Maximum Spatial Perturbation Consistency for Unpaired Image-to-Image Translation

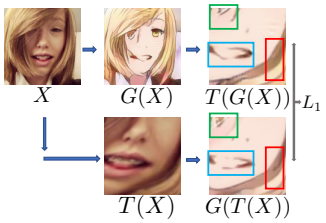
Yanwu Xu[1], Shaoan Xie[2], Wenhao Wu[4], Kun Zhang[2,5], Mingming Gong[3†], Kayhan Batmanghelich[1†]
[1]Department of Biomedical Informatics, University of Pittsburgh,
{yanwuxu,kayhan}@pitt.edu
[2]Department of Philosophy, Carnegie Mellon University
[3]School of Mathematics and Statistics, The University of Melbourne
[4]Department of Computer Vision Technology (VIS), Baidu Inc
[5]Mohamed bin Zayed University of Artificial Intelligence

Abstract

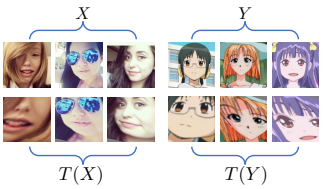
Unpaired image-to-image translation (I2I) is an ill-posed problem, as an infinite number of translation functions can map the source domain distribution to the target distribution. Therefore, much effort has been put into designing suitable constraints, e.g., cycle consistency (CycleGAN), geometry consistency (GCGAN), and contrastive learning-based constraints (CUTGAN), that help better pose the problem. However, these well-known constraints have limitations: (1) they are either too restrictive or too weak for specific I2I tasks; (2) these methods result in content distortion when there is a significant spatial variation between the source and target domains. This paper proposes a universal regularization technique called maximum spatial perturbation consistency (MSPC), which enforces a spatial perturbation function (T) and the translation operator (G) to be commutative (i.e., $T \circ G = G \circ T$). In addition, we introduce two adversarial training components for learning the spatial perturbation function. The first one lets T compete with G to achieve maximum perturbation. The second one lets G and T compete with discriminators to align the spatial variations caused by the change of object size, object distortion, background interruptions, etc. Our method outperforms the state-of-the-art methods on most I2I benchmarks. We also introduce a new benchmark, namely the front face to profile face dataset, to emphasize the underlying challenges of I2I for real-world applications. We finally perform ablation experiments to study the sensitivity of our method to the severity of spatial perturbation and its effectiveness for distribution alignment.

1. Introduction

In unpaired image-to-to image translation (I2I), one aims to translate images from a source domain $\mathcal{X}$ to a target domain $\mathcal{Y}$, with data drawn from the marginal distribution of the source domain (P_X) and that of the target domain (P_Y). Unpaired I2I has many applications, such as super-resolution [12, 15], image editing [13, 49], and image denoising [4, 41]. However, it is an ill-posed problem, as there is an infinite choice of translators G that can map P_X to P_Y.



(a) Consistency regularization with spatial perturbation function T



(b) Spatial alignment of spatial perturbation function T

Figure 1. In this figure, we illustrate the the proposed MSPC on (a) consistency regularization under maximum spatial perturbation and (b) aligning the spatial distributions between source X_T and Y_T via spatial perturbation function T.

Various constraints on the translation function G have been proposed to remedy the ill-posedness of the problem. For example, cycle consistency (CycleGAN) [50] enforces the cyclic reconstruction consistency: $X \rightarrow G(X) \rightarrow X$, which means G and its inverse are bijections. CUTGAN [37] maximizes the mutual information between an input image and the translated image via constrastive learning on the patch-level features. The GCGAN [16], on the other hand, effectively uses geometric consistency by applying a

† Equal Contribution. Code is released at https://github.com/batmanlab/MSPC.

predefined geometry transformation g, i.e., fixed rotation, encouraging G to be robust to geometry transformation. The underlying assumption of the GCGAN is that the G and g are commutative (i.e., $g \circ G = G \circ g$). However, CycleGAN assumes that the relationship of bijection between source and target, which is limited for most real-life applications [37]. For instance, the translation function is non-invertible in the *Cityscapes* $\rightarrow$ *Parsing* task. Though geometry consistency used in GCGAN is a general I2I constraint, it is too weak in the sense that the model would easily memorize the pattern of a fixed transformation. CUTGAN enforces the strong correlation between the input images and the translated images at the corresponded patches; thus it would fail when the patches at the same spatial location do not contain the same content, *e.g.*, in the *Front Face* $\rightarrow$ *Profile* task (shown in Figure 5). Thus, the above models are either too restrictive or too weak for specific I2I tasks. Besides, all of them overlook the extra spatial variations in image translation, which are caused by the change of object size, object distortion, background interruptions, *etc*.

