Unpaired Image-to-Image Translation with Shortest Path Regularization

Shaoan Xie [1] , Yanwu Xu[2], Mingming Gong[4,3], Kun Zhang[1,3]
[1]Carnegie Mellon University
[2]Boston University
[3]Mohamed bin Zayed University of Artificial Intelligence
[4]University of Melbourne
shaoan@cmu.edu, yanwu@bu.edu, mingming.gong@unimelb.edu.au, kunz1@cmu.edu

Abstract

Unpaired image-to-image translation aims to learn proper mappings that can map images from one domain to another domain while preserving the content of the input image. However, with large enough capacities, the network can learn to map the inputs to any random permutation of images in another domain. Existing methods treat two domains as discrete and propose different assumptions to address this problem. In this paper, we start from a different perspective and consider the paths connecting the two domains. We assume that the optimal path length between the input and output image should be the shortest among all possible paths. Based on this assumption, we propose a new method to allow generating images along the path and present a simple way to encourage the network to find the shortest path without pair information. Extensive experiments on various tasks demonstrate the superiority of our approach. The code is available at https://github.com/MidPush/santa.

1. Introduction

Many important problems in computer vision can be viewed as image-to-image translation problems, including domain adaptation [22, 46], super-resolution [7, 66] and medical image analysis [2]. Let $\mathcal{X}$ and $\mathcal{Y}$ represent two domains, respectively. In unpaired image-to-image translation, we are given two collections of images from the two domains with distributions $\{P_{\mathcal{X}}, P_{\mathcal{Y}}\}$ without pair information. Our goal is to find the true conditional (joint) distribution of two domains $P_{\mathcal{Y}|\mathcal{X}}$ ($P_{\mathcal{X},\mathcal{Y}}$); with the true conditional (joint) distribution, we are able to translate the input images in one domain such that the outputs look like images in another domain while the semantic information of the input images is preserved. For example, given old human faces in one domain and young human faces in another domain, we want to learn the true joint distribution across the two domains. Then we can translate the old face into young face image while preserving the important information (e.g., the identity) of the input face (see Fig. 1).

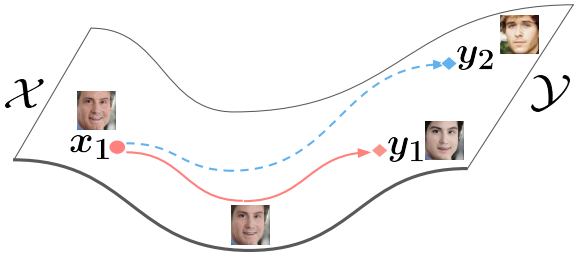

Figure 1. Illustration of our shortest path assumption. Almost existing methods only use two discrete domains $\{\mathcal{X}, \mathcal{Y}\}$. Instead, we consider the paths connecting two domains and assume that the optimal mapping generates shortest path, e.g., $x_1 \rightarrow y_1$ rather than $x_1 \rightarrow y_2$.

However, there can exist an infinite number of joint distributions corresponding to the given two marginal distributions [39]. It means that the problem is highly ill-posed and we may not derive meaningful results without any additional assumptions. As one of the most popular assumptions, cycle consistency [71] assumes that the optimal mapping should be one-to-one and has been achieving impressive performance in many tasks. However, the one-to-one mapping assumption can be restrictive sometimes [48], especially when images from one domain have additional information compared to the other domain. As an alternative, contrastive learning based method [48] has become popular recently. It assumes that the mutual information of patches in the same location of the input and translated image should be maximized. Then it employs the infoNCE loss [57] to associate corresponding patches and disassociate them from others. However, it has been shown that the choices of samples in the contrastive learning can have large impact on the results and most of recent image translation

methods are trying to improve it, e.g., negative sample mining [30, 58, 68] and positive sample mining [24].

Departing from existing unpaired image translation methods, we consider the paths connecting images in the first domain to images in the second domain and propose a shortest path assumption to address the ill-posed joint distribution learning problem. Specifically, we assume that the path, connecting one image in the first domain to its corresponding paired image in another domain, should be the shortest one compared to other paths (see Fig. 1). However, as we are only given images from two discrete domains, we have no access to real images on the paths. To address this problem, we first make a shared generating assumption to allow synthesizing images along the path: images are generated with the same function from the shared latent code space and the domain variable (as a surrogate of the domain-specific information). Then by changing the value of the domain variable, we obtain a path connecting the input and output image. In order to find a proper mapping, we need to minimize the path length, which can be formulated as expectation of the norm of the Jacobian matrix with respect to the domain variable. To reduce the cost of computing the high-dimensional Jacobian matrix computation, we further propose to use finite difference method to approximate the Jacobian matrix and penalize the squared norm. The contributions of this paper lie in the following aspects:

1. We propose a shortest path assumption to address the ill-posed unpaired image translation problem. We also conduct experiments to justify our assumption.

2. We propose a simple and efficient way to allow synthesizing images on the path connecting two domains and penalize the path length to find a proper mapping. Our method is the fastest one among all unpaired image translation methods.

