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

Bowen Zhang\textsuperscript{1*}, Chenyang Qi\textsuperscript{2*}, Pan Zhang\textsuperscript{1}, Bo Zhang\textsuperscript{3†}, HsiangTao Wu\textsuperscript{3}
Dong Chen\textsuperscript{2}, Qifeng Chen\textsuperscript{2†}, Yong Wang\textsuperscript{1}, Fang Wen\textsuperscript{3}
\textsuperscript{1}USTC, \textsuperscript{2}HKUST, \textsuperscript{3}Microsoft

Figure 1. Our method yields identity-preserving talking head generation. See the webpage for video demos.

Abstract

In this work, we propose an ID-preserving talking head generation framework, which advances previous methods in two aspects. First, as opposed to interpolating from sparse flow, we claim that dense landmarks are crucial to achieving accurate geometry-aware flow fields. Second, inspired by face-swapping methods, we adaptively fuse the source identity during synthesis, so that the network better preserves the key characteristics of the image portrait. Although the proposed model surpasses prior generation fidelity on established benchmarks, personalized fine-tuning is still needed to further make the talking head generation qualified for real usage. However, this process is rather computationally demanding that is unaffordable to standard users. To alleviate this, we propose a fast adaptation model using a meta-learning approach. The learned model can be adapted to a high-quality personalized model as fast as 30 seconds. Last but not least, a spatial-temporal enhancement module is proposed to improve the fine details while ensuring temporal coherency. Extensive experiments prove the significant superiority of our approach over the state of the arts in both one-shot and personalized settings.

1. Introduction

Talking head generation\textsuperscript{[2,8,9,28,31,35,38,47,49,56]} has found extensive applications in face-to-face live chat, virtual reality and virtual avatars in games and videos. In this paper, we aim to synthesize a realistic talking head with a single source image (one-shot) that provides the appearance of a given person while being animatable according to the motion of the driving person. Recently, considerable progress has been made with neural rendering techniques, bypassing the sophisticated 3D human modeling process and expensive driving sensors. While these works attain increasing fidelity and higher rendering resolution, identity preserving remains a challenging issue since the human vision system is particularly sensitive to any nuanced deviation from the person’s facial geometry.

Prior arts mainly focus on learning a geometry-aware warping field, either by interpolating from sparse 2D/3D landmarks or leveraging 3D face prior, e.g., 3D morphable
face model (3DMM) [3,4]. However, fine-grained facial geometry may not be well described by a set of sparse landmarks or inaccurate face reconstruction. Indeed, the warping field, trained in a self-supervised manner rather than using accurate flow ground truth, can only model coarse geometry deformation, lacking the expressivity that captures the subtle semantic characteristics of the portrait.

In this paper, we propose to better preserve the portrait identity in two ways. First, we claim that dense facial landmarks are sufficient for an accurate warping field prediction without the need for local affine transformation. Specifically, we adopt a landmark prediction model [51] trained on synthetic data [50], yielding 669 head landmarks that offer significantly richer information on facial geometry. In addition, we build upon the face-swapping approach [27] and propose to enhance the perceptual identity by attentionally fusing the identity feature of the source portrait while retaining the pose and expression of the intermediate warping. Equipped with these two improvements, our one-shot model demonstrates a significant advantage over prior works in terms of both image quality and perceptual identity preservation when animating in-the-wild portraits.

While our one-shot talking head model has achieved state-of-the-art quality, it is still infeasible to guarantee satisfactory synthesis results because such a one-shot setting is inherently ill-posed—one may never hallucinate the person-specific facial shape and occluded content from a single photo. Hence, ultimately we encounter the uncanny valley [37] that a user becomes uncomfortable as the synthesis results approach to realism. To circumvent this, one workaround is to finetune the model using several minutes of personal data. Such personalized training has been widely adopted in industry to ensure product-level quality, while requiring less amount of personal data.

Specifically, we optimize the model for specific personal identity, expression, and pose, which provides an easy tool to edit and render portrait images [12, 13, 23, 24, 40, 52]. Some attempts [23, 24] render the animated faces using predefined facial landmarks [14] is also a straightforward approach [15, 43, 57, 62] to represent the motion of driving images. For example, PFLD [14] uses 98 facial landmarks. However, they could not generate an accurate warping flow since the landmarks are