To tackle the issues above, we propose a novel regularization called the maximum spatial perturbation consistency (MSPC), which enforces a new type of constraint and aligns the content's spatial distribution content across domains. Our MSPC generalizes GCGAN by learning a spatial perturbation function T, which adaptively transforms each image with an image-dependent spatial perturbation. Moreover, MSPC is based on the new insight that consistency on hard spatial perturbation would boost the robustness of translator G. Thus, MSPC enforces the maximum spatial perturbation function (T) and the translation operator (G) to be commutative (i.e., $T \circ G = G \circ T$). To generate the maximum spatial perturbation, we introduce a differentiable spatial transformer T [24] to compete with the translation network G in a mini-max game, which we mark as the perturbation branch. More specifically, T tries to maximize the distance between $T(G(X))$ and $G(T(X))$, and G minimizes the difference between them. In this way, our method dynamically generates the hardest spatial transformation for each image, avoiding overfitting G to specific spatial transformations. The Figure 1a give a simple illustration of how the image-dependent spatial perturbation works on the I2I framework.

To align the spatial distribution of the content, T and G cooperate to compete with a discriminator D_{pert} in another mini-max game, which we mark as an alignment branch. In the alignment branch, T participates in aligning the distribution between the translated images and the target images by alleviating the spatial discrepancy, *i.e*. adjusting the object's size, cropping out the noisy background, and further reducing undesired distortions in the translation network G. We evaluate our model on several widely studied benchmarks, and additionally, we construct a *Front Face* $\rightarrow$ *Profile* dataset with significant domain gaps to emphasize the challenges in real-world applications. The experimental results show that the proposed MSPC outperforms its competitors on most I2I tasks. More importantly, MSPC performs the most stable across various I2I tasks, demonstrating the universality of our constraint. The Figure 1b shows the visual examples the alignment effect on source and target images via dynamic spatial transformation function.

2. Related Work

2.1. Generative Adversarial Network

Generative adversarial networks (GANs) [19] train a min-max game between the generator G and the discriminator D, where D tries to discriminate between the data distribution and the generated distribution. When G and D reach a equilibrium, the generated distribution will exactly match the data distribution. In recent years, GANs have been explored in many image synthesis tasks, such as supervised and unsupervised image generation [3, 11, 18, 33, 34], domain adaptation [2, 17, 47], image inpainting [36, 40, 42], *etc*.

2.2. Image-to-Image Translation

The paired image-to-image translation task can be traced back to [14], which proposes a non-parametric texture model. With the development of deep learning, the recent Pix2Pix model [23] expands the conditional GAN model to the image translation and learns a conditional mapping from source images to the target images with paired data. There are also other works in this line of research, such as [25, 39]. However, paired images are expensive to collect, and thus the latest works focus on the setting with semi-supervised and unsupervised settings. Compared to existing unpaired setting, [44] considers a more challenging setting where contents of two domains are unaligned and proposes to address this issue with importance re-weighting. As a semi-supervised method, [35] performs image translation with the combined limited paired images and sufficient unpaired images. Furthermore, [1, 5–8, 16, 26, 28, 29, 31, 37, 43, 46, 48, 50] focus on the unsupervised image translation tasks. In these works, CycleGAN [50] proposes a cycle consistency between the input images and the translated images. GCGAN [16] minimizes the error translated images via the rotation on the input images. CUTGAN [37] maximizes the mutual information between the input and the translated images via contrastive learning. UNIT [28] proposes a strong assumption of content sharing and style change between two image domains in the latent space. To obtain diverse translation results, MUNIT [22] and DRIT [27] disentangle the content and the style and generate diverse outputs by combing the same content with different styles. In this

paper, we focus on the unsupervised task with deterministic output of image translation.

2.3. Consistency Regularization of Semi-Supervised Learning

Among various methods for the semi-supervised classification, clustering, or regression task, consistency regularization has attracted much attention, as discussed in a recent survey paper on deep semi-supervised learning [45]. The constraint of consistency regularization assumes that the manifold of data is smooth and that the model is robust to the realistic perturbation on the data points. In other words, consistency regularization can force the model to learn a smooth manifold via incorporating the unlabeled data. Though GCGAN was proposed from a different perspective, it can be considered as a variation of Π model [38], which enforces consistent model prediction on two random augmentations on a labeled or unlabeled sample.

