Pacific Graphics 2021
E. Eisemann, K. Singh, and F.-L Zhang
(Guest Editors)

Volume 40 (2021), Number 7

Deep Learning-Based Unsupervised Human Facial Retargeting

Seonghyeon Kim[1], Sunjin Jung[1], Kwanggyoon Seo[1], Roger Blanco i Ribera[2], Junyong Noh[1]

[1]KAIST, Visual Media Lab [2]C-JeS Gulliver Studios

1. Architecture

Table 1 and Table 2 show the architecture of ReenactNet and BPNet, respectively.

Table 1: *Overview of architecture of ReenactNet. Convolutional filters are specified in the format of "k(#kernel size)s(#stride)". PS2 indicates a pixel shuffle layer [SCH* 16] with an upscale factor of 2. The two decoders D_s and D_t of the autoencoder share the same structure.*

Encoder E	Filter	Activation function	Output
Conv	k3s1	ReLU	$16 \times 128 \times 128$
Conv	k3s2	ReLU	$32 \times 64 \times 64$
Conv	k3s2	ReLU	$64 \times 32 \times 32$
Conv	k3s2	ReLU	$128 \times 16 \times 16$
Conv	k3s2	ReLU	$256 \times 8 \times 8$
Conv	k3s2	ReLU	$512 \times 4 \times 4$
FC	-	-	512
FC	-	-	8192
Conv	k3s1	-	$512 \times 4 \times 4$
PS2	-	LReLU ($\alpha = 0.2$)	$512 \times 8 \times 8$

Decoder D	Filter	Activation function	Output
Conv	k3s1	LReLU ($\alpha = 0.2$)	$512 \times 8 \times 8$
PS2	-	-	$512 \times 16 \times 16$
Conv	k3s1	LReLU ($\alpha = 0.2$)	$256 \times 16 \times 16$
PS2	-	-	$256 \times 32 \times 32$
Conv	k3s1	LReLU ($\alpha = 0.2$)	$128 \times 32 \times 32$
PS2	-	-	$128 \times 64 \times 64$
Conv	k3s1	LReLU ($\alpha = 0.2$)	$64 \times 64 \times 64$
PS2	-	-	$64 \times 128 \times 128$
Conv	k3s1	LReLU ($\alpha = 0.2$)	$32 \times 128 \times 128$
PS2	-	-	$32 \times 256 \times 256$
Conv	k7s1	tanh	$3 \times 128 \times 128$

2. Additional Results

The following Figures 1, 2, 3, and 4 show additional results.

Table 2: *Overview of the architecture of BPNet. Convolutional filters are specified in the format of "k(#kernel size)s(#stride)".*

Encoder E	Filter	Activation function	Output
Conv	k3s1	ReLU	$16 \times 128 \times 128$
Conv	k3s2	ReLU	$32 \times 64 \times 64$
Conv	k3s2	ReLU	$64 \times 32 \times 32$
Conv	k3s2	ReLU	$128 \times 16 \times 16$
Conv	k3s2	ReLU	$256 \times 8 \times 8$
Conv	k3s2	ReLU	$512 \times 4 \times 4$
FC	-	ReLU	2048
FC	-	ReLU	512
FC	-	ReLU	256
FC	-	ReLU	128
FC	-	-	52

References

[SCH*16] Shi W., Caballero J., Huszár F., Totz J., Aitken A. P., Bishop R., Rueckert D., Wang Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1874–1883.

