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

CVPR 2022

Haneol Lee

Vision Study

25th July, 2022

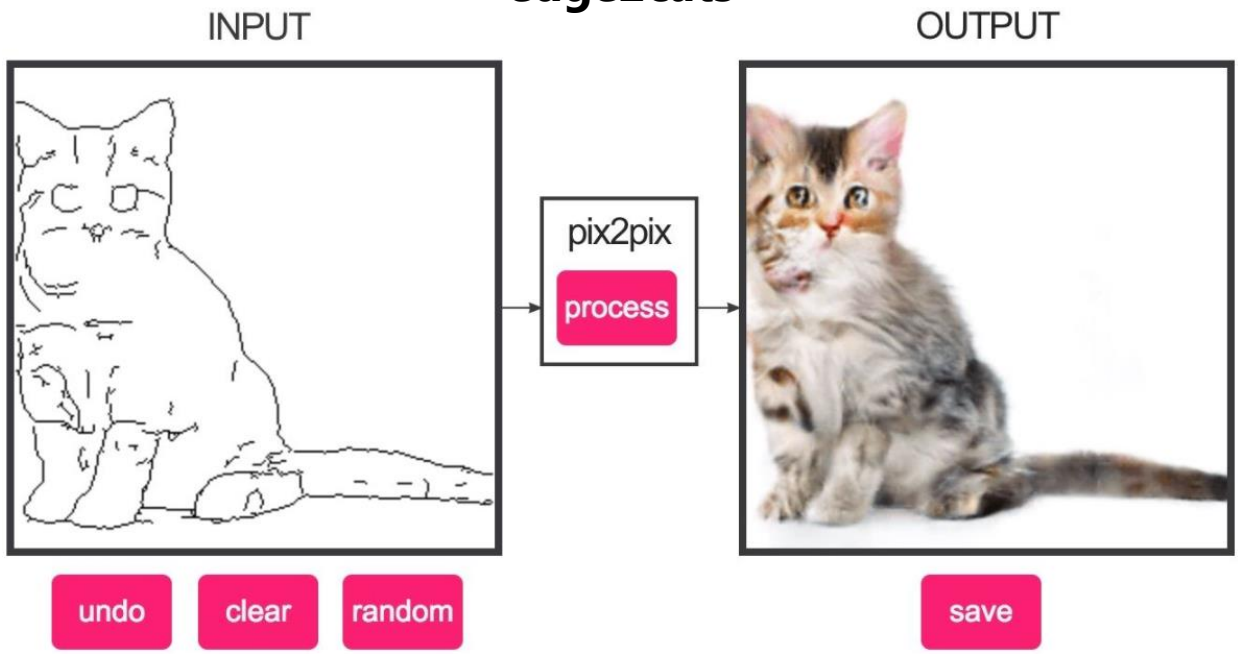
1. Image to Image Translation (I2I)?

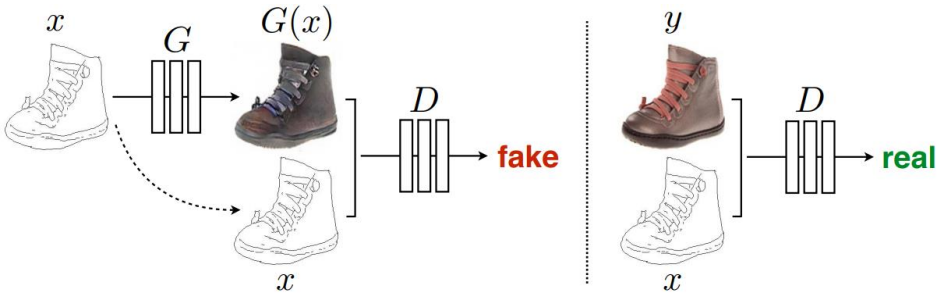
Cycle Consistency Regularization?