The regularization method closely related to the proposed MSPC is virtual adversarial training (VAT) [32]. VAT introduced the concept of adversarial attack [20] as a consistency regularization in semi-supervised classification. This method learns a maximum adversarial perturbation as a additive noise on the data-level. To be more specific, it finds an optimal perturbation γ on a input sample x under the constraint of $\gamma < \delta$. Letting $\mathcal{R}$ and f denote the estimation of distance between two vectors and the predicted model respectively, we can formulate it as:

$$\min_{f} \max_{\gamma;\|\gamma\|\leqslant\delta} \mathbb{E}_{x\in P_X} \mathcal{R}(f(\theta, x), f(\theta, x+\gamma)). \tag{1}$$

3. Proposed Method

In unsupervised I2I, one has access to the unpaired images $\mathcal{X}, \mathcal{Y} \subseteq \mathbf{R}^{C\times H\times W}$, which are from the source and target domains, respectively. The goal is to translate image of $\{x; x \in \mathcal{X}\}$ to $\{y; y \in \mathcal{Y}\}$. Our proposed MSPC has four components and three branches. For the components, we have an image translator G, a spatial perturbation function T and two image discriminators D and D_T. As the three branches, a) G and D are for regular adversarial training for the image translation; b) G and T compete with each other in the maximum spatial perturbation branch; c) G and T cooperate together to compete with D_T in the spatial alignment branch. The overall architecture of our method is shown in Figure 2a. Below we will explain our method in the order of the branches.

3.1. Adversarial Constraint on Image Translation

A straightforward way of building the translation framework (branch a) is to utilize generative adversarial training [19], which forces the translated images to be similar to the target images.

$$\min_{G}\max_{D} \mathbb{E}_{y\sim P_Y} \log D(y) + \mathbb{E}_{x\sim P_X} \log(1 - D(G(x))),$$

which is exactly branch a) of our method and has been widely adopted in most I2I approaches [16, 37, 50].

3.2. Maximum Spatial Perturbation Consistency

In the maximum spatial perturbation branch (branch b), we specify the proposed maximum spatial perturbation consistency (MSPC) for regularizing the unsupervised translation network. Concisely, we propose an adversarial spatial perturbation network T that is to be trained together with the translator G. The formulation is as follows:

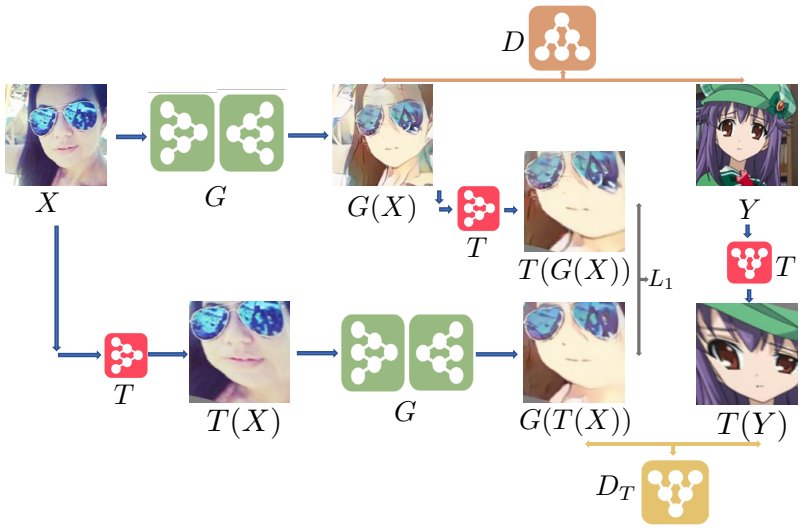
$$\min_{G}\max_{T} \mathbb{E}_{x\sim P_X} \|T(G(x)) - G(T(x))\|_1, \tag{2}$$

where T aims to maximize the L_1 distance between the translated image from original input x and the spatial perturbed image $T(x)$, and G learns to minimize the divergence caused by T, which is the effect of of spatial perturbation. It is worth noting that T is a parameterized and differentiable network, thanks to [24]; details will be introduced later. Thus, for each image x_i, the learned spatial perturbation T_i is specific to the image. In other words, T generates different spatial perturbations for different images, while in GCGAN, T only represents a fixed spatial transformation. Moreover, our spatial perturbation function T changes as training proceeds. To design the consistency loss, we construct the correspondence between the translated image $G(x_i)$ and the perturbed translated image $G(T_i(x_i))$ via applying the learned T_i on the translated images, which is $T_i(G(x_i))$. A graphic illustration of this branch is given in Figure 2b.

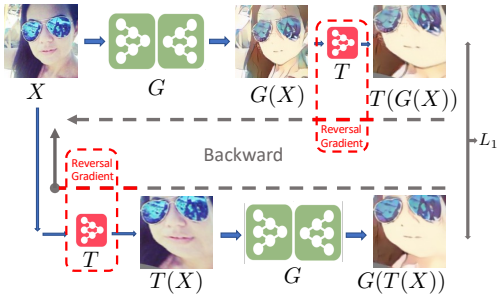
3.3. Spatial Alignment of the Transformer T

