

Image-to-Image Translation with Multi-Path Consistency Regularization

Jianxin Lin^{1*}, Yingce Xia^{3*}, Yijun Wang², Tao Qin³ and Zhibo Chen^{1†}

¹CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System,

University of Science and Technology of China

²University of Science and Technology of China

³Microsoft Research Asia

{linjx, wyjun}@mail.ustc.edu.cn, {yingce.xia, taoqin}@microsoft.com, chenzhibo@ustc.edu.cn

Abstract

Image translation across different domains has attracted much attention in both machine learning and computer vision communities. Taking the translation from source domain \mathcal{D}_s to target domain \mathcal{D}_t as an example, existing algorithms mainly rely on two kinds of loss for training: One is the discrimination loss, which is used to differentiate images generated by the models and natural images; the other is the reconstruction loss, which measures the difference between an original image and the reconstructed version through $\mathcal{D}_s \rightarrow \mathcal{D}_t \rightarrow \mathcal{D}_s$ translation. In this work, we introduce a new kind of loss, multi-path consistency loss, which evaluates the differences between direct translation $\mathcal{D}_s \rightarrow \mathcal{D}_t$ and indirect translation $\mathcal{D}_s \rightarrow \mathcal{D}_a \rightarrow \mathcal{D}_t$ with \mathcal{D}_a as an auxiliary domain, to regularize training. For multi-domain translation (at least, three) which focuses on building translation models between any two domains, at each training iteration, we randomly select three domains, set them respectively as the source, auxiliary and target domains, build the multi-path consistency loss and optimize the network. For two-domain translation, we need to introduce an additional auxiliary domain and construct the multi-path consistency loss. We conduct various experiments to demonstrate the effectiveness of our proposed methods, including face-to-face translation, paint-to-photo translation, and de-raining/de-noising translation.

1 Introduction

Image-to-image translation aims at learning a mapping that can transfer an image from a source domain to a target one, while maintaining the main representations of the input image from the source domain. Many computer vision problems can be viewed as image-to-image translation tasks, including image stylization [Gatys *et al.*, 2016], image restoration [Mao *et al.*, 2016], image segmentation [Girshick, 2015] and so on. Since a large amount of parallel data is costly to collect in

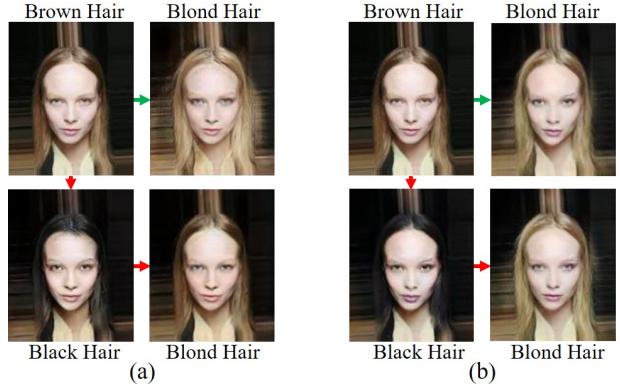


Figure 1: Illustration of multi-path consistency regularization on the translation between different hair colors. (a) Results of StarGAN. (b) Our results.

practice, most of recent works have focused on unsupervised image-to-image translation algorithms. Two kinds of algorithms are widely adopted for this problem. The first one is generative adversarial networks (briefly, GAN), which consists of an image generator used to produce images and an image discriminator used to verify whether an image is a fake one from a machine or a natural one. Ideally, the training will reach an equilibrium where the generator should generate “real” images that the discriminator cannot distinguish from natural images [Goodfellow *et al.*, 2014]. The other one is dual learning [He *et al.*, 2016], which is first proposed for neural machine translation and then successfully adapted into image translation. In dual learning, a pair of dual tasks is involved, like man-to-woman translation v.s. woman-to-man translation. The reconstruction loss of the two tasks are minimized during optimization. The combination of GAN algorithms and dual learning leads to many algorithms for two-domain image translation like CycleGAN [Zhu *et al.*, 2017], DualGAN [Yi *et al.*, 2017], DiscoGAN [Kim *et al.*, 2017], conditional DualGAN [Lin *et al.*, 2018a], and for multi-domain image translation like StarGAN [Choi *et al.*, 2018].