Pix2Pix, CycleGAN

Backgrounds: Image to image translation

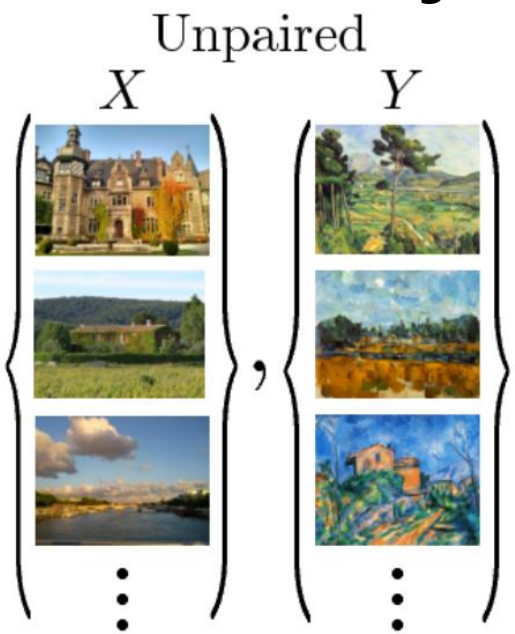
edge2cats

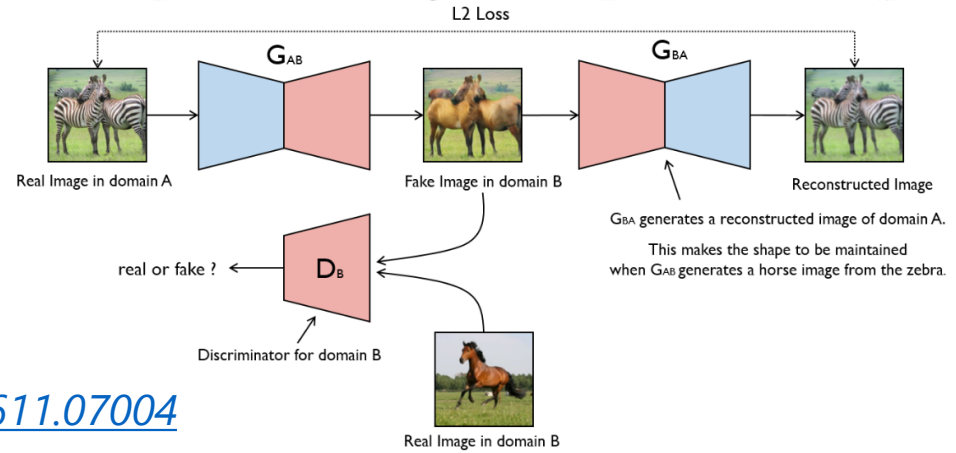




Pix2Pix
edge2shoes

CycleGAN
Photos⇔Paintings



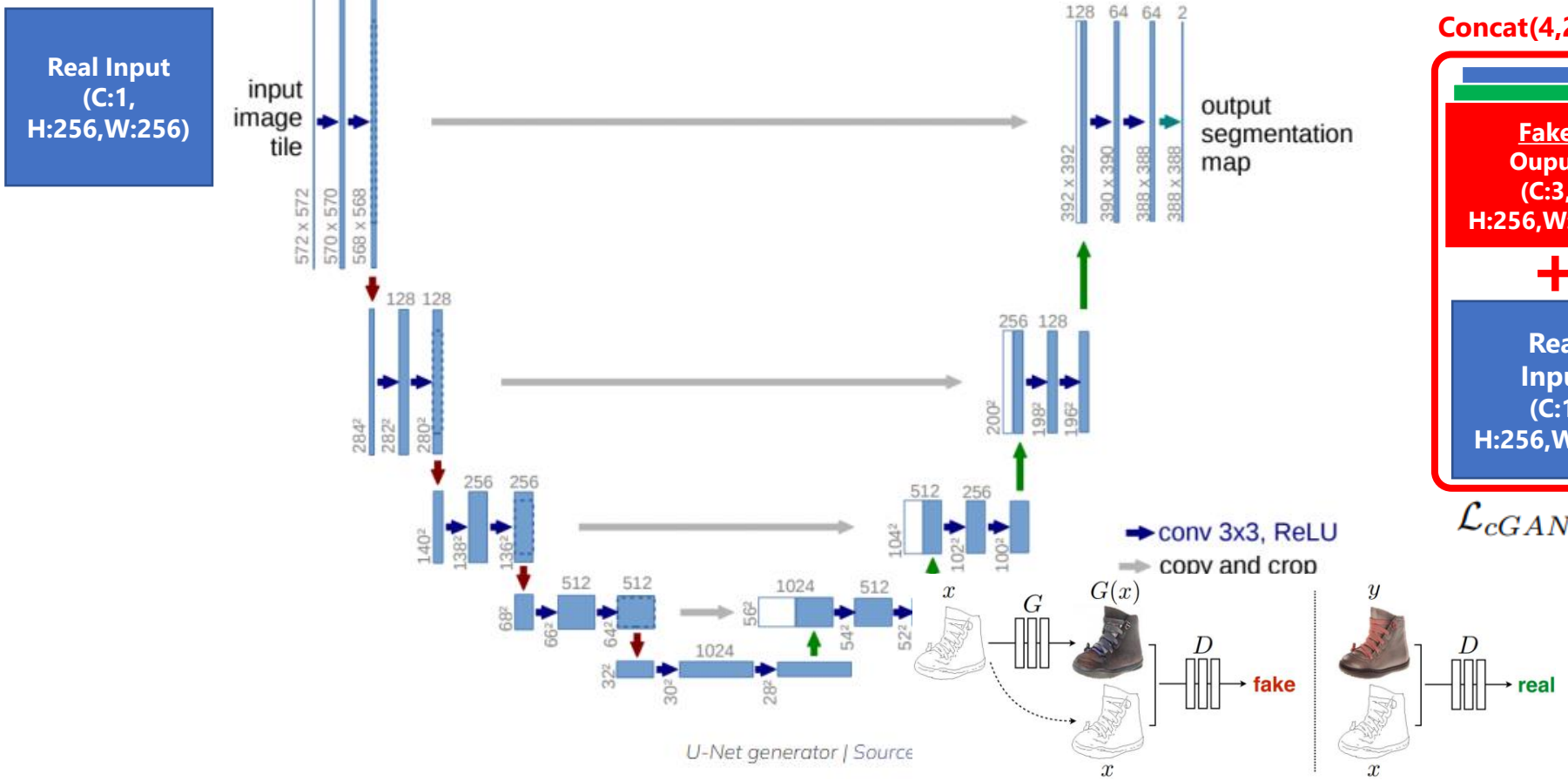


Link: https://arxiv.org/abs/1611.07004

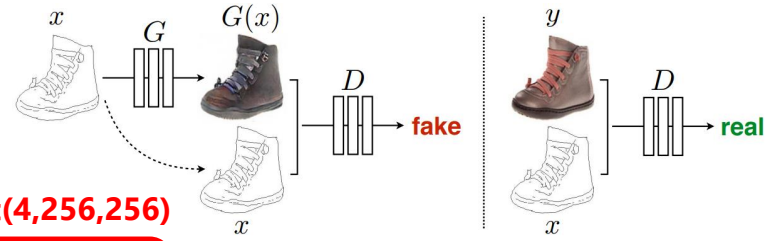
Link: https://arxiv.org/abs/1703.10593

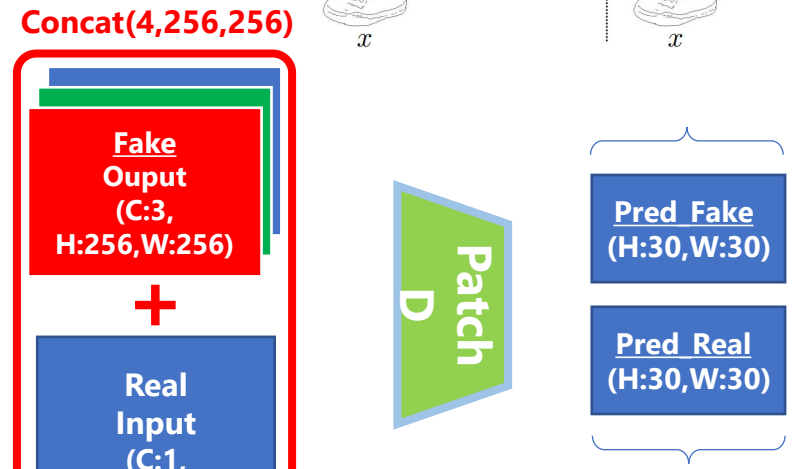
Backgrounds: Pix2Pix Architecture

- *Generator: **Unet256**, ResNet*



- *Discriminator*





$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z))]$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

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

Abstract

- **Unpaired I2I는 ill-posed problem이며, 소스 도메인에서 타겟 도메인 분포로의 맵핑에 있어서 무한에 가깝게 많은 경우의 수가 있으므로 풀기 어려운 문제.**
- **따라서 적절한 constraints(제약)을 주는 게 중요한데, 가령, cycle consistency, geometry consistency, 그리고 contrastive learning-based consistency가 있음. 그러나 (1) 기존 방식들은 too restrictive하거나, 특정 I2I tasks에서는 그 효과가 너무 약함. (2) 소스와 타겟 도메인 사이에 spatial perturbation이 있을 경우, content distortion을 유발시킨다는 문제가 있음.**
- **위 논문에서는 universal regularization technique인 maximum spatial perturbation consistency (MSPC)을 제안하고, spatial perturbation function (T)와 Translation operator (G)가 서로 commutative하도록 설정하는 universal regularization을 적용하여 기존 문제를 해결하는 제약을 줌. (i.e. $T \circ G = G \circ T$)**
- **I2I 벤치마크 모델에서 SOTA 성능 및 comparable한 성능 달성 (정량, 정성 평가)**
- **제안한 method의 Spatial perturbation에 대한 severity와 distribution alignment의 효과에 대한 Ablation study 진행**

