Pose Adapted Shape Learning for Large-Pose Face Reenactment

Supplementary Material

This report includes the following sections ([L #] refers to the line number in the paper where we ask the reader to refer to the Supplementary document.):

1 Additional details of the network settings for the major modules.
2 Additional details of the six PAEs (Pose Adapted face Encoders) [L 381] .
3 Additional details and specifications of the MPIE-LP and VoxCeleb2-LP datasets [L 431].
4 Details of the evaluation metrics [L 440].
5 Additional ablation study on different loss settings.
6 Comparison on VoxCeleb2 for regular pose variation.
7 Computational Cost.
8 More comparisons with other approaches on the MPIE-LP and VoxCeleb2-LP datasets.
9 Link to the code and model. The code and pretrained model are available on https://github.com/AvLab-CV/PASL.

1. Network Settings of Major Modules

Table 1 presents the dimensions of the DECA parameters, including the identity shape parameter β, expression shape parameter ψ, and pose parameter θ, as well as dimensions of the recomposed shap s_{rc}, the source shape s_s and the output shape s_o in the Cycle-consistent Shape Generator (CSG) [L 223, 238, 472].

Notations	β	ψ	θ	s_{rc}	s_s	s_o
Dim	100	50	6	$256^2 \times 3$	$256^2 \times 3$	$256^2 \times 3$

Table 1. Dimensions of identity shape parameter β, expression shape parameter ψ, pose parameter θ, recomposed shap s_{rc}, source shape s_s, and output shape s_o in the CSG (Cycle-consistent Shape Generator).

Table 2 shows the network settings of the encoder V_e and the decoder V_d, and Table 3 are the network settings of the style encoder E_s and the discriminator D_f. These modules are all in the Attention Embedded Generator (AEG) [L 262∼264].

Table 4 shows the dimensions of the latent code c_{rc}, the attention feature code F^t_{at}, the style code c_s, the feature sequence f_{m-1}, the self-attention feature sequence f_M, the mapping weights of query $\boldsymbol{W}_q$, the mapping weights of key $\boldsymbol{W}_k$, the mapping weights of value $\boldsymbol{W}_v$, and the triplet (query Q_m, key K_m, value V_m) in the transformer T. The number of multi-headed attention decoding layers N and the number of heads N_h are also given in Table 4.

Layer	Resample	Norm	Output Dim
	V_e		
Input	-	-	$256^2 \times 3$
$Conv1 \times 1$	-	-	$256^2 \times 64$
ResBlk 1	AvgPool	IN	128^3
ResBlk 2	AvgPool	IN	$64^2 \times 256$
ResBlk 3	AvgPool	IN	$32^2 \times 512$
ResBlk 4	AvgPool	IN	$16^2 \times 512$
ResBlk 5	AvgPool	IN	$8^2 \times 512$
	V_d		
ResBlk 1	Upsample	AdaIN	$16^2 \times 512$
ResBlk 2	Upsample	AdaIN	$32^2 \times 512$
ResBlk 3	Upsample	AdaIN	$64^2 \times 256$
ResBlk 4	Upsample	AdaIN	128^3
ResBlk 5	Upsample	AdaIN	$256^2 \times 64$
$Conv1 \times 1$	-	-	$256^2 \times 3$

Table 2. The network settings of the encoder V_e and the decoder V_d in the AEG (Attention Embedded Generator).

Layer	Resample	Output Dim
Input	-	$256^2 \times 3$
$Conv1 \times 1$	-	$256^2 \times 64$
ResBlk 1	AvgPool	128^3
ResBlk 2	AvgPool	$64^2 \times 256$
ResBlk 3	AvgPool	$32^2 \times 512$
ResBlk 4	AvgPool	$16^2 \times 512$
ResBlk 5	AvgPool	$8^2 \times 512$
ResBlk 6	AvgPool	$4^2 \times 512$
Leaky Relu	-	$16^2 \times 512$
$Conv4 \times 4$	-	$1^2 \times 512$
Leaky Relu	-	$1^2 \times 512$
Reshape	-	512
FC	-	128/1

Table 3. The network settings of the style encoder E_s and the discriminator D_f.

Notations	c_{rc}	F^t_{at}	c_s	f_{m-1}	f_M	N	N_h
Dim	$8^2 \times 512$	$8^2 \times 512$	128	512	64×512	64×512	8
Notations	W_q	W_k	W_v	Q_m	K_m	V_m	Q
Dim	512×512	512×512	512×512	512	512	512	512

Table 4. Dimensions of latent code c_{rc}, attention feature code F^t_{at}, style code c_s, feature sequence f_{m-1}, self-attention feature sequence f_M, the triplet (Q_m, K_m, V_m), mapping weights of query $\boldsymbol{W}_q$, the mapping weights of key $\boldsymbol{W}_k$, the mapping weights of value $\boldsymbol{W}_v$, and the number of the multi-headed attention decoding layers N and the number of heads N_h in transformer T.