In branch b), T plays an important role to generate maximum perturbation that tries to confuse G and enable G to be more robust across different I2I tasks. Furthermore, the deforming property of T can help align the spatial distribution in an unsupervised manner between the source images X and the target images Y by scaling, rotating, cropping noisy background, etc. As shown in Figure 2c, G and T try to force the distribution of $G(T(X))$ to approach the distribution of the transformed target images $T(Y)$ via adversarial training with another discriminator D_T. In this process, the target distribution of $P(T(Y))$ is also deformed to be close to the generated distribution, which is different from the regular generative adversarial training with a fixed target distribution. Thus, in the process of c), the adversarial training process can be formulated as the following min-max game,

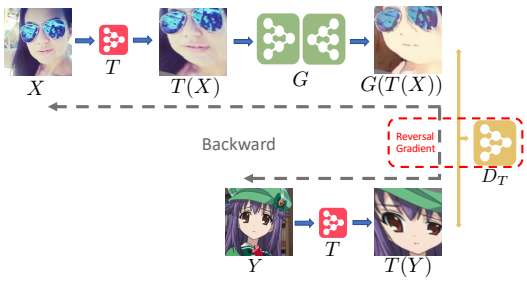
$$\min_{G,T}\max_{D_T} \mathbb{E}_{y\sim P_Y} \log D(T(y)) + \mathbb{E}_{x\sim P_X} \log(1 - D(G(T(x)))). \tag{3}$$



(a) Complete Model of MSPC.



(b) Maximum Spatial Perturbation Consistency



(c) Spatial Alignment of the Transformer T.

Figure 2. Illustration of proposed MSPC model, (a)we can summarize our model as three branches of learning 1) $X \rightarrow G \rightarrow D \leftarrow Y$; 2) $X \rightarrow G \rightarrow T \rightarrow L_1 \leftarrow G \leftarrow T \leftarrow X$; 3) $X \rightarrow T \rightarrow G \rightarrow D_T \leftarrow T \leftarrow Y$. 1), 2), 3) specify the regular adversarial training, maximum spatial perturbation and spatial alignment respectively. To be more specific, we show the adversarial training between G, T in (b) and G, T, D_T in (c) via the forward and backward flow.

3.4. Differentiable T

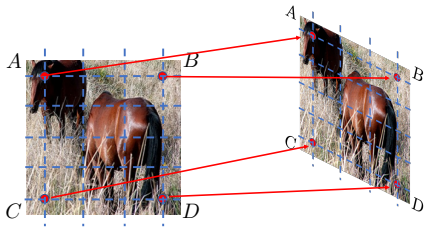


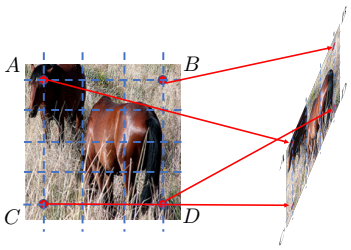
Figure 3. Illustration of spatial transformation network (STN). The network T outputs the coordinates of the deformed grids over the images and then the new images are generated via interpolating in these grids; it is differentiable and can be optimized with stochastic gradient decent.

All of these functionalities of T in the above sections are based on the nice property that T is differentiable and can be optimized with stochastic gradient decent. According to [24], it can be modeled in two steps. In the first step of transforming image, we construct a grid over the image, and the transformation network T outputs the coordinates of the transformed grids. Assuming the image size is $H \times W$, we can simply formulate the process of transformation as

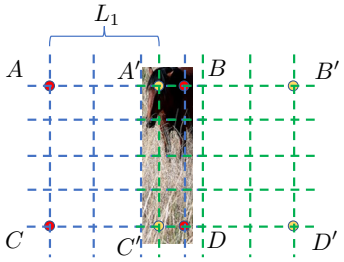
$$\{(p_i^1, p_j^2); i = 1, 2, 3, ..., n, j = 1, 2, 3, ..., m\} = T(x),$$
$$V_{i,j}^c = \sum_n^H \sum_m^W U_{nm}^c k(p_i^1 - q_m^1; \Phi_{p^1}) k(p_j^2 - q_n^2; \Phi_{p^2}),$$
$$\forall i, m \in [1 \dots H]; \ \forall j, n \in [1 \dots W]; \ \forall c \in [1 \dots C], \quad (4)$$

where (q_i^1, q_i^2) represent the coordinate of original grid, U is the pixel value of original image, c is the indicator of image channel, (p_i^1, p_i^2) denotes the new coordinates of transformed grids, $k(;\Phi_{p^1}), k(;\Phi_{p^2})$ represents the kernel of the interpolating image, and we use V_i to denote the transformed pixel value in location (p_i^1, p_i^2). See Figure 3 for an graphic illustration. For the convenience of later formulation, we simply refer to $T(x)$ as the learned transformed image.

3.5. Constraint on T



(a) Image get distorted heavily without scaling constraint.



(b) Majority of Image is cropped out.