3. Extensive experiments are conducted to demonstrate the superiority of our proposed method.

2. Related Work

Image-to-Image Translation With paired data, one can employ the Generative Adversarial Network (GAN) [19] to generate high-fidelity images while enforcing the consistency between the result and target [28]. With unpaired data, there can be an infinite number of mappings between two domains. To address this issue, cycle consistency is proposed to enforce the network to be a one-to-one mapping and is shown to achieve impressive visual performance [32, 35, 63, 71]. Cycle consistency becomes an important factor to image translation methods [9, 14, 31, 43, 54–56]. However, the one-to-one assumption admits multiple solutions [45] and may be too restrictive for some datasets [48]. As an alternative to cycle consistency, relationship preservation methods encourage relationships present in the input be analogously reflected in the output [5, 17, 21, 48, 58, 60, 70]. For example, DistanceGAN [5] enforces the network to preserve the distance order of two random inputs after the translation. Decent [60] encourages the density changes to be close for all patches. Shared latent space assumption is also a popular assumption in image translation which states that for any given pair of images, there exists a shared latent code in the latent space [27, 34, 36, 37]. For instance, UNIT [37] employs two weight-sharing VAEs for image reconstruction and translation. MUNIT [27] and DRIT [34] take a step further and assume that the representation can be decomposed into shared content code and domain specific style code. Multimodal image translation can be achieved by combing same content with different style codes. In contrast to UNIT, UFDN [36] employs a shared VAE and apply adversarial training to obtain domain invariant representations. Contrastive learning is gaining more attention in image-to-image translation recently. They assumes that the two corresponding patches in the input and output images should have larger mutual information that others [21, 29, 30, 48, 58, 70]. Recently multimodal [1, 25, 27, 34, 40, 41, 44, 47, 52, 65, 72] and multi-domain [6, 10, 11, 36], few-shot translation [38, 50] are also gaining wide popularity. [4] proposed a new setting where no domain label is available. [59] considers the case where content of images in two domains are not aligned.

Latent Space Interpolations HomoGAN [8] proposed to interpolate two different samples in the latent space and proposed to minimize the homomorphic gap between the latent space and the attribute space for the face attribute translation task. DLOW [18] applies weighted adversarial training on the intermediate domains for the domain adaptation task. CoMoGAN [49] proposes the functional instance normalization layer to help continuous mapping from source domain with the guidance of the physical models, e.g., a tone mapping for continuous translation from day to night. The good empirical performances of HomoGAN, DLOW, and ComoGAN highlight the effectiveness of intermediate domains, but they all need additional supervision (e.g., attribute or class labels), which makes them unsuitable for our task. [3, 51, 62] propose to minimize the arc length to find meaningful interpolations of unconditional generative models given two fixed random noises. For multi-modal and multi-domain image translation, we may interpolate between two style codes within the same domain [11, 27, 40], but they do not support interpolation between two different domains.

Optimal Transport Our method is also deeply connected to the Monge problem in optimal transport: $\inf_f \int_{\mathcal{X}} c(x, f(x))p(x)dx$, where $f(x) \in \mathcal{Y}$, and the Kantonorovich formulation: $\inf_\pi \int_{\mathcal{X},\mathcal{Y}} c(x, y)\pi(x, y)dxdy$,

where π is the joint distribution measure. One important problem is to define the cost function c. Many methods consider the cost $c(x, y) = |x - y|$ and $c(x, y) = |x - y|^2$. Unfortunately, these cost functions may not be suitable for our image-to-image translation task since the Euclidean distance in image space across two domains may not be meaningful. If labels are given, one may consider computing the cost in the feature space [13, 15, 61]. [16] uses different cost functions c for different image translation tasks. Optimal transport has also been applied in finding the correspondence between two sets of images [53, 67]. However, they are also not applicable in our setting since there is no guarantee that the data in two domains are paired.

3. Shortest Path Regularized Unpaired Image Translation

In this section, we first introduce the shortest path assumption. Then, we present our generative model that builds the path connecting the two domains. Finally, we give the exact formulation of our shortest path regularization and its efficient approximation.

3.1. Shortest Path Assumption

Given samples from two marginal distributions $\{x_i\} \sim P_{\mathcal{X}}$ and $\{y_j\} \sim P_{\mathcal{Y}}$, our goal is to infer the true conditional distribution $P_{\mathcal{Y}|\mathcal{X}}$ or the joint distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}} = P_{\mathcal{Y}|\mathcal{X}} P_{\mathcal{X}}$. Since there can be infinite number of possible joint distributions that can yield the given marginals, it is impossible to infer the true joint distribution without additional assumptions.

To tackle this issue, we make the *shortest path assumption*. Specifically, we assume that for any image $x_1 \in \mathcal{X}$, the path from x_1 to its true paired image $y_1 \in \mathcal{Y}$ is the shortest one on the manifold. As shown in Figure 1, there can be many paths of translating x_1 to the domain $\mathcal{Y}$, e.g., $x_1 \to y_1$ and $x_1 \to y_2$. We assume the optimal path $x_1 \to y_1$ is the shortest one, i.e., the length of curve γ_1 is less than the lengths of other curves including γ_2. The intuition is that given pair (x_1, y_1), they share the same latent content z despite in different domains. When connecting x_1 to another point y_2, they share less information and the content information in x_1 may not be kept in x_2, which is undesirable in image-to-image translation. Therefore, transforming x_1 to y_1 should be easier than transforming x_1 to y_2. As a consequence, the path $x_1 \to y_1$ should be shorter than the path $x_1 \to y_2$ since no additional latent change is needed. We provide experimental justification of this assumption in Section 4.3.