2. Additional Details of Pose-Adapted face Encoders (PAEs)

[L 381] The procedure of dividing the MS1Mv3 [4] into training and validating subsets is described as follows. We divided the subjects into 85% for training and 15% for validation for each pose subset. As the number of images with large poses varies from one subject to another, we took this

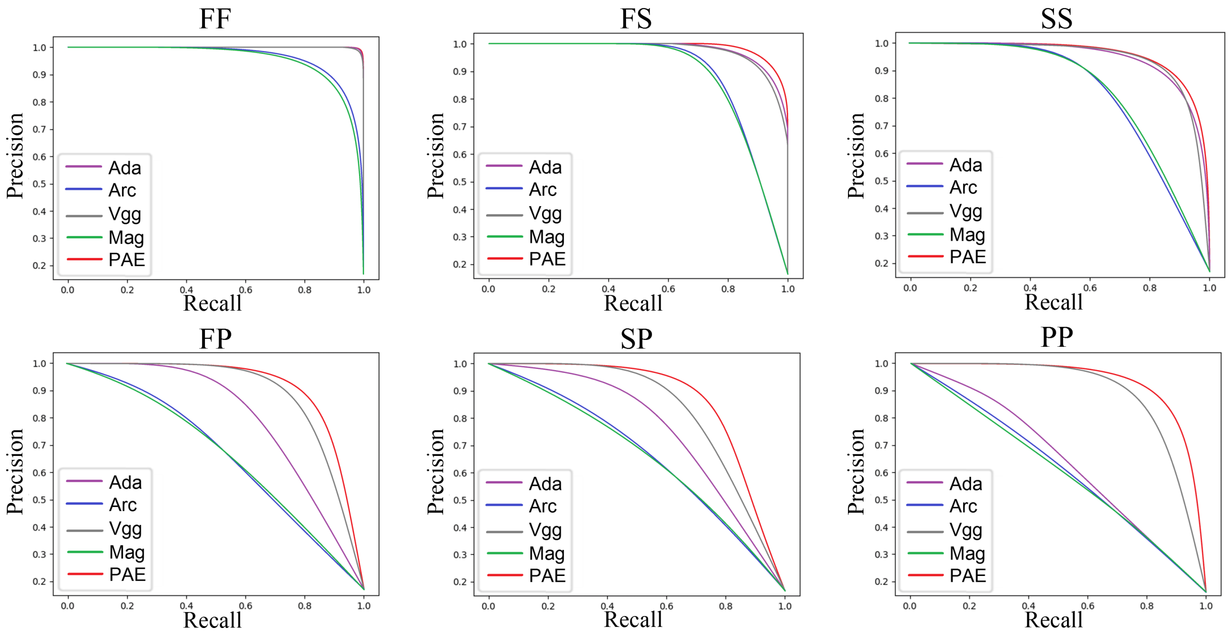

Figure 1. Precision and recall of PAEs, Magface (Mag) [7], Arcface (Arc) [3], Adaface (Ada) [6], and VGGFace2 (Vgg) [2] tested on IJB-C database.

	Training					
PAEs	E_{ff}	E_{ss}	E_{pp}	E_{fs}	E_{sp}	E_{fp}
No. images	72,615	172,444	59,264	98,278	98,296	132,256
No. subjects	12,469	12,134	13,453	11,164	11,179	12,578
	Testing					
No. images	12,814	12,765	10,438	17,343	17,324	23,280
No. subjects	2,208	2,148	2,378	1,972	1,965	2,213

Table 5. Numbers of images and subjects for six PAEs, upper part is for the training set and lower part is for the testing set.

	E_{ss}	E_{pp}	E_{sp}	E_{fp}	Total
	Training (pairs)				
MPIE-LP	414,720	898,560	1,520,640	1,935,360	4,769,280
	Testing (pairs)				
	205,800	411,600	843,680	1,058,410	2,519,490
	Training (pairs)				
VoxCeleb2-LP	1,269,610	1,233,590	684,320	1,314,260	4,501,780
	Testing (pairs)				
	182,620	180,730	136,940	196,850	697,140

Table 6. Numbers of training and testing pairs of MPIE-LP with 80/47 subjects in training/test sets, and of VoxCeleb2-LP with 1259/196 subjects.

fact into account when organizing the training and testing splits. Table 5 shows the numbers of subjects and images in the training and testing sets for the six PAEs. Figure 1 shows the precision and recall of the PAEs compared to off-the-shelf pre-trained face encoders when testing on the IJB-C dataset [8], with a magnified view to each subfigure in Figure 2. Figure 3 illustrates the comparisons with the same encoders but fine-tuned on our training set, with zoom-in views in Figure 4.

3. Dataset Specification

[L 431] Table 6 gives the numbers of pose pairs for the four large-pose subsets (ss, pp, sp, fp) in the training and testing sets of the MPIE-LP and VoxCeleb2-LP. The pose pairs in each large-pose subset are evenly spread out within specific face/head orientation boundaries, so the large-pose reenactment performance can be better evaluated. Figure 5 shows the Pie chart of the percentages of four large-pose subsets in the training set (left) and testing set (right) of VoxCeleb2-LP. As the VoxCeleb2 is an in-the-wild collection, it is impossible to make the face orientation distributed as evenly as the MPIE. However, for each subset, the data is made as evenly distributed as possible for the specification for that subset. In terms of naming, different videos of the same persons are arranged in sequence according to the serial numbers, and consecutive images are arranged in order.

