Supplementary of "Maximum Spatial Perturbation Consistency for Unpaired Image-to-Image Translation"

Algorithm 1 Two steps of optimization of mini-batch MSP for Eq. 7, α denotes the learning rate, $\max(\cdot,\cdot)$ and $\min(\cdot,\cdot)$ represent the hinge loss.

Choose M samples of $x^{(i)}(i = 1,\dots,M)$ and N samples of $y^{(j)}(j = 1,\dots,N)$ from $\mathcal{X},\mathcal{Y}$ respectively.
Optimizing D, D_{pert}, T

$\mathcal{R}_1 = \frac{1}{N}\sum_{j=1}^{N}\log D(y_j) + \frac{1}{M}\sum_{i=1}^{M}\log(1-D(G(x_i)))$

$\mathcal{R}_2 = \frac{1}{N}\sum_{j=1}^{N}\log D_T(T(y_j)) + \frac{1}{M}\sum_{i=1}^{M}\log(1 - D_T(G(T(x_i))))$

$\mathcal{R}_3 = \frac{1}{M}\sum_{i=1}^{M}\|T(G(x)), G(T(x))\|_1$

$p = T_p(x), \mathcal{R}_4 = \frac{1}{M}\sum_{i=1}^{M}[\max(\frac{|p_{jk}p_{jl}|}{|q_{jk}q_{jl}|}, a)$

$-\min(\frac{|p_{jk}p_{jl}|}{|q_{jk}q_{jl}|}, \frac{1}{a}) + |(\sum_k^K p_{jk}) - b|]$

$\theta_D := \theta_D + \alpha\nabla_{\theta_D}\mathcal{R}_1, \theta_{D_T} := \theta_{D_T} + \nabla_{\theta_{D_T}}\mathcal{R}_2,$
$\theta_T := \theta_T - \alpha\nabla_{\theta_T}(\mathcal{R}_2 - \mathcal{R}_3 + \mathcal{R}_4)$
Optimizing G

$\mathcal{R}_1 = \frac{1}{N}\sum_{j=1}^{N}\log D(y_j) + \frac{1}{M}\sum_{i=1}^{M}\log(1-D(G(x_i)))$

$\mathcal{R}_2 = \frac{1}{M}\sum_{i=1}^{M}\|T(G(x)), G(T(x))\|_1$

$\mathcal{R}_3 = \frac{1}{N}\sum_{j=1}^{N}\log D_T(T(y_j)) + \frac{1}{M}\sum_{i=1}^{M}$

$\theta_G := \theta_G - \alpha\nabla_{\theta_G}(\mathcal{R}_1 + \mathcal{R}_2 + \mathcal{R}_3)$

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

1. Implementation of modified VAT and MT

1.1. Modified Virtual Adversarial Training (VAT)

VAT [4] introduced the concept of adversarial attack [3] 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:

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

To apply the VAT, we adapt the semi-supervised framework to the the I2I task. Similar to our proposed MSPC, we introduce another noisy perturbation branch with additional discriminator D_V. Then, we can reconstruct the framework as follows,

$$\begin{aligned}&\min_G \max_{D,D_T} \mathbb{E}_{y\sim P_Y}\log D(y) + \mathbb{E}_{x\sim P_X}\log(1 - D(G(x)))\\ &+ \mathbb{E}_{y\sim P_Y}\log D_V(y) + \mathbb{E}_{x\sim P_X}\log(1 - D_V(G(x+\gamma))),\\ &\min_G \max_{\gamma;\|\gamma\|\leqslant\delta} \mathbb{E}_{x\sim P_X}\|G(x), G(x+\gamma)\|_1. \quad (2)\end{aligned}$$

Referring to [4], the optimal $\hat{\gamma}$ can be derived from the first-order derivative w.r.t. $\epsilon\gamma$ and ϵ is a very small positive constant, which is $\hat{\gamma} = \frac{\partial\|G(x), G(x+\epsilon\gamma)\|_1}{\partial\epsilon\gamma}$. The intuition is that, the direction of maximum perturbation is exactly the same as the current derivative. But VAT is trivious due to that VAT is often unstable when the task is becoming more complex.

1.2. Modified Mean Teacher (MT)

MT [5] is a simple yet nonmethod, which has been successfully applied in many applications [1,2,6]. It utilizes the exponential moving average (EMA) of the learned model as the teacher reference for correction. The modified MT can be formulated as,

$$\begin{aligned}&\min_G \max_D \mathbb{E}_{y\sim P_Y}\log D(y) + \mathbb{E}_{x\sim P_X}\log(1 - D(G(x)))\\ &+ \mathbb{E}_{x\sim P_X}\|G(x), G_{EMA}(x)\|_1, \quad (3)\end{aligned}$$

where G_{EMA} is the EMA of G and will not participate in the gradient back-propagation.