Introduction

Goal: Image translation $\{x; x \in \mathcal{X}\}$ **to** $\{y; y \in \mathcal{Y}\}$

$\mathcal{X}, \mathcal{Y} \subseteq \mathbf{R}^{C \times H \times W}$ **: Unpaired images**

Consistency Regularization of Semi-Supervised Learning

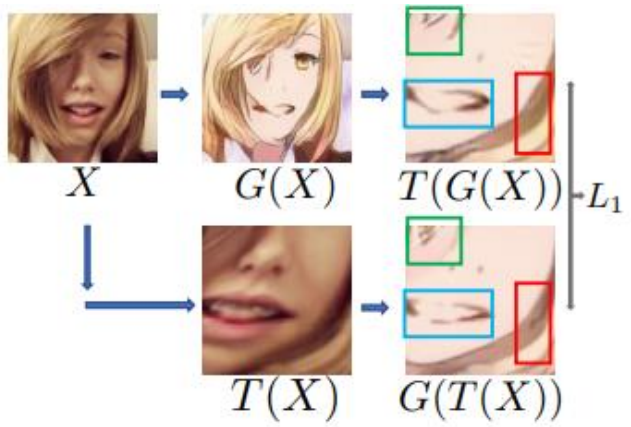
$$\min_{f} \max_{\gamma; \|\gamma\| \leqslant \delta} \mathbb{E}_{x \in P_X} \mathcal{R}(f(\theta, x), f(\theta, x + \gamma)). \quad (1)$$

$\mathcal{R}, f$ **: the estimation of distance between two vectors and the predicted model respectively**

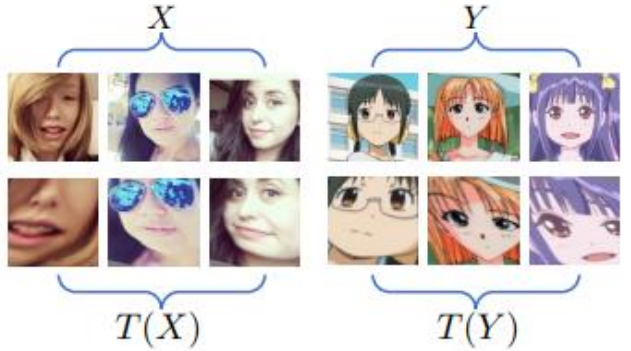
The Representatives of Consistency Regularization:

- **GCGAN**
- **CUTGAN**

Virtual adversarial training (VAT): the concept of adversarial attack as a consistency regularization in semi-supervised classification.



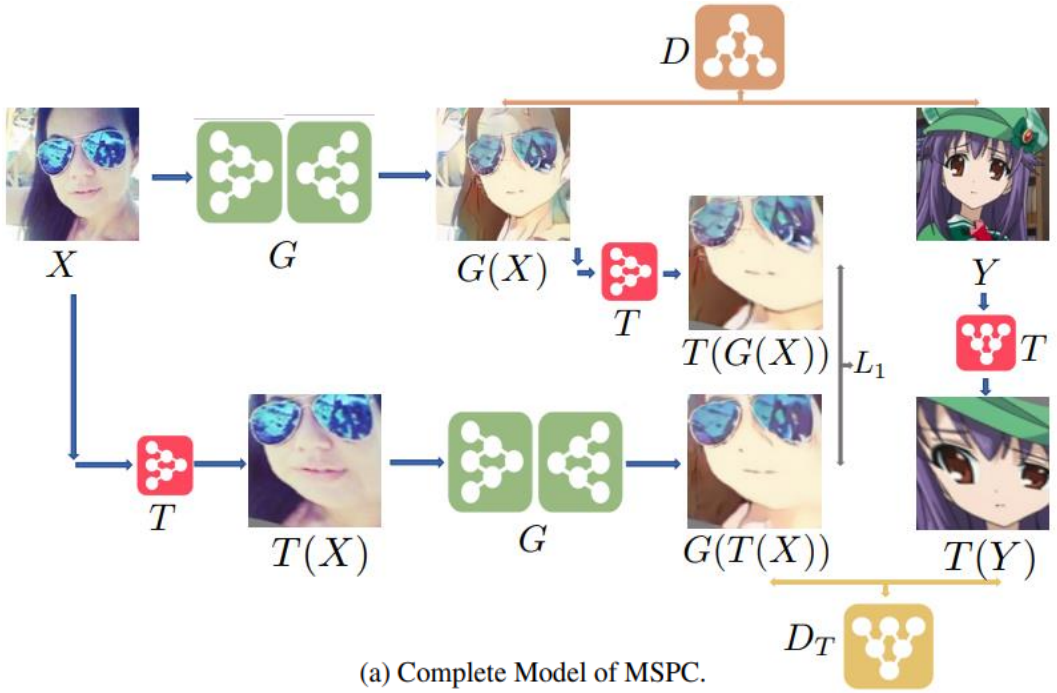
(a) Consistency regularization with spatial perturbation function T



(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.

Method: proposed MSPC model



(a) Complete Model of MSPC.

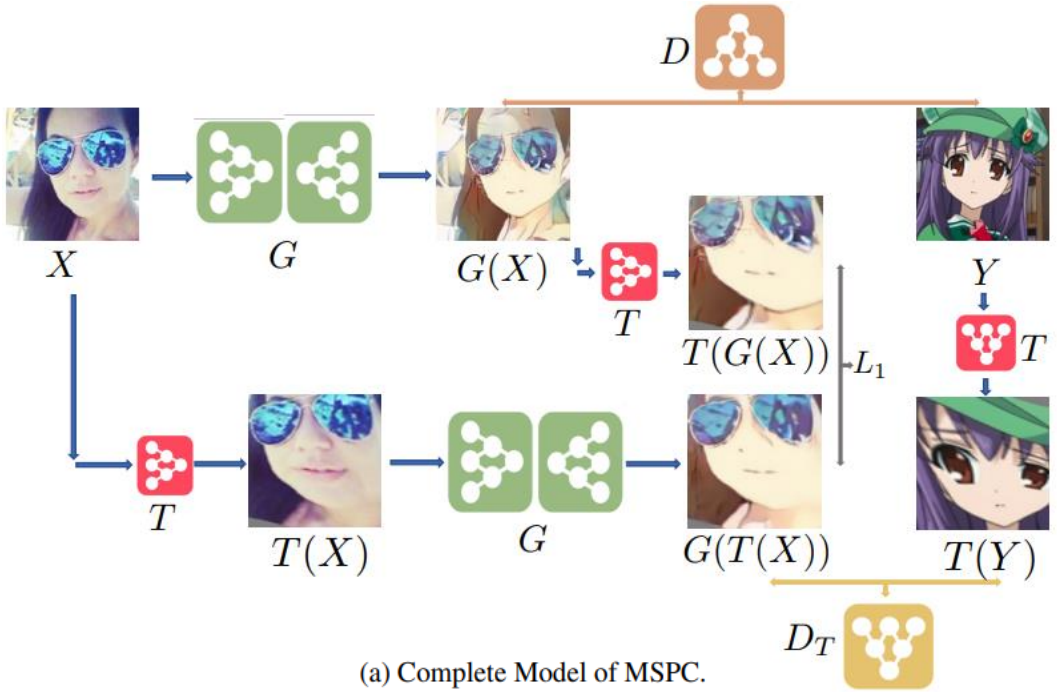
Model consists of three branches of learning

