Supplementary File for Particularity beyond Commonality: Unpaired Identity Transfer with Multiple References

Anonymous ECCV submission

Paper ID 2323

1 More Visual Results

More visual results on RaFD [3], Multi-PIE [2], CelebA [7] and DeepFashion [6] datasets are shown in Fig. 1, 2, 3 & 4 respectively.

2 User Study for Evaluation

We conduct user study among 68 subjects for method comparison, with random 24 groups of generated samples on RaFD dataset. Corresponding to our quantitative evaluation metrics, each subject is instructed to choose the best item on translation accuracy (**Acc**), content preserving (**Con**), perceptual realism (**Per**) and overall transform performance (**Overall**). The results in Tab. 1 demonstrate that our method achieves the best performance among different approaches on the metrics, especially on translation accuracy and overall transform performance, which indicates that our method can better translate the input faces to correct identities and produce the best translated results visually.

Methods	Acc (%)	Con (%)	Per (%)	Overall (%)
FUNIT [4]	7.84	3.92	31.37	9.31
Star-F [1]	0.49	36.27	1.96	8.82
Star-U [1]	11.27	1.96	1.96	4.41
Ours	**80.39**	**57.84**	**64.71**	**77.45**

Table 1. User study of generated results among different methods. The value refers to the ratio of selecting as best item.

3 Effects of Different Reference Images

We further study the effects of references on the translated results in detail. We generate final results with different reference numbers, as well as different expressions of references, which is shown in Fig. 5. In the case of only 1 reference image provided, the generated results vary a lot according to the changes of references and they still preserve part of facial expressions from their corresponding references. However, with the reference number increased, the generated results become more stable with the most suitable facial expression.

4 Network Architecture

Encoder Structure In the encoder, we extract three level features from images for alignment and fusion. Its architecture is shown in Table. 2. Three level features are extracted from the output of 'Conv1', 'Conv2' and 'ResBlock', which correspond to low-, median- and high-level features.

Layer	Output Size	(kernel, stride)
Inputs	$H \times W \times 3$	(- , -)
Conv1	$H \times W \times 64$	(7, 1)
Conv2	$\frac{H}{2} \times \frac{W}{2} \times 128$	(4, 2)
Conv3	$\frac{H}{4} \times \frac{W}{4} \times 256$	(4, 2)
ResBlock $\times 2$	$\frac{H}{4} \times \frac{W}{4} \times 256$	(3, 1)

Table 2. Network architecture of Encoder. All convolution layers in 'Conv' blocks and 'Resblocks' are followed by Instance normalization [9] and ReLU.

Layer	Output Size	(kernel, stride)
Inputs	$\frac{H}{4} \times \frac{W}{4} \times 256$	(- , -)
ResBlock $\times 2$	$\frac{H}{4} \times \frac{W}{4} \times 256$	(3, 1)
Upsample	$\frac{H}{2} \times \frac{W}{2} \times 256$	(-, -)
Conv1	$\frac{H}{2} \times \frac{W}{2} \times 128$	(5, 1)
Upsample (Skip)	$H \times W \times 128$	(-, -)
Conv2	$H \times W \times 64$	(5, 1)
Conv3_1 (Skip)	$H \times W \times 64$	(7, 1)
Conv3_2	$H \times W \times 64$	(7, 1)

Table 3. Network architecture of Decoder. All convolution layers in 'Conv' blocks and 'ResBlocks' are followed by Instance normalization [9] and ReLU except 'Conv3_2' layer, while 'Skip' indicates the corresponding aligned and fused features from encoder are concatenated with features in current layers as skip connection.

Alignment Network Structure We design alignment networks for different level features from encoder. For higher level feature, we adopt less down-sample operations. Alignment network structure for each level feature is shown in Table. 4. For each alignment network, the reference image and content image feature are concatenated as input, and the network produce a 3-channel output, with 2 channels as optical flow map and an extra channel as confidence map.