3.2. Building the Path

Motivated by the shortest path assumption, we would like to build paths that connecting two domains. However, most existing image translation methods train a single network G to translate $x \in \mathcal{X}$ into another domain, i.e., $G(x) \in \mathcal{Y}$ [48, 71], which is incapable of generating images along the path. We first present a way to allow networks synthesizing the paths by making assumption on the data generation process.

We assume that images from different domains are generated with the same function G^* from a latent code space $\mathcal{Z}$ which is shared across domains, and a continuous domain variable θ. In other words, a ground truth pair data $(x, y) \sim P_{\mathcal{X},\mathcal{Y}}$ are generated as

$$x = G^*(z, 0), y = G^*(z, 1),$$

where $z \in \mathcal{Z}$ is the shared latent code. Now, we need to find a shared latent space first and learn the unknown optimal mapping G^*.

Shared Latent Space. Given any two images $x \sim P_{\mathcal{X}}$ and $y \sim P_{\mathcal{Y}}$, we employ a shared encoder E to extract the latent codes:

$$z_x = E(x), \quad z_y = E(y).$$

To encourage the latent codes lie in the same space, we propose to match their distributions $q(z_x), q(z_y)$ with Kullback-Leibler (KL) Divergence

$$\mathcal{L}_{\text{kl}} = \text{KL}(q(z_x)||p(z)) + \text{KL}(q(z_y)||p(z)), \tag{1}$$

where $p(z)$ is the prior distribution and we assume it as isotropic Gaussian $\mathcal{N}(0, I)$.

Reconstruction $G(z_x, 0)$. After obtaining the shared latent space, we need to build the paths from first domain to another domain. To this end, we train the decoder G to reconstruct the input image x

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{x \sim P_{\mathcal{X}}} \|G(z_x, 0) - x\|_1, \tag{2}$$

where $G(z_x, 0)$ serves as the starting point of the path. The combination of the reconstruction loss $\mathcal{L}_{\text{rec}}$ and the KL divergence loss $\mathcal{L}_{\text{kl}}$ can be viewed as a variational autoencoder [33], which is famous for its powerful inference ability of the latent variable z_x.

Translation $G(z_x, 1)$. We now train the decoder to find an ending point of the path, i.e., translating the images $x \in \mathcal{X}$ to another domain $\mathcal{Y}$. We would like to have $\tilde{x} = G(z, 1)$ to look like images in domain $\mathcal{Y}$. To match the distribution between $P_{G(z_x,1)}$ and $P_{\mathcal{Y}}$, we adopt the generative adversarial network (GAN) [71]. In detail, we employ a discriminator D and D is trained to distinguish the generated images $G(z, 1)$ from the real images $y \in \mathcal{Y}$, and the encoder and decoder E, G are trained to fool D. The loss of GAN is defined as follows:

$$\mathcal{L}_{\text{gan}} = \mathbb{E}_{x \sim P_{\mathcal{X}}, y \sim P_{\mathcal{Y}}}[(D(\tilde{x}) - 1)^2 + D(y)^2]. \tag{3}$$

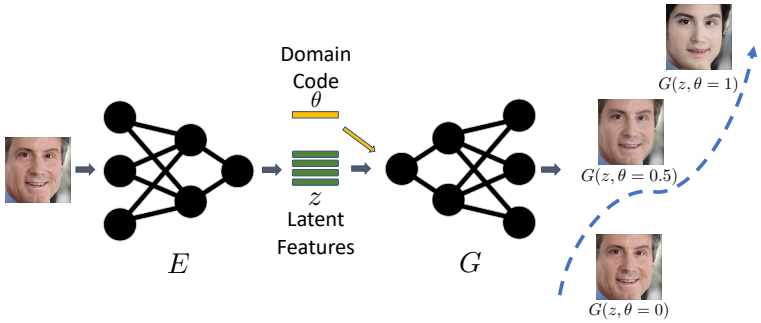

Figure 2. The diagram of our model. We employ a shared encoder E to extract latent code from the shared space . Then we assume that images in two domains are generated with the same function and we use a shared decoder G to perform reconstruction and translation. We can generate images along the path via changing the value of domain code θ. However, without any regularization, the path may be very long and leads to distortion in the translation results. Therefore, we apply our path length regularization to find a proper mapping between two domains.

3.3. Path Length Regularization

Now we can synthesize the images along the path by moving θ from 0 to 1 and we obtain a path $\gamma_x : [0, 1] \rightarrow G(z_x, \theta)$, where the start $G(z_x, 0)$ is the reconstruction of the input image and the end $G(z_x, 1)$ is the translation result. Without any regularization, the translation $G(z_x, 1)$ can be mapped to any image in domain $\mathcal{Y}$ which means that the semantic information of input x can be distorted or even discarded. Therefore, we propose to regularize the path length of γ_x according to our shortest path assumption.