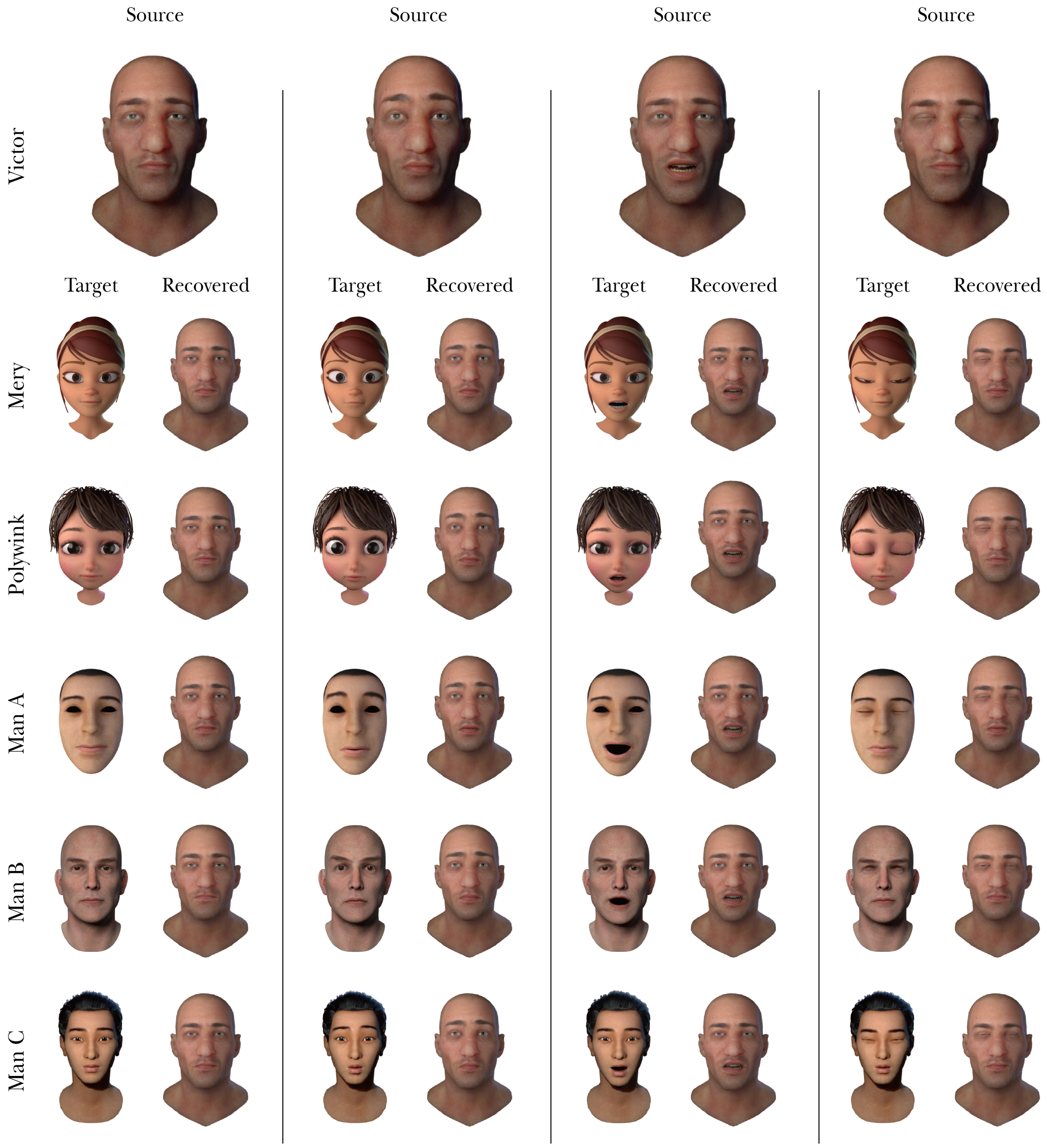

Figure 1: *Cyclic retargeting of expressions to verify robustness of our method. Each expression of the source model is retargeted to different models as shown (Target) and then retargeted back to the source model (Recovered).*