Decoder Structure At the stage of decoding, the aligned and fused features are fed to corresponding layers of decoder. These features are gradually decoded to the final generated images.

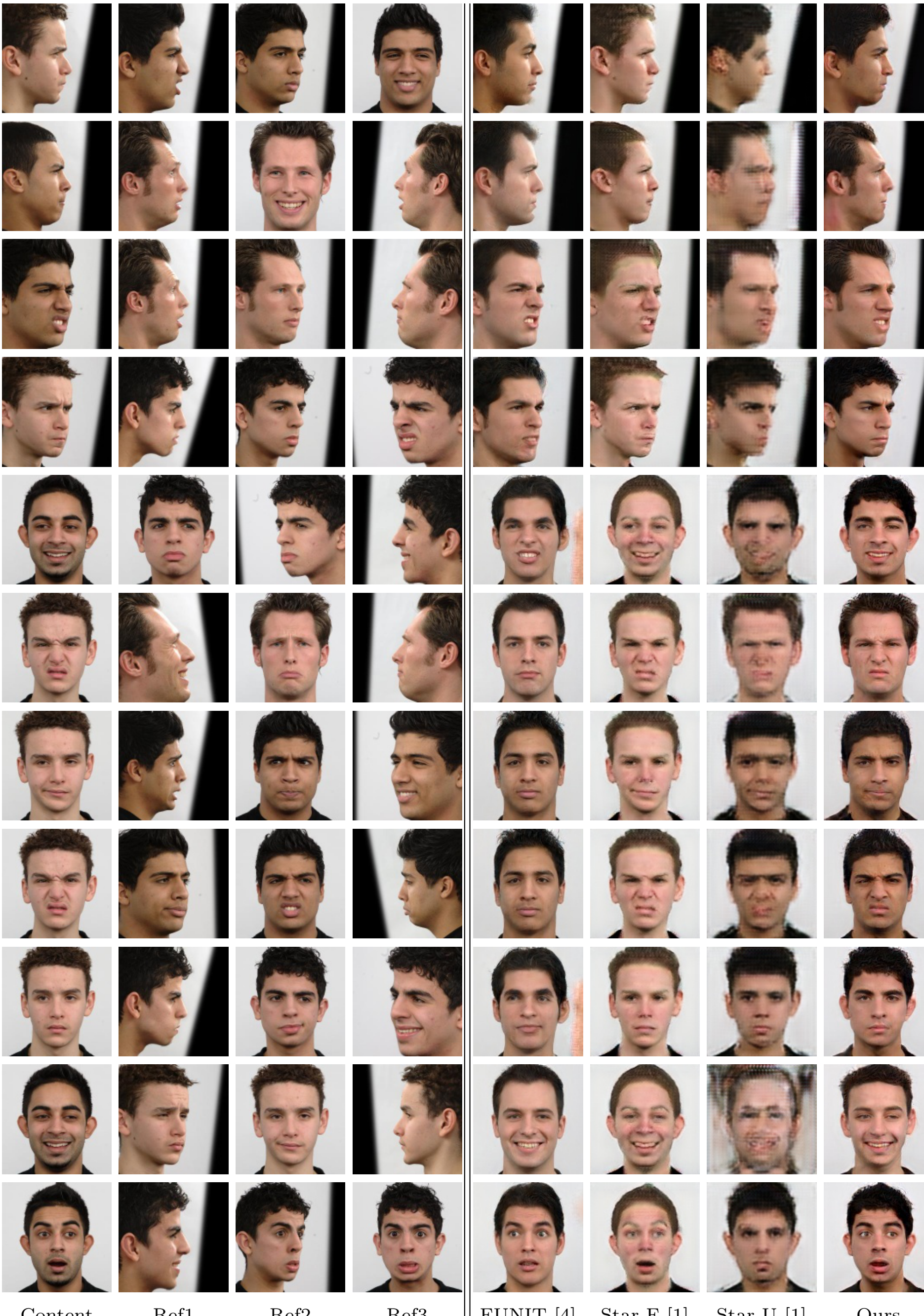

Fig. 1. More visual comparison on RaFD dataset.