Figure 4. Illustration of information loss on perturbed images caused by unconstrained T.

However, without suitable constraints enforced on T, T would produce trivial transformation on images, which may worsen the performance of G, leading to information loss of images as illustrated in Figure 4. One can naturally come up with an immediate, straightforward way to impose this constraint, for instance, by using

$$\|T(G(x)) - G(T(x))\|_1 < \epsilon,\ w.r.t.\ T. \tag{5}$$

However, G is a flexible function that can gradually adapt to whatever transformation learned from T, and thus T would still produce transformation beyond the given image distribution. To solve this issue, we directly design a relative scaling constraint of T on the original and transformed coordinates, which is designed to tackle the issue shown in Figure 4a. Besides, the major proportion of images would be moved out of the original grids as illustrated in Figure 4b, thus we also enforce an absolute constraint on T, which restricts the average translation of target coordinates in a reasonable range. According to the property of T as explained in Section 3.4, the spatial transformation is based on system of coordinates. Thus, we can directly enforce the relative scaling and the absolute translation constraint on the transformed coordinates, which can be formulated as

$$\frac{1}{a} < \frac{|p_i p_j|}{|q_i q_j|} < a,\ i \neq j\ \&\ -b < \sum_{i=1}^{n} p_i < b, \tag{6}$$

where q_i, p_i are the grid coordinates of original and transformed images, respectively, and a, b are constants. The intuition is that, we do not allow the image to be severely distorted beyond a certain scaling and the average translation of coordinates should also be controlled in a reasonable range. The overall formulation of our model can be summarized as follow:

$$\begin{aligned}
&\min_{G,T}\max_{D,D_T} \mathbb{E}_{y\sim P_Y} \log D(y) + \mathbb{E}_{x\sim P_X} \log(1 - D(G(x))) \\
&+ \mathbb{E}_{y\sim P_Y} \log D_T(T(y)) + \mathbb{E}_{x\sim P_X} \log(1 - D_T(G(T(x)))), \\
&\min_{G}\max_{T} \mathbb{E}_{x\sim P_X} \|T(G(x)), G(T(x))\|_1, \\
&s.t.\ \frac{1}{a} < \frac{|p_i p_j|}{|q_i q_j|} < a,\ i \neq j\ \&\ -b < \sum_{i=1}^{n} p_i < b.
\end{aligned} \tag{7}$$

4. Experiment

We conduct quantitative experiments in different settings on front face→profile, Cityscapes [9], Google Map [23], horse→zebra translations. For face→profile, we aim to simulate the real-world application, in which we do not have any paired training identities from source to target but evaluate the performance on the held-out front and profile faces with the paired identities. The Cityscapes and Google Map datasets contain paired images in the training datasets, but all the models are trained in an unpaired manner and also tested on paired held-out testing set.s Additionally, we also test the model the on the popular horse→zebra where paired data are not available.

4.1. Training Configuration

We unify the model training configuration in this section. We Compare our MSPC, the modified virtual adversarial training (VAT), and the modified mean teacher (MT) models with the recently proposed, popular CycleGAN, GCGAN and CUTGAN, where "modified" means transferred from semi-supervised framework to I2I. Please refer to the Section 1 in supplementary for the detailed implementation of modified VAT and MT. We choose the 9-layers of ResNet-Generator with encoder-decoder style [50] and the PatchGAN-Discriminator [23] for all of the models. Besides, we choose the Resnet-19 as our T network structure. For all of the model optimization, we set the batch-size to 4 and optimizer to Adam with learning rate 2×10^{-4} and $\beta = [0.5, 0.999]$. On all of the dataset, to be fair, we train each model with 200 epoches and we report the performance of the model from the last epoch because of no validation is provided.

Additionally, for our MSPC model, we have three minimax game between G, T, D, D_{pert}. Thus, we separate the model training procedure into two steps, $\{D, D_{pert}, T\} - step$ and $G - step$. In each step, we only optimize the corresponding networks and fix others. The size of the spatial transformation grid is 2×2. For all the experiments, we set

Figure 5. Examples on dataset with paired source and target images, all examples are held out from training dataset. The front face→profile task does not include any paired identity, which is a difficult setting and CycleGAN, GCGAN and CUT cannot be stably trained and collapse in the early training stage. Our model shows a stability across all tasks of image translation.

