MetaPortrait: Identity-Preserving Talking Head Generation with Fast Personalized Adaptation
Supplementary Material

Summary

This supplementary material is organized as follows:

- Section 1 introduces the implementation of our architecture and training details.
- Section 2 discusses more ablation studies of our designs.
- Section 3 shows more comparisons with previous works.
- Section 4 discusses the limitation of the proposed method.

1. Implementation Details

Dataset. Following previous work [4], we train our warping and refinement networks on cropped VoxCeleb2 dataset [3], which consists of 145k videos from 6k different identities. We preprocess the videos by cropping the faces with bounding boxes containing the landmarks from the first frame and resize each video sequence to 256×256 resolution. We randomly select 500 videos from the VoxCeleb2 for evaluation. We use the source and driving frames from the same identity for training and same-identity reenactment evaluation, where the driving frames is also the ground truth image. For cross-identity reenactment evaluation, we randomly shuffle the identity in the previous test set, where the source and driving frames have different identities.

Training details. We train the 256×256 base model on the VoxCeleb2 dataset with batch size of 48 using Adam optimizer of learning rate 2×10^{-4} on $8 \times$Tesla V100 GPUs. We set hyperparameters of losses as: $\lambda_{\text{r}} = 10, \lambda_{\text{id}} = 20, \lambda_{\text{eye}} = 50, \lambda_{\text{mouth}} = 50$ and $\lambda_{\text{adv}} = 1$. We first train the warping network for 200,000 iterations, then the warping and refinement network jointly for 200,000 more iterations.

We further conduct our meta-learning stage for $N = 14,000$ outer iterations and meta-learning rate $\beta = 2 \times 10^{-5}$. In each iteration, we train an inner loop for $K = 24$ iteration with inner loop learning rate $\alpha = 2 \times 10^{-4}$ on 48 images per identity.

We train our temporal super-resolution module on the HDTF dataset for 20,000 iterations with batch size 8 and video sequence length 7 using Adam optimizer of learning rate 1×10^{-4}.

Metrics. Following previous works [8, 9], we evaluate our temporal consistency using warping error E_{warp}. For each frame O_t, we calculate the warping error with previous frame O_{t-1} as:

$$E_{\text{pair}}(t) = \frac{\sum_{i=1}^{N} M_t(i) \left\| y_t(i) - W\left(y_{t-1}\right)(i) \right\|_1}{\sum_{i=1}^{N} M_t(i)}, \quad (1)$$

$$E_{\text{warp}}\left(\{t\}_{t=1}^{T}\right) = \frac{1}{T-1} \sum_{t=2}^{T} \left\{E_{\text{pair}}(t)\right\}, \quad (2)$$

where M_t is the occlusion map [11] for a pair of images y_t and y_{t-1}, N is the number of pixels, and W is backward warping operation with optical flow [14]. The averages warping error $E_{\text{warp}}\left(\{t\}_{t=1}^{T}\right)$ is used to evaluate our temporal consistency.

Approach to perform Cross ID reenactment. We can reconstruct an accurate 3D face by fitting a morphable face model [2] based on the dense facial landmarks [17], which well disentangles identity with expression and motion. Benefiting from this, when performing challenging cross-identity reenactment, we simply combine the identity coefficients from the source 3D face with expression and head motion coefficients from the driving face and obtain a new 3D face. Then we project the resultant 3D face to 2D landmarks to serve as the driving target. In this way, there is no leakage of the driving identity so that the source identity could be well preserved.