Fig. 2. More visual comparison on Multi-PIE dataset.

Fig. 3. More visual comparison on CelebA dataset. Noted that the references of the same person on CelebA are quite different, and thus our method obtain an average identity for results among three references.

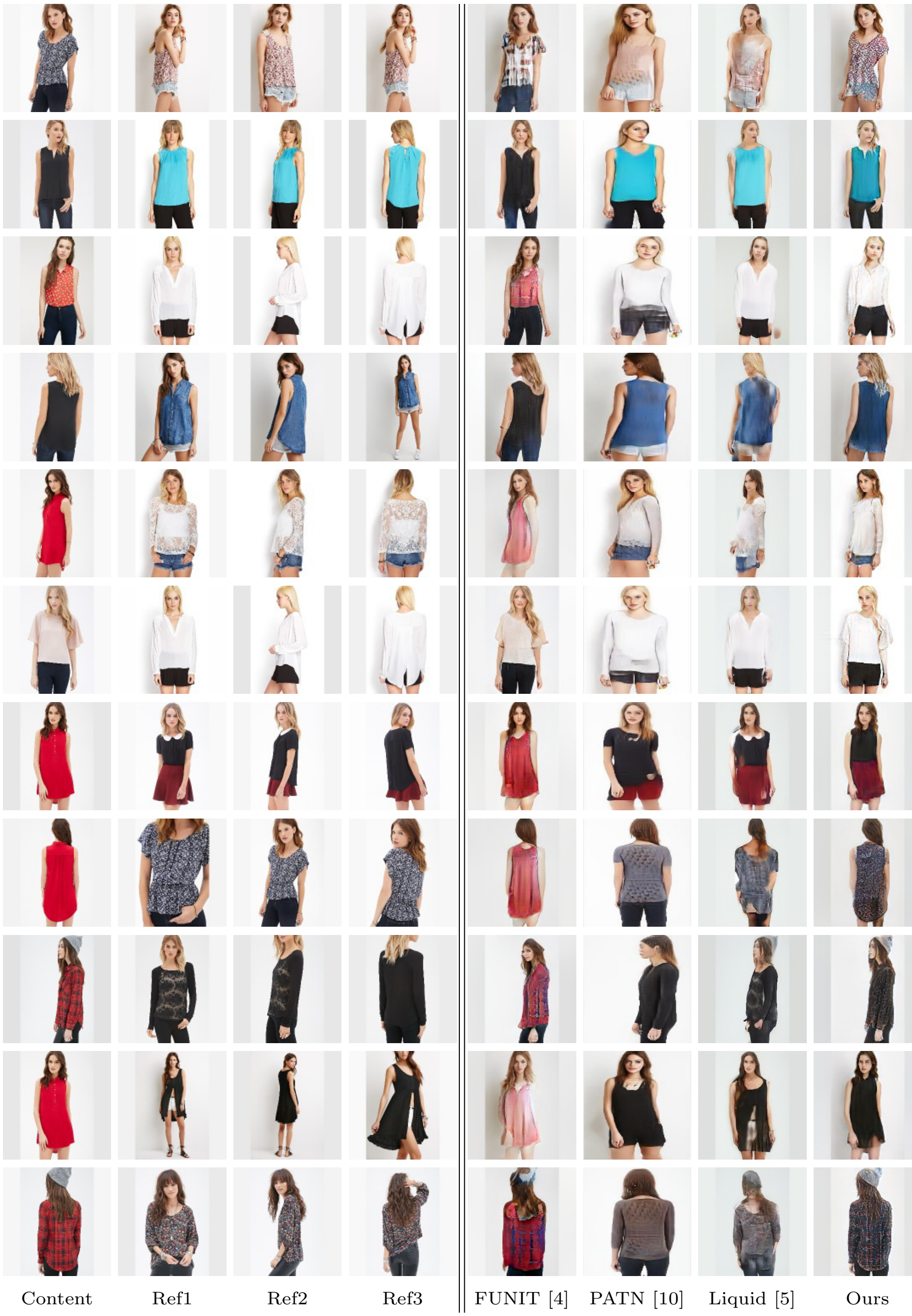

Fig. 4. More visual comparison on DeepFashion dataset.