Method	**Cityscapes→Parsing**			**Front Face→Profile**	**Horse→Zebra**
	pixAcc↑	**classAcc↑**	**mAP↑**	**FID↓**	**FID↓**
CycleGAN [50]	0.595	0.234	0.171	107.70	69.40
GCGAN [16]	0.563	0.195	0.143	128.31	74.89
CUTGAN [37]	0.587	0.225	0.166	244.50	84.26
MT Modified	0.121	0.055	0.018	52.95	62.28
VAT Modified	0.484	0.100	0.064	145.54	70.21
MSPC (ours)	**0.740**	**0.296**	**0.226**	**37.01**	**61.2**
Method	**Parsing→Cityscapes**			**Aerial Photograph→Map**	
	pixAcc↑	**classAcc↑**	**mAP ↑**	**RMSE↓**	**PixACC ↑**
CycleGAN [50]	0.508	0.184	0.117	**32.70**	**0.265**
GCGAN [16]	0.583	0.201	0.128	33.12	0.264
CUTGAN [37]	**0.681**	**0.243**	**0.172**	35.45	0.222
MT Modified	0.455	0.145	0.086	35.43	0.216
VAT Modified	0.281	0.109	0.053	63.38	0.042
MSPC (ours)	0.612	0.214	0.156	32.97	**0.265**

Table 1. **Comparison with baselines** on four dataset with quantitative results, they are conducted on the translation settings of cityscapes→parsing, parsing→cityscapes, front face→profile, horse→zebra and aerial photograph→map respectively. The best scores are bold. Our model shows overall competitive results and robust performance across different settings.

the maximum scale of perturbation to be $a = \frac{1}{3}, b = 3$ and the translation factors to be $c = -0.25, d = 0.25$.

4.2. Dataset Configuration and Results

Front Face→Profile In this new dataset, we aim to have an unbiased evaluation metric in real-life applications and explore the possibility of performing the image translation task under a big gap between the source and target domains. To construct such a front face→profile image translation dataset, we sample from CMU Multi-PIE Face [21], which consists of 250 identities with different camera angles and the conditions of illumination. We extract two angles of the front and the profile from the dataset and divide them into training and testing sets by different identities. All face images are resized to 128×128.. In the training set, we have 200 identities, 100 in the source and 100 in the target, which do not overlap. For the testing division, we set the source and the target to be paired and calculate the FID score between the translated profile faces and the ground truth of the profile faces. it is worth mentioning that the FID socore is unbiased in this setting due to the paired idenity in the

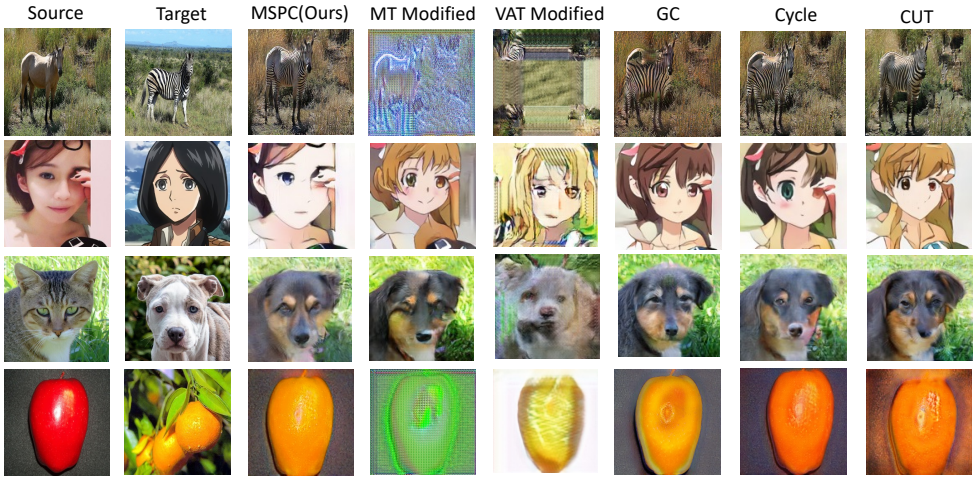


Figure 6. Examples on dataset with unpaired source and target images, all examples are held out from training dataset.

testing set. The lower FID score on the testing set indicates better performance of models.

Quantitative results are shown in Table 1 and some of the qualitative results are shown in Figure 5. More qualitative results will be listed in the Section 2 of the supplementary. As we can see from the table and the generated faces, the CycleGan, GCGAN, and CUTGAN failed to stably generate profile from front faces and that our model of the MSPC and the modified MT can generate the faces with high fidelity. Furthermore, except our model, all the remaining models fail to translate front face to the profile while keeping the identity. This illustrates that our model is robust to the large domain gap in the image translation task.

Cityscapes consists of city scene images and the mask-level annotation, which can be used to test the ability of model to discover the correspondence between data and labels. There are 3,975 images with paired segmentation mask, 19 categories, and 1 ignored class. We follow the standard training setting of [16, 37, 50]: the dataset is separated into the 2,975 and 500 samples for training and testing. The original resolution of the image is 1024×2048. During the training, the images are resized to 128×128 for city→parsing direction. For the parsing→image synthesis, we first resize images into 144×144 and then randomly crop images to be 128×128. In this experiment, we are trying to explore how well the models can discover the semantics without paired labels.