Detailed architecture. The detailed architecture of our warping and refinement network is shown in Figure 1 and Figure 2. "3×3-Conv-k-1" indicates a convolutional layer with kernel size of 3, channel dimensions of k and stride of 1. "LReLU, ReLU" indicates LeakyReLU [19] and ReLU [1] activation function respectively. Figure 3 illustrates the architecture of our temporal super-resolution network, where "Conv3d-k-1" represents a 3D convolution over temporal and spatial dimensions with k feature dimen-

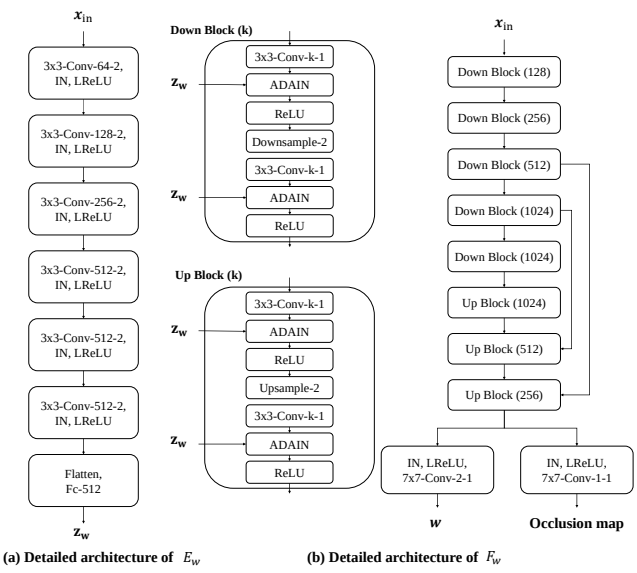

Figure 1. Detailed architecture of our warping network.

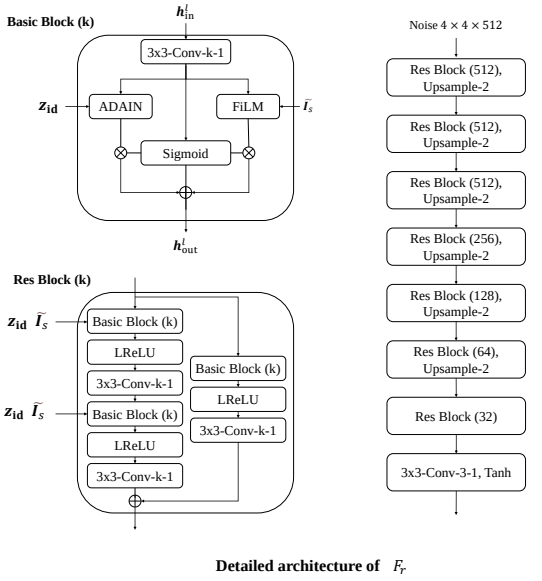

Figure 2. Detailed architecture of our refinement network.

sions and stride of 1.

2. Additional Ablation

2.1. Visualization of ID, Landmark Ablation

Figure 4 illustrates the warped images using the flow field produced by the warping network to evaluate the effectiveness of our dense landmark. The results guided by our dense landmarks are more accurate without obvious artifacts. In Figure 5, we show the visual changes brought by our ID-preserving refinement. Our source identity is better preserved, especially in the area of eye makeup and dimple.

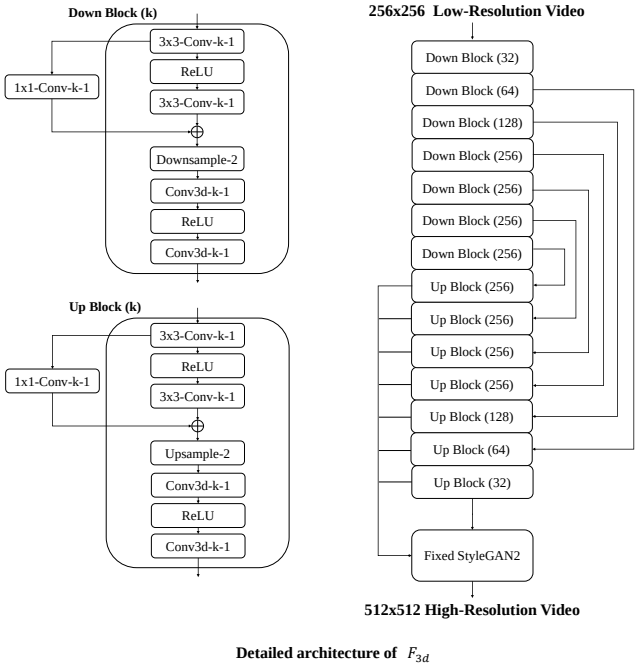

Figure 3. Detailed architecture of our temporal super-resolution network.