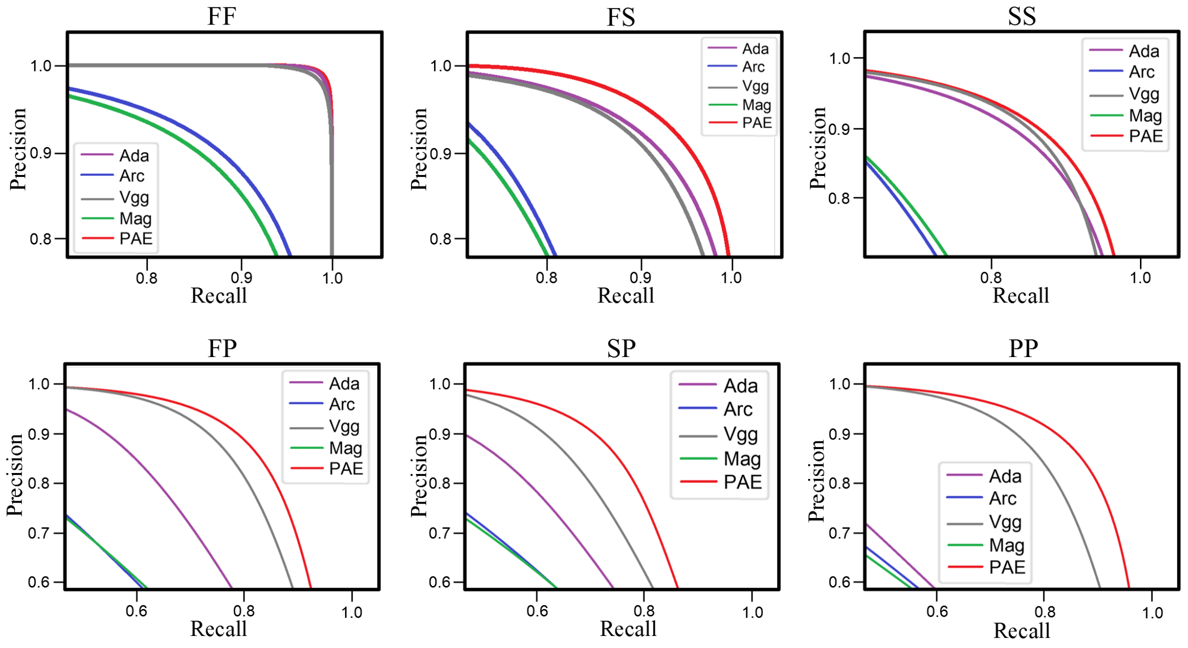

Figure 2. Zoomed-in views of Figure 1

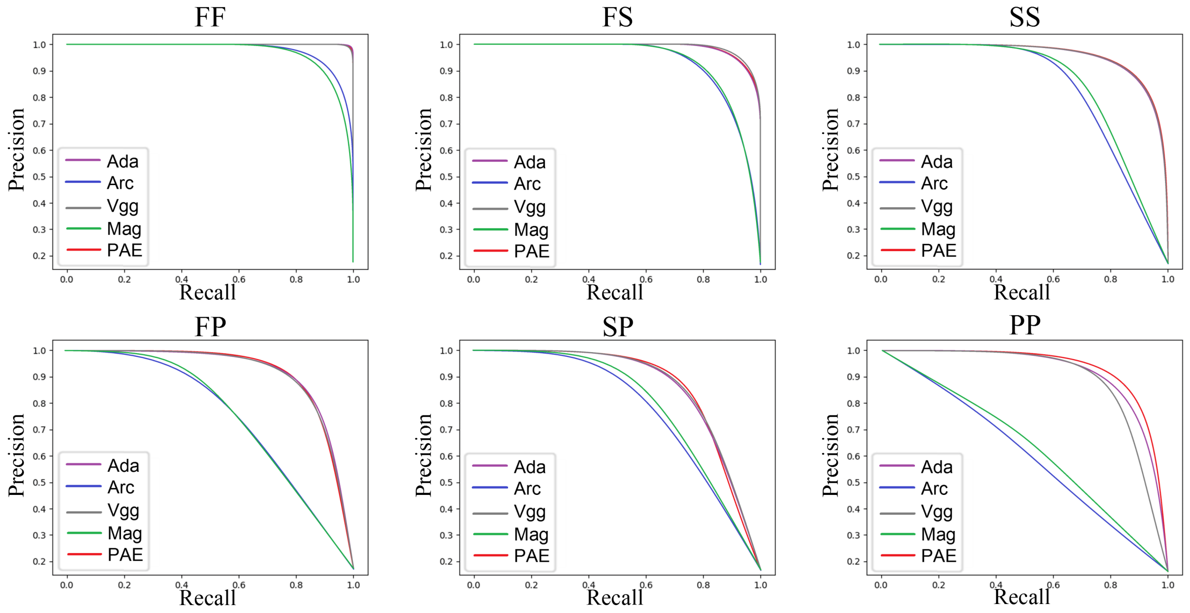

Figure 3. Precision and recall of PAEs, Magface (Mag) [7], Arcface (Arc) [3], Adaface (Ada) [6], and VGGFace2 (Vgg) [2] tested on IJB-C database. All selected encoders are fine-tuned on the training sets of PAEs.