1) **X → G → D ← Y**
2) **X → G → T → L1 ← G ← T ← X**
3) **X → T → G → D_T ← T ← Y**

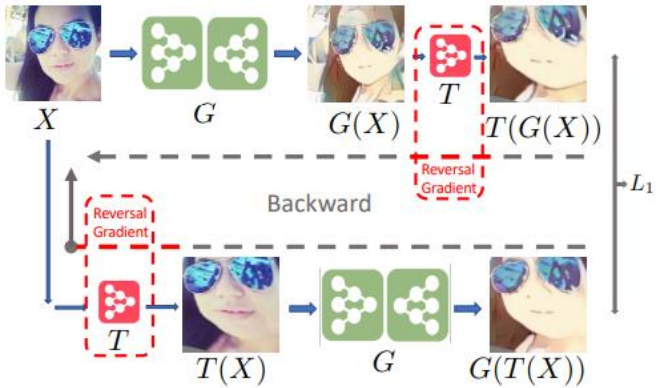
3.1. Adversarial Constraint on Image Translation

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

Method: proposed MSPC model



(a) Complete Model of MSPC.



(b) Maximum Spatial Perturbation Consistency

Model consists of three branches of learning

1) **X → G → D ← Y**
2) **X → G → T → L1 ← G ← T ← X**
3) **X → T → G → D_T ← T ← Y**

3.1. Adversarial Constraint on Image Translation

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

3.2. Maximum Spatial Perturbation Consistency

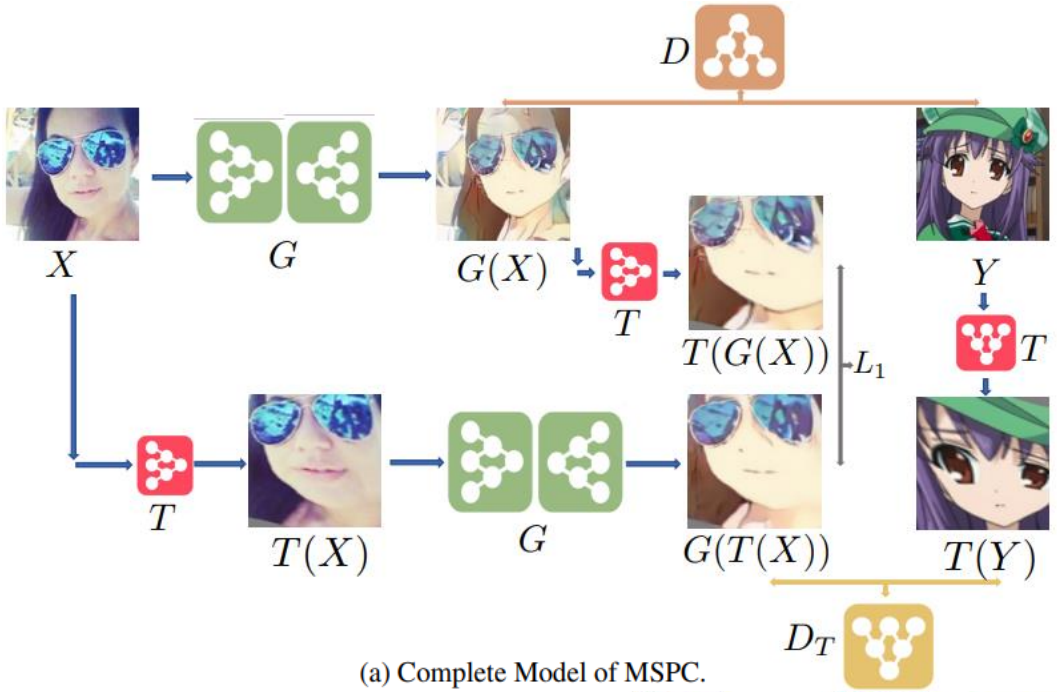
$$\min_G \max_T \mathbb{E}_{x\sim P_X} \|T(G(x)) - G(T(x))\|_1, \quad (2)$$

T_i : **Adversarial spatial perturbation network, jointly learned with** G

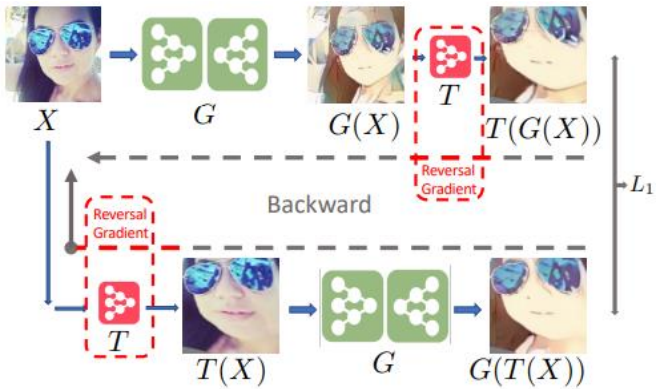
$$G(T_i(x_i)) \approx T_i(G(x_i))$$

***The architecture of T is based on ResNet19**

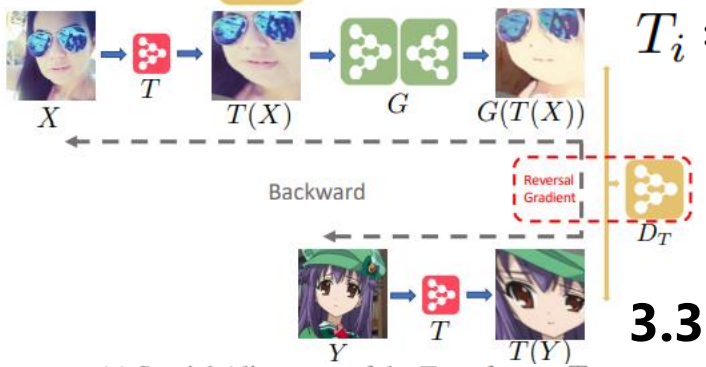
Method: proposed MSPC model



(a) Complete Model of MSPC.



(b) Maximum Spatial Perturbation Consistency



(c) Spatial Alignment of the Transformer T.

