OTAvatar: One-shot Talking Face Avatar with Controllable Tri-plane Rendering

Zhiyuan Ma[1,2*] Xiangyu Zhu[3*] Guojun Qi[4] Zhen Lei[1,2,3†] Lei Zhang[1]
[1]The Hong Kong Polytechnic University
[2]Center for Artificial Intelligence and Robotics, HKISI CAS
[3]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[4]OPPO Research
zm2354.ma@connect.polyu.hk, xiangyu.zhu@nlpr.ia.ac.cn
guojunq@gmail.com, zlei@nlpr.ia.ac.cn, cslzhang@comp.polyu.edu.hk

1. Supplementary Material

1.1. Pesudo Code

The pseudo-code for our suggested inverting-by-decoupling training scheme is appended in Algo. 1.

1.2. Identity code interpolation

In this paper, we use a motion controller C and decoupling-by-inverting strategy to disentangle the latent code of face generator to identity code $\mathbf{w}_{id}$ and motion code $\mathbf{w}_x$. In this section, we examine the disentanglement by performing the task of identity interpolation. The interpolated identity is generated by:

$$\mathbf{w}_{id}^{'} = \alpha \mathbf{w}_{id\_a} + (1-\alpha)\mathbf{w}_{id\_b} \tag{1}$$

where $\mathbf{w}_{id\_a}$ and $\mathbf{w}_{id\_b}$ are the identity codes estimated from two source reference image using decoupling-by-inverting strategy. The interpolated latent code $\mathbf{w}_{id}^{'}$ is then combined with any motion and generate animations. The animation result in shown in Fig. 1. It shows that our model can animate consistent motion with smooth-varying identity attributes. And it validates that our method achieves the disentanglement of motion and identity in the latent space of pre-trained face generator. The animation result is also appended in the supplementary video.

1.3. More Qualitative Comparison

Multi-view dataset. We show more qualitative comparison result on the multi-view stereo dataset Multiface [6] in Fig. 2. To further compare the robustness against the pose variation, we choose an overhead view as the single reference.

Monocular dataset. To evaluate the 3D consistency in monocular talking dataset, we change the pose coefficients to be rotating while keeping the expression coefficients in sync with the driving frames, as shown in Fig. 3.

From both comparisons, we observe PIRenderer [4] and StyleHEAT [8] suffer from drastic unnatural image distortion, while our methods can maintain the multi-view consistency and depict natural motion on the expression; com-

Algorithm 1 Training Scheme of Decoupling-by-Inverting

Input: Talking Face Dataset $\mathcal{D}$,
3D face animator $G(\cdot,\cdot;\Theta) = G_{eg}(\cdot + C(\cdot;\Theta_c);\Theta_{eg})$
where $\Theta = \Theta_{eg} \cup \Theta_c$

$\Theta_c \leftarrow$ random initialization
for $i \leftarrow 1$ **to** T **do**
  /* collect source and target data point */
  $\mathcal{V} \leftarrow$ random video clip sampled from $\mathcal{D}$
  $\mathbf{I}_s, \mathbf{x}_s, \mathbf{p}_s \leftarrow$ data of random frame sampled from $\mathcal{V}$
  $\mathbf{I}_d, \mathbf{x}_d, \mathbf{p}_d \leftarrow$ data of random frame sampled from $\mathcal{V}$
  $\mathbf{w}_{id} \leftarrow \mathbf{w}_{avg}$ // initialize $\mathbf{w}_{id}$ using average
  $\theta_c \leftarrow \Theta_c$ // initialize θ_c using EMA weights
  $\theta \leftarrow \Theta_{eg} \cup \theta_c$ // assemble G with trainable θ_c
  for $n \leftarrow 1$ **to** $N_{id} + N_{mo}$ **do**
    /* calculate optimization objectives */
    $\mathcal{L}_s \leftarrow \mathcal{L}(\mathbf{I}_s, \mathcal{R}(G(\mathbf{w}_{id}, \mathbf{x}_s; \theta), \mathbf{p}_s))$
    $\mathcal{L}_d \leftarrow \mathcal{L}(\mathbf{I}_d, \mathcal{R}(G(\mathbf{w}_{id}, \mathbf{x}_d; \theta), \mathbf{p}_d))$
    if $n < N_{id}$ **then**
      /* optimize identity code */
      Update $\mathbf{w}_{id}$ using $\nabla_{\mathbf{w}_{id}}(\mathcal{L}_s + \mathcal{L}_d)$
    else
      /* train motion calibration */
      Update θ_c using $\nabla_{\theta_c}(\mathcal{L}_s + \mathcal{L}_d)$
    end
  end
  Calculate $\mathcal{L}_s$, $\mathcal{L}_t$ using line.7, line.8
  Finetune Θ_{eg} on $\mathcal{L}_s + \mathcal{L}_t$
  $\Theta_{ca} \leftarrow \beta\Theta_c + (1-\beta)\theta_c$ // Update EMA weights
end