We know that multiple domains are bridged by multi-path consistency. Take the pictures in Figure 1 as an example. We want to work on a three-domain image translation problem,

*The first two authors contributed equally to this work

†Corresponding author

which targets at changing the hair color of the input image to a specific one. Ideally, the direct translation (i.e., one-hop translation) from brown hair to blond should be the same as the indirect translation (i.e., two-hop translation) from brown to black to blond. However, such an important property is ignored in current literature. As shown in Figure 1(a), without multi-path consistency regularization, the direct translation and indirect translation are not consistent in terms of hair color. Besides, on the right of the face in the one-hop translation, there is much horizontal noise. To keep the two generated images consistent, in this paper, we propose a new loss, *multi-path consistency loss*, which explicitly models the relationship among three domains. We require that the differences between direct translation from source to target domain and indirect translation from source to auxiliary to target domain should be minimized. For example, in Figure 1, the L_1 -norm loss of the two translated blond hair pictures should be minimized. After applying this constraint, as shown in Figure 1(b), the direct and indirect translations are much similar, and both the direct and indirect translations are of less noise.

Multi-path consistency loss can be generally applied in image translation tasks. For multi-domain (≥ 3) translation, during each training iteration, we can randomly select three domains, apply the multi-path consistency loss to each translation task, and eventually obtain models that can generate better images. For the two-domain image translation problem, we need to introduce a third auxiliary domain to help establish the multi-path consistency relation.

Our contributions can be summarized as follows: (1) We propose a new learning framework with multi-path consistency loss that can leverage the information among multiple domains. Such a loss function regularizes the training of each task and leads to better performance. We provide an efficient algorithm to optimize such a framework. (2) We conduct rich experiments to verify the proposed method. Specifically, we work on face-to-face translation, paint-to-photo translation, and de-raining/de-noising translation. For qualitative analysis, the models after applying multi-path consistency loss can generate clearer images with less blocks artifacts. For quantitative analysis, we calculate the classification errors and PSNR for tasks, which all outperform the baselines. We also conduct user study on the multi-domain translation task and 59.14%/89.85% users vote for that our proposed method is better on face-to-face and paint-to-photo translations.

2 Related works

In this section, we summarize the literature about GAN and unsupervised image-to-image translation.

GAN GAN [Goodfellow *et al.*, 2014] was firstly proposed to generate images in an unsupervised manner. A GAN is made up of a generator and a discriminator. The generator maps a random noise to an image and the discriminator verifies whether the image is a natural one or a fake one. The training of GAN is formulated as a two-player minmax game. Various versions of GAN have been proposed to exploit its capability for different image generation tasks [Arjovsky *et al.*, 2017; Huang *et al.*, 2017; Lin *et al.*, 2018bl]. InfoGAN [Chen *et al.*, 2016] learns to disentangle latent representations by maxi-

mizing the mutual information between a small subset of the latent variables and the observation. [Radford *et al.*, 2015] presented a series of deep convolutional generative networks (DCGANs) for high-quality image generation and unsupervised image classification tasks, which bridges the convolutional neural networks and unsupervised image generation together. SRGAN [Ledig *et al.*, 2017] maps low-resolution images to high resolution images. Isola *et al.* [Isola *et al.*, 2017] proposed a general conditional GAN for image-to-image translation tasks, which could be used to solve label-to-street scene and aerial-to-map translation problems.

Unsupervised image-to-image translation Since it is usually hard to collect a large amount of parallel data for supervised image-to-image translation tasks, unsupervised learning based algorithms have been widely adopted. Based on adversarial training, Dumoulin *et al.* [Dumoulin *et al.*, 2016] and Donahue *et al.* [Donahue *et al.*, 2016] proposed algorithms to jointly learn mappings between the latent space and data space bidirectionally. Taigman *et al.* [Taigman *et al.*, 2016] presented a domain transfer network (DTN) for unsupervised cross-domain image generation under the assumption that a constant latent space between two domains exists, which could generate images of target domain’s style and preserve their identity. Inspired by the idea of dual learning [He *et al.*, 2016], DualGAN [Yi *et al.*, 2017], DiscoGAN [Kim *et al.*, 2017] and CycleGAN [Zhu *et al.*, 2017] were proposed to tackle the unpaired image translation problem by jointly training two cross-domain translation models. Meanwhile, several works [Choi *et al.*, 2018; Liu *et al.*, 2018] have been further proposed for multiple domain image-to-image translation with a single model only.

3 Framework

In this section, we introduce our proposed framework built on multi-path consistency loss. Suppose we have N different image domains $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_N\}$ where $N \geq 2$. A domain can be seen as a collection of images. Generally, the image translation task aims at learning $N(N - 1)$ mappings $f_{i,j} : \mathcal{D}_i \mapsto \mathcal{D}_j$ where $i \neq j$. Also, we might come across the cases that we are interested in a subset of the $N(N - 1)$ mappings. We first show how to build translation models between \mathcal{D}_i and \mathcal{D}_j with \mathcal{D}_k as an auxiliary domain and then present the general framework for multi-domain image translation with consistency loss. Note that in our framework, $i, j, k \in [N]$ and they are different to each other.

