Supplementary Material of "Alleviating Semantics Distortion in Unsupervised Low-Level Image-to-Image Translation via Structure Consistency Constraint"

A. Appendix

A.1. Details of Solving α

To solve the $rSMI(V^{x_i}, V^{\hat{y}_i})$, we directly estimate the density ratio using a linear combination of kernel functions of $\{v_j^{x_i}\}_{j=1}^M \in V^{x_i}$ and $\{v_j^{\hat{y}_i}\}_{j=1}^M \in V^{\hat{y}_i}$:

$$\frac{S_i}{\beta S_i + (1-\beta)Q_i} = \omega_\alpha(v^{x_i}, v^{\hat{y}_i}) = \sum_{l=1}^{m} \alpha_l \phi_l(v^{x_i}, v^{\hat{y}_i}) = \alpha^T \phi(v^{x_i}, v^{\hat{y}_i}), \tag{1}$$

where $\phi \in \mathbb{R}^m$ is the kernel function, $\alpha \in \mathbb{R}^m$ is the parameter vector we need to solve, and m is the number of kernels. α is learned so that the following squared error $J(\alpha)$ [9] is minimized:

$$J(\alpha) = \mathbb{E}_{\beta S_i + (1-\beta)Q_i}[(\omega_\alpha(v^{x_i}, v^{\hat{y}_i}) - \omega^*(v^{x_i}, v^{\hat{y}_i}))^2] = \mathbb{E}_Q[(1-\beta)\omega_\alpha^2] + \mathbb{E}_S[\beta\omega_\alpha^2 - 2\omega_\alpha] + J_0,$$

where J_0 is a constant number respect to α, and therefore can be safely ignored. Thus, the optimization problem is given as:

$$\min_\alpha [\alpha^T H \alpha - 2\alpha^T h],$$

where

$$H = (1-\beta)\mathbb{E}_Q[\phi\phi^T] + \beta\mathbb{E}_S[\phi\phi^T], \qquad h = \mathbb{E}_S[\phi].$$

For computational efficiency, we define the kernel function $\phi(v^{x_i}, v^{\hat{y}_i})$ as the product of $K(v^{x_i}; k_c) \in \mathbb{R}^m$ and $L(v^{x_i}; l_c) \in \mathbb{R}^m$, which are kernel functions of v^{x_i} and $v^{\hat{y}_i}$ respectively:

$$\phi(v^{x_i}, v^{\hat{y}_i}) = K(v^{x_i}) \circ L(v^{\hat{y}_i}),$$

where $\circ$ denotes the Hadamard product. Approximating the expectations in H and h by empirical averages, and adding a quadratic regularizer $\alpha^T R\alpha$ to avoid over-fitting, the objective function in our optimize problem becomes:

$$\hat{J}(\alpha) = [\alpha^T \hat{H}\alpha - 2\hat{h}^T\alpha + \lambda\alpha^T R\alpha], \tag{2}$$

where R is the positive semi-definite regularization matrix, and

$$\hat{H} = \frac{1-\beta}{n}(K \circ L)(K \circ L)^T + \frac{\beta}{n^2}(KK^T) \circ (LL^T), \qquad \hat{h} = \frac{1}{n^2}(K1_n) \circ (L1_n),$$

where n is the number of samples, 1_n is the n-dimensional vector filled by ones, and K and L are two $m \times n$ matrices composed by kernel functions. The equation 2 is a unconstrained quadratic problem, and thus could be solved by analytically and the optimal solution of $\hat{\alpha}$ is:

$$\hat{\alpha} = (\hat{H} + \lambda R)^{-1}\hat{h}.$$

A.2. Experimental Analysis

A.2.1 β Analysis

We conduct the sensitive analysis of β on the digits datasets (each experiment is repeated 3 times) and the results are shown as Figure 1 (b). We can see the performance of translation models are all improved with varied β, and we use 0.5 for convenience.