4. Evaluation Metrics

[L 440] Evaluation metrics were selected to test the source identity preservation, reference action transformation and photo-realistic quality of the generated target faces, including the Frechet-Inception Distance (FID), Cosine Similarity (CSIM), Average Rotation Distance (ARD), Learned Perceptual Image Patch Similarity (LPIPS) . FID evaluates the photo-realistic quality by measuring the distribution dis-

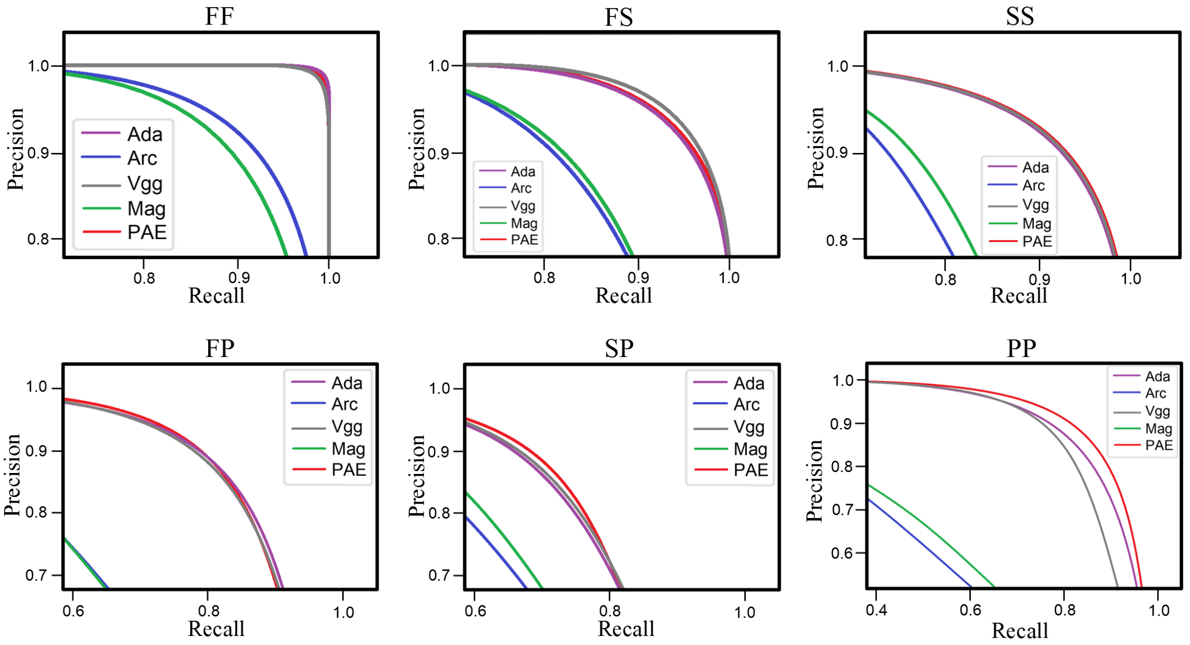

Figure 4. Zoomed-in views of Figure 3

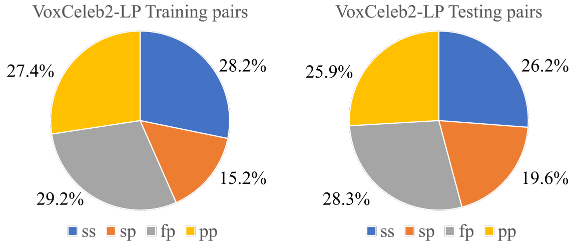

Figure 5. Percentages of four large-pose subsets in the training set (left) and testing set (right) of VoxCeleb2-LP.

tance between the features extracted from the real and generated images. CSIM measures the identity preservation in the generated images by computing the cosine similarity between the facial features extracted from the source and generated images. ARD evaluates the pose transformation and can be computed by using the rotation matrix obtained from the 3DMM of the reference and the generated faces. LPIPS computes the similarity between the activations of two image patches, and is shown to match human perception well. It can only be computed for self-reenactment or when the ground-truth target face is available, such as MPIE-LP.

5. Additional Ablation Study

Table 7 shows the FID, CSIM, ARD and LPIPS for the cross-reenactment on the MPIE-LP testing set with different settings. The third row from the bottom shows the performance of the PASL with the complete set of losses. The top row shows the same model but without the PAE-based identity loss L_{id}, and the performance is the worst, demonstrating the effectiveness of the loss function. The other rows show the performance without other component losses. It shows that the perceptual loss L_{per} is more important than the style loss L_{sty} , and L_{cc} is the second most important loss, behind L_{id}. Figure 6 shows a qualitative comparison of the reenacted faces for the settings in Table 7.

Metrics	FID↓	CSIM↑	ARD↓	LPIPS↓
w/o $\mathcal{L}_{id}$	42.79	0.153	2.653	0.292
w/o $\mathcal{L}_{sty}$	29.54	0.348	2.461	0.254
w/o $\mathcal{L}_{per}$	27.63	0.367	2.493	0.249
w/o $\mathcal{L}_{cc}$	26.25	0.381	2.537	0.242
PASL	**18.1**	**0.46**	**2.24**	**0.21**

Table 7. Cross-reenactment performance on MPIE-LP for different settings on losses.

6. Comparison on VoxCeleb2 for Regular Pose Variation

The table in Figure 8 shows the performance of PASL compared to state-of-the-art approaches on the VoxCeleb2 dataset in regular pose. PASL outperforms the selected SOTA approaches in all metrics on this benchmark dataset.