*Equal contribution.
†Corresponding author.

Figure 1. **The interpolation results of identity code**. Our model can generate smooth-varying identity attribute beyond motion control.

pared to the 3D method of HeadNeRF [3], we achieve a more faithful reconstruction of the subject on the skin color and torso. For more detail, please refer to the supplementary video.

	CSIM	AED	APD	AKD	FID
$\mathbf{w}_x \in \mathcal{W}$	**0.719**	3.352	0.453	4.783	104.6
w/o code book	0.662	3.342	0.457	4.974	103.2
Ours	0.694	**2.850**	**0.405**	**4.307**	**101.8**

Table 1. **Ablation study on the controller architecture**. Experiments are conducted on the cross-identity reenactment.

1.4. The Network Architecture of the Controller

We proceed to describe the controller architecture in this section. Fig. 4 shows the details of the motion controller C. We input the window of adjacent 27 frames of expression and pose coefficients, they form up the motion signal of size 73×27. We use three 1D convolution layers, noted as F_T to compress the noises in the sequential 3DMM coefficients. After that, a five-layer MLP is implemented to transform the feature into the magnitudes of 20 orthogonal bases of size 512 in the code book. We calculate 20×14 of such magnitudes, by first outputting 20×15 scalars and then adding 20 of them to the others. Then the 20×14 scalars are multiplied with the orthogonal bases in the code book D and transformed to the motion code of size 14×512 . Here 14 is the maximum number of latent codes $\mathbf{w} \in \mathcal{W}$ as input to the face generator G [1]. The 14×512 latent features form up the $\mathcal{W}^+$ space [7]. In our implementation, it is formulated as the summation of the identity code $\mathbf{w}_{id}$ and the motion code $\mathbf{w}_x$.

we conduct experiments to evaluate the effectiveness of the proposed controller architecture. One ablation model is constructed by reducing the number of output scalars to 1×20. It is multiplied with the code book D to make up the single motion code of size 1×512, it is further repeated 14 times to fit the required shape of 14×512. This model