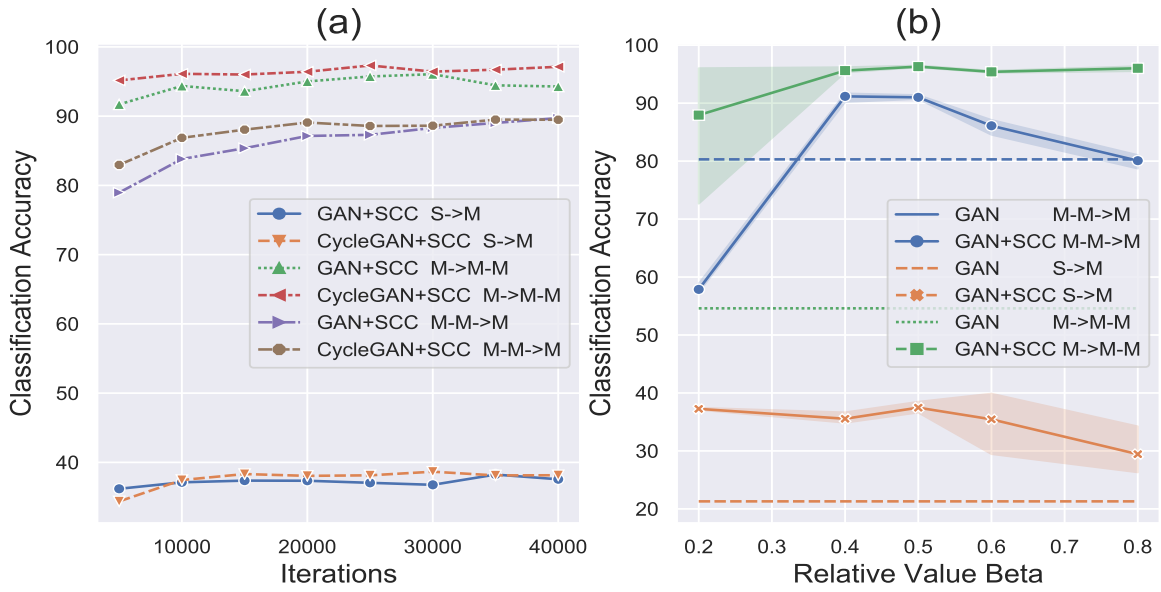


Figure 1. The training curves and the sensitive analysis about β on Digits datasets

A.2.2 Generation Diversity Analysis

We conduct the generation diversity experiments on the edge2shoes dataset. Following MUNIT [4], we calculate the average LPIPS distance between 1900 pairs of randomly generated images (sampled from 100 input images). MUNIT with SCC has the average LPIPS of 0.120, improving the diversity of original MUNIT model with 0.104 LPIPS score. Therefore, our SCC has no negative impact on generation diversity. Some generation examples are given as Figure 2.

Figure 2. The generation example of MUNIT+SCC on the edge2shoes. Specifically, images at first two rows are source domain images and the others are translated images by MUNIT+SCC.

Figure 3. The generation example of MUNIT on the edge2shoes. Specifically, images at first two rows are source domain images and the others are translated images by MUNIT.

A.2.3 Stability Analysis

We conduct the training stability analysis of our SCC on the digits datasets and the results are shown as Figure 1 (a). We can see the training procedure is stable with our SCC.

A.3. Experiments

A.3.1 KID scores of Qualitative evluation

Following the recent work [5], we use KID score [1] as the evaluation metric to evaluate the . The results are reported as Table 1, and we can see that the vanilla GAN method coupled with our SCC can achieve the comparable results with those methods with larger model size. In addition, a simple generator based on res-blocks trained by the combination of cycle, geometry and our SCC constraint can achieve SOTA performance on almost all datasets.

Table 1. KID scores for style transfer tasks. The results of baselines (AGGAN [10] , DRIT [6] , UNIT [7] , MUNIT [4]) are from [5]. Here U (light) is the light version of U-GAT-IT. Specifically, VGG(cosine)/VGG(L2) refer to the Contextual loss [8] and Content loss [3], respectively, and they optimize contextual and L2 distance of input and translated images' **VGG** features, respectively.

	Params	selfie2anime	horse2zebra	photo2por	anime2selfie	zebra2horse	por2photo
AGGAN	\	14.63±0.55	7.58±0.71	2.33±0.36	12.72±1.03	8.80±0.66	2.19±0.40
DRIT	65.0M	15.08±0.62	9.79±0.62	5.85±0.54	14.85±0.60	10.98±0.55	4.76±0.72
UNIT	\	14.71±0.59	10.44±0.67	**1.20±0.31**	26.32±0.92	14.93±0.75	1.42±0.24
MUNIT	46.6M	13.85±0.41	11.41±0.83	4.75±0.52	13.94±0.72	16.47±0.954	3.30±0.47
U-GAT-IT(full)	134.0M	11.61±0.57	7.06±0.8	1.79±0.34	11.52±0.57	7.47±0.71	1.69±0.53
U-GAT-IT(light)	74.0M	12.31±0.50	7.25±0.8	3.43±0.28	15.22±0.51	9.83±0.58	2.67±0.33
U (light)+SCC	74.0M	**10.37±0.32**	5.19±0.46	3.19±0.26	**10.30±0.47**	7.80±0.48	1.85±0.26
GAN+VGG(cosine)	588.1M	12.77±0.38	9.39±0.39	3.95±0.26	14.81±0.41	10.36±0.51	3.05±0.25
GAN+VGG(L2)	588.1M	11.42±0.42	6.87±0.58	1.87±0.25	12.28±0.45	9.15±0.49	1.77±0.27
GAN+VGG(L1)	588.1M	11.32±0.45	8.71±0.39	2.59±0.27	13.18±0.39	9.76±0.53	2.31±0.28
GAN + SCC	14.1M	11.37±0.41	7.28±0.52	3.86±0.39	11.61±0.40	7.15±0.46	1.88±0.25
CycleGAN	28.3M	13.08±0.49	8.05±0.72	1.84±0.34	11.84±0.74	8.0±0.66	1.82±0.36
Cycle + SCC	28.3M	11.66±0.41	6.59±0.49	2.91±0.22	10.83±0.44	6.77±0.52	1.62±0.15
GcGAN-rot	16.9M	11.89±0.42	7.05±0.45	2.24±0.26	13.28±0.35	7.67±0.47	1.84±0.28
GcGAN + SCC	16.9M	10.75±0.42	5.12±0.44	1.97±0.24	10.96±0.40	7.10±0.50	1.64±0.22
CUT	18.1M	12.1±0.42	8.45±0.45	2.85±0.33	12.45±0.54	8.99±0.5	2.23±0.31
CUT + SCC	18.1M	11.75±0.41	6.26±0.44	2.31±0.3	12.05±0.44	8.4±0.43	2.11±0.26
Gc+Cycle+SCC	45.2M	10.61±0.44	**4.82±0.68**	1.64±0.24	10.92±0.35	**6.28±0.52**	**1.31±0.27**