For both modified VAT and MT, we use the same networks and training configuration as other models.

References

[1] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[2] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 1

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015. 1

[4] 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. 1

[5] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. 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. 1

[6] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5941–5950, June 2021. 1

2. More Qualitative Results

In this section, we show additional qualitative results from the held-out testing dataset.

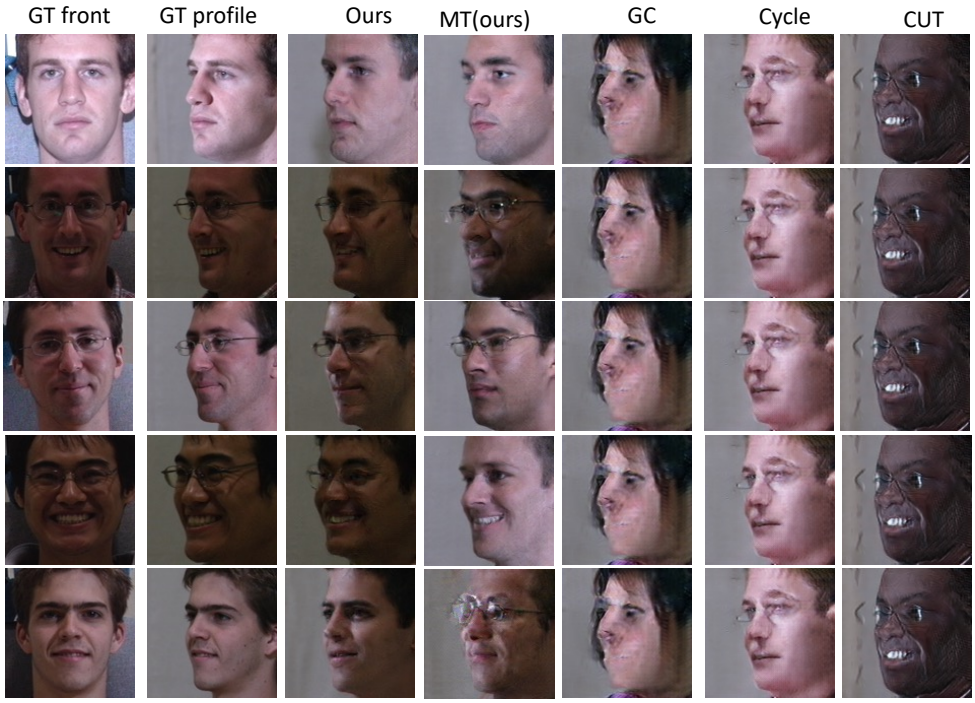

Figure 1. front face2profile.

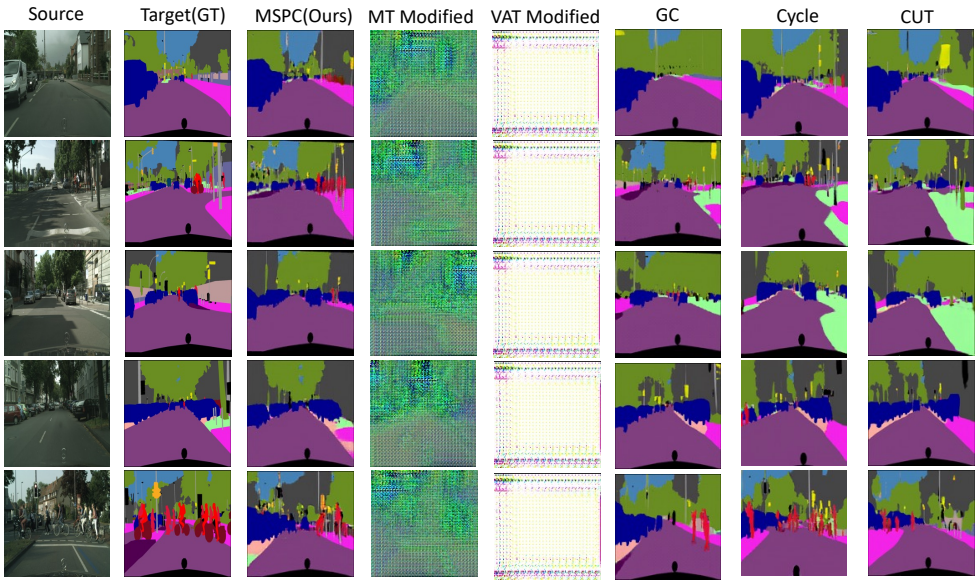

Figure 2. city2parsing.