Model consists of three branches of learning

1) X → G → D ← Y
2) X → G → T → L1 ← G ← T ← X
3) X → T → G → D_T ← T ← Y

3.1. Adversarial Constraint on Image Translation

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

3.2. Maximum Spatial Perturbation Consistency

$$\min_G \max_T \mathbb{E}_{x\sim P_X} \|T(G(x)) - G(T(x))\|_1, \quad (2)$$

T_i : **Adversarial spatial perturbation network, jointly learned with** G

$$G(T_i(x_i)) \approx T_i(G(x_i))$$

3.3. Spatial Alignment of the Transformer T

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

***The architecture of T is based on ResNet19**

Method: proposed MSPC model

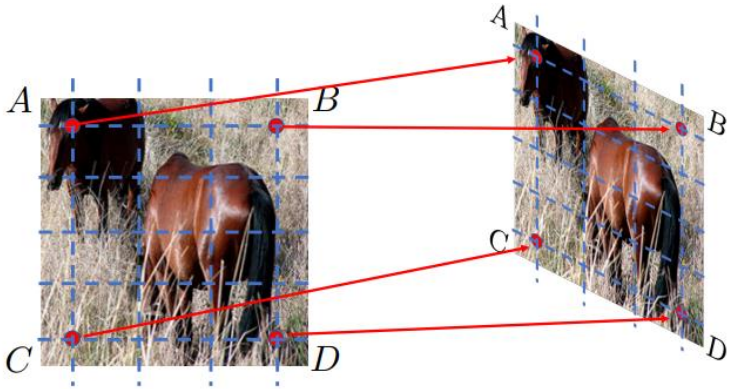


Figure 3. Illustration of spatial transformation network (STN). The network T outputs the coordinates of the deformed grids over the images and then the new images are generated via interpolating in these grids; it is differentiable and can be optimized with stochastic gradient decent.

3.4. Differentiable *T*

$$\{(p_i^1, p_j^2); i = 1, 2, 3, ..., n, j = 1, 2, 3, ..., m\} = T(x),$$

$$V_{i,j}^c = \sum_n^H \sum_m^W U_{nm}^c k(p_i^1 - q_m^1; \Phi_{p^1}) k(p_j^2 - q_n^2; \Phi_{p^2}),$$

$$\forall i, m \in [1 \dots H];\ \forall j, n \in [1 \dots W];\ \forall c \in [1 \dots C], \quad (4)$$

(q_i^1, q_i^2) : the coordinate of original grid

(p_i^1, p_i^2) : the new coordinates of transformed grids

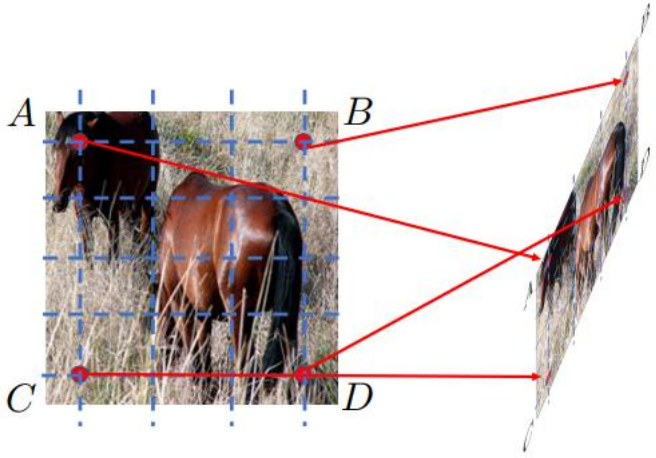
U, V_i : the pixel value of original image, transformed image

$k(;\Phi_{p^1}), k(;\Phi_{p^2})$: the kernel of the interpolating image

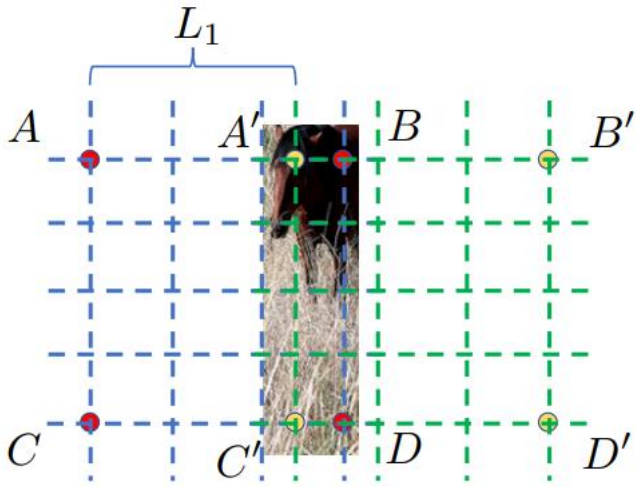
***Differentiable T is referenced by 'Spatial transformer networks',** Max et al.

****The architecture of T is based on ResNet19** T is **differentiable**, **can be optimized** with stochastic gradient decent.

Method: proposed MSPC model



(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.

3.5. Constraint on T

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

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

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

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

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

$$s.t. \; \frac{1}{a} < \frac{|p_i p_j|}{|q_i q_j|} < a, \; i \neq j \; \& \; -b < \sum_{i=1}^{n} p_i < b. \tag{7}$$

(q_i^1, q_i^2) : the coordinate of original grid

(p_i^1, p_i^2) : the new coordinates of transformed grids

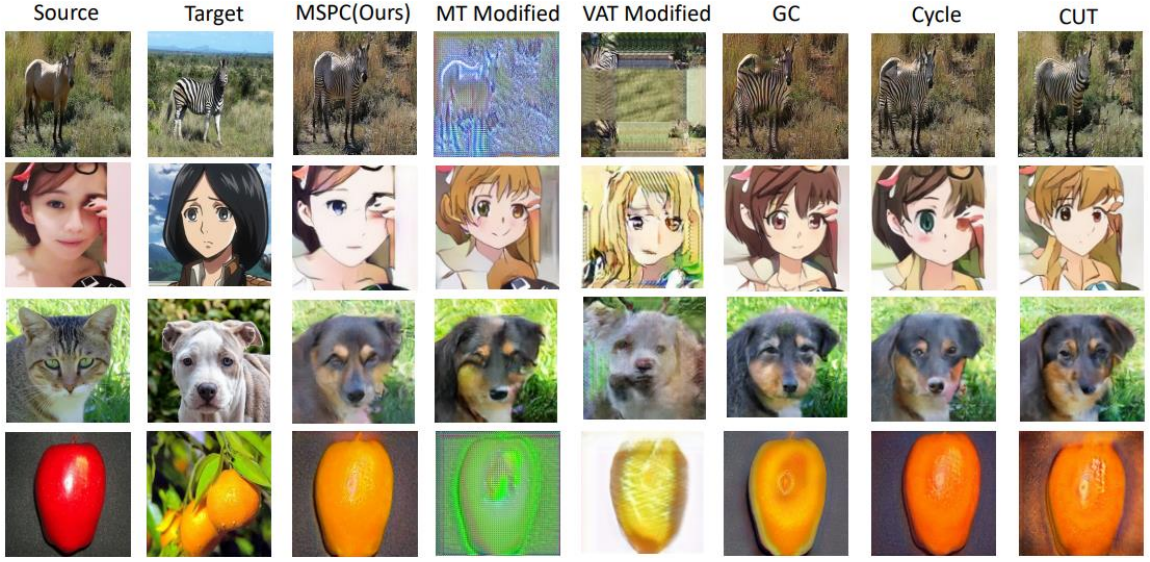
T is **differentiable**, **can be optimized** with stochastic gradient decent.