A.4. Experimental Details

We will public codes and experimental setting for the convenience of reproducing results in our paper.

A.4.1 Digits

All digits images are resized to 32×32 resolution. Following [2], the network details of this experiment are given in Table 2.

Following all settings of the original models, the learning rate for generator and discriminator is 0.0002, the training epochs is 40000 and the batch size is 64.

A.4.2 Cityscapes

All images are resized to 128×128 resolution. Following [2, 11], the network details of this experiment are given in Table 3.

Following all settings of the original models, the learning rate for all generators and discriminators is 0.0002, the batch size is 1 and the training epochs for CUT is 400 and other models is 200.

Table 2. The network details of digits translation tasks, where C = Feature channel, K = Kernel size, S = Stride size, Deconv/Conv = Deconvolutional/Convolutional layer and "channels" donotes the image channels of target domain, such as 1 for MNIST, 3 for MNIST-M.

Generator				
index	Layers	C	K	S
1	Conv + LeakyReLU	64	4	2
2	Conv + LeakyReLU	128	4	2
3	Conv + LeakyReLU	128	3	1
4	Conv + LeakyReLU	128	3	1
5	Deconv + LeakyReLU	64	4	2
6	Deconv + LeakyReLU	channels	4	2
7	Tanh	-	-	-
Discriminator				
index	Layers	C	K	S
1	Conv + LeakyReLU	64	4	2
2	Conv + LeakyReLU	128	4	2
3	Conv + LeakyReLU	256	4	2
4	Conv + LeakyReLU	512	4	2
5	Conv	512	4	2

Table 3. The network details of digits translation tasks, where C = Feature channel, K = Kernel size, S = Stride size, Deconv/Conv = Deconvolutional/Convolutional layer and ResBlk = A residual block

Generator				
index	Layers	C	K	S
1	Conv + ReLU	64	7	1
2	Conv + ReLU	128	3	2
3	Conv + ReLU	256	3	3
4-9	ResBlk + ReLU	256	3	1
10	Deconv + ReLU	128	3	2
11	Deconv + ReLU	64	3	2
12	Conv	3	7	1
13	Tanh	-	-	-
Discriminator				
index	Layers	C	K	S
1	Conv + LeakyReLU	64	4	2
2	Conv + LeakyReLU	128	4	2
3	Conv + LeakyReLU	256	4	2
4	Conv + LeakyReLU	512	4	1
5	Conv	512	4	1

A.4.3 Maps

All images are resized to 256×256 resolution. Following [2, 11], the network details is similar to the details of Cityscape, but the generator contains 9 res-blocks for images with 256×256 resolution. Following all settings of the original models, the learning rate for all generators and discriminators is 0.0002, the batch size is 1 and the training epochs for CUT is 400 and other models is 200.

A.4.4 Style Transfer

All settings are same with Maps A.4.3. The details of datasets as follows:
selfie2anime This dataset is from U-GAT-IT [5], which contains 3400 training images and 100 images for test.
horse2zebra This dataset is from CycleGAN [11], whose training sets contains 1,067 horse images and 1,334 zebra images. The test set consists of 120 horse images and 140 zebra images.

portrait2photo This dataset is from DRIT [6], whose training sets contains 6,452 photo images and 1,811 portrait images. The test set consists of 751 photo images and 400 portrait images. Following all settings of the original models, the learning rate for all generators and discriminators is 0.0002 and the training epochs for CUT is 400 and other models is 200.

A.5. Analysis on the Cat2Dog Dataset

To analyze the performance of our SCC on geometry-variant datasets, we incorporate our SCC constraint into CycleGAN model and train it on the cat → dog dataset. The results are shown as Figure 4 , we can see that the trained translation model can successfully translate dog images at the top row to cat images and preserve the basic image content (*i.e.* locations of eyes, mouth, directions of faces), even if there are some changes of geometric structure. However, as images at the bottom row show, the translation model fails to translate the dog images to cat images in a meaningful way, as the mouth of dogs block the background but the mouth of cats do not, and so the translation model need to "imagine" some background area that be blocked, which needs us to propose more constraints.