For the evaluation on cityscapes dataset, we follow the same protocol of [9, 30, 50]. We report the average pixel accuracy, class accuracy, and the mean IOU with respect to the ground truth. To evaluate the quality of parsing→image synthesis, we utilize the pre-trained FCN [23] to extract the predicted segmentation map.

Aerial photo→Map The setting of the dataset is similar to Cityscapes and is obtained from the Google Map [23]. It contains 1096 training images and 1098 testing images. We conduct the translation in direction Aerial photo→Map. The images are resized to 256×256. The RMSE and the pixel accuracy are reported across different models.

Horse→Zebra For the Horse→Zebra translation scenario, we test if the model is capable of handling the case of real-life applications. The dataset is re-sampled from ImageNet [10]. The source dataset includes 939 horse images and the target includes 1177 zebra images from the wild. The images are resized to 256×256. Because there are no paired images in the testing set, the FID score is biased and reported for reference only.

Overall, our model gain a competitive performance on all dataset settings and shows a very robust generality. We found that CUT achieves high scores of semantic segmentation on the Parsing→Cityscapes task and that CycleGAN has the best results on Aerial photo→Map. On the remaining datasets, our model always achieves the best results under the same settings. CUT owns the feature of maximizing the mutual information, which can translate images well on a setting without changing much semantic information. The bijective assumption of CycleGAN is suitable for the Map dataset. More qualitative results are shown in Figure 6, which are operated on horse→zebra, selfie→anime, cat → dog, and apple→orange. One can see that the proposed MSPC can preserve the image features well and does not cause unnecessary change of the background, which shows the ability of the spatial alignment of the proposed MSPC.

4.3. Ablation Study

Front Face → Profile, changing scaling factor a. FID ↓.					
$a = 1$	$a = 2$	$a = 3$	$a = 5$	$a = 8$	RSP
42.19	41.82	37.01	38.72	60.21	67.33

Table 2. This tables shows the results of the proposed MSPC under different scales of perturbation by changing the scaling factor of a as well as the random spatial perturbation (RSP) for comparison.

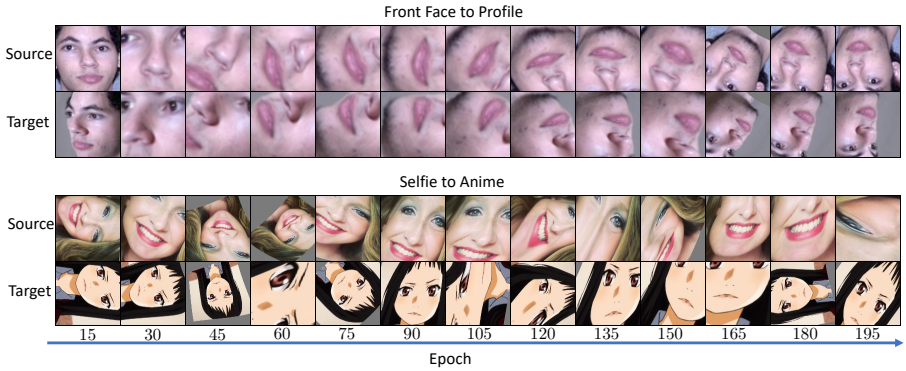


Figure 7. Perturbation changes as epoch grows.

Effect of scale of perturbation To study the effect of perturbation on model performance, we change the scaling factor of the proposed scaling constraint and conduct the experiments on the front face→profile setting and report the FIDs. To show the effectiveness of maximum perturbation, we also compare with the model of random spatial perturbation (RSP) in Table 2, in which the spatial transformation is randomly sampled from the fixed and predefined spatial transformation of rotation, cropping, zoom in, zoom out, stretching, and squeezing. The results in Table 2 shows that in a certain range of perturbation, more severe perturbation leads to better performance. However, if the perturbation goes beyond image distribution, e.g., images get unreasonably distorted, the performance of MSPC would be cut back. Also the visualization of perturbation without the constraint in Equation 6 is shown in the Section 3 of supplementary. Also, we show the dynamic changing of perturbation during training in Figure 7.

Front Face → Profile, divergence between distributions. FID ↓			
X, Y	$T(X), T(Y)$	$G(X), Y$	$G(T(X)), T(Y)$
112.69	65.81	37.01	30.85

Table 3. This tables quantifies the effect of spatial alignment by transformer T. Each row reports the divergence between listed pairs. $X, Y, T(X), T(Y)$ denote the source images, target images, transformed source images by T, and transformed target images by T. $G(X)$ is the translated images and $G(T(X))$ represents the translated transformed images.