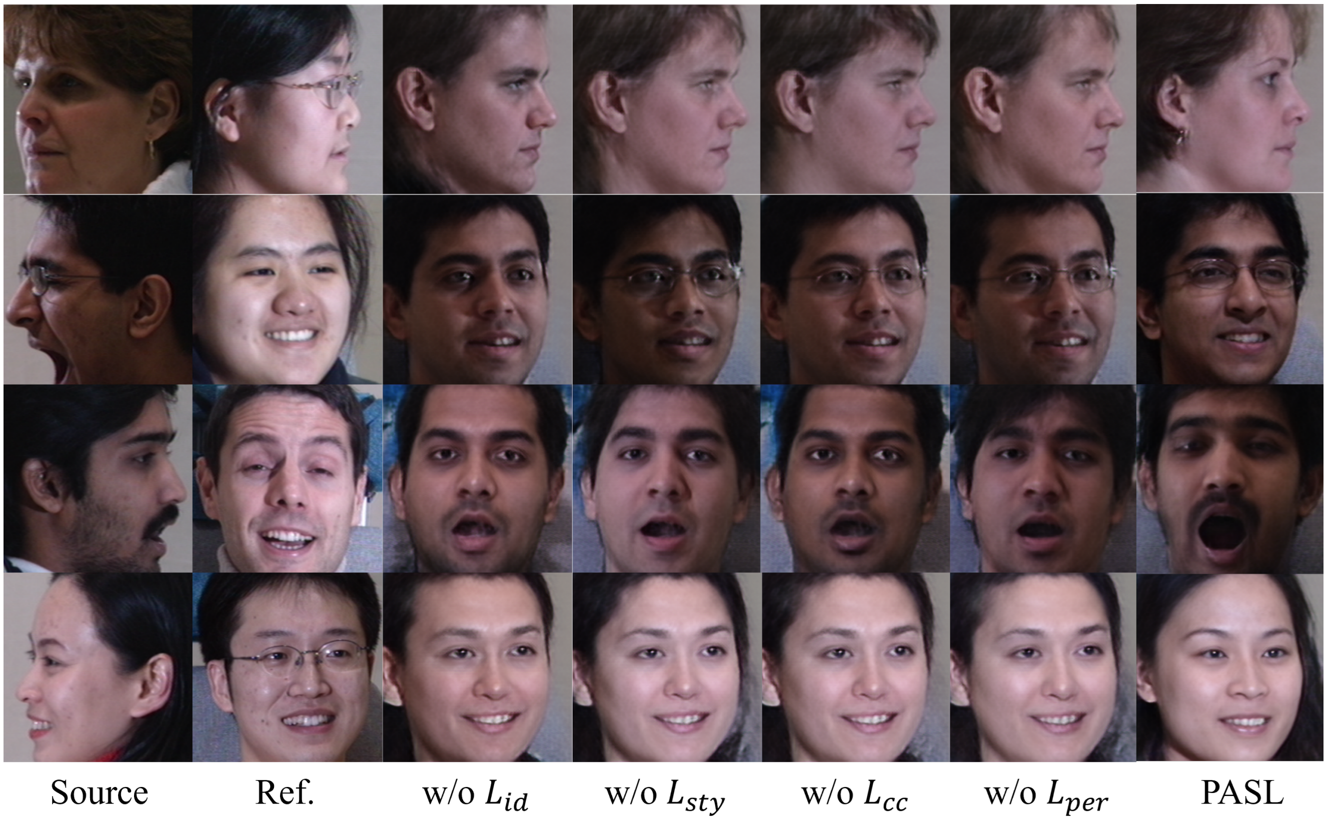

Figure 6. Samples of target face for the different settings in Table 7.

Metrics	FID↓	CSIM↑	ARD↓	LPIPS↓
Bi-layer[9]	88.72	0.38/0.2/0.42	3.01	0.51
DG[5]	36.7	0.43/0.45/0.58	2.89	0.22
HyperReenact[1]	59.8	0.55/0.53/0.61	2.93	0.21
PASL	**32.4**	0.57/0.52/**0.65**	**2.65**	**0.2**

Table 8. Self-reenactment performance on regular pose VoxCeleb2

	Parameters (M)	FLOPS (G)
Baseline	194.7	121.2
Baseline+CSG	220.5	133.5
Baseline+PAE	268.2	121.2
PASL	294.1	133.5

Table 9. The computing cost with various settings of PASL

7. Computational Cost

In this section, we will be discussing the computing cost and the experiments that were conducted using an NVIDIA 3090 GPU. We have included a table (Table 9) which displays the parameters and FLOPS for different settings of PASL. The PAE utilizes six face encoders to extract facial features for calculating L_{id}. This only increases the number of parameters and does not affect computational complexity. To compare our approach with state-of-the-art (SotA) approaches, we have included a table (Table 10) that shows the FLOPS and inference time (images/sec). Some of the SotA approaches only provide demo code, so we compared them during inference. HyperReenact[1] is the most computationally expensive approach, while Bi-layer[9] has the lowest computational complexity.

Metrics	FLOPS (G)	Inference time (imgs/s)
Bi-layer[9]	77.1	5.02
DG[5]	85.5	4.32
HyperReenact[1]	103.7	3.48
PASL	97.3	3.71

Table 10. The computing cost with different settings of PASL

8. More Comparisons with Other Approaches

Figure 7, 8, 9 and 10 present more comparisons of the self- and cross-reenactment results with other approaches on the MPIE-LP and VoxCeleb2-LP. *Note the large pose differences between the sources and reference, which are not seen in previous work.*

9. Code and Model

The code and pretrained model are available on https://github.com/AvLab-CV/PASL.