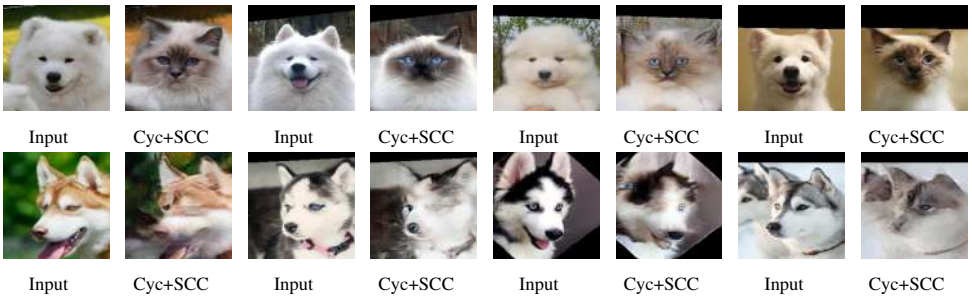
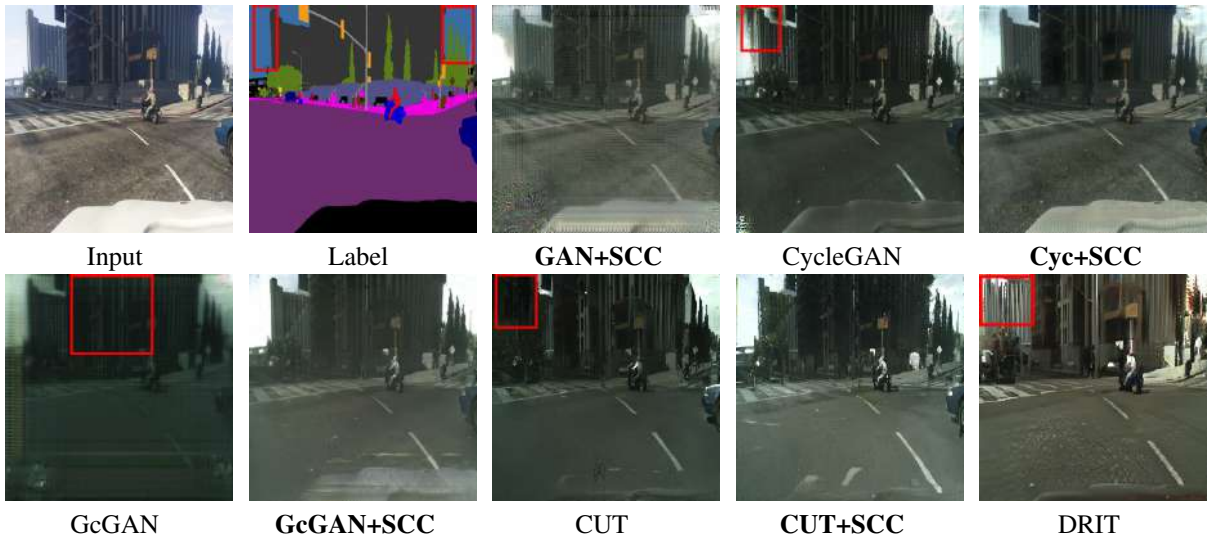


Figure 4. Qualitative results on a geometry-variant dataset, including Dog → Cat. Images at the top row are successful cases, while images at the bottom row are failure cases.

A.6. Generated Samples

A.7. GTA → Cityscapes




Input
Label
GAN+SCC
CycleGAN
Cyc+SCC
GcGAN
GcGAN+SCC
CUT
CUT+SCC
DRIT
Input
Label
GAN+SCC
CycleGAN
Cyc+SCC
GcGAN
GcGAN+SCC
CUT
CUT+SCC
DRIT
Input
Label
GAN+SCC
CycleGAN
Cyc+SCC
GcGAN
GcGAN+SCC
CUT
CUT+SCC
DRIT