Training Configuration

- **G: ResNet layer encoder-decoder, T: ResNet-19.**
- **We choose the 9-layers of ResNet-Generator with encoder-decoder style [54] and the PatchGAN-Discriminator [25] for all of the models.**
- **Adam optimizer with learning rate $2 \times 10-4$ and $\beta = [0.5, 0.999]$, batch size 4.**
- **Training every model until 200 epoches.**
- **Three min-max game between G, T, D, D_{Pert}.**
- **When separating the model training procedure into two steps, {D, D_{Pert}, T} − step and G − step, only optimize the corresponding networks and fix others.**
- **$a = 1/3$, $b = 3$ and the translation factors to be $c = -0.25$, $d = 0.25$.**

Experiments

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→Profle	Horse→Zebra
	pixAcc↑	classAcc↑	mAP↑	FID↓	FID↓
CycleGAN [54]	0.595	0.234	0.171	107.70	69.40
GCGAN [18]	0.563	0.195	0.143	128.31	74.89
CUTGAN [39]	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 [54]	0.508	0.184	0.117	**32.70**	**0.265**
GCGAN [18]	0.583	0.201	0.128	33.12	0.264
CUTGAN [39]	**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**

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

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.

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.

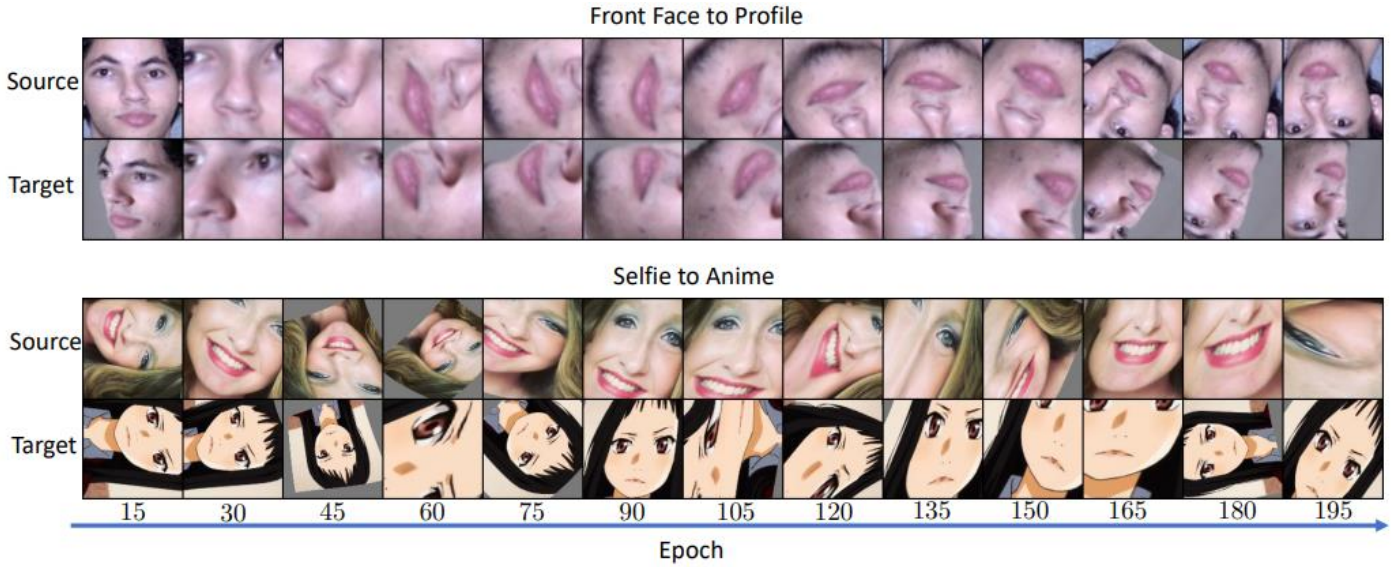


Figure 7. Perturbation changes as epoch grows.

(X, Y) : Divergence between X, Y.

$\big(T(X), T(Y)\big)$, Divergence between $T(X), T(Y)$, is smaller than (X, Y), because of the effect of spatial alignment by T only.

Qualitative Results

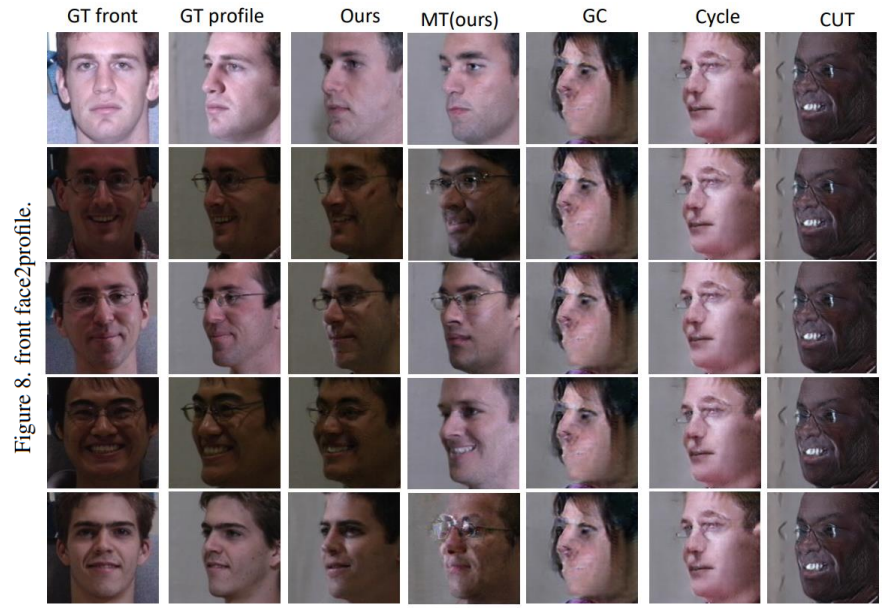


Figure 8. front face2profile.