Effect of Spatial Alignment of T As we have mentioned in Section 3.3, the spatial perturbation function also plays a role in aligning the image distributions. We conduct an experiment on front face→profile to demonstrate this effect by comparing the FID score between different data pairs. We listed all the controlling pairs in Table 3. (X, Y) denotes the divergence between the original source and target images without image translation or spatial transformation. $(T(X), T(Y))$ is the pair of images of spatial transformed source and target images. $G(X)$ and $G(T(X))$ represent the translated images and the translated spatial transformed images. The divergence of pair of $(T(X), T(Y))$ is smaller than (X, Y), because of the effect of spatial alignment by T only. The divergence is further reduced after both the spatial alignment and the image translation, compared to the pair of $(G(X), Y)$ with only image translation. The result clearly shows that the transformer T is capable of alleviating the discrepancy in distribution between the source and the target via the spatial transformation.

5. Conclusion

This paper proposes a general regularization method of maximum spatial perturbation consistency (MSPC) to address the limitations of the popular models for image-to-image translation (I2I), including [16, 37, 50]. We demonstrate 1) that the proposed MSPC is more robust to different applications; 2) that MSPC can help alleviate the spatial discrepancy between domains, such as the discrepancy caused adjusting the object's size and cropping out the noisy background, and further reduce undesired distortions for the translation network. Our method outperforms the state-of-the-art methods on most of of the I2I benchmarks. We also introduce a new benchmark, namely, the front face to profile face dataset, to emphasize the underlying challenges of I2I for real-world applications. We finally perform ablation experiments to investigate the sensitivity of our method to the severity of spatial perturbation and its effectiveness for distribution alignment.

6. Acknowledge

This work was partially supported by NIH Award Number 1R01HL141813-01, NSF 1839332 Tripod+X, SAP SE, and Pennsylvania Department of Health. We are grateful for the computational resources provided by Pittsburgh SuperComputing grant number TG-ASC170024. MG is supported by Australian Research Council Project DE210101624. KZ would like to acknowledge the support by the National Institutes of Health (NIH) under Contract R01HL159805, by the NSF-Convergence Accelerator Track-D award #2134901, and by the United States Air Force under Contract No. FA8650-17-C7715.

References

[1] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14154–14163, 2021. 2

[2] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 2

[4] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[5] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8168–8177, 2020. 2

[6] Wonwoong Cho, Sungha Choi, David Keetae Park, Inkyu Shin, and Jaegul Choo. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10639–10647, 2019. 2

[7] 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

[8] 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

[9] 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 (CVPR)*, June 2016. 5, 7

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7

[11] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 2

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 1

[13] D. Ponsa E. Rublee E. Riba, D. Mishkin and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Winter Conference on Applications of Computer Vision*, 2020. 1

[14] Alexei Efros and Thomas Leung. Texture synthesis by non-parametric sampling. In *In International Conference on Computer Vision*, pages 1033–1038, 1999. 2

[15] Francesco Cardinale et al. Isr. `https://github.com/idealo/image-super-resolution`, 2018. 1

[16] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 6, 7, 8

[17] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2

[18] Mingming Gong, Yanwu Xu, Chunyuan Li, Kun Zhang, and Kayhan Batmanghelich. Twin auxilary classifiers gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 1330–1339. Curran Associates, Inc., 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 Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2, 3

[20] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015. 3

[21] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–8, 2008. 6

[22] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2, 5, 7

[24] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2, 3, 4

[25] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *ArXiv*, abs/1612.00215, 2016. 2

[26] 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

[27] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 2

[28] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[29] 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/CVF International Conference on Computer Vision*, pages 10551–10560, 2019. 2

[30] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society. 7

[31] Youssef A Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image to image translation. *arXiv preprint arXiv:1806.02311*, 2018. 2

[32] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2019. 3

[33] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 2

[34] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018. 2

[35] Aamir Mustafa and Rafał K. Mantiuk. Transformation consistency regularization- a semi-supervised paradigm for image-to-image translation. In *ECCV*, 2020. 2

[36] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2

[37] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020. 1, 2, 3, 6, 7, 8

[38] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 3

[39] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6836–6845, 2017. 2

[40] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions, 2020. 2

[41] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *arXiv:1711.10925*, 2017. 1

[42] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *arXiv:1711.10925*, 2017. 2

[43] 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. *CoRR*, abs/2108.04547, 2021. 2

[44] 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 (ICCV)*, pages 14174–14184, October 2021. 2

[45] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *CoRR*, abs/2103.00550, 2021. 3

[46] Xiaoming Yu, Yuanqi Chen, Thomas Li, Shan Liu, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. *arXiv preprint arXiv:1909.07877*, 2019. 2

[47] Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, Qingsong Liu, and Clark Glymour. Domain adaptation as a problem of inference on graphical models. *Advances in Neural Information Processing Systems*, 33, 2020. 2

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

[49] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 1

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3, 5, 6, 7, 8