Fig. 5. Effects of references on generated results. The image at the left top is the content image. For 'k reference(s)' rows, the first k rows are various combinations of different reference images ('Ref$i$'), and the last row is the corresponding generated result ('Result').

	Layer	Output Size	(kernel, stride)
Low-level alignment	Inputs	$H \times W \times 64 \times 2$	(- , -)
	Conv1	$\frac{H}{2} \times \frac{W}{2} \times 64$	(4, 2)
	Conv2	$\frac{H}{4} \times \frac{W}{4} \times 64$	(4, 2)
	Conv3	$\frac{H}{8} \times \frac{W}{8} \times 64$	(4, 2)
	ResBlock $\times 2$	$\frac{H}{8} \times \frac{W}{8} \times 64$	(3, 1)
	Upsample	$\frac{H}{4} \times \frac{W}{4} \times 64$	(-, -)
	Conv4	$\frac{H}{4} \times \frac{W}{4} \times 32$	(5, 1)
	Upsample	$\frac{H}{2} \times \frac{W}{2} \times 32$	(-, -)
	Conv5	$\frac{H}{2} \times \frac{W}{2} \times 32$	(5, 1)
	Upsample	$H \times W \times 32$	(-, -)
	Conv6_1	$H \times W \times 16$	(5, 1)
	Conv6_2	$H \times W \times 3$	(3, 1)
Median-level alignment	Inputs	$\frac{H}{2} \times \frac{W}{2} \times 128 \times 2$	(- , -)
	Conv1	$\frac{H}{4} \times \frac{W}{4} \times 128$	(4, 2)
	Conv2	$\frac{H}{8} \times \frac{W}{8} \times 128$	(4, 2)
	ResBlock $\times 2$	$\frac{H}{8} \times \frac{W}{8} \times 128$	(3, 1)
	Upsample	$\frac{H}{4} \times \frac{W}{4} \times 128$	(-, -)
	Conv3	$\frac{H}{4} \times \frac{W}{4} \times 64$	(5, 1)
	Upsample	$\frac{H}{2} \times \frac{W}{2} \times 64$	(-, -)
	Conv4_1	$\frac{H}{2} \times \frac{W}{2} \times 64$	(5, 1)
	Conv4_2	$\frac{H}{2} \times \frac{W}{2} \times 3$	(3, 1)
High-level alignment	Inputs	$\frac{H}{4} \times \frac{W}{4} \times 256 \times 2$	(- , -)
	Conv1	$\frac{H}{8} \times \frac{W}{8} \times 256$	(4, 2)
	ResBlock $\times 2$	$\frac{H}{8} \times \frac{W}{8} \times 256$	(3, 1)
	Upsample	$\frac{H}{4} \times \frac{W}{4} \times 256$	(-, -)
	Conv2_1	$\frac{H}{4} \times \frac{W}{4} \times 128$	(5, 1)
	Conv2_2	$\frac{H}{4} \times \frac{W}{4} \times 3$	(3, 1)

Table 4. Network architecture of multiple level alignment networks. All convolution layers in 'Conv' blocks and 'Resblocks' are followed by Batch normalization and ReLU except the final convolution layer in each level. Besides, there are skip connections inside the network structure like U-net [8].

References

1. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR (2018)
2. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing (2010)
3. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and emotion (2010)
4. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: ICCV (2019)
5. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: ICCV (2019)
6. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016)
7. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
9. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
10. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR (2019)