Supplementary Document for "SPatchGAN: A Statistical Feature Based Discriminator for Unsupervised Image-to-Image Translation"

Xuning Shao, Weidong Zhang
NetEase Games AI Lab
599 Wangshang Road, Binjiang District, Hangzhou, P.R. China
{shaoxuning, zhangweidong02}@corp.netease.com

1. Supplementary Material

In this document, we provide the detailed network architecture of SPatchGAN, the data augmentation method, the study about the weak cycle constraint, the experimental results of applying the SPatchGAN discriminator to other image translation frameworks, as well as additional comparison results for SPatchGAN and the baselines.

1.1. Implementation Details

We describe the details of our network architecture in this section. A convolutional layer with kernel size $p \times p$, stride q and number of output channels w is denoted as Kp-Sq-Cw. A fully connected layer with number of output channels w is denoted as FCw. A $2\times$ nearest-neighbor upsampling layer is denoted as U2. A residual block with a shortcut branch and a residual branch b is denoted as RES(b). A block b repeated z times is denoted as $b \times z$.

Discriminator. The discriminator consists of a feature extraction block and four scales. Each scale has a downsampling block, an adaptation block and three MLPs. Spectral normalization (SN) and Leaky-ReLU (LReLU) with a slope of 0.2 are used in the discriminator.

- Feature extraction block: K4-S2-C256-SN-LReLU, K4-S2-C512-SN-LReLU.
- Downsampling block: K4-S2-C1024-SN-LReLU.
- Adaptation block: (K1-S1-C1024-SN-LReLU) $\times 2$.
- MLP: (FC1024-SN-LReLU) $\times 2$, FC1-SN.

Given an input tensor $a \in \mathbb{R}^{H\times W\times C}$ with height H, width W and number of channels C, the statistical feature of uncorrected standard deviation is calculated as

$$s_k = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(a_{i,j,k} - \bar{a_k})^2}, \tag{1}$$

where s_k is the k-th element of the output feature vector $s \in \mathbb{R}^C$, $\bar{a_k}$ is the average value of the k-th input feature map, and $a_{i,j,k}$ is the input element at the i-th row, j-th column of the k-th feature map.

Forward generator. The forward generator consists of a downsampling module, a residual module, and an upsampling module. We use instance normalization (IN) in the downsampling and residual modules, and use layer normalization (LN) in the upsampling module. ReLU is utilized as the activation function except for the output layer, which uses Tanh.

- Downsampling module: K3-S2-C128-IN-ReLU, K3-S2-C256-IN-ReLU, K3-S2-C512-IN-ReLU.
- Residual module: RES(K3-S1-C512-IN-ReLU, K3-S1-C512-IN) $\times 8$.
- Upsampling module: U2, (K3-S1-C512-LN-ReLU) $\times 2$, U2, K3-S1-C256-LN-ReLU, U2, K3-S1-C128-LN-ReLU, K3-S1-C3-Tanh.

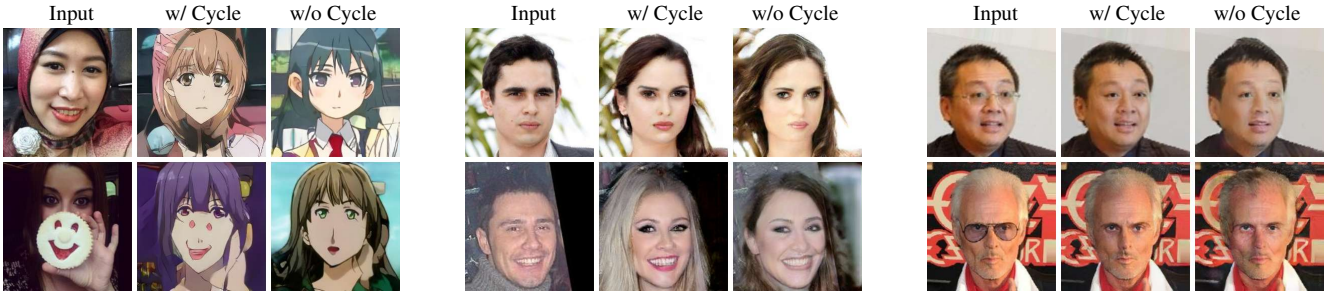
Backward generator. The backward generator has a pre-mixing module, a residual module, and a post-mixing module. The residual module has the same structure as the forward generator.

- Pre-mixing module: K3-S1-C512-IN.
- Post-mixing module: K3-S1-C512-LN-ReLU, K3-S1-C3-Tanh.

1.2. Data Augmentation

For selfie-to-anime, we adopt the data augmentation method in U-GAT-IT that first resizes the images to 286×286, then randomly crops the images to 256×256. For male-to-female and glasses removal, the images are center cropped to 178×178, resized to 256×256, and randomly shifted by up to 13 pixels horizontally and vertically.

We apply color jittering with random brightness offset in $[-0.125, 0.125]$, random hue offset in $[-0.02, 0.02]$, random saturation factor in $[0.8, 1.2]$, and random contrast factor in $[0.8, 1.2]$. All images are also randomly flipped horizontally.