Given a curve $\gamma : [a, b] \rightarrow \mathcal{M}$, the arc length of the curve γ [20] is defined as $L(\gamma) = \int_a^b \|\gamma'(\theta)\| d\theta, where \gamma'(\theta) = \frac{d}{d\theta}\gamma(\theta)$ is the velocity of the curve. In our case, we have the arc length of

$$L(\gamma_x) = \int_0^1 \|J_\theta\| d\theta,$$

where $J_\theta = \frac{d}{d\theta} G(z_x, \theta) = \frac{d}{d\theta}\gamma_x(\theta)$ is the Jacobian matrix of the decoder $G(z_x, \theta)$ with respect to θ.

To find the minima of the path length $L(\gamma_x)$, we optimize the energy functional of the curve as follows:

$$\gamma^{\text{shortest}} = \arg\min \frac{1}{2} \int_a^b \|J_t\|^2 d\theta. \tag{4}$$

The minima γ^{shortest} also has constant speed parameterization [20], which implies that we can obtain a smoothly changing path by minimizing the energy functional.

However, the cost of computing the Jacobian matrix J_θ is prohibitively expensive due to the high dimensionality of the network output $G(z, \theta)$. To address this issue, we propose to use the classical central finite difference method to approximate the Jacobian vector

$$\hat{J}_\theta = \frac{G(z, \theta + \frac{h}{2}) - G(z, \theta - \frac{h}{2})}{h}, \tag{5}$$

where h is a hyper-paramter that control the granularity of the estimated Jacobian matrix. In practice, we randomly sample $h \sim U(0.1, 0.2)$.

Multi-layer Feature Path Length Inspired by the multi-layer patchwise learning in CUT [48], we can also penalize the path length on multi-layer features. Specifically, we select L layers of interest and can obtain a set of features of two domains as $G^l(z_x, 0), G^l(z_x, 1)$, where l is the l th chosen layer of the decoder G. Then we can compute the Jacobian matrix at layer l as

$$\hat{J}^l_\theta = \frac{G^l(z, \theta + \frac{h}{2}) - G^l(z, \theta - \frac{h}{2})}{h}.$$

Given features from L layers, our path length regularization loss is defined as

$$\mathcal{L}_{\text{path}} = \mathbb{E}_{x \sim P_{\mathcal{X}}} \mathbb{E}_{\theta \sim U(0,1)} \frac{1}{L} \sum_{l=1}^{L} \|\hat{J}^l_\theta\|^2. \tag{6}$$

3.4. Additional Regularization

We additionally introduce the reconstruction (identity) loss on the domain $\mathcal{Y}$ to further regularize the networks, which is commonly used by previous translation methods [48, 71]. The identity loss is defined as follows

$$\mathcal{L}_{\text{idt}} = \mathbb{E}_{y \sim P_{\mathcal{Y}}} \|y - G(z_y, 1)\|. \tag{7}$$

3.5. Full Objective

Our full objective is as follows.

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{gan}} + \lambda_{\text{idt}}\mathcal{L}_{\text{idt}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}}, \tag{8}$$

where $\lambda_{\text{idt}}, \lambda_{\text{rec}}, \lambda_{\text{kl}}, \lambda_{path}$ are hyper-parameters that balances different loss functions.

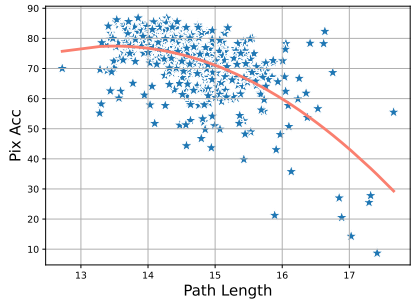

Figure 3. Justification of our shortest path assumption. We observe that there is strong negative correlation between the path length and the pixel accuracy. Therefore, we need to minimize the path length to find a proper mapping.

We aim to solve

$$E^*, G^* = \arg\min_{E,G}\max_{D}\mathcal{L}_{\text{full}}. \tag{9}$$

In the inference stage, given an input image x from domain $\mathcal{X}$, we can get its translation in domain $\mathcal{Y}$ by first encoding it to the latent space by the encoder E and then using the generator G with $\theta = 1$ to decode the latent information into the translated image, i.e., $y = G(E(x), 1)$.

3.6. Discussion

As introduced above, patch mutual information maximization [21, 48, 58] encourages that the representations of patches on the same location, from two domains, should be close. However, patches on the same location in two domains can have totally different meanings and the large domain difference may lead to failure [21]. Our method can also be viewed as encouraging the features of two domains to be close by penalizing the Jacobian norm. But the major difference is that we are encouraging the features of two close domains to be close. We are not enforcing the features of $G(z, 0)$ and translation $G(z, 1)$ to be close brutely. We are encouraging features of $G(z, \theta_1)$ and $G(z, \theta_1 + \epsilon)$ to be close and ϵ is a small value while allowing necessary changes between the features of $G(z, 0)$ and $G(z, 1)$. Through the lens of intermediate domains, our model is able to preserve semantic information in the input while allowing necessary changes.