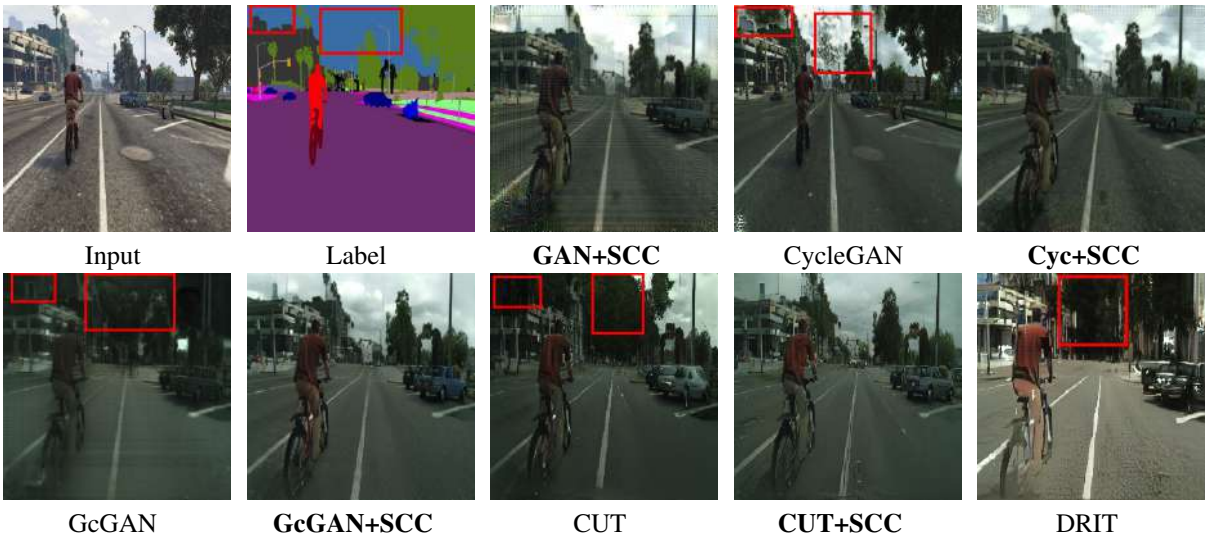


Table 5. Qualitative results on GTA → Cityscapes. Obviously, the semantic information, such as sky, is better preserved by the translation model further constrained by our SCC.