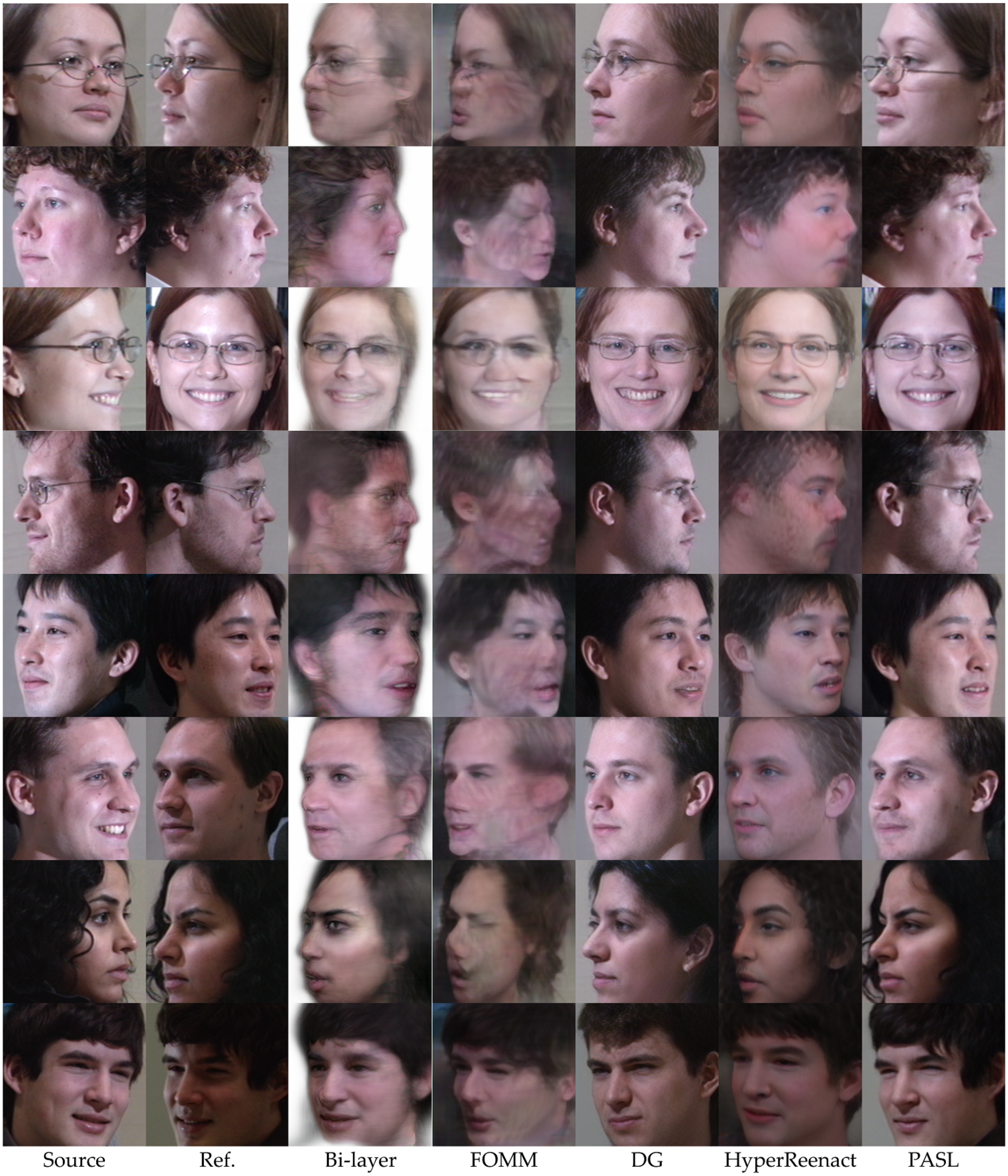

Figure 7. Comparison with other approaches for self-reenactment on MPIE-LP.