Figure 3. city2parsing.

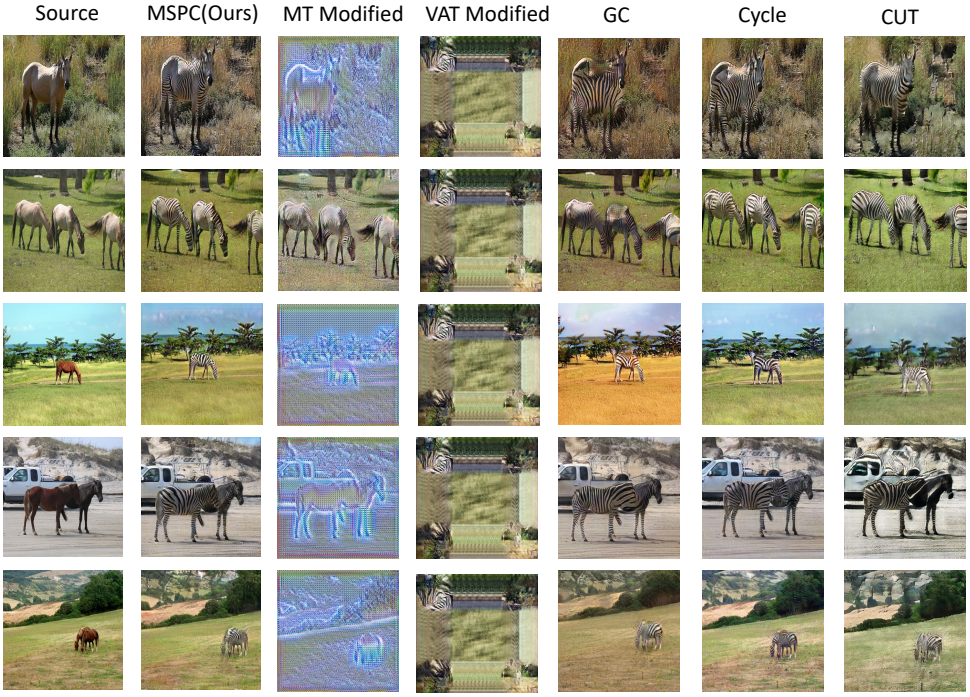

Figure 4. horse2zebra.

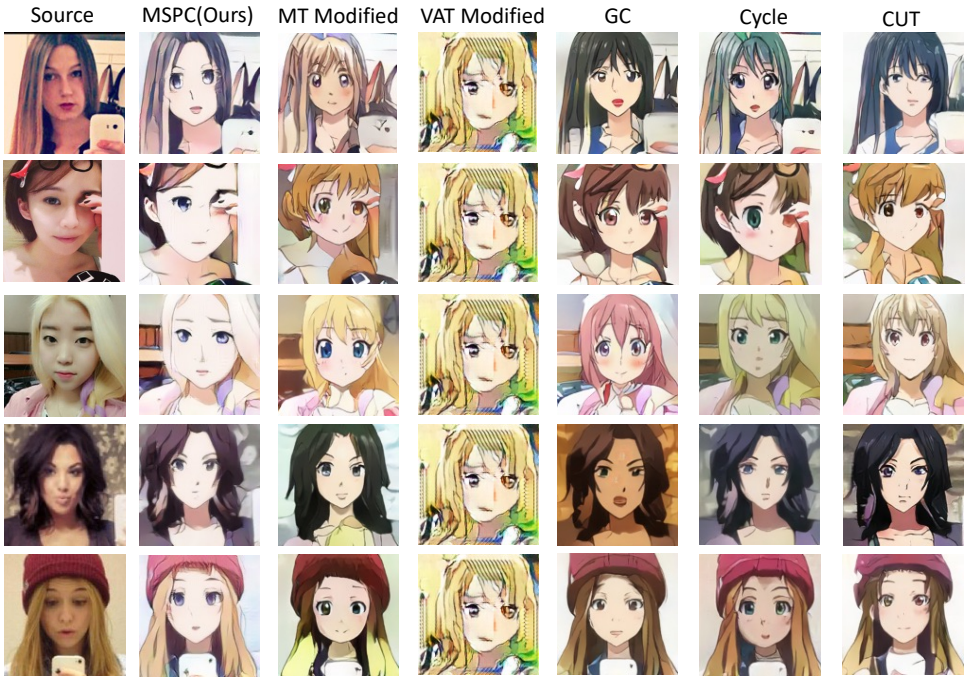

Figure 5. selfie2anime.