4. Experiments

We test across several datasets. We first present the implementation details, dataset, metrics and baseline methods. We then provide justification of our assumption on real dataset. We then compare against the baseline methods quantitatively and qualitatively. We finally perform ablation studies and analyze our method.

4.1. Implementation Details

We mostly follow the setting of [48, 71]. In detail, we use LSGAN objective [42] and a 9-resnet-block based Generator. Since we have to implement the shared generating process assumption, we use the first 4 resnet blocks as the encoder E and the rest 5 resnet blocks as the decoder G. Our decoder G needs to generate images in different domains to generate the path by moving the value of θ. Therefore, we adopt the AdaIN [26] to introduce the influence of θ. We use a domain embedding to generate the parameters for AdaIN with input θ and then we perform classical forward process to get images along the path. The total number of generator is 11.428M. As a reference, recent methods [24, 48, 58] use a generator 11.378M and an additional MLP with 0.560 M. We use the Adam optimizer with learning rate 2e-4. We set $\lambda_{\text{rec}} = 5, \lambda_{kl} = 0.01, \lambda_{idt} = 5$. For λ_{path}, we choose from [0.05, 0.1, 0.2]. We train all tasks with 400 epochs.

4.2. Datasets, Evaluation Metric and Baselines

Datasets We conduct experiments on the benchmark datasets: label→city [12], cat→dog [11], horse→zebra [71]. Then we further test the model on winter→summer dataset. To further testify the effectiveness of our method, we build a high-resolution aging dataset. We apply super-resolution model to images in the public UTKFace dataset [69] and split the dataset to young domain and old domain according to ages. The dataset contains around 1500 training and 500 testing images.

Evaluation All tasks are trained at 256×256 resolution. We have two main goals for image translation tasks: semantic information preservation and high visual quality. Since label→city have ground truth labels, we can use it measure the semantic information preservation ability. Following [48], we use a pretrained DRN segmentation network [64] to map the generated city photos to segmentations. Then we compute mAP, PixAcc and clsAcc by comparing with the input segmentation images. We use the evaluation script * provided by [30]. It is also worth noting that the evaluation protocols of MoNCE and QS-Attn are different from [30]. So, we re-evaluate their results with the script. As for other datasets that don't have ground truth pairs, we adopt the commonly used Frechet Inception Distance (FID) to measure the visual quality of the generated images. FID computes the divergence between the generated images and the real images in the feature space.

4.3. Justification of Our Assumption

Although our shortest path assumption is quite intuitive, it is still necessary to test whether this assumption holds in real dataset. Therefore, we use the paired dataset label→city to justify our assumption. We first run our