3.1 Translation between \mathcal{D}_i and \mathcal{D}_j with an auxiliary domain \mathcal{D}_k

To effectively obtain the two translation models $f_{i,j}$ and $f_{j,i}$ with an auxiliary domain \mathcal{D}_k , we need the following additional components in the system: (1) Three discriminators d_i , d_j and d_k , which are used to classify whether an input image is a natural one or an image generated by the machine. Mathematically, $d_l : \mathcal{D}_l \mapsto [0, 1]$ models the probability that the input image is a natural image in domain \mathcal{D}_l , $l \in \{i, j, k\}$. (2) Four auxiliary mappings $f_{i,k}$, $f_{k,i}$, $f_{j,k}$ and $f_{k,j}$, which are all related to \mathcal{D}_k .

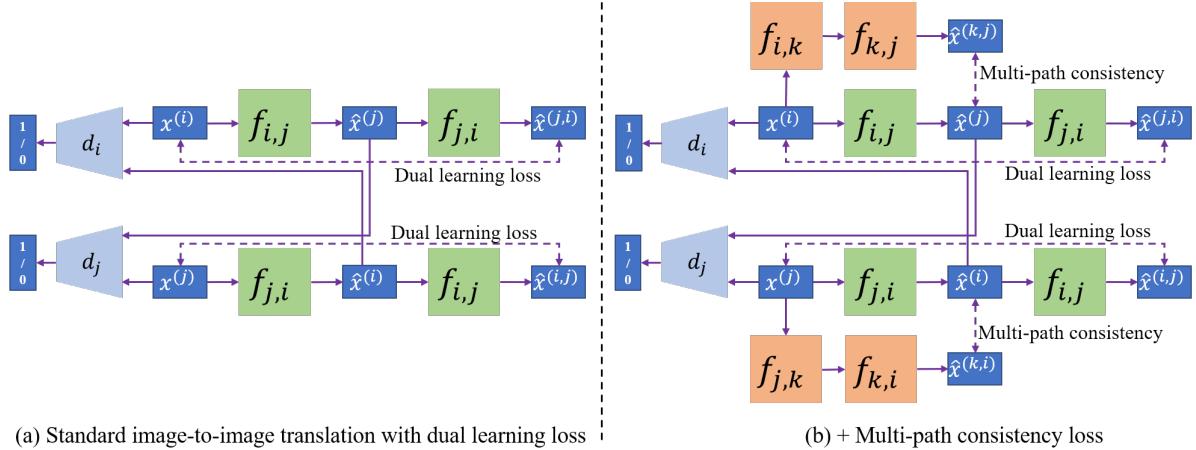


Figure 2: The standard and our proposed frameworks of image-to-image translation, where $\hat{x}^{(j)} = f_{i,j}(x^{(i)})$, $\hat{x}^{(i)} = f_{j,i}(x^{(j)})$, $\hat{x}^{(j,i)} = f_{j,i}(\hat{x}^{(j)})$, $\hat{x}^{(i,j)} = f_{i,j}(\hat{x}^{(i)})$, $\hat{x}^{(k,j)} = f_{k,j}(f_{i,k}(x^{(i)}))$, $\hat{x}^{(k,i)} = f_{k,i}(f_{j,k}(x^{(j)}))$.

Considering that deep learning algorithms usually iteratively work on mini-batches of data instead of the whole training datasets at the same time, in the remaining part of this section, we describe how the models are updated on the batches $\mathcal{B}_l \subseteq \mathcal{D}_l$, $l \in \{i, j, k\}$, where \mathcal{B}_l is a minibatch of data in \mathcal{D}_l .

The training loss consists of three parts:

(1) *Dual learning loss* between \mathcal{D}_i and \mathcal{D}_j , which models the reconstruction duality between $f_{i,j}$ and $f_{j,i}$. Mathematically,

$$\ell_d^{i,j} = \frac{1}{|\mathcal{B}_i|} \sum_{x^{(i)} \in \mathcal{B}_i} \|x^{(i)} - f_{j,i}(f_{i,j}(x^{(i)}))\|_1 + \frac{1}{|\mathcal{B}_j|} \sum_{x^{(j)} \in \mathcal{B}_j} \|x^{(j)} - f_{i,j}(f_{j,i}(x^{(j)}))\|_1, \quad (1)$$

where $|\mathcal{B}_i|$ and $|\mathcal{B}_j|$ are the numbers of images in mini-batch \mathcal{B}_i and mini-batch \mathcal{B}_j .

