi-domain translation results of Art in Figure 16, Season in Figure 17 and Weather in Figure 18.

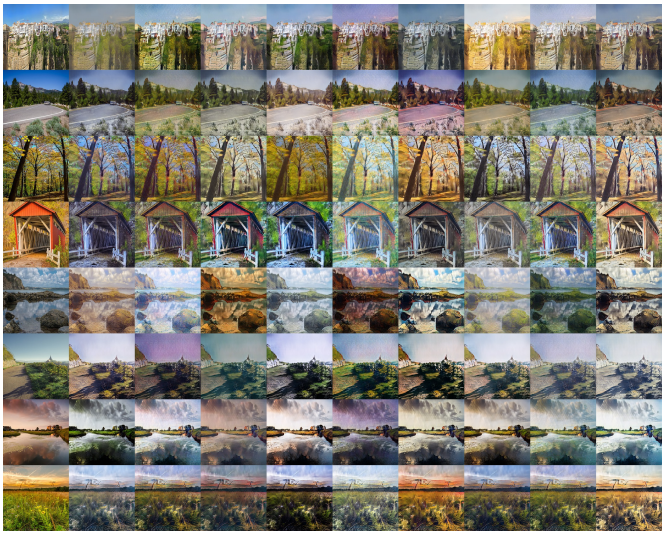

Figure 16: Art result. Better look by zooming in.

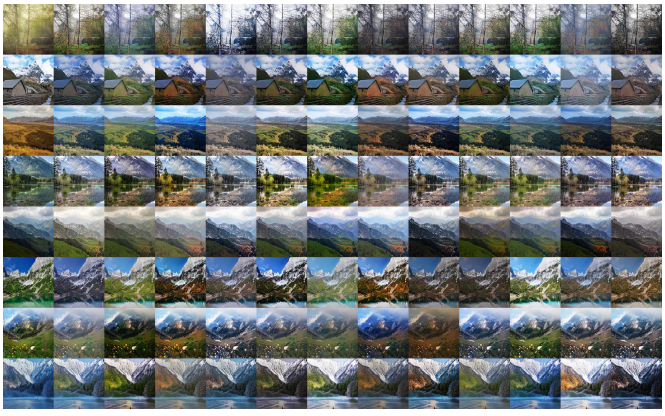

Figure 17: Season result. Better look by zooming in.

Figure 18: Weather result. Better look by zooming in.