ReMix: Towards Image-to-Image Translation with Limited Data Supplementary Material

Jie Cao[1,2], Luanxuan Hou[1,2], Ming-Hsuan Yang[3,4,5], Ran He[1,2], Zhenan Sun[1,2]
[1]NLPR, CRIPAC & CEBSIT, CASIA [2]AIR, UCAS
[3]University of California at Merced [4]Google Research [5]Yonsei University
{jie.cao, luanxuan.hou}@cripac.ia.ac.cn, mhyang@ucmerced.edu
{rhe, znsun}@nlpr.ia.ac.cn

1. Additional Implementation Details

In this work, we use the machine learning library Pytorch. We implement the baselines, *i.e.*, StarGAN v2 [1], StyleGAN 2 [5], and SPADE Net [7], using the open-source codes[1][2][3].

Loss functions. For the StarGAN model, we have $\mathbf{s} = G(\mathbf{x}, \mathbf{t})$. That is, the generator G transforms the source image $\mathbf{x}$ based on the representation $\mathbf{t}$. The reconstruction loss $\mathcal{L}_{rec}$ is computed as follows:

$$\mathcal{L}_{rec}(\mathbf{s}, \mathbf{t}) = \|E(\mathbf{s}) - \mathbf{t}\|_1 , \tag{1}$$

where E denotes the encoder in the StarGAN v2 model, and $\|\cdot\|_1$ denotes L1 Norm. Then we compute the relative form of the reconstruction loss by instantiating $\mathcal{L}_{con}$ in Equations (8) and (9) of the paper with $\mathcal{L}_{rec}$.

For the StyleGAN model, we have the NIR input $\mathbf{x}$ and the corresponding VIS face $\mathbf{t}$. Let $\mathbf{s} = G(\mathbf{x})$, and we compute the L1 distance loss [3] as:

$$\mathcal{L}_{l1}(\mathbf{s}, \mathbf{t}) = \|\mathbf{s} - \mathbf{t}\|_1 . \tag{2}$$

For the SPADE model, we have the semantic map $\mathbf{x}$ and the corresponding city scene $\mathbf{t}$. Let $\mathbf{s} = G(\mathbf{x})$, and we compute the perceptual loss [4] as:

$$\mathcal{L}_{perc}(\mathbf{s}, \mathbf{t}) = \sum_{j=1}^{N} \frac{10}{M_j} \cdot \left\| f^j(\mathbf{s}) - f^j(\mathbf{t}) \right\|_1 , \tag{3}$$

where f^j denotes the j-th layer with M_j elements of the VGGNet [8]. Similarly, we compute the relative form of losses, $\mathcal{L}'_{perc}$ and $\mathcal{L}'_{l1}$, by applying Equations (8) and (9) in our paper. We only modify the mentioned losses for the ReMix method, and the other losses remain the same.

[1]https://github.com/clovaai/stargan-v2
[2]https://github.com/rosinality/stylegan2-pytorch
[3]https://github.com/NVlabs/SPADE

Table 1: Results of the ablation study on the AFHQ dataset [1]. We use Fréchet Inception Distance (FID, lower is better) as the metric. "iter" denotes the number of the training iterations. We compute the FID scores using the models with different augmentation probability p.

iter	p	FID (latent-guided)	FID (refference-guided)
50k	0.00	19.23	22.69
	0.25	17.72	18.10
	0.50	20.80	21.33
	0.75	27.95	22.65
	1.00	27.86	25.04
100k	0.00	16.18	19.78
	0.25	**15.22**	**15.56**
	0.50	15.77	18.76
	0.75	26.47	23.19
	1.00	28.58	26.49

FID scores. For the animal face translation task, we sample 10 style representations to produce 10 different results for each source image. Then, we compute the FID scores based on the translated results and the training images in the target domain. We compute the FID scores for every pair of the image domains, including dog→cat, dog→wildlife, cat→wildlife, cat→wildlife, wildlife→cat, and wildlife→dog. We report the average values in the table of our paper. For the image synthesis from semantic label maps task, we compute the FID scores using the translated results and the corresponding ground truths in the testing set. We do not sample multiple results for each input in this case.

2. Additional Results

Ablation study. We conduct a simple grid search for the augmentation probability p on the AFHQ dataset [1]. The

search space is $\{0, 0.25, 0.5, 0.75, 1\}$, and we use the FID score as the metric. Table 1 reports the FID scores of the models with different p. We find that $p = 0.25$ has the best performance in the ablation study. We use this found value for all the experiments in this work.

Visualization results. Figures 1 and 2 show additional reference-guided translation results on the AFHQ dataset [1]. Our model translates the input source image based on the given reference image, generating diverse results. Moreover, Figure 3 shows additional latent-guided translation results. In this case, our model translates the input into diverse results using the style representations randomly sampled from Gaussian distribution.