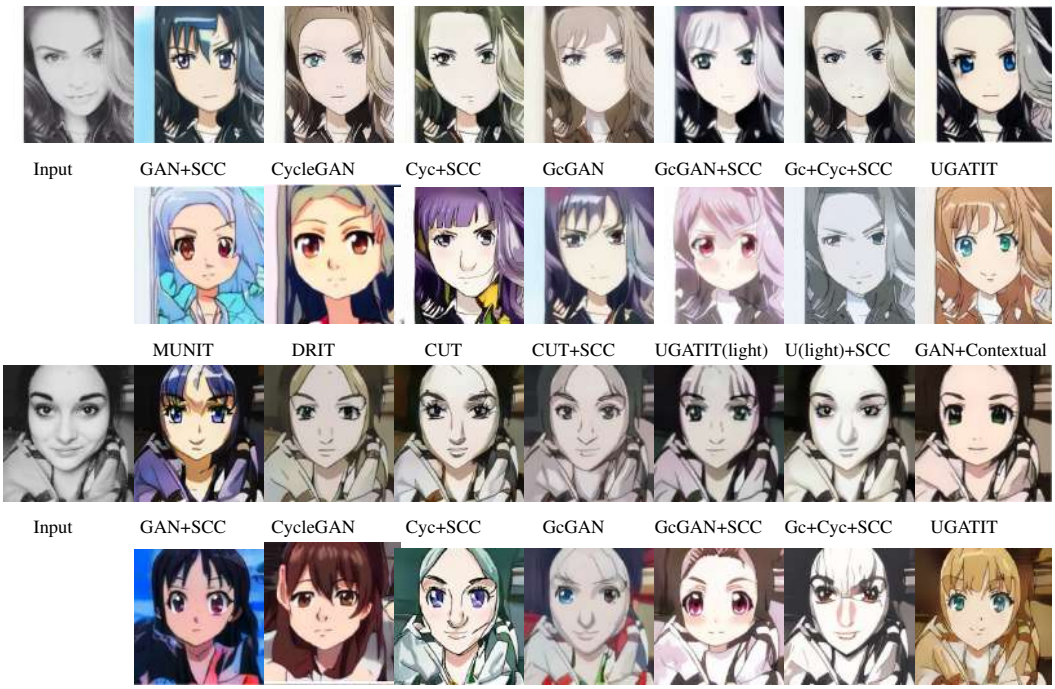
A.7.1 Maps

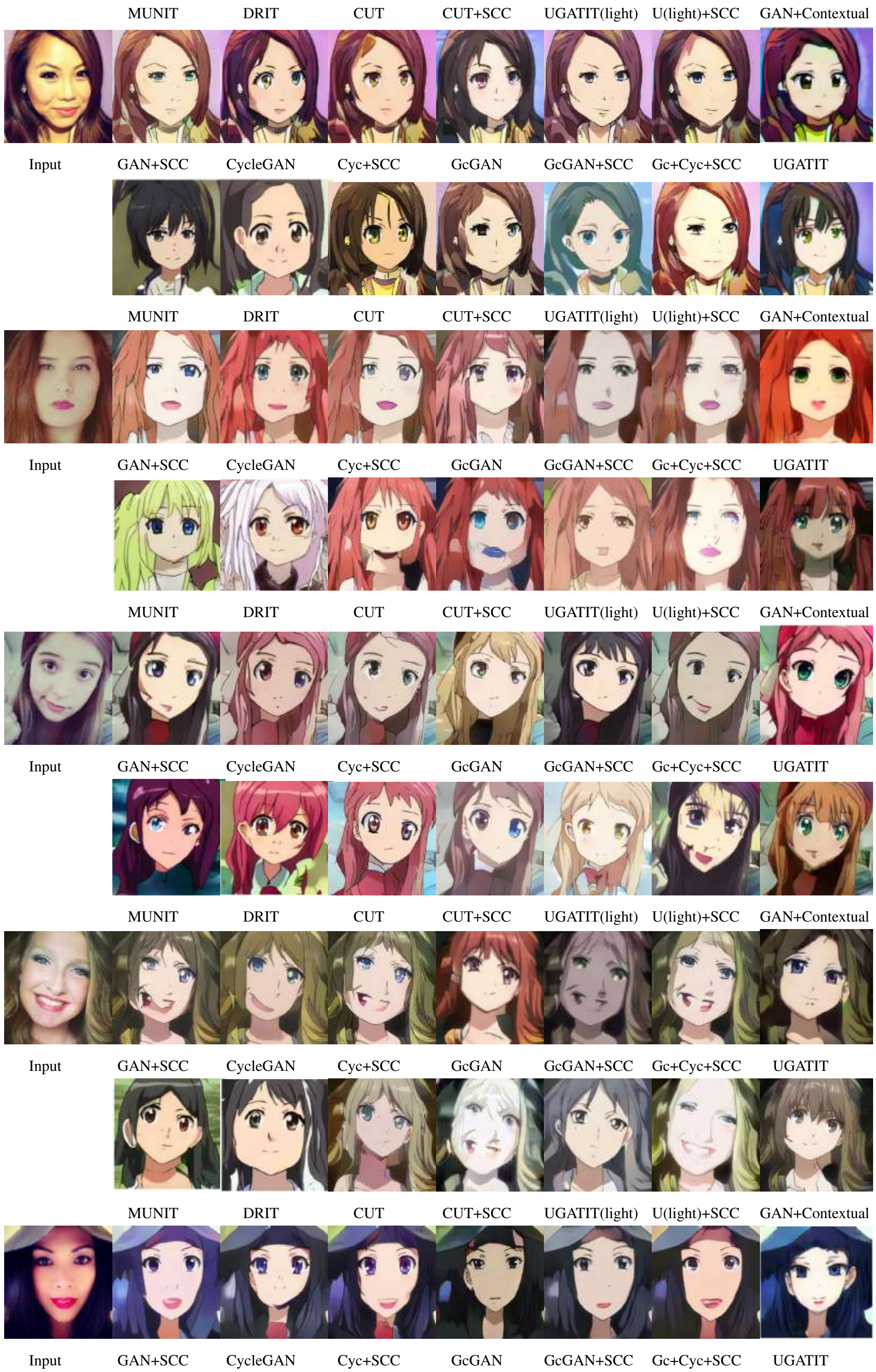
Table 6. Qualitative results on the Maps dataset.

A.7.2 Cityscapes

Table 7. Qualitative results on the Cityscape Dataset.

A.7.3 Qualitative Results




MUNIT
DRIT
CUT
CUT+SCC
UGATIT(light)
U(light)+SCC
GAN+Contextual
Input
GAN+SCC
CycleGAN
Cyc+SCC
GcGAN
GcGAN+SCC
Gc+Cyc+SCC
UGATIT
MUNIT
DRIT
CUT
CUT+SCC
UGATIT(light)
U(light)+SCC
GAN+Contextual
Input
GAN+SCC
CycleGAN
Cyc+SCC
GcGAN
GcGAN+SCC
Gc+Cyc+SCC
UGATIT
MUNIT
DRIT
CUT
CUT+SCC
UGATIT(light)
U(light)+SCC
GAN+Contextual
Input
GAN+SCC
CycleGAN
Cyc+SCC
GcGAN
GcGAN+SCC
Gc+Cyc+SCC
UGATIT
MUNIT
DRIT
CUT
CUT+SCC
UGATIT(light)
U(light)+SCC
GAN+Contextual
Input
GAN+SCC
CycleGAN
Cyc+SCC
GcGAN
GcGAN+SCC
Gc+Cyc+SCC
UGATIT
MUNIT
DRIT
CUT
CUT+SCC
UGATIT(light)
U(light)+SCC
GAN+Contextual
Input
GAN+SCC
CycleGAN
Cyc+SCC
GcGAN
GcGAN+SCC
Gc+Cyc+SCC
UGATIT

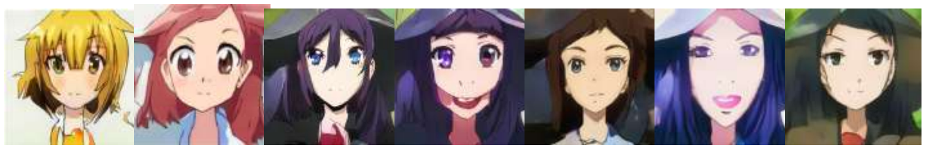


Table 8. Qualitative results on Selfie $\rightarrow$ Anime. Obviously, the geometry structure, such as face shape, is better preserved by the translation model further constrained by our SCC.