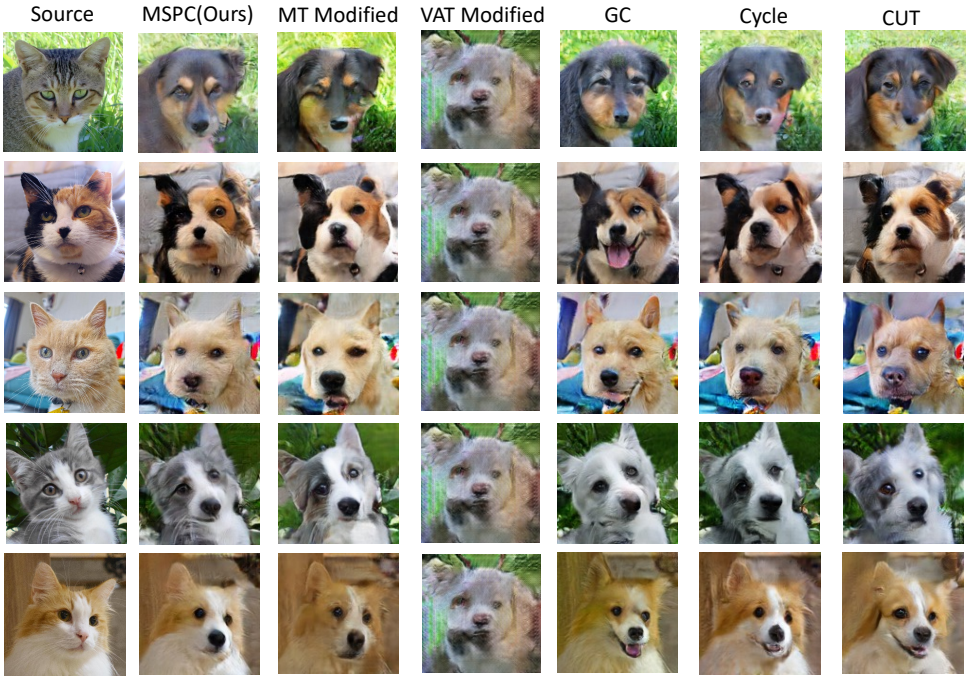

Figure 6. cat2dog.

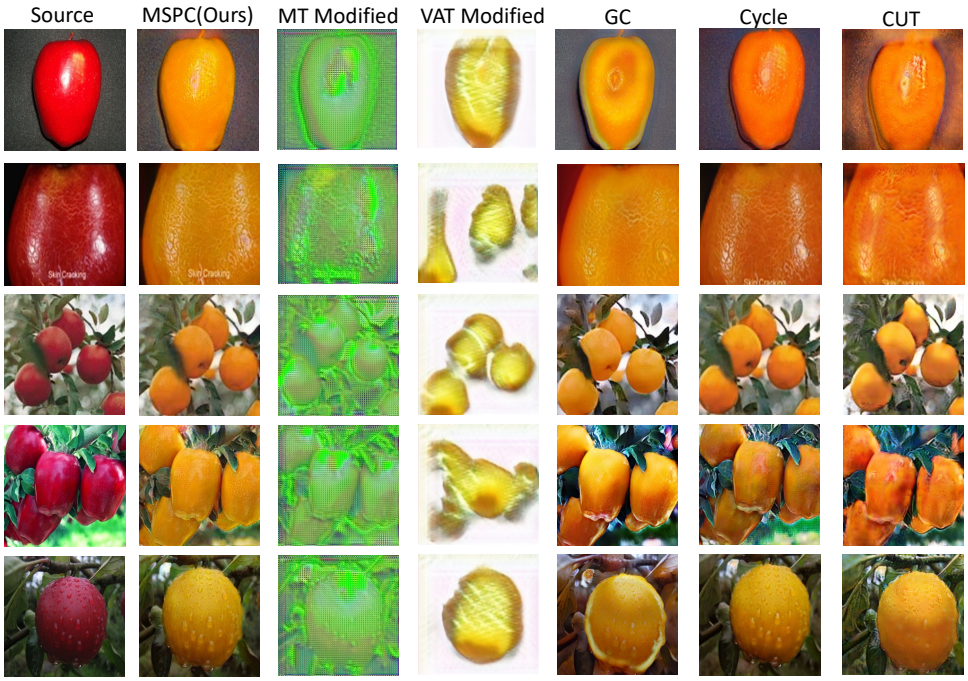

Figure 7. cat2dog.