Figure 4 shows synthesized identities from the model trained on the CASIA dataset [6]. Given two NIR faces, we average the representations extracted by the encoder [9]. Then, the decoder [5] uses the averaged representation to generate a new VIS face. Furthermore, Figure 5 shows additional synthesis results from the proposed method on the Cityscapes dataset [2] with comparison to those from the mixup method [10].

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8188–8197, 2020.

[2] 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 *CVPR*, 2016.

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

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

[6] Stan Li, Dong Yi, Zhen Lei, and Shengcai Liao. The casia nir-vis 2.0 face database. In *CVPRW*, 2013.

[7] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE TIFS*, 2018.

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Figure 1: Results of the reference-guided translation on the AFHQ dataset [1]. We train our model under the 10% data settings. Our model translates the source images based on the given reference images.

Figure 2: Results of the reference-guided translation on the AFHQ dataset [1]. We train our model under the 10% data settings. Our model translates the source images based on the given reference images.

Figure 3: Results of the latent-guided translation on the AFHQ dataset [1]. We train our model under the 10% data settings. Our model translates the source images with randomly sampled style representations.

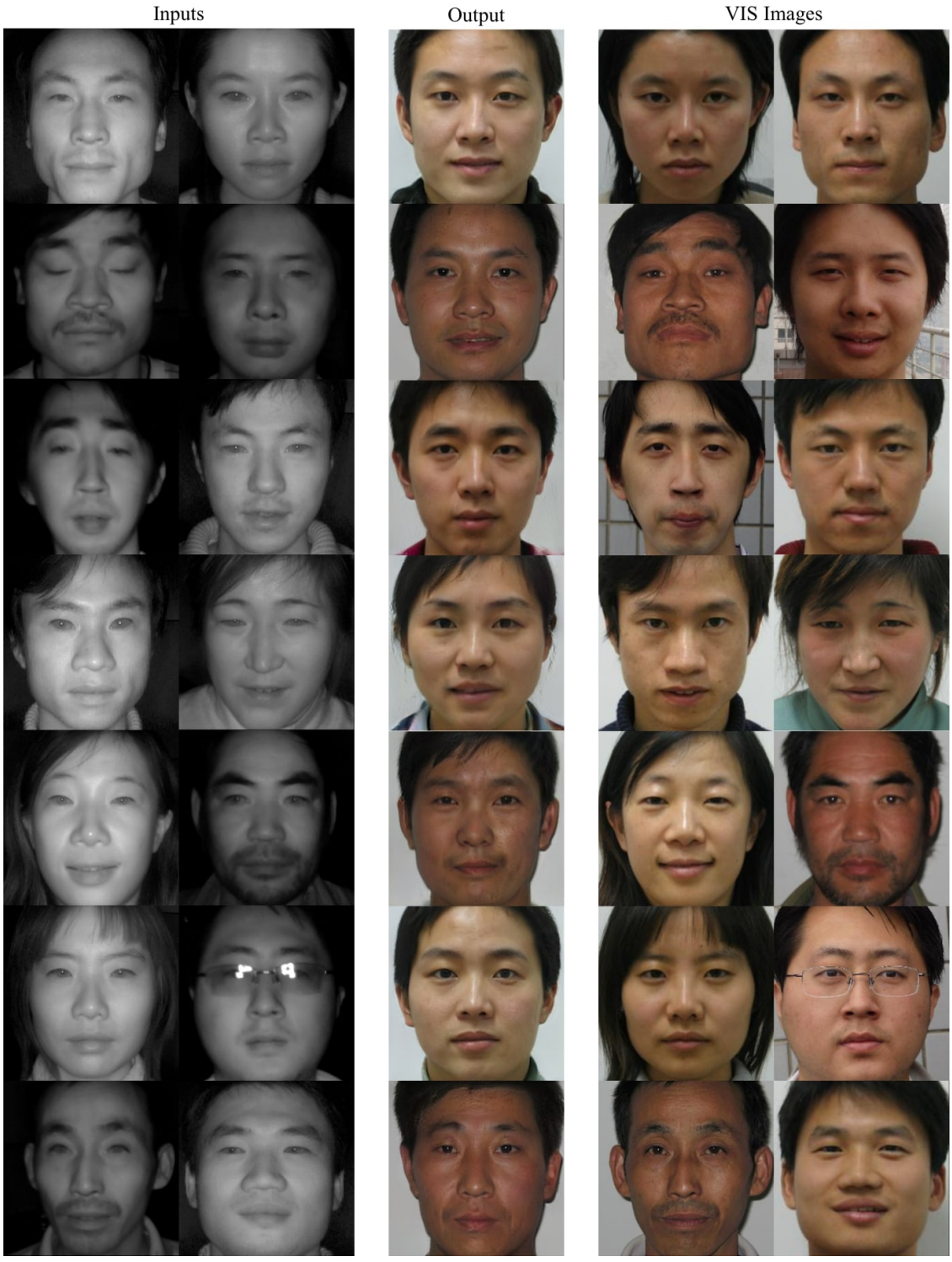

Figure 4: Visual examples of the synthesized identities from the model trained on the CASIA dataset [6]. We mix the representations of the two NIR faces and then generate a VIS face with the mixed input. The corresponding VIS faces of the inputs are shown as references. Each output has a new identity different from either of the input.