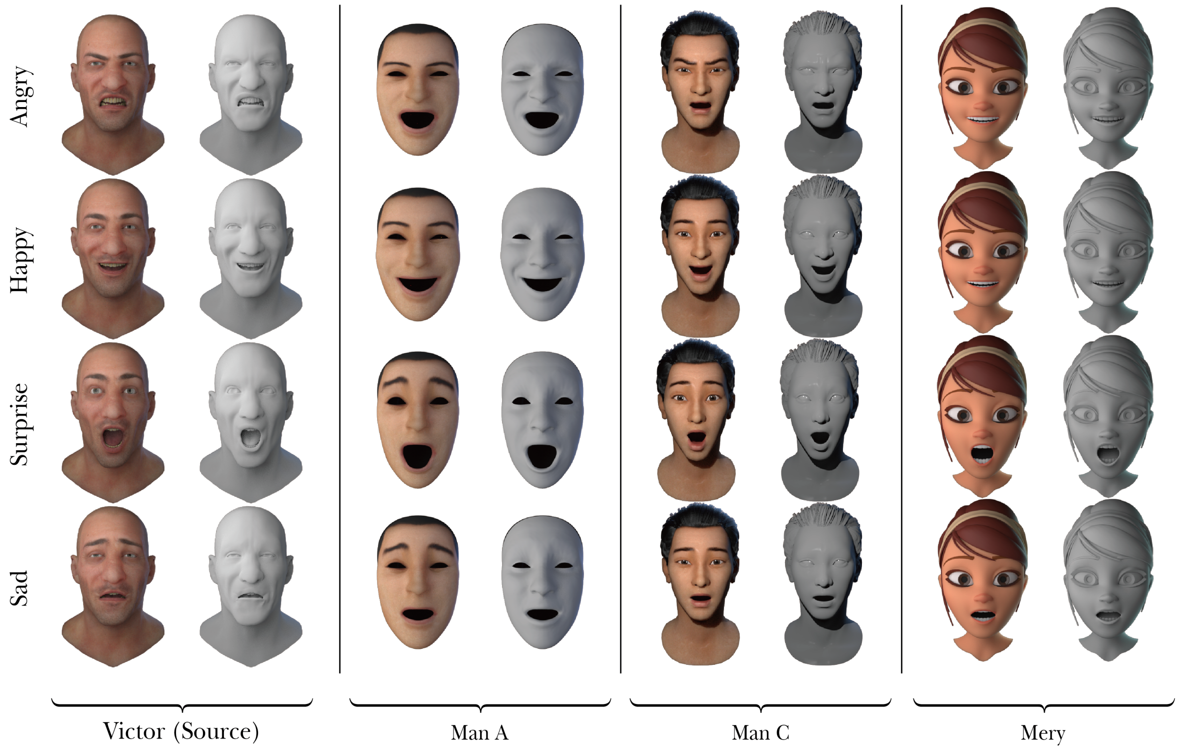

Figure 2: *Results of our method on extreme expressions: anger, happiness, surprise, and sadness. We added 1000 more frames of animation to the training datasets of source and target models because the original dataset does not cover an extreme range of expressions. The expressions of the source model are reproduced well on Man A and Man C. In case of Mery, we observed that the angry and sad expressions are not convincingly transferred compared to the other expressions due to the large difference in facial proportion between the source and target models. However, the other two expressions are retargeted well.*

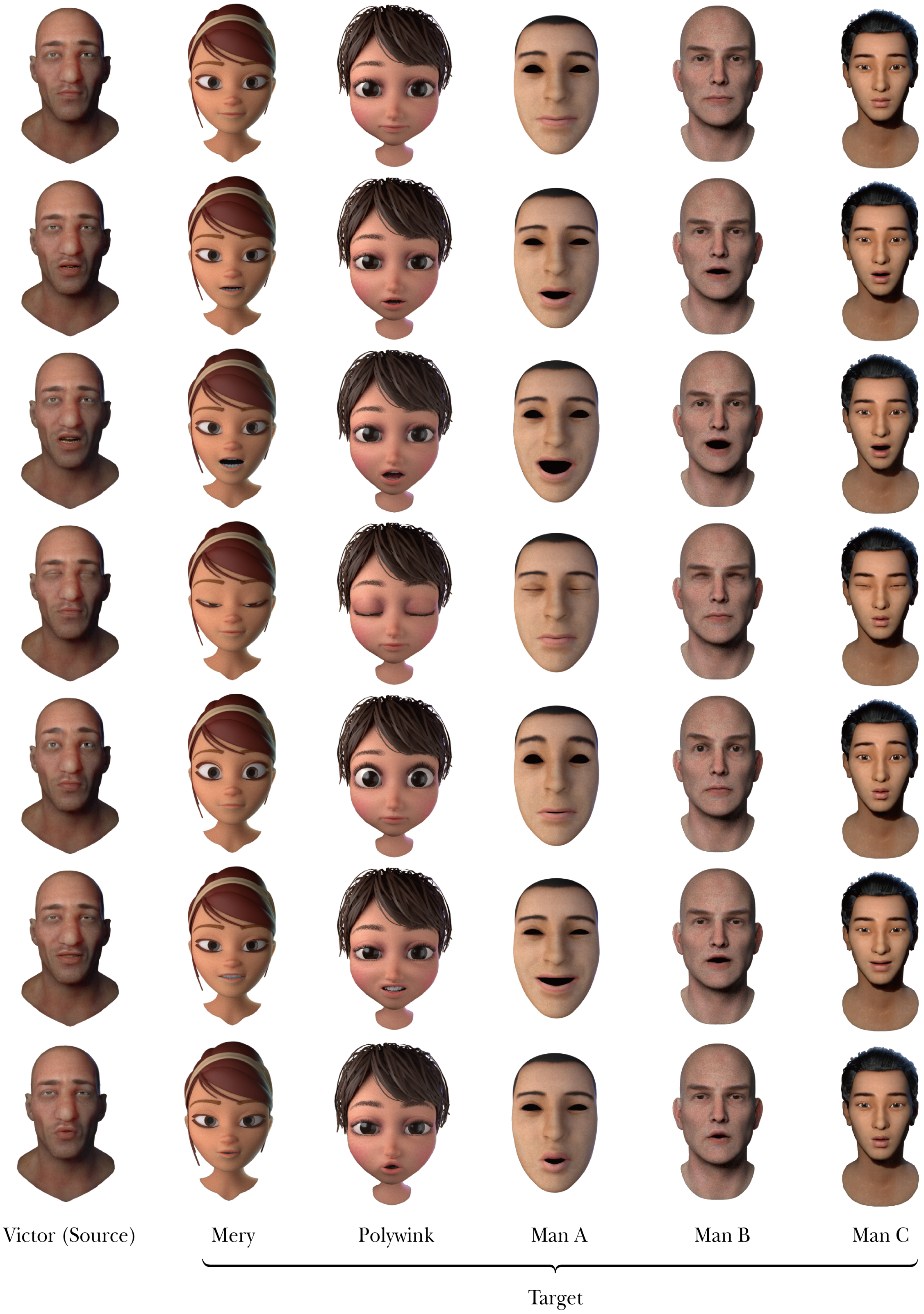

Figure 3: *Additional results of our method retargeted to various models with different shapes, genders, and styles.*