*https://github.com/jcy132/Hneg_SRC/tree/main/Single-modal

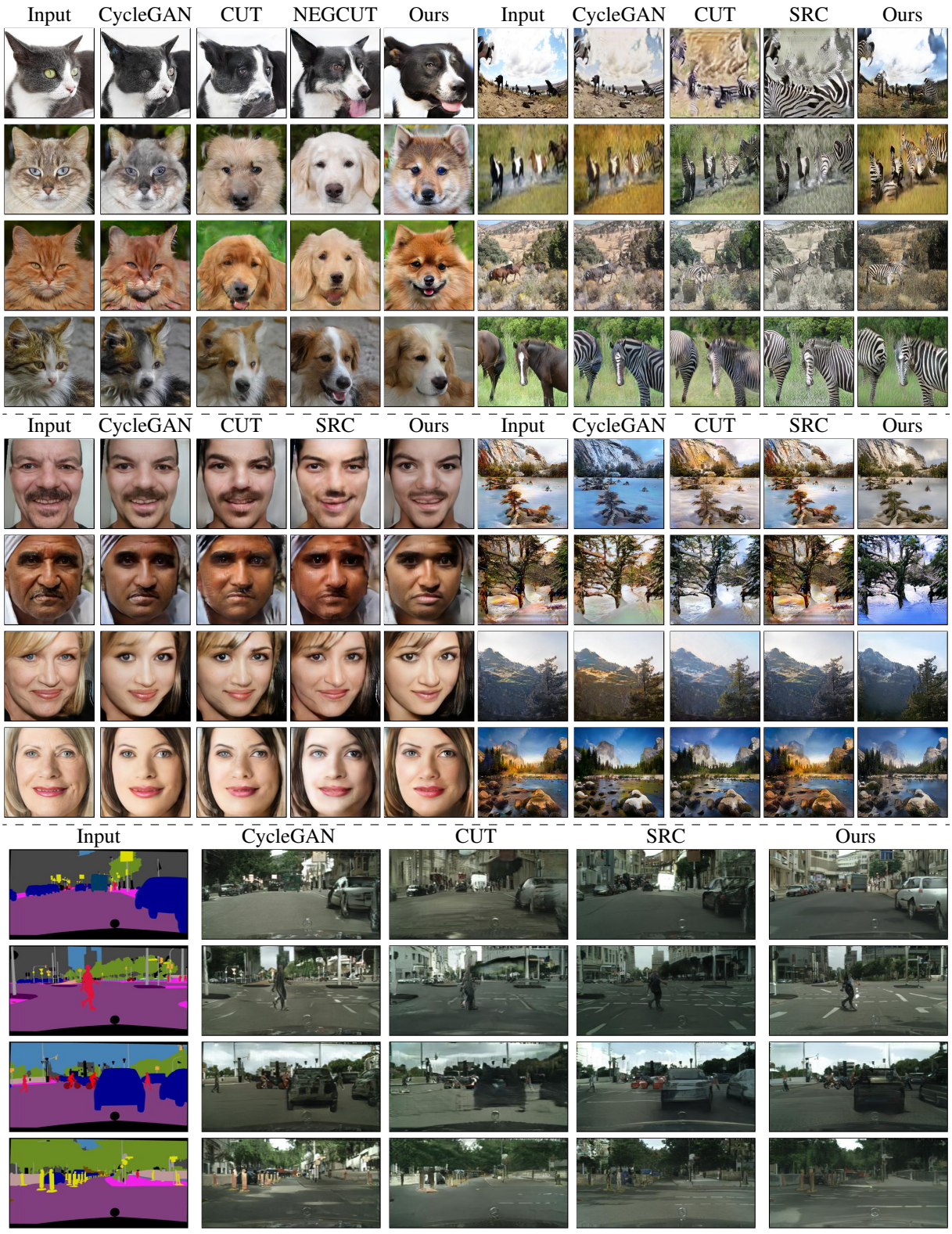

Figure 4. The generated samples on five tasks: cat→dog, horse→zebra, old→young, winter→summer and label→city. Cycle consistency [71] sometimes can be over-restrictive, e.g., the cat→dog task. Contrastive learning based methods (CUT [48], SRC [30]) sometimes are less-regularized, e.g., the horse→zebra tasks, which may be caused by the sample selection problem. By contrast, our method considers the paths connecting the two domains and learn to generate high-quality images while preserving important problems.

Table 1. Quantitative results on benchmark datasets

Method	Label→City				Cat → Dog	Horse → Zebra	Speed
	mAP ↑	pAcc ↑	cAcc ↑	FID ↓	FID ↓	FID ↓	Sec/iter ↓
CycleGAN [71]	20.4	55.9	25.4	76.3	85.9	77.2	0.171
CUT [48]	24.7	68.8	30.7	56.4	76.2	45.5	0.138
NEGCUT [58]	27.6	71.4	35.0	48.5	55.9	39.6	0.275
MoNCE [68]	26.4	72.4	32.5	54.7	-	41.9	0.231
QS-Attn [23]	27.8	72.3	34.4	50.2	80.0	42.3	0.182
SRC [30]	29.0	73.5	35.6	46.4	-	**34.4**	0.139
Ours	**31.0**	**73.6**	**37.4**	**46.1**	**52.1**	36.2	**0.136**

Table 2. Quantitative results on additional datasets.

Method	old→young	winter→summer
	FID ↓ ↓	FID ↓
CycleGAN [71]	43.5	75.1
CUT [48]	44.2	80.3
NEGCUT [58]	45.8	75.8
MoNCE	42.8	78.2
QS-Attn [23]	45.2	77.2
SRC [30]	42.7	71.6
Ours	**41.9**	**70.9**

Figure 5. Samples of interpolations without (first and third) and with our proposed regularization (second and fourth). Without our path length regularization, the model may suddenly change the output. By contrast, our method generates images smoothly along the path.

model without enforcing the path length regularization, i.e., we set $\lambda_{\text{path}} = 0$. Then we use this model to generate city photos from the input segmentation labels. For each generated city photo, we feed it into the DRN network and compare against the input label. So, we have the pixel accuracy for each image. We also compute the path length by traversing θ from 0 to 1 with 0.1 interval. The Pearson Correlation between path length and Accuracy on cityscapes dataset is -0.53 with p-value 4e-38. The very small p-value indicates that the correlations between path lengths and the performances are statistically significant. We present the scatter plot in Fig. 3 and we can observe a strong negative correlation between the path length and the pixel accuracy. This result suggest that we need to minimize the path length to improve the pixel accuracy (or find a proper mapping), which aligns with our shortest path assumption. Therefore, we argue that our shortest path assumption holds in real dataset.

4.4. Comparison with Baselines

We present the quantitative results in Table. 1 and 2. We observe that our method achieves best results in four out of five tasks. In particular, our method has achieved a very high mAP 31.0 on the label→city dataset while the previous best method can only achieve 29.0. The encouraging results suggest that our method is able to preserve semantic information in the input image. As for the cat→dog dataset, our method achieves best FID and NEGCUT [58] achieves the second best performance. It is worth noting that our method is the fastest one and only needs half of the training time of NEGCUT.

We provide the generated samples in Fig. 4. For visualization purpose, we only compare with the classical CycleGAN [71], CUT [48] and the best baseline model for each dataset. On the cat→dog task, we observe that CycleGAN fails in translating the cat images into dog images, implying that the cycle consistency may be over restrictive on such tasks. we can observe that NEGCUT and our methods are able to generate realistic dog images from the cat images. However, our shortest path regularization encourages the model to preserve more information about the input. On the horse→zebra dataset, SRC [30] model achieves best FID 34.4 and our model achieves the second best FID as 36.2. We can find that SRC generates more distortion than our method. For example, it generates unncessary zebra strips on the sky and the grass.