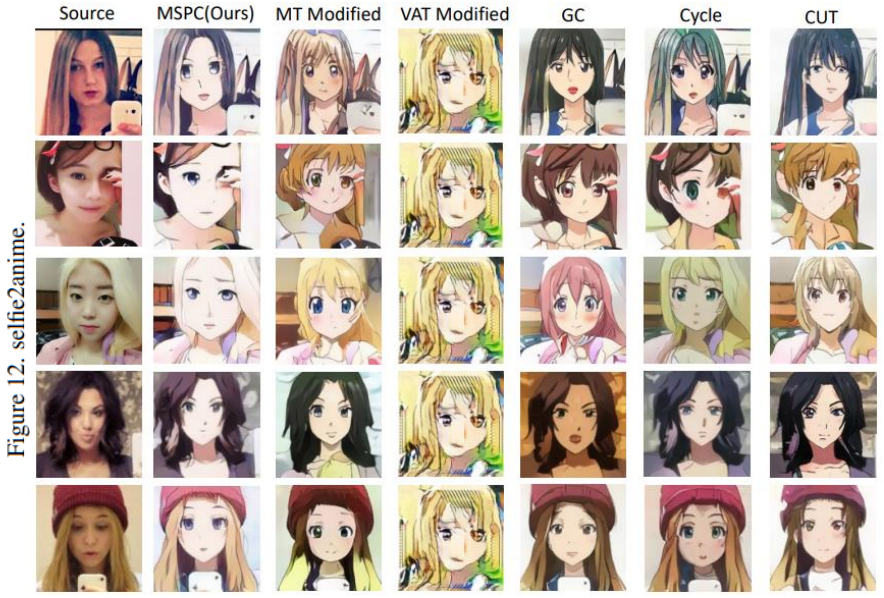


Figure 12. selfie2anime.

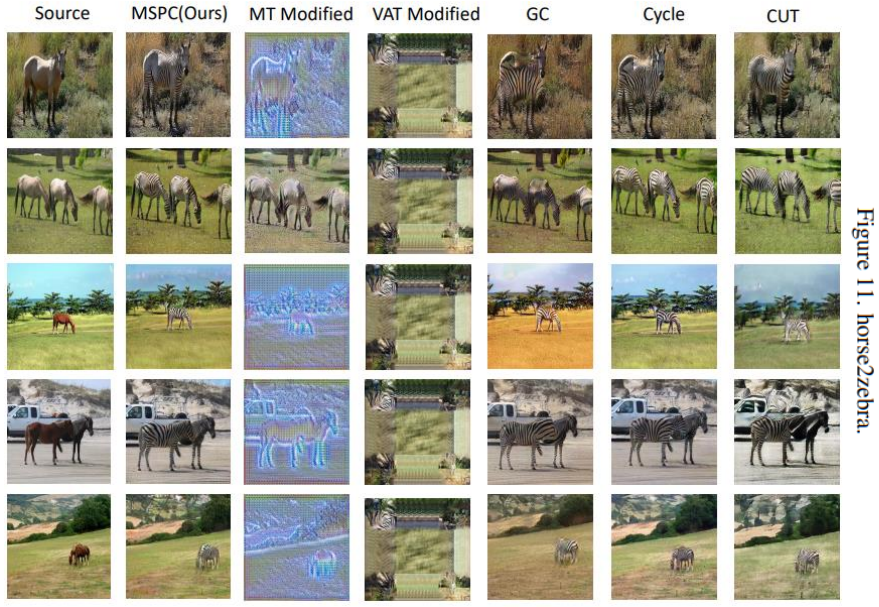


Figure 11. horse2zebra.

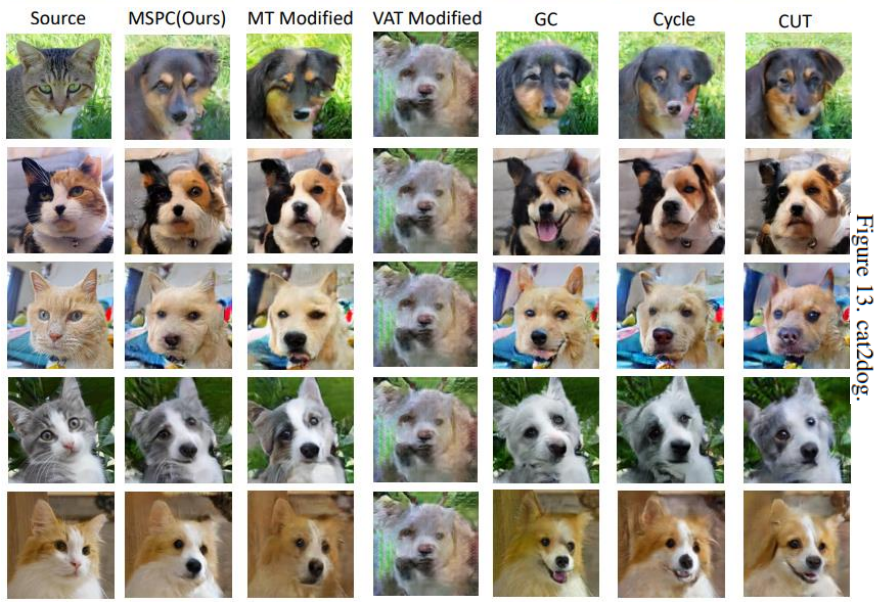


Figure 13. cat2dog.

Qualitative Results

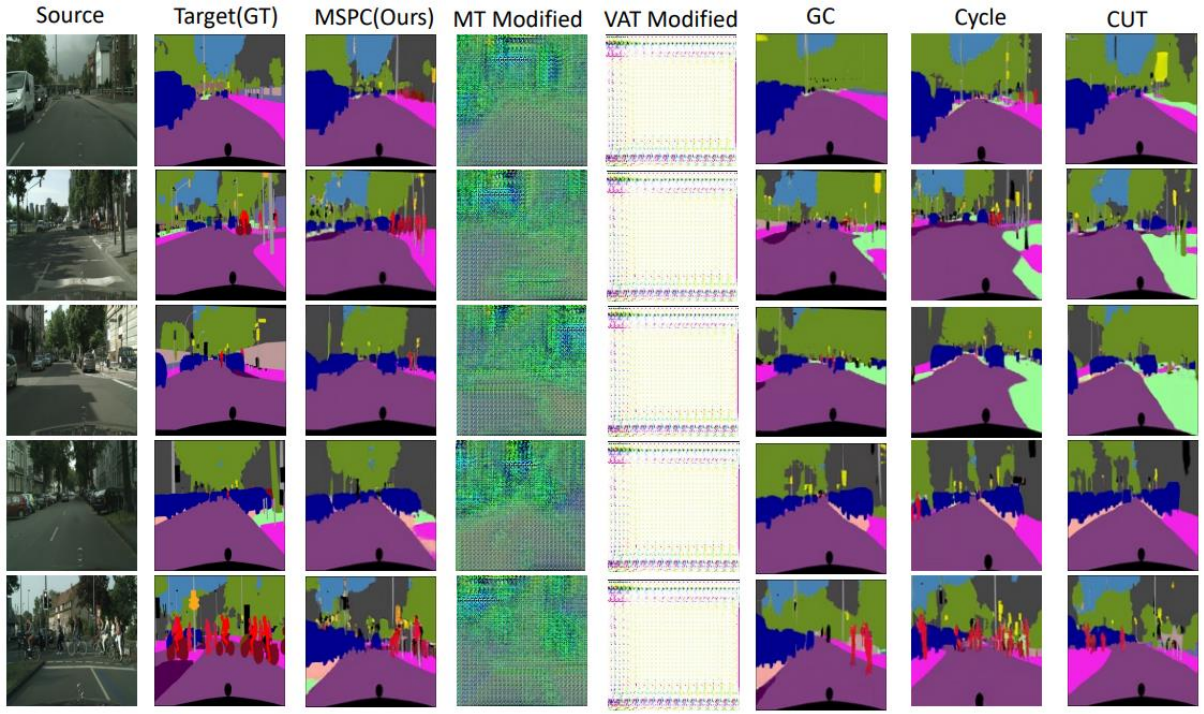


Figure 9. city2parsing.