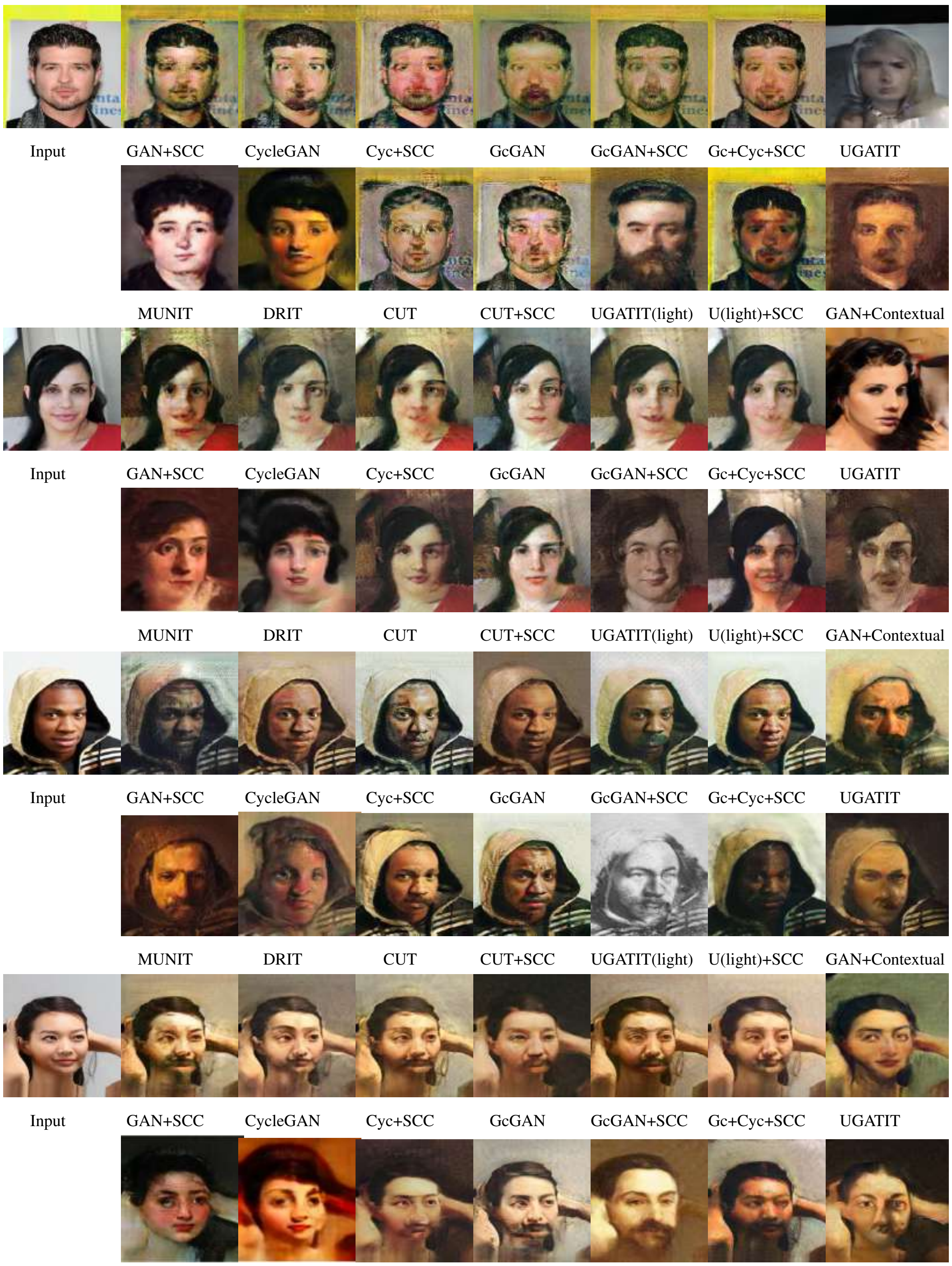


Table 9. Qualitative results on photo → portrait. Obviously, the semantic information, such as face shape, is better preserved by the translation model further constrained by our SCC.

Table 10. Qualitative results on Horse → Zebra. Obviously, the semantic information, such as horse shape, is better preserved by the translation model further constrained by our SCC.

A.7.4 Digits

GAN

GAN + SCC

CycleGAN

CycleGAN + SCC

GcGAN-rot

GcGAN-rot + SCC

GcGAN-vf

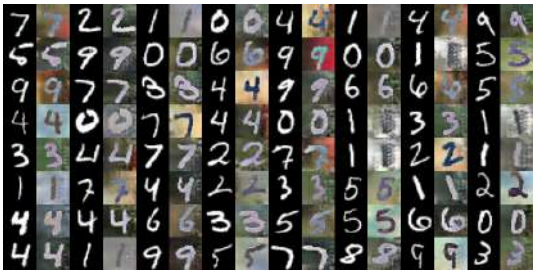
GcGAN-vf + SCC

Gc-rot+Cycle+SCC

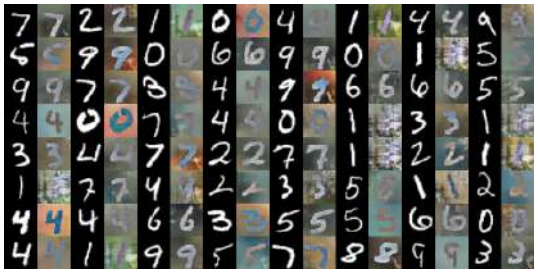
Gc-vf+Cycle

Table 11. Qualitative comparisons on SVHN→MNIST.