4.5. Ablation Study

We conducted ablation experiments to verify the effectiveness of each module in our loss function. The quantitative results are provided in Table. 3. We provide qualitative ablation results in the supplementary. We observe that the baseline method with only $\mathcal{L}_{\text{gan}}\&\mathcal{L}_{\text{idt}}$ has the worst per-

Table 3. Quantitative results of ablation study. Config D,E,F are based on config C.

Config	Settings				label→city				old→young
	$\mathcal{L}_{\text{gan}}$&$\mathcal{L}_{\text{idt}}$	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{kl}}$	$\mathcal{L}_{\text{path}}$	**mAP** ↑	**pAcc** ↑	**cAcc** ↑	FID ↓	FID ↓
A	✓				23.1	61.6	29.2	48.1	46.3
B	✓	✓			26.0	67.0	32.7	46.1	44.6
C	✓	✓	✓		28.3	70.9	34.8	47.8	44.0
D	✓	✓	✓	✓	31.0	73.6	37.4	46.1	41.9
E	Euclidean Distance Cross Domain				29.5	71.8	35.6	47.7	44.5
F	VGG Distance Cross Domain				22.8	65.5	29.0	49.7	50.0

Table 4. Ablation study on interpolation performance. With the path length regulariztaion, our method is able to generate images that along the path, which is proved by the lowest value of FID.

Config	Settings				old→young
	$\mathcal{L}_{\text{gan}}$&$\mathcal{L}_{\text{idt}}$	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{kl}}$	$\mathcal{L}_{\text{path}}$	Interpolation FID ↓
A	✓				86.77
B	✓	✓			39.33
C	✓	✓	✓		41.27
D	✓	✓	✓	✓	28.62

formance. It suggests that we need further regularization to find a proper mapping between two domains. Adding $\mathcal{L}_{\text{rec}}$ serves a good regularization as the mAP is increased and FID is decreased. Then we add $\mathcal{L}_{\text{kl}}$ to the model and we can find that it also helps semantic preservation. However, without our path length regularization, the paths are not well regulalrized and may suffer label flipping on the label→city dataset. After we apply our regularization, the mAP jumps from 28.3 to 31.0 and the pixel Acc also increases from 70.9 to 73.6. The encouraging improvement highlight the importance to regularize the paths. In addition, our path length regularization also help matching the distribution as demonstrated by the FID improvement from 44.0 to 41.9.

An interesting question is that do we really need intermediate domains since we are enforcing our path length regularization on multi-layer features. So, we build method E based on method C. We can notice that it brings small improvement over method C on label→city task but it hurts the FID on the old→young task. The comparison between method D and E demonstrate that we need to use intermediate domains as the domain difference can be quite large.

We also explore the option that using the commonly used pretrained VGG model to extract meaningful features so we can directly minimizing the difference across domains. However, the results of method F are worse than the base method C on both label→city task and old→young task. The performance degradation is expected as the VGG is only trained on ImageNet dataset. However, the images in label→city and old→young datasets lie out of the domain of ImageNet classes. Therefore, using VGG to extract features may not serve as a good way to preserve semantic information on these two datasets.

In addition, we further examine the effectiveness of our regularization on the images along the path for old→young dataset in Table. 4. Specifically, we apply the super-resolution model to the faces that are labeled between young and old. The intermediate domain contains 12405 images. If we are able to generate the true images along the path, the FID between our interpolations and the real faces should be low. For each testing image, we generate five images by traversing the value of θ in [0.2, 0.4, 0.6, 0.8]. So we generate 2000 images for each config. We present samples of interpolation in Fig. 5.

5. Conclusions, Limitation and Future Work

In this paper, we have proposed the shortest path assumption where the path connecting two corresponding points in the two domains is the shortest one. Then we propose a generative model to allow generating the images along the path. Finally, we propose an efficient way to enforce the shortest path constraint by regularizing the Jacobian of generators. We have conducted thorough experiments on various benchmarks and the high quality of generated images demonstrates the effectiveness of our method.

Nevertheless, one main limitation of our method could be: our discrete approximation of the path length may cause some inaccuracies in the estimation and better approximation methods could be sought in future work. Another interesting direction would be extending our method to multi-domain image-to-image translation.