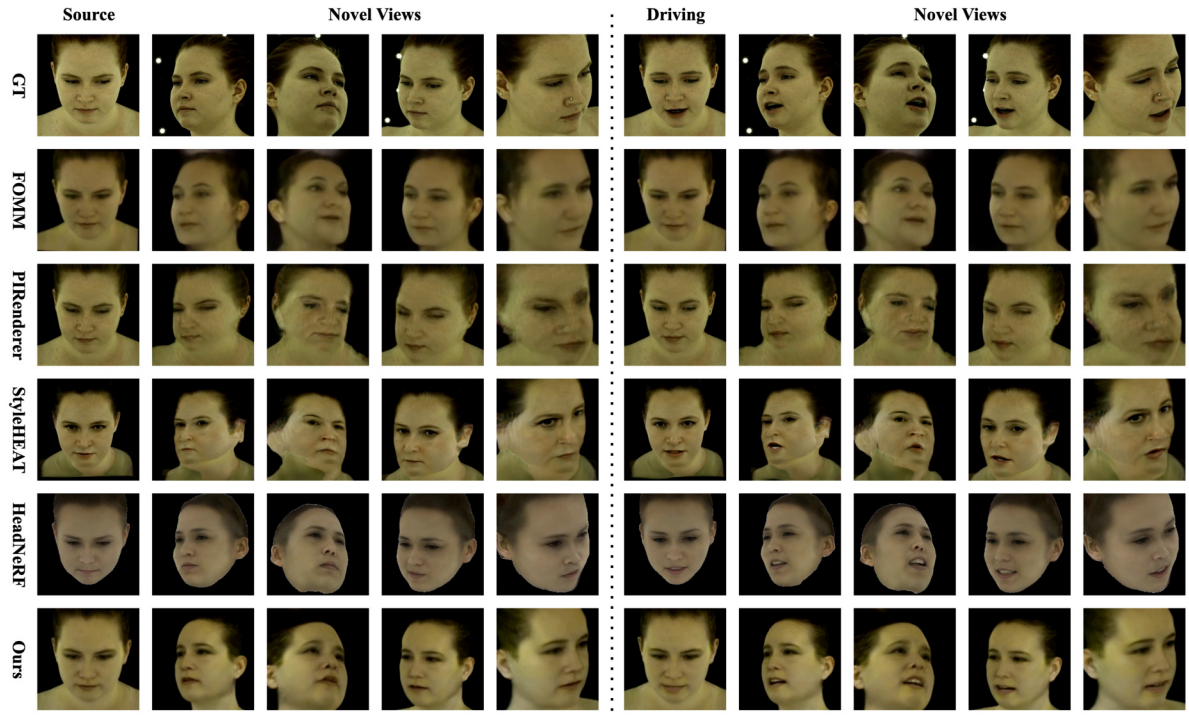

Figure 2. **3D Consistency on multi-view dataset**. We demonstrate additional visualization comparison on the Multiface dataset [6], with more drastic camera view variation. All methods use the source frame to extract the identity feature, then extract 3DMM coefficients of pose and expression from the driving frame to generate the talking face. This subject is not included in the training set of any methods.