(2) *Multi-path consistency loss* with an auxiliary domain \mathcal{D}_k , which regularizes the training by leveraging the information provided by the third domain. Mathematically,

$$\ell_c^{i,j|k} = \frac{1}{|\mathcal{B}_i|} \sum_{x^{(i)} \in \mathcal{B}_i} \|f_{i,j}(x^{(i)}) - f_{k,j}(f_{i,k}(x^{(i)}))\|_1 + \frac{1}{|\mathcal{B}_j|} \sum_{x^{(j)} \in \mathcal{B}_j} \|f_{j,i}(x^{(j)}) - f_{k,i}(f_{j,k}(x^{(j)}))\|_1. \quad (2)$$

(3) *GAN loss*, which enforces the generated images to be natural enough. Let $\hat{\mathcal{B}}_l$ denote a collection of all generated/fake images to domain \mathcal{D}_l . When $l = k$, $\hat{\mathcal{B}}_l = \{f_{i,k}(x^{(i)}) | x^{(i)} \in \mathcal{B}_i\} \cup \{f_{j,k}(x^{(j)}) | x^{(j)} \in \mathcal{B}_j\}$. When $l \in \{i, j\}$, $\hat{\mathcal{B}}_l$ is the combination of one-hop translation and two-hop translation, defined as $\{f_{p,l}(x^{(p)}) | x^{(p)} \in \mathcal{B}_p\} \cup \{f_{k,l}(f_{p,k}(x^{(p)})) | x^{(p)} \in \mathcal{B}_p\}$ where $p = \{i, j\} \setminus \{l\}$. The GAN loss is defined as fol-

lows:

$$\ell_{\text{GAN}}^{i,j|k} = \sum_{l \in \{i, j, k\}} \left\{ \frac{1}{|\mathcal{B}_l|} \sum_{x^{(l)} \in \mathcal{B}_l} [\log d_l(x^{(l)})] + \frac{1}{|\hat{\mathcal{B}}_l|} \sum_{\hat{x}^{(l)} \in \hat{\mathcal{B}}_l} [\log(1 - d_l(\hat{x}^{(l)}))] \right\}. \quad (3)$$

All $f_{.,.}$'s work together to minimize the GAN loss, while all $d_{.}$'s try to enlarge the GAN loss.

Given the aforementioned three kinds of loss, the overall loss can be defined as follows:

$$\ell_{\text{total}}^{i,j|k}(\mathcal{B}_i, \mathcal{B}_j, \mathcal{B}_k) = \ell_d^{i,j} + \ell_c^{i,j|k} + \alpha \ell_{\text{GAN}}^{i,j|k}, \quad (4)$$

where α is a hyper-parameter balancing the tradeoff between the GAN loss and other losses. All the six generators $f_{.,.}$'s work on minimizing Eqn. (4) while all three discriminators $d_{.}$'s work on maximizing Eqn. (4).

3.2 Multi-domain image translation

For an N -domain translation problem, when $N \geq 3$, at each training iteration, it is too costly to build the consistency loss for each three domains. Alternatively, we can randomly select three domains \mathcal{D}_i , \mathcal{D}_j , \mathcal{D}_k and build the consistency loss as follows:

$$\ell_{\text{total}}^{i,j,k}(\mathcal{B}_i, \mathcal{B}_j, \mathcal{B}_k) = \ell_{\text{total}}^{i,j|k}(\mathcal{B}_i, \mathcal{B}_j, \mathcal{B}_k) + \ell_{\text{total}}^{i,k|j}(\mathcal{B}_i, \mathcal{B}_k, \mathcal{B}_j) + \ell_{\text{total}}^{j,k|i}(\mathcal{B}_j, \mathcal{B}_k, \mathcal{B}_i), \quad (5)$$

where the $\ell_{\text{total}}^{i,j|k}$ is defined in Eqn. (4) and the other notations can be similarly defined.

Discussion

(1) When $N = 2$, we need to find the third domain as a auxiliary domain to help establish consistency. In this case, we can use Eqn. (4) as the training objective, without applying consistency loss on the third domain. We work on the de-raining and de-noising tasks to verify such a case. (See Section 5.)

(2) We can build the consistency loss with longer paths, e.g., the translation from $\mathcal{D}_1 \rightarrow \mathcal{D}_3 \dots \mathcal{D}_N \rightarrow \mathcal{D}_2$ should be consistent with $\mathcal{D}_1 \rightarrow \mathcal{D}_2$. Considering computation resource limitation, we leave this study to future work.