Figure 4. Qualitative comparison of warping quality of sparse landmarks and dense landmark encoding.

2.2. Ablation of Temporal Super-resolution

In Figure 6, we select a column of the generated frame and visualize its temporal change. The bicubic-upsampled video lack texture of hair. The naive 2D face restoration baseline [16] generates more flickering artifacts. Our results have clear and stable temporal motion, which is close to

Figure 5. Qualitative comparison of identity-preserving architecture.

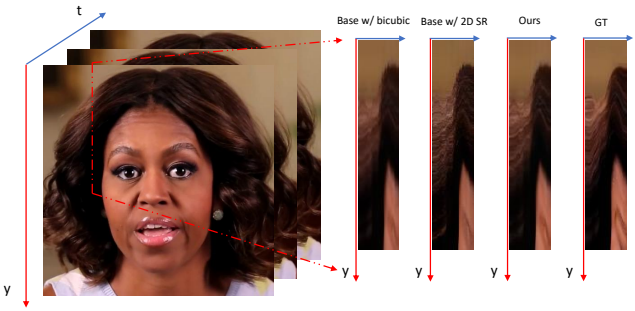

Figure 6. Comparison of temporal profile. We select a column and observe its changes with time index. The result of our temporal super-resolution is more stable and consistent without flickering noise.

ground truth. In Table 1, we provide an additional comparison with StyleHEAT [20] at 512×512 resolution and evaluate baselines using FVD [15], which is a popular metrics in video generation. Our method achieves the lowest FVD, which demonstrates our high temporal fidelity. Since StyleHEAT generates unrealistic over-smooth and flickering images, its FID, LPIPS, and FVD are much worse than ours. Note that L1 loss E_{warp} is biased towards over-smoothed results. Thus, the E_{warp} of StyleHEAT and bicubic upsampled video can be lower than ours.

3. Additional Comparison Results

Methods	FID↓	LPIPS↓	FVD↓	E_{warp} ↓
Ground Truth	-	-	-	**0.0182**
StyleHEAT [20]	44.5207	0.2840	572.363	0.0207
Base w/ bicubic	25.5762	0.2285	248.413	0.0184
Base w/ GFPGAN [16]	22.6351	0.2178	172.754	0.0242
Ours	**21.4974**	**0.2079**	**162.685**	0.0213

Table 1. Quantitative evaluation of our temporal super-resolution on self-reconstruction at 512×512 resolution.

Methods	Quality↓	Identity↓	Motion↓
FOMM [12]	3.48	3.25	2.99
PIRender [10]	3.12	2.92	3.28
DaGAN [5]	3.48	3.55	3.01
DAM [13]	3.65	3.58	3.02
ROME [7]	2.77	2.87	3.14
StyleHEAT [20]	2.99	3.26	3.71
Ours	**1.51**	**1.57**	**1.84**

Table 2. Average ranking score of user study. User prefer ours the best in both three aspects.

3.1. Additional Qualitative Comparison with Recent Methods on a Larger Scale Test Set

Main results. We perform our evaluation including recent methods [5, 7, 13] on a larger test set, which contains 20 test videos from HDTF [21] and VFHQ [18] following the setting of StyleHEAT [20]. For the same-id case, we evaluate using 500 frames of each video with 10k frames in total, while for the cross-id case, we use 1000 source images from CelebA-HQ [6] as source images and use 100 frames of each video to drive 50 source images with 100k frames in total. The results are shown in Table 3, in which our method achieves the best scores in almost all the metrics. Moreover, the FVD score of our full model improves significantly compared with the base model, which illustrates the effectiveness of the proposed temporal super-resolution network. Note that the compared methods target the one-shot setting and inevitably exhibit artifacts. In contrast, we are the first to study a personalized model which is of practical significance, and the proposed fast personalization is orthogonal to prior techniques and can be generally applied.