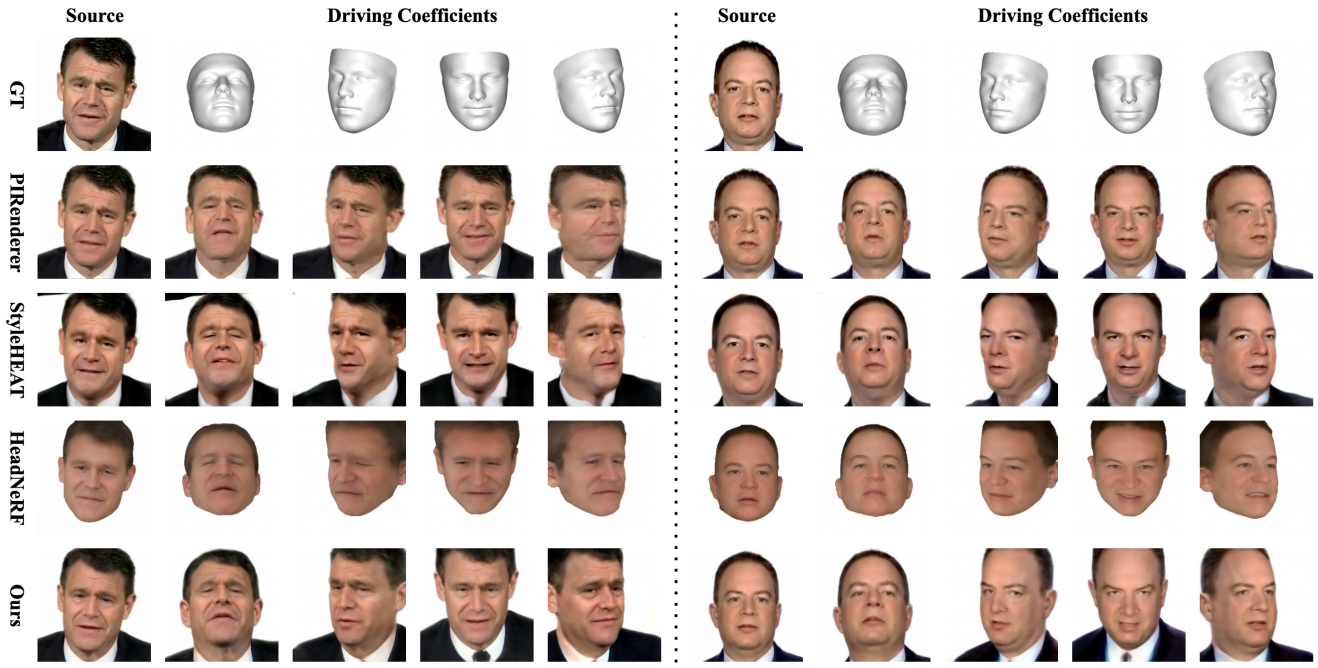

Figure 3. **3D Consistency on monocular dataset**. We demonstrate additional visualization comparison on the HDTF dataset [9], with more drastic camera view variation. All methods use the source frame to extract the identity feature, then use the coefficients of pose and expression as visualized by the face meshes to generate talking faces. This subject is not included in the training set of any methods.

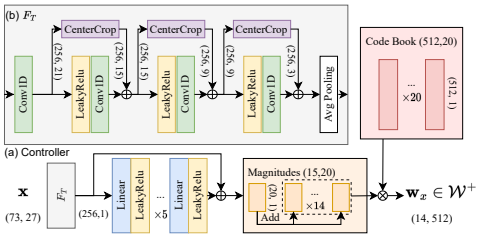

Figure 4. **Controller architecture**. We show the details of the architecture of C in (a) and the sub-module F_T in (b). The controller takes as input 3DMM coefficients and output motion code in $\mathcal{W}^+$ space [7].

	Motion ↑	Identity ↑
FOMM [5]	4.1	5.3
PIRender [4]	3.5	4.3
StyleHEAT [8]	7.4	5.7
HeadNeRF [3]	4.6	3.8
Ours	**9.0**	**8.5**

Table 2. **User study on the animation quaility**.

is written as $\mathbf{w}_x \in \mathcal{W}$ in Table. 1. In another ablation model, we replace the code book with vanilla 20×512 linear weights which can also transform every 20 scalars to 512-dimensional latent code. From Table. 1, we observe the performance deterioration on both ablation models on the motion controllability, which is evaluated by AED, APD and AKD, and also on the image quality as quantified by FID. Even though representing motion code in $\mathcal{W}$ facilitates higher identity consistency as measured by CSIM, it neglects the fact that each of the 14 latent codes contributes differently to the motion deformation, as is testified in 2D GANs [2], therefore the motion controllability is reduced.

1.5. User Study

To assess the quality of the animation, we conduct a user study. We sent the results animated by our method and baselines to 20 people (the students in the university). Users are asked to evaluate the animation based on 1) motion that is in sync with driving videos and, and 2) identity similarity compared with the source portrait. Their ratings are averaged, scaled to a maximum of 10, and shown in Table. 2. Our method is the most desired in motion controllability and identity consistency.

References

[1] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022.

[2] Min Jin Chong, Hsin-Ying Lee, and David Forsyth. Stylegan of all trades: Image manipulation with only pretrained stylegan. *arXiv preprint arXiv:2111.01619*, 2021.

[3] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022.

[4] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021.

[5] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022.

[7] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[8] 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. *arXiv preprint arXiv:2203.04036*, 2022.

[9] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.