Figure 10. city2parsing.

Conclusion

- **Propose a general regularization method of maximum spatial perturbation consistency (MSPC) to address limitations of previous I2I models.**
    1) **MSPC is more robust to different applications.**
    2) **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.**
    3) **MSPC can reduce undesired distortions for the translation network.**
- **Our method outperforms the state-of8 the-art methods on most of the I2I benchmarks.**
- **Investigate the sensitivity of our method to the severity of spatial perturbation and its effectiveness for distribution alignment.**

Q&A

Supplementary: Additional details of MSPC

Table A. This tables shows the results of the proposed MSPC and MSPC without the spatial alignment branch in Fugure 2(c) for comparison. To show the stability, we run each setting for 5 times and calculate the mean and std.

Front Face → Profile. FID ↓.	
MSPC	MSPC without spatial alignment
38.61 ± 2.57	53.41 ± 4.83

Table A. This tables shows the results of the proposed MSPC and MSPC without the spatial alignment branch in Fugure 2(c) for comparison. To show the stability, we run each setting for 5 times and calculate the mean and std.

Supplementary: Modified Virtual Adversarial Training (VAT)

VAT [34] introduced the concept of adversarial attack [22] as a consistency regularization in semi-supervised classification. This method learns a maximum adversarial perturbation as a additive , which is on the data-level. To be more specific, it finds an optimal perturbation γ on an 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:

$$\min_{f} \max_{\gamma;\|\gamma\|\leqslant\delta} \mathbb{E}_{x\in P_X} \mathcal{R}(f(\theta, x), f(\theta, x+\gamma)). \tag{8}$$

$$\begin{aligned}
&\min_{G} \max_{D,D_T} \mathbb{E}_{y\sim P_Y} \log D(y) + \mathbb{E}_{x\sim P_X} \log(1 - D(G(x))) \\
&+ \mathbb{E}_{y\sim P_Y} \log D_V(y) + \mathbb{E}_{x\sim P_X} \log(1 - D_V(G(x+\gamma))), \\
&\min_{G} \max_{\gamma;\|\gamma\|\leqslant\delta} \mathbb{E}_{x\sim P_X} \|G(x), G(x+\gamma)\|_1.
\end{aligned} \tag{9}$$

Supplementary: Visualization of Transformer T without constraints

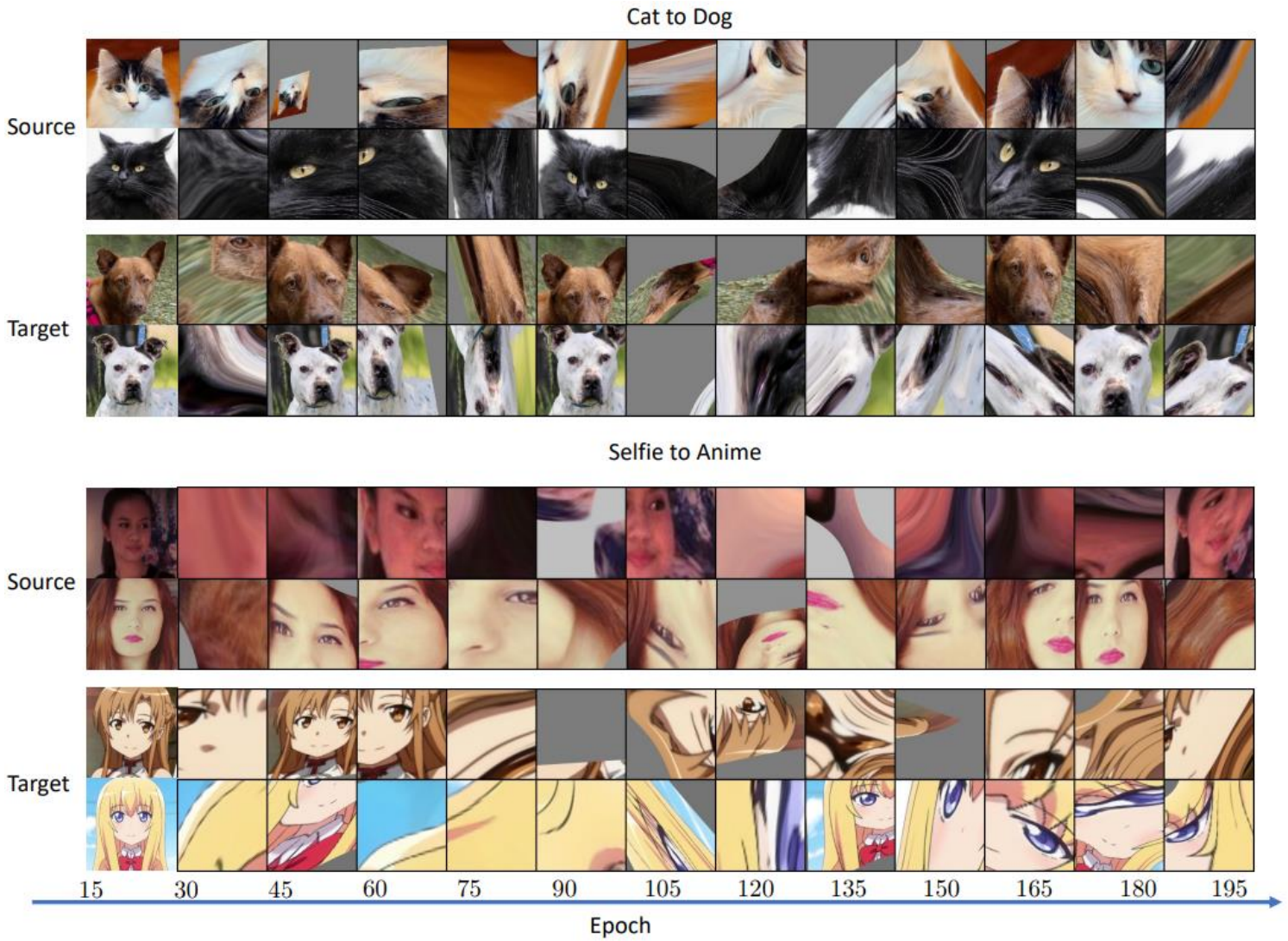


Figure 15. Perturbation changes as epoch grows. In this figure, we do not add the constraint to the T.