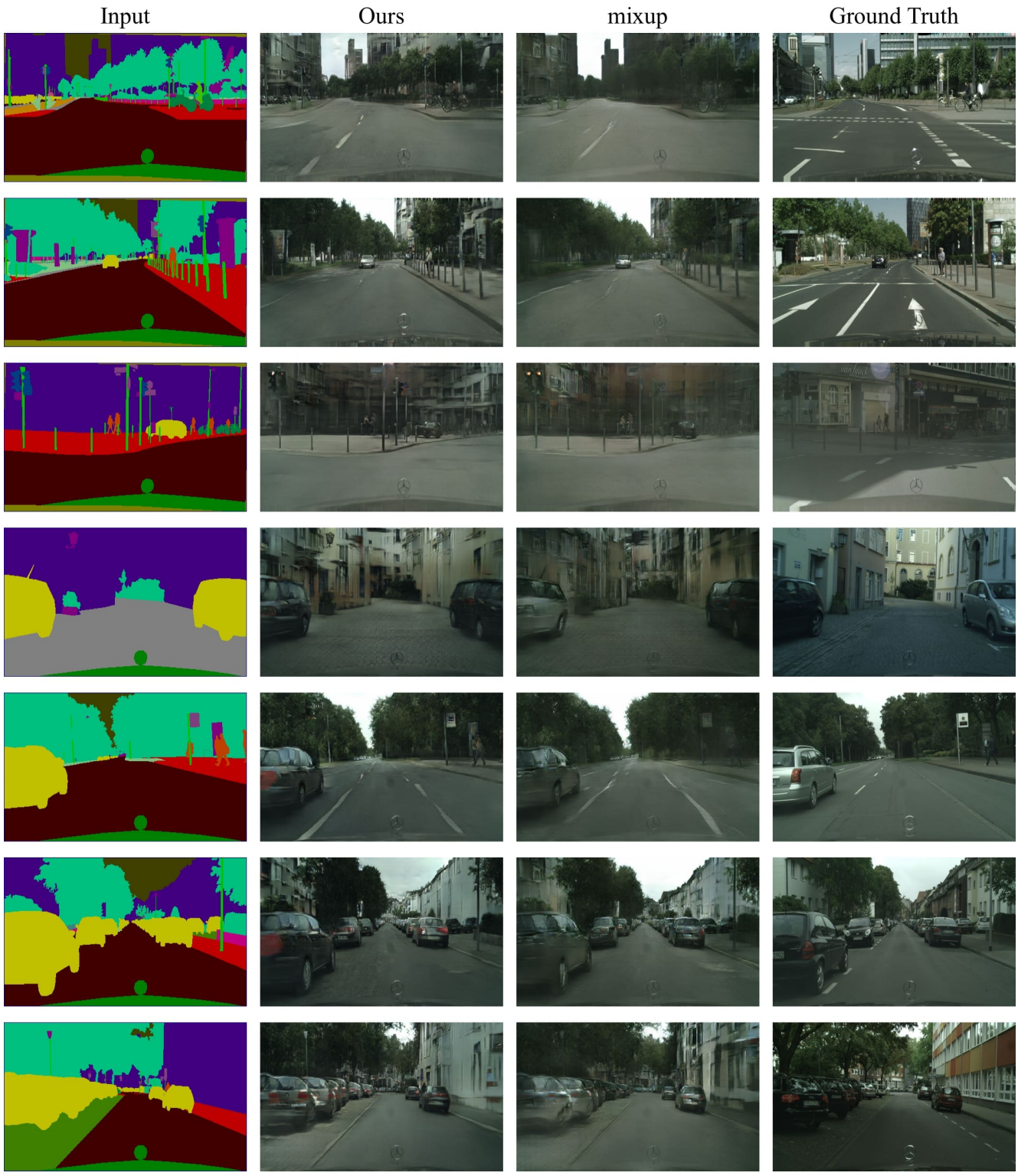

Figure 5: Visual examples synthesized by different methods with 10% training data on the Cityscapes dataset [2]. From left to right, the columns are the inputs, the results of our method, the results of the mixup method [10], and the ground truths.