Number of parameters and runtime (FPS). We also provide the comparison of the model size and throughput between our model and other methods in Table 3.

User study. We also conduct a user study to obtain the user's subject evaluation of different approaches. We present all the results produced by each comparing method to the participants and ask them to rank the score from 1 to 7 (1 is the best, 7 is the worst) on three perspectives independently: the image quality, the identity preservation and the motion drivability. 20 subjects are asked to rank different methods with 15 sets of comparisons in each study. The

Methods	Params (M)	FPS	**Same-ID** 256^2				**Cross-ID** 256^2	
			FID↓	FVD↓	LPIPS↓	ID Loss↓	FID↓	ID Loss↓
FOMM [12]	59.80	51.57	22.7112	136.4454	0.1577	0.0848	37.4306	0.4368
PIRender [10]	22.52	9.72	28.2376	367.3942	0.1881	0.1200	40.4600	0.3600
DaGAN [5]	60.36	29.04	21.0879	**108.5139**	0.1427	0.0912	34.1784	0.4640
DAM [13]	59.75	43.74	23.7192	140.8459	0.1363	0.0832	40.4675	0.4400
ROME [7]	123.85	2.63	119.9319	1204.52	0.5422	0.3376	102.9575	0.5360
Ours	130.28	16.78	**18.1581**	219.6183	**0.1335**	**0.0496**	**25.1646**	**0.1920**
Methods	Params (M)	FPS	**Same-ID** 512^2				**Cross-ID** 512^2	
StyleHEAT [20]	367.70	0.03	41.3364	244.6287	0.2957	0.2560	136.3959	0.4960
Ours	284.97	1.22	**21.1314**	**131.8511**	**0.2150**	**0.0880**	**127.3204**	**0.2544**

Table 3. Evaluation against more baselines on a larger scale test set

Figure 7. Failure case of our framework.

average ranking is shown in Table 2. Our method earns user preferences the best in both three aspects.

3.2. Additional Qualitative Comparison

We also provide additional videos on the webpage to evaluate our results qualitatively. "Ours-Base" in the video denotes our base model in Sec 3.1, while "Ours-Full" denotes our full model with temporal-consistent super-resolution network. Our model is able to provide state-of-the-art generation quality with high temporal fidelity on both self-reconstruction and cross-reenactment tasks. Moreover, the videos of fast personalization illustrate the strong adaptation capability of our meta-learned model. The in-the-wild examples also demonstrate the generalized ability of the proposed model.

4. Limitation

Our one-shot model may not handle occlusions well. As shown in Figure 7, the occluded text in the background appears blurry in the output result. One possible solution is to inpaint the background from pretrained matting and combined it with the generation results using alpha-blending following [4], which we leave for future work.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 1

[2] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniahy, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016. 1

[3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *conference of the international speech communication association*, 2018. 1

[4] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *ACM Multimedia*, 2022. 1, 4

[5] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 2022. 3, 4

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 3

[7] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *ECCV*, 2022. 3, 4

[8] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 1

[9] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In *NeurIPS*, 2020. 1

[10] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, 2021. 3, 4

[11] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *German Conference on Pattern Recognition*, pages 26–36, 2016. 1

[12] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 3, 4

[13] Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *CVPR*, 2022. 3, 4

[14] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 1

[15] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *ICLR*, pages 694–711. Springer, 2019. 3

[16] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 2, 3

[17] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3d face reconstruction with dense landmarks. In *ECCV*, 2022. 1

[18] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 3

[19] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network, 2015. 1

[20] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. In *ECCV*, 2022. 3, 4

[21] Zhimeng Zhang, Lincheng Li, and Yu Ding. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 3