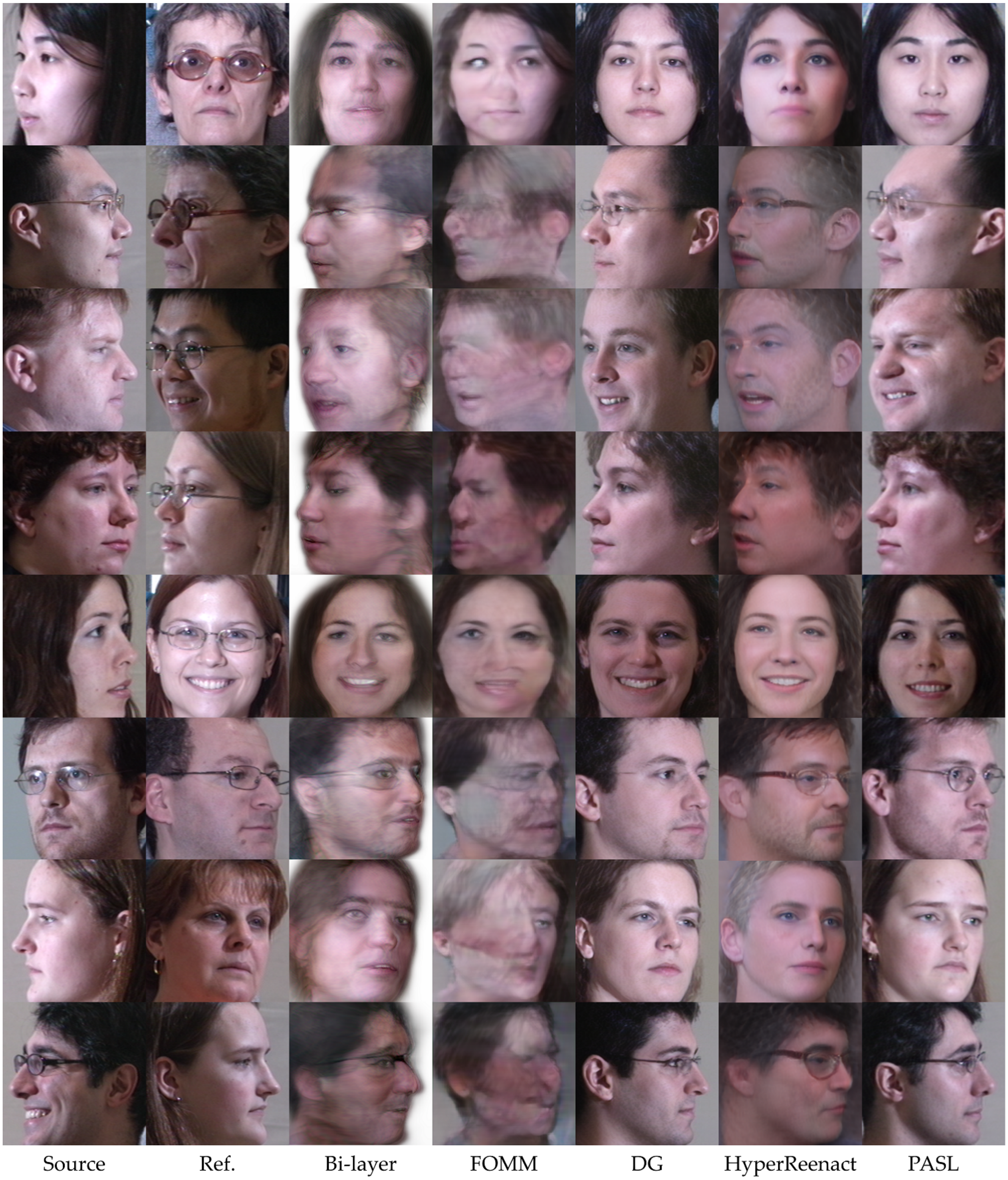

Figure 8. Comparison with other approaches for cross-reenactment on MPIE-LP.