(a) Selfie-to-anime (b) Male-to-female (c) Glasses removal

Figure 1: Generated images of SPatchGAN with and without the weak cycle constraint. Some redundant changes can be observed after removing the weak cycle constraint.

Model	Selfie-to-Anime		Male-to-Female		Glasses Removal	
	FID	KID	FID	KID	FID	KID
SPatchGAN w/ Weak Cycle	83.3	0.0214	**8.73**	**0.0056**	**13.9**	**0.0031**
SPatchGAN w/o Weak Cycle	**82.4**	**0.0168**	11.4	0.0080	16.2	0.0047

Table 1: Quantitative results of SPatchGAN with and without the weak cycle constraint. Lower is better.

1.3. Study of the Weak Cycle Constraint

To further evaluate the stability of SPatchGAN, we try to completely remove the weak cycle constraint. The qualitative results are shown in Figure 1. The generated images still have a good overall quality, verifying that the network has been stabilized to a large extent by the discriminator itself. However, the results without the weak cycle sometimes become too disconnected from the source images. *E.g.*, the headscarf and the object in front of the face completely disappear in Figure 1a. There are also some undesirable changes of the background in Figure 1b, and some redundant changes of the hair in Figure 1c.

The quantitative results are summarized in Table 1. The FID and KID actually improve for selfie-to-anime after removing the weak cycle. This is partially due to the fact that the similarity between the source and generated images is *not* considered by the metrics. Without the constraint, the images can be translated more freely to match the distributions. We enable the weak cycle in the default setting, since it is desirable to keep the generated image correlated to the source image.

The weak cycle is beneficial for FID and KID in the male-to-female and glasses removal cases. For the applications which aim to adjust only a part of the image, the constraint helps to exclude the unnecessary changes and make the training process more efficient.

Generally speaking, our method helps to separate the need for keeping the source and target images correlated from the need for stabilizing the network. The former is ensured by the weak cycle constraint, while the latter is mainly guaranteed by the SPatchGAN discriminator. Therefore, we can optimize the cycle weight λ^{cyc} on an application basis without worrying too much about the stability issues. In contrast, the flexibility of the original cycle based framework is much more limited, since the cycle constraints have to be strict enough to stabilize the network.

1.4. Applicability to Other Frameworks

The SPatchGAN discriminator is generally agnostic of the architecture and constraints for the generator, and can be potentially leveraged to enhance other image translation frameworks. To study its applicability to other frameworks, we directly replace the default discriminators of CycleGAN and MUNIT with the discriminator of SPatchGAN, and evaluate their performance with the male-to-female dataset. The other modules and hyperparameters are unchanged. The qualitative and quantitative results are shown in Figure 2 and Table 2. CycleGAN and MUNIT with the SPatchGAN discriminator are denoted as S-CycleGAN and S-MUNIT respectively. Multimodal results are shown for MUNIT and S-MUNIT.

With the full cycle constraints, the main problem of CycleGAN is the limited shape deformation. In contrast, MUNIT introduces additional stochasticity for multimodality, and suffers more from the instability issue. It can be seen from Figure 2 that S-CycleGAN helps to make the hairstyle and face more feminine than CycleGAN. S-MUNIT helps to alleviate the blurriness of the generated images compared to MUNIT. The quantitative results of S-CycleGAN and S-MUNIT are also better than CycleGAN and MUNIT according to Table 2.

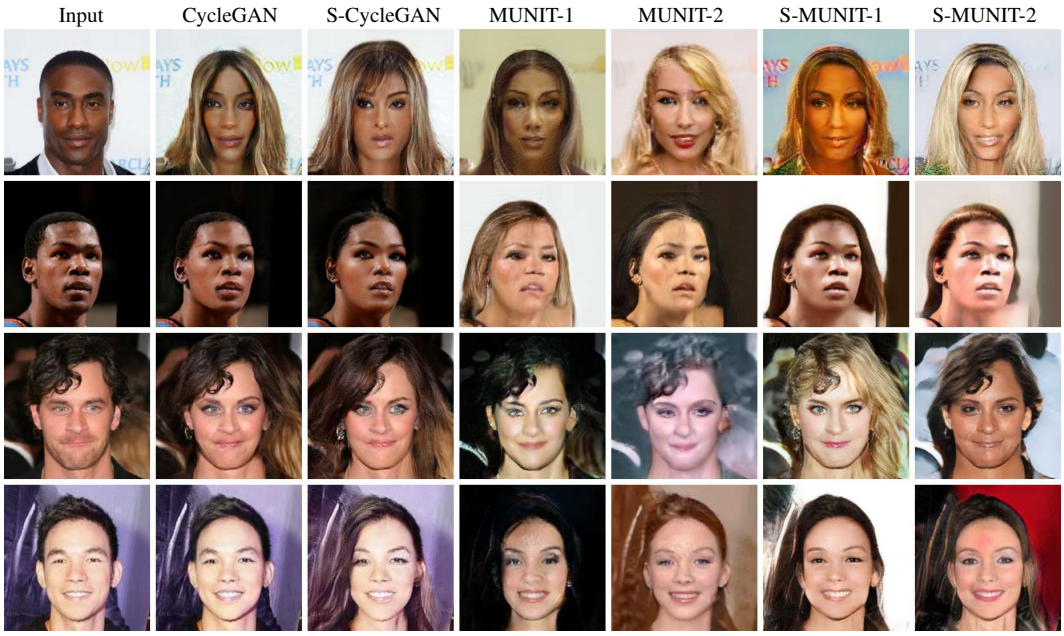


Figure 2: Generated images of the applicability studies for male-to-female.