3. Visualization of Transformer T without constraints

In this section, we visualize the effect of the spatial transformer T on both the source and target images. As we can see in below figure, the spatial transformation generates perturbed images without keeping the information of images without the constraint on T. If there is much information lost on images, the T will hurt the performance of the I2I.

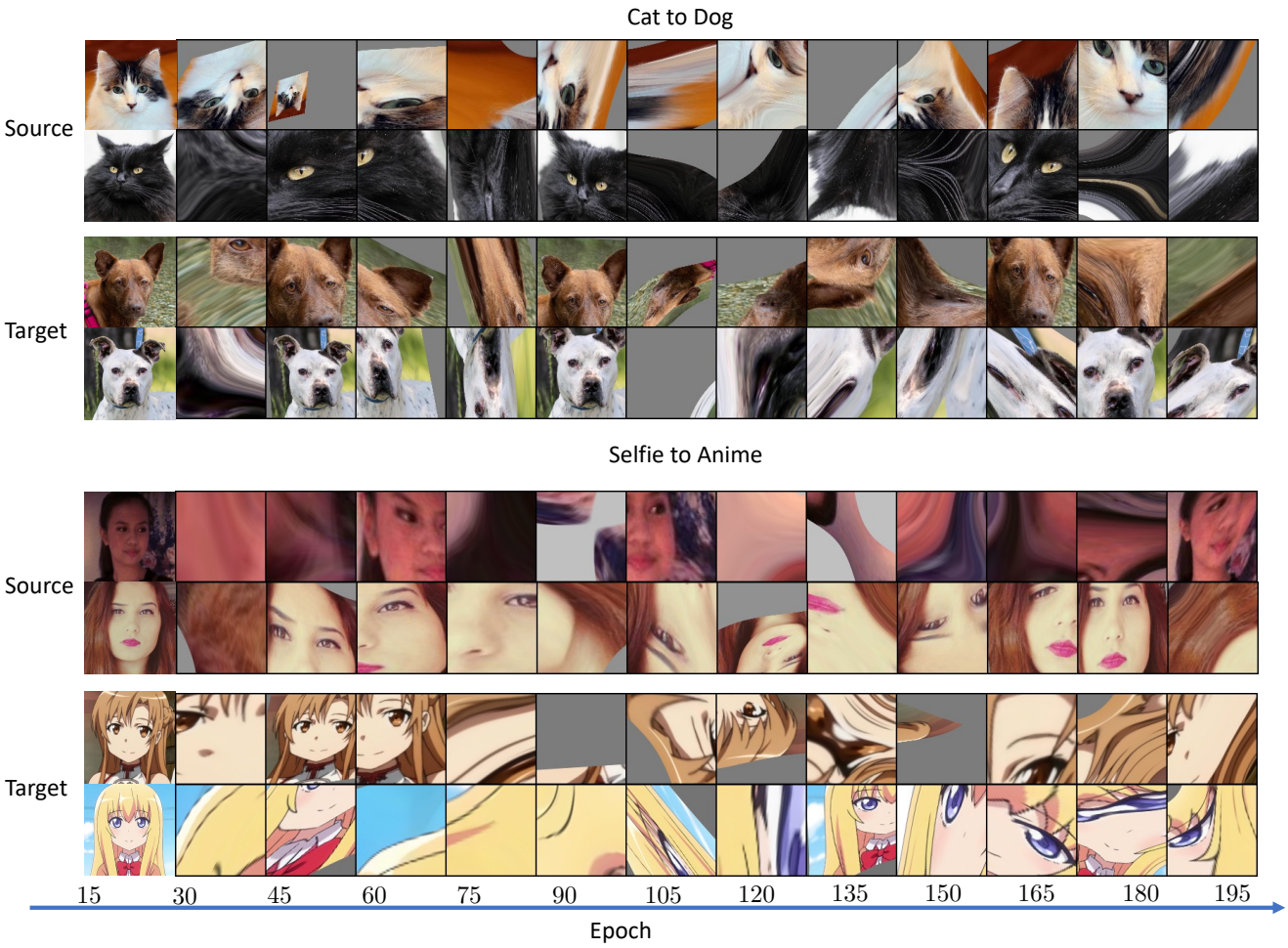

Figure 8. Perturbation changes as epoch grows. In this figure, we do not add the constraint to the T.