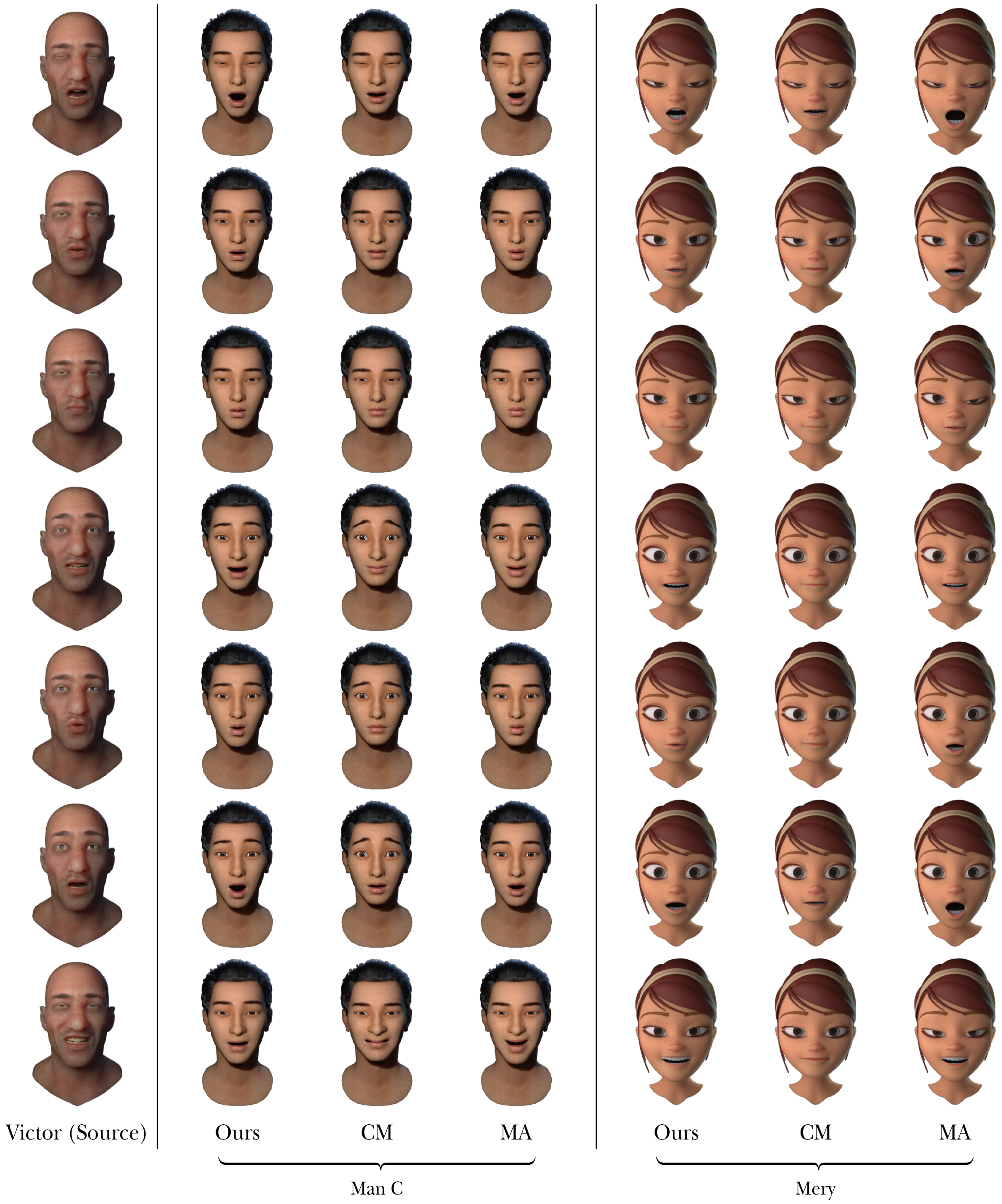

Figure 4: *Comparison of retargeting results produced by our method (Ours), cross-mapping (CM), and manifold alignment (MA). In all cases, our method generates results superior or comparable to those of other methods.*