Acknowledgement

This project was partially supported by the National Institutes of Health (NIH) under Contract R01HL159805, by the NSF-Convergence Accelerator Track-D award 2134901, by a grant from Apple Inc., a grant from KDDI Research Inc, and generous gifts from Salesforce Inc., Microsoft Research, and Amazon Research. MG was supported by ARC DE210101624.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018. 2

[2] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020. 1

[3] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017. 2

[4] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. *arXiv preprint arXiv:2006.06500*, 2020. 2

[5] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in neural information processing systems*, pages 752–762, 2017. 2

[6] Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan. Multi-marginal wasserstein gan. In *Advances in Neural Information Processing Systems*, pages 1774–1784, 2019. 2

[7] Shuaijun Chen, Zhen Han, Enyan Dai, Xu Jia, Ziluan Liu, Liu Xing, Xueyi Zou, Chunjing Xu, Jianzhuang Liu, and Qi Tian. Unsupervised image super-resolution with an indirect supervised path. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 468–469, 2020. 1

[8] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2408–2416, 2019. 2

[9] Yu-Jie Chen, Shin-I Cheng, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Vector quantized image-to-image translation. In *European Conference on Computer Vision*, pages 440–456. Springer, 2022. 2

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 2, 5

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[13] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[14] Yusuf Dalva, Said Fahri Altındiş, and Aysegul Dundar. Vecgan: Image-to-image translation with interpretable latent directions. In *European Conference on Computer Vision*, pages 153–169. Springer, 2022. 2

[15] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018. 3

[16] Emmanuel de Bézenac, Ibrahim Ayed, and Patrick Gallinari. Optimal unsupervised domain translation. *arXiv preprint arXiv:1906.01292*, 2019. 3

[17] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 2

[18] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2477–2486, 2019. 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[20] A Ben Hamza and Hamid Krim. Geodesic matching of triangulated surfaces. *IEEE transactions on image processing*, 15(8):2249–2258, 2006. 4

[21] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 746–755, 2021. 2, 5

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 1

[23] Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18291–18300, 2022. 7

[24] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017. 2, 5

[25] Qiusheng Huang, Zhilin Zheng, Xueqi Hu, Li Sun, and Qingli Li. Bridging the gap between label-and reference-based synthesis in multi-attribute image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14628–14637, 2021. 2

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceed-*

ings of the IEEE international conference on computer vision, pages 1501–1510, 2017. 5

[27] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 2

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[29] Zhiwei Jia, Bodi Yuan, Kangkang Wang, Hong Wu, David Clifford, Zhiqiang Yuan, and Hao Su. Semantically robust unpaired image translation for data with unmatched semantics statistics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14273–14283, 2021. 2

[30] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18260–18269, 2022. 2, 5, 6, 7

[31] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 2

[32] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017. 2

[33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[34] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[35] Wallace Lira, Johannes Merz, Daniel Ritchie, Daniel Cohen-Or, and Hao Zhang. Ganhopper: Multi-hop gan for unsupervised image-to-image translation. In *European conference on computer vision*, pages 363–379. Springer, 2020. 2

[36] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in neural information processing systems*, pages 2590–2599, 2018. 2

[37] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 2

[38] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019. 2

[39] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *Advances in neural information processing systems*, 29:469–477, 2016. 1

[40] Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10785–10794, 2021. 2

[41] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145*, 2018. 2

[42] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 5

[43] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 3693–3703, 2018. 2

[44] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018. 2

[45] Nikita Moriakov, Jonas Adler, and Jonas Teuwen. Kernel of cyclegan as a principle homogeneous space. *arXiv preprint arXiv:2001.09061*, 2020. 2

[46] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 1

[47] Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7860–7869, 2020. 2

[48] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. *arXiv preprint arXiv:2007.15651*, 2020. 1, 2, 3, 4, 5, 6, 7

[49] Fabio Pizzati, Pietro Cerri, and Raoul de Charette. Comogan: continuous model-guided image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14288–14298, 2021. 2

[50] Fabio Pizzati, Jean-François Lalonde, and Raoul de Charette. Manifest: Manifold deformation for few-shot image translation. *arXiv preprint arXiv:2111.13681*, 2021. 2

[51] Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323, 2018. 2

[52] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S Huang. Towards instance-level image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3683–3692, 2019. 2

[53] Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 3

[54] Yuda Song, Hui Qian, and Xin Du. Multi-curve translator for real-time high-resolution image-to-image translation. *arXiv preprint arXiv:2203.07756*, 2022. 2

[55] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[56] Justin Theiss, Jay Leverett, Daeil Kim, and Aayush Prakash. Unpaired image translation via vector symbolic architectures. *arXiv preprint arXiv:2209.02686*, 2022. 2

[57] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018. 1

[58] Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. *arXiv preprint arXiv:2108.04547*, 2021. 2, 5, 7

[59] Shaoan Xie, Mingming Gong, Yanwu Xu, and Kun Zhang. Unaligned image-to-image translation by learning to reweight. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14174–14184, 2021. 2

[60] Shaoan Xie, Qirong Ho, and Kun Zhang. Unsupervised image-to-image translation with density changing regularization. *Advances in Neural Information Processing Systems*, 35:28545–28558, 2022. 2

[61] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4403, 2020. 3

[62] Mengyu Yang, David Rokeby, and Xavier Snelgrove. Mask-guided discovery of semantic manifolds in generative models. *arXiv preprint arXiv:2105.07273*, 2021. 2

[63] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[64] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[65] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *Advances in Neural Information Processing Systems*, pages 2990–2999, 2019. 2

[66] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018. 1

[67] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021. 3

[68] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18280–18290, 2022. 2, 7

[69] Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 5

[70] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16407–16417, 2021. 2

[71] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2, 3, 4, 5, 6, 7

[72] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 2