3.3 Connection with StarGAN

For an N -domain translation, when N is large, it is impractical to learn $N(N - 1)$ mappings. StarGAN [Choi *et al.*, 2018] is a recently proposed method which uses a single model with different target labels to achieve image translation. With StarGAN, the mapping from \mathcal{D}_i to \mathcal{D}_j could be specified as $f(x^{(i)}, c_j)$ where $x^{(i)} \in \mathcal{D}_i$, f is shared among all tasks and c_j is a learnable vector used to identify \mathcal{D}_j . All the generators share a same copy of the parameters except the target domain labels.

In terms of the discriminator, StarGAN only consists of one network which is not only used for justifying whether an image is a natural or fake one, but also serving as a classifier that distinguishes which domain does the input belong to. We also adopt such a kind of discriminator when using StarGAN as the basic model architecture. In this case, let $d_{\text{cls}}(l|x)$ denote the probability that the input x is categorized as an image from domain l , $l \in [N]$. Following the notations in Section 3.1, the classification cost $\ell_{\text{cls},r}^{i,j|k}$ of real images and fake images $\ell_{\text{cls},f}^{i,j|k}$ for StarGAN can be formulated as follows:

$$\begin{aligned}\ell_{\text{cls},r}^{i,j|k} &= \sum_{l \in \{i,j,k\}} \frac{-1}{|\mathcal{B}_l|} \sum_{x^{(l)} \in \mathcal{B}_l} [\log d_{\text{cls}}(l|x^{(l)})], \\ \ell_{\text{cls},f}^{i,j|k} &= \sum_{l \in \{i,j,k\}} \frac{-1}{|\hat{\mathcal{B}}_l|} \sum_{\hat{x}^{(l)} \in \hat{\mathcal{B}}_l} [\log d_{\text{cls}}(l|\hat{x}^{(l)})].\end{aligned}\quad (6)$$

When using StarGAN with the aforementioned classification errors, the image generators and discriminators cannot share a common objective function. Therefore, the loss function with multi-path consistency regularization, i.e., Eqn. (4), should be split and re-formulated as follows:

$$\begin{aligned}\ell_{\text{total},G}^{i,j|k}(\mathcal{B}_i, \mathcal{B}_j, \mathcal{B}_k) &= \ell_d^{i,j} + \ell_c^{i,j|k} + \alpha \ell_{\text{GAN}}^{i,j|k} + \beta \ell_{\text{cls},f}^{i,j|k}, \\ \ell_{\text{total},D}^{i,j|k}(\mathcal{B}_i, \mathcal{B}_j, \mathcal{B}_k) &= -\alpha \ell_{\text{GAN}}^{i,j|k} + \beta \ell_{\text{cls},r}^{i,j|k},\end{aligned}\quad (7)$$

where both α and β are the hyper-parameters. The generator and discriminator should try to minimize $\ell_{\text{total},G}^{i,j|k}$ and $\ell_{\text{total},D}^{i,j|k}$ respectively. Also, Eqn (5) should be re-defined accordingly.

4 Experiments on multi-domain translation

For multi-domain translation, we carry out two groups of experiments to verify our proposed framework, which are face-to-face translation with different attributes and paint-to-photo translation with different art styles. We choose StarGAN [Choi *et al.*, 2018], a state-of-the-art algorithm on multi-domain image translation as our baseline.

4.1 Setting

Datasets For multi-domain face-to-face translation, we use the CelebA dataset [Liu *et al.*, 2015], which consists of 202,599 face images of celebrities. Following [Choi *et al.*, 2018], we select seven attributes and build seven domains correspondingly. Among these attributes, three of them represent hair color, including black hair, blond hair, brown hair; two of them represent the gender, i.e., male and female; the left two represent age, including old and young. Note that these

seven features are not disjoint. That is, a man can both have blond hair and be young.

For multi-domain paint-to-photo translation, we use the paintings and photographs collected by [Zhu *et al.*, 2017], where we construct five domains including Cezanne, Monet, Ukiyo-e, Vangogh and photographs.

Architecture For multi-domain translation tasks, we choose StarGAN [Choi *et al.*, 2018] as our basic structure. One reason is that for an N -domain translation, we need $N(N - 1)$ independent models to achieve translations between any two domains; with StarGAN, we only need one model. Another reason is that [Choi *et al.*, 2018] claim that on face-to-face translation, StarGAN achieves better performance for multi-domain translation compared to simply using multiple CycleGANs since multi-tasks are involved in the same model and the common knowledge among different tasks can be shared to achieve better performance.

Optimization We use Adam optimizer [Kingma and Ba, 2014] with learning rate 0.0001 for the first 10 epochs and linearly decay the learning rate every 10 epochs. All the models are trained on one NVIDIA K40 GPU for one day. The α in Eqn. (4) and Eqn. (7) is set to 0.1, and β in Eqn. (7) is also set to 0.1.