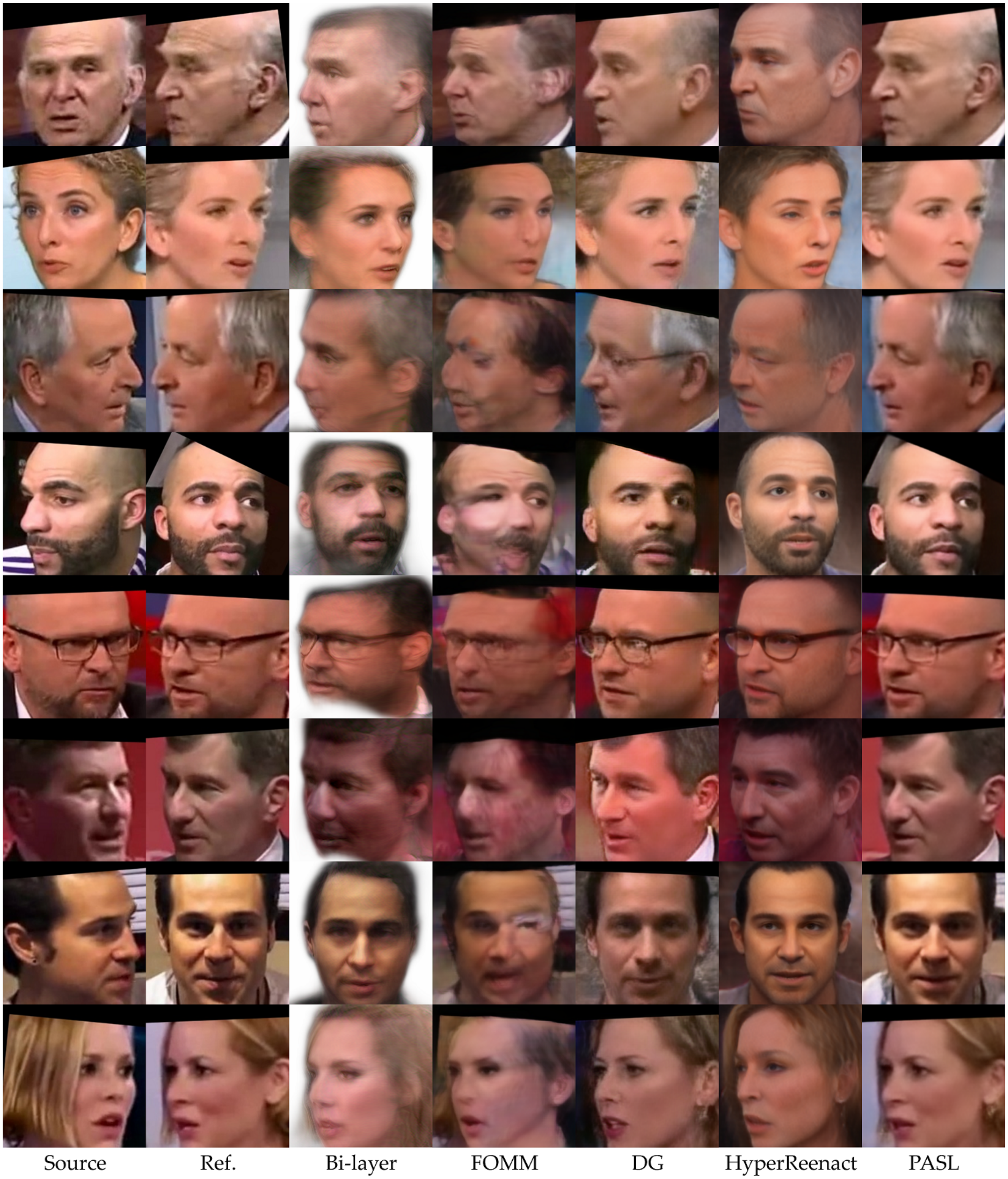

Figure 9. Comparison with other approaches for self-reenactment on VoxCeleb2-LP.

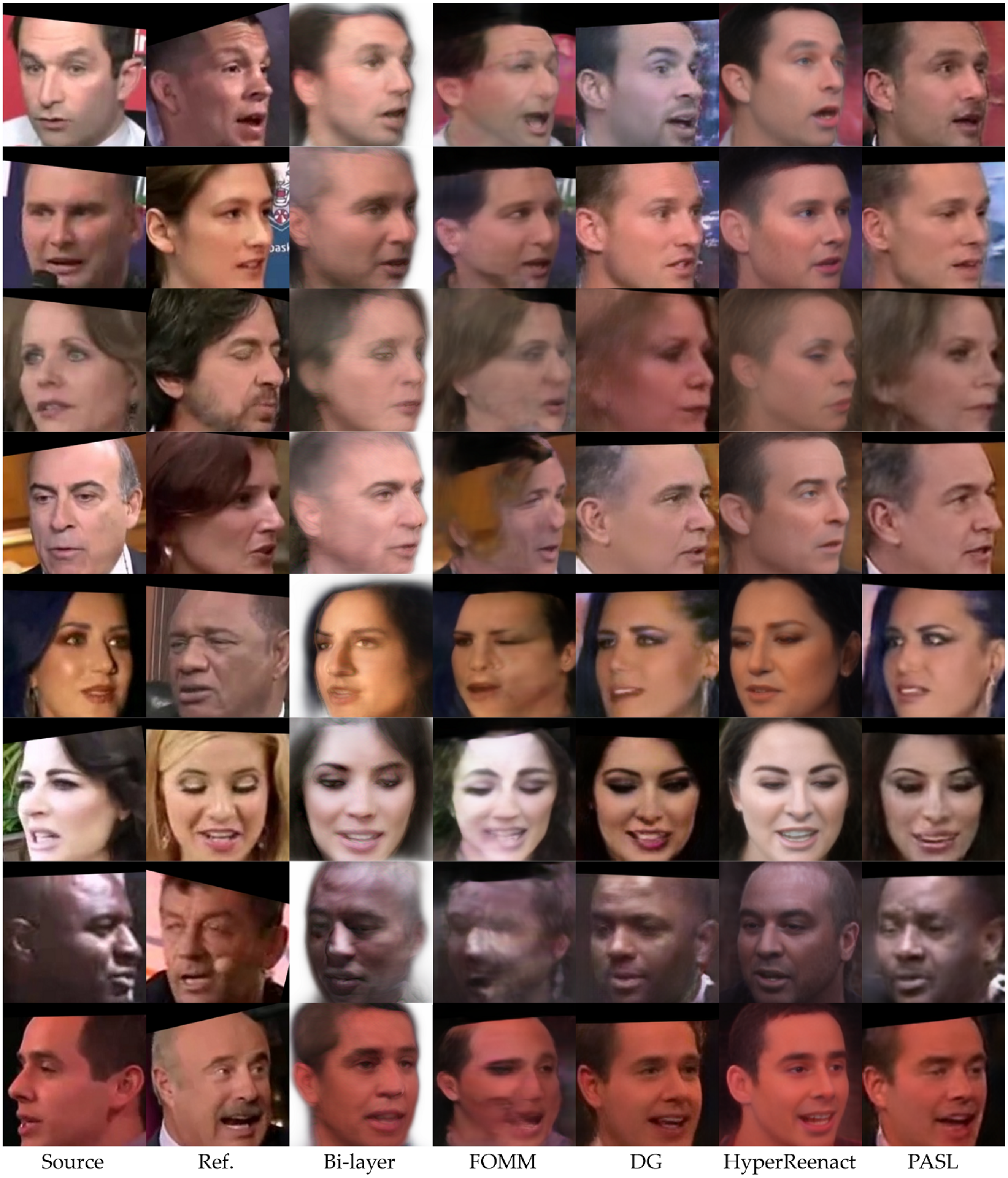

Figure 10. Comparison with other approaches for cross-reenactment on VoxCeleb2-LP.

References

[1] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2, 3

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 3

[4] Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 1

[5] Gee-Sern Hsu, Chun-Hung Tsai, and Hung-Yi Wu. Dual-generator face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 642–650, 2022. 5

[6] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[7] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 2, 3

[8] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn C. Adams, Tim Miller, Nathan D. Kalka, Anil K. Jain, James A. Duncan, Kristen E Allen, Jordan Cheney, and Patrick Grother. Iarpa janus benchmark-b face dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017. 2

[9] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference of Computer vision (ECCV)*, 2020. 5