Model	FID	KID
CycleGAN	24.5	0.0240
S-CycleGAN	13.4	0.0107
MUNIT	20.8	0.0161
S-MUNIT	17.0	0.0123

Table 2: Quantitative results of the applicability studies for male-to-female. Lower is better.

1.5. Additional Experimental Results

We show additional results for selfie-to-anime, male-to-female and glasses removal in Figure 3, Figure 4 and Figure 5, respectively.

Figure 3: Additional results of selfie-to-anime translation.

Figure 4: Additional results of male-to-female translation.

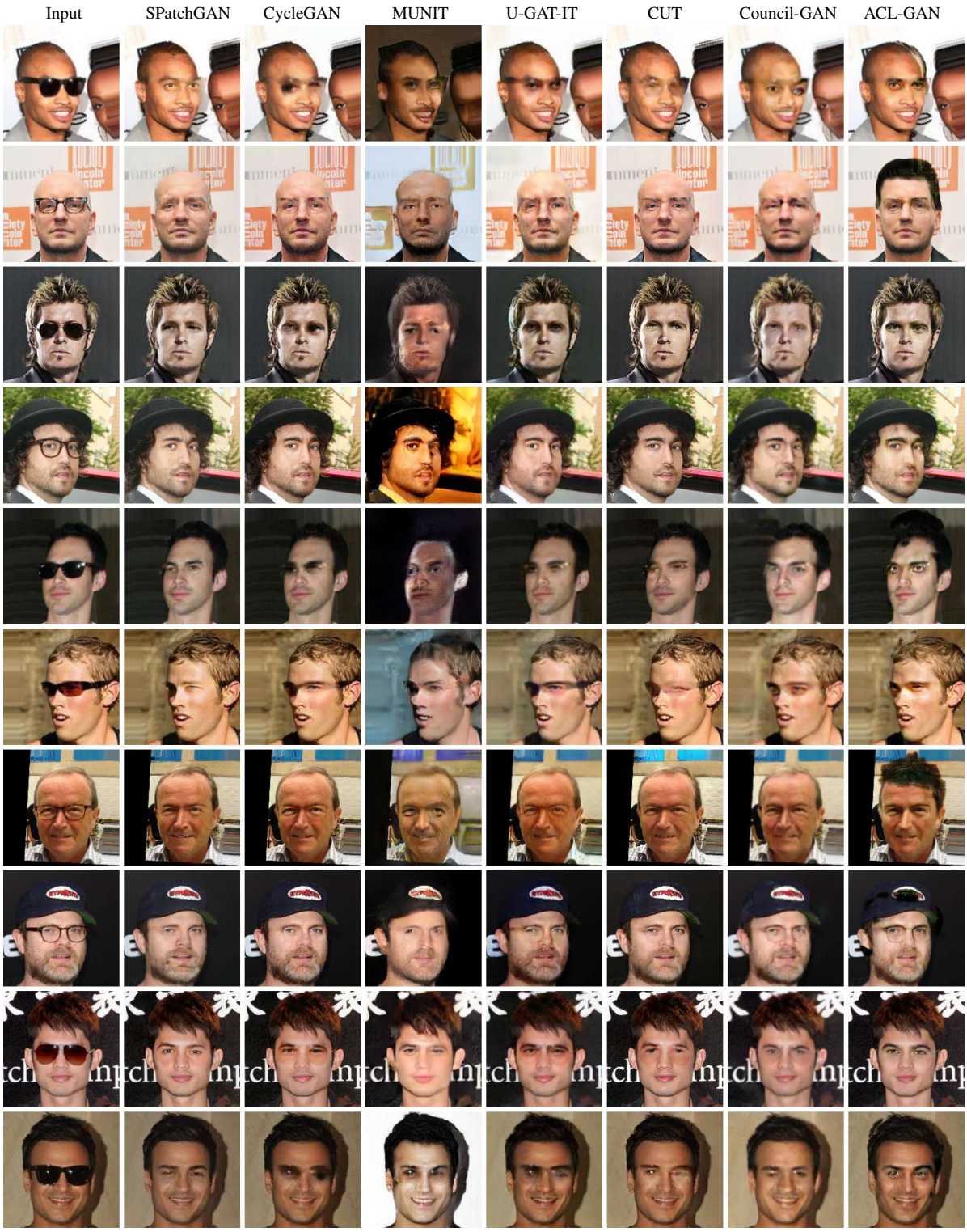


Figure 5: Additional results of glasses removal.