Evaluation We take both qualitative and quantitative analysis to verify the experiment results. For qualitative analysis, we visualize the results of both the baseline and our algorithm, and compare their differences. For quantitative analysis, following [Choi *et al.*, 2018], we perform classification experiments on generation synthesis. We train the classifiers on the image-translation training data using the same architecture as that for the discriminator, resulting in near-perfect accuracies, and then compute the classification error rates of generated images based on the classifiers. The Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] that measures similarity between generated image dataset and real image dataset is used to evaluate translated results quality. The lower the FID is, the better the translation results are. We also carry out user study for the generated images.

4.2 Face-to-face translation

The results of face-to-face translation are shown in Figure 3. In general, both the baseline and our proposed method can successfully transfer the images to the target domain. But there are many places where our proposed method outperforms the baseline:

(1) Our proposed method could preserve more information of the input images. Take the translations in the upper part as an example. To translate the input to black hair and blond hair domain, the domain-specific feature [Lin *et al.*, 2018a] to be changed is the hair color, while the other domain-independent features should be kept as many as possible. The baseline algorithm changes the beard of input images, while our algorithm keeps such a feature due to the multi-path consistency regularization. Since a consistency regularization requires both the one-hop translation and two-hop translation to preserve enough similarity, and the two-hop translation path is independent of one-hop translation path, errors/distortions in the translation results are more likely to be avoided from

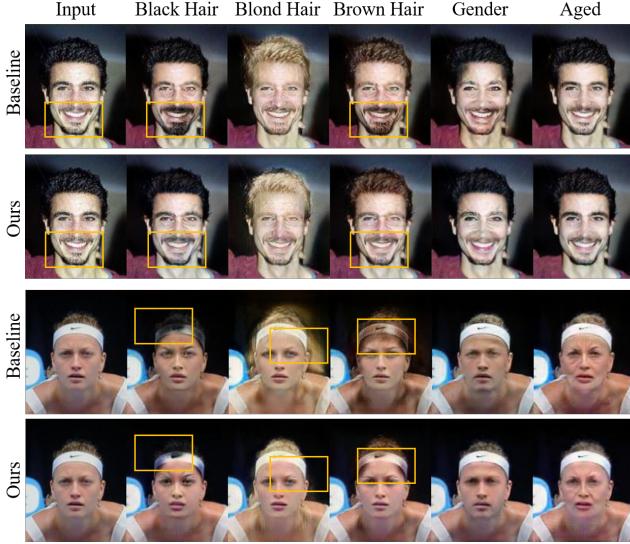


Figure 3: Two groups of multi-domain face-to-face translation results. The rows started with “Baseline” and “Ours” represent the results of the baseline and our method.

the aspect of probability. As a result, our proposed method can carry more information of the original image.

(2) After applying multi-path consistency regularization, our algorithm could generate clearer images with less noise than the baseline, like the images with black hair and blond hair in the bottom part. One possible reason is that multi-path consistency pushes the models to generate consist images. The random noise would affect the consistency and our proposed regularization way could reduce such effects.

For quantitative evaluation, the results are shown in Table 1 and Table 2. We can see that our proposed method achieves lower classification error rates on three different domains, which demonstrates that our generator could produce images with more significant features in the target domain. For FID score, our algorithm also achieves 1.79 improvement, which suggests that our translation results’s distribution are more similar to real ones’.

	Hair Color	Gender	Age
Baseline	19.01%	11.60%	25.52%
Ours	17.08%	10.23%	24.39 %
Improvements	1.93%	1.37%	1.13%

Table 1: Classification error rates of face-to-face translation.

Baseline	Ours	Improvement
20.15	18.36	1.79

Table 2: FID scores of face-to-face translation.

4.3 Paint-to-photo translation

The results of multi-domain paint-to-photo translation are shown in Figure 4. Again, the models trained with multi-path consistency loss outperform the baselines. For example, our model can generate more domain-specific paintings than the baseline method as shown in the generated Vangogh paintings. We also observe that our model effectively reduces the block artifact in the translation results, such as generated Cezanne, Monet and Vangogh paintings. Besides, our model prefers generating images with clearer edges and context. As shown in the upper-left corner of Figure 4, we could generate images with obvious edges and content, while the baseline algorithm fails with unclear and messy generations.

Similar to face-to-face translation, we also show the classification errors and FID. The results are in Table 3 and Table 4. Our algorithm achieves significantly better results than the baseline, which demonstrates the effectiveness of our method.

Baseline	Ours	Improvement
35.52%	30.17%	5.35%

Table 3: Classification error rates of paint-to-photo translation.