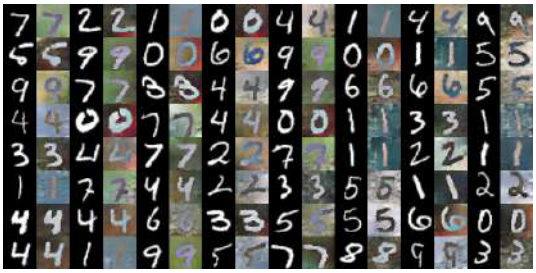
GAN



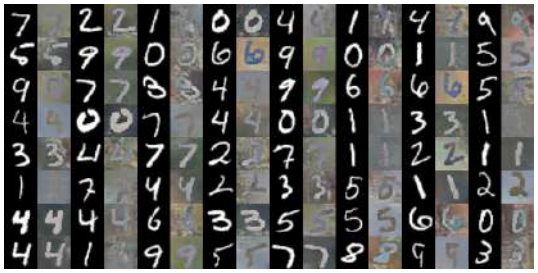
GAN + SCC



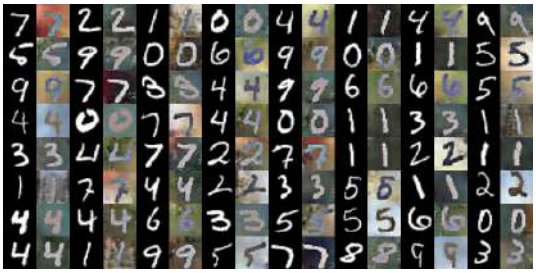
CycleGAN



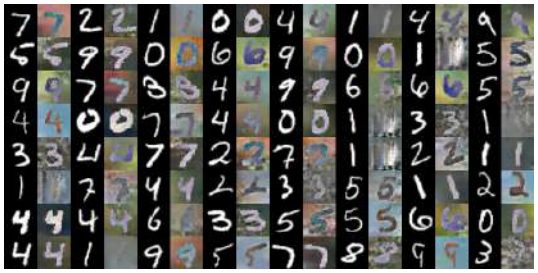
CycleGAN + SCC



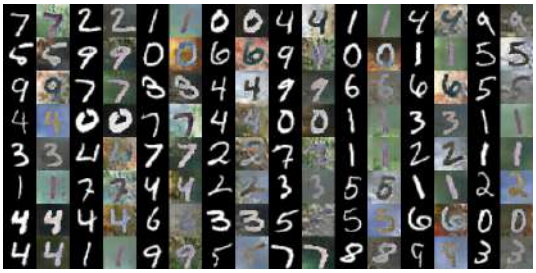
GcGAN-rot



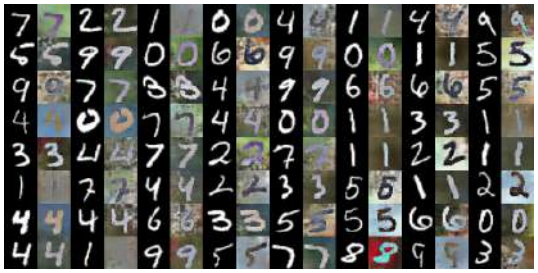
GcGAN-rot + SCC



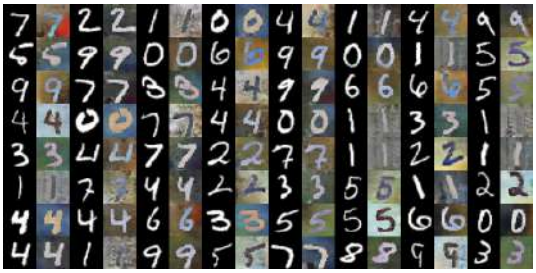
GcGAN-vf



GcGAN-vf + SCC



Gc-rot+Cycle+SCC



Gc-vf+Cycle

Table 12. Qualitative comparisons on MNIST→MNIST-M.

A.7.5 Ablation Study

Figure 5. The overlarge λ_{SCC} example on SVHN→MNIST.

An example of SVHN to MNIST translation when λ_{SCC} is set to 25 is shown as Figure 5. The images are almost translated without any changes in geometry structures. However, the overlarge λ_{SCC} causes the translation model neglect the style information from adversarial loss, resulting in some images with opposite color. This phenomenon indicates that our SCC has good performance on the preservation of geometry structure but should be appropriate with style information.

References

[1] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

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

[3] 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.

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

[5] 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.

[6] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1905.01270*, 2019.

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

[8] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018.

[9] Masashi Sugiyama. Machine learning with squared-loss mutual information. *Entropy*, 15(1):80–112, 2013.

[10] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

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