	Cezanne	Monet	Ukiyo-e	Vangogh	Photograph
Baseline	219.43	199.77	163.46	226.77	79.33
Ours	210.82	170.52	154.29	216.78	64.10
Improvements	8.61	29.25	9.17	9.99	15.23

Table 4: FID scores of paint-to-photo translation.

4.4 User study

We carry out user study to further evaluate our results. 20 users with diverse education backgrounds are chosen as reviewers. We randomly select 35 groups of generated images for the face-to-face and paint-to-photo translation, where each group contains the translation results from the same input image to different categories of both the baseline and our algorithm. For any group of experiments, reviewers have to choose the better one without knowing which algorithm the images are generated from.

The statistics are shown in Table 5. Among the 700 votes for face-to-face translation, 59.14% belongs to our proposed algorithm, and for paint-to-photo, 89.85% belongs to ours. The user study shows that we achieve better performance than the baseline, especially for paint-to-photo translation.

face-to-face		paint-to-photo	
Baseline	Ours	Baseline	Ours
40.86%	59.14%	10.14%	89.85%

Table 5: Statistics of user study.

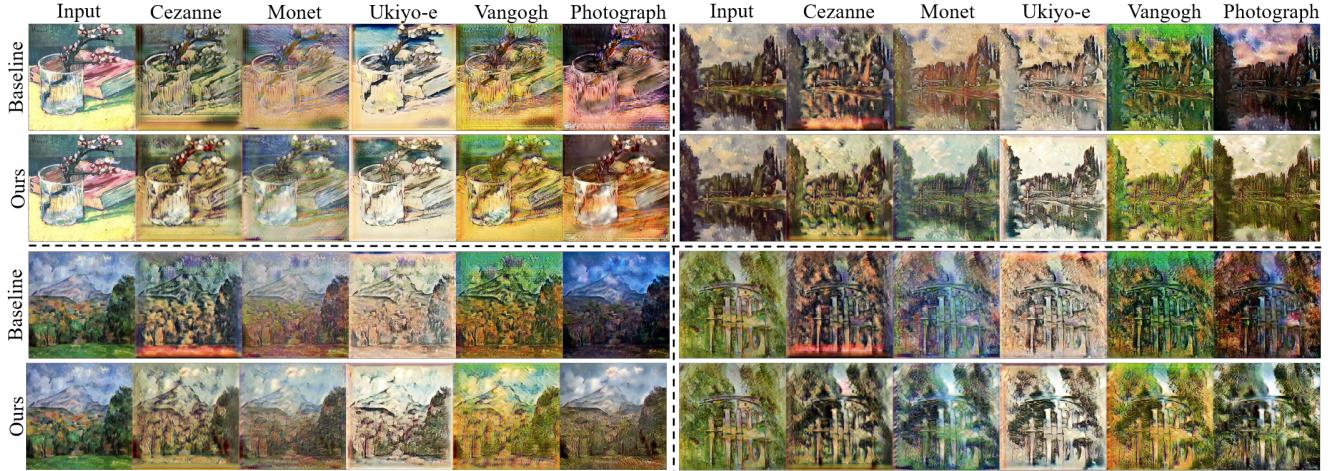


Figure 4: Multi-domain paint-to-photo translation results.



Figure 5: Unsupervised de-raining (first two rows) and de-noising (last two rows) results. From left to right, the columns represent the rainy/noisy input, the original clean image, the results of StarGAN (St) and CycleGAN (Cy) without multi-path consistency loss, and the corresponding results with our method (St+ours, Cy+ours) respectively.

5 Experiments on two-domain translation

In this section, we work on two-domain translations with auxiliary domains. We choose two different tasks, unsupervised de-raining and unsupervised de-noising, which means to remove the rain or noise from the input images.

5.1 Setting

Datasets We use the raining images and original images collected by [Fu *et al.*, 2017; Yang *et al.*, 2017]. For unsupervised translation, we randomly shuffle the raining images and the original ones to obtain an unaligned dataset. As for the unaligned dataset for de-noising, we add uniform noise to each original image and then shuffle them. For the de-raining and de-noising experiments, we choose the noise image domain and raining image domain as auxiliary ones respectively.

Architecture We first choose StarGAN as the basic network architecture. The model architecture is the same as that used in Section 4.1. In addition, to verify the generality of our framework, we also apply CycleGAN [Zhu *et al.*, 2017] to this task. To combine with our framework, a total of six generators and three discriminators are implemented. We follow Section 3.1 to jointly optimize the de-rain and add-rain networks for the de-raining task with consistency loss built upon the domains with images of random noise. Similar method is also applied to the de-noising task.

Evaluation For qualitative analysis, again we compare the images generated by the baseline and our algorithm. For quantitative analysis, except for the classification errors, we check the Peak Signal-to-Noise Ratio (briefly, PSNR), of the generated images with the original images. The larger PSNR is, the better the restoration quality is.

5.2 Results

The unsupervised de-raining and unsupervised de-noising results are shown in Figure 5. On the two tasks, our proposed method can improve both StarGAN and CycleGAN and generate cleaner images with less block artifacts, smoother colors and clearer facial expressions. We also find that in de-raining and de-noising tasks, CycleGAN outperforms StarGAN and can generate images with less rain and noise. One reason is that unlike face-to-face translation whose domain-independent features are centralized and easy to capture, natural scenes are usually diverse and complex, in which a single StarGAN might not have enough capacity to model. In comparison, a CycleGAN works for two-direction translation only, which has enough capacity to model and de-rain/de-noise the images.

We report the classification error rates and PSNR (dB) of de-raining and de-noising in Table 6. The classification error rates of StarGAN and CycleGAN before using multi-path consistency loss are 2.91% and 1.93% respectively, while after applying multi-path consistency loss the numbers are 1.70% and 1.65%, which shows the efficiency of our method. In terms of PSNR, as shown in Table 6, our method achieves higher scores, which means that our model has better restoration abilities. That is, for the two-domain translation, our framework still works. We also plot PSNR curves of the StarGAN based models on the test set w.r.t. training steps.

The results are shown in Figure 6. On the two tasks, training with multi-path consistency regularization could always achieve higher PSNR than the corresponding baseline. This shows that our proposed method can achieve not only higher PSNR values, but also faster convergence speed.

Method	Classification Error	r→c (dB)	n→c (dB)
StarGAN	2.91%	19.43	20.39
CycleGAN	1.93%	20.87	21.99
Ours(St)	1.70%	21.13	23.25
Ours(Cy)	1.65%	21.21	23.28

Table 6: Classification error rates and PSNR (dB) of de-raining and de-noising translation results.

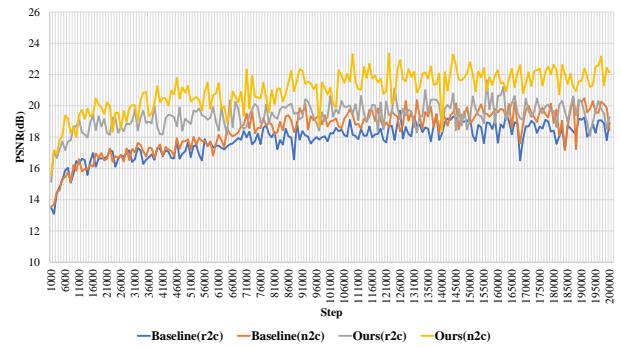


Figure 6: PSNR curve w.r.t training steps.

6 Conclusion and future work

In this paper, we propose a new kind of loss, multi-path consistency loss, which can leverage the information of multiple domains to regularize the training. We provide an effective way to optimize such a framework under multi-domain translation environments. Qualitative and quantitative results on multiple tasks demonstrate the effectiveness of our method.

For future work, it is worth studying what would happen if more than three paths are included. In addition, we will generalize multi-path consistency by using stochastic latent variables as the auxiliary domain

7 Acknowledgement

This work was supported in part by NSFC under Grant 61571413, 61632001.

References

- [Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.
- [Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

- [Choi *et al.*, 2018] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [Donahue *et al.*, 2016] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.
- [Dumoulin *et al.*, 2016] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.
- [Fu *et al.*, 2017] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1715–1723, 2017.
- [Gatys *et al.*, 2016] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.
- [Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [Goodfellow *et al.*, 2014] 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.
- [He *et al.*, 2016] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016.
- [Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017.
- [Huang *et al.*, 2017] Xun Huang, Yixuan Li, Omid Poursaeed, John E Hopcroft, and Serge J Belongie. Stacked generative adversarial networks. In *CVPR*, volume 2, page 3, 2017.
- [Isola *et al.*, 2017] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, July 2017.
- [Kim *et al.*, 2017] 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*, pages 1857–1865, 2017.
- [Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [Ledig *et al.*, 2017] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, July 2017.
- [Lin *et al.*, 2018a] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [Lin *et al.*, 2018b] Jianxin Lin, Tiankuang Zhou, and Zhibo Chen. Multi-scale face restoration with sequential gating ensemble network. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7122–7129, 2018.
- [Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.
- [Liu *et al.*, 2018] 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.
- [Mao *et al.*, 2016] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems 29*, pages 2802–2810, 2016.
- [Radford *et al.*, 2015] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [Taigman *et al.*, 2016] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.
- [Yang *et al.*, 2017] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1357–1366, 2017.
- [Yi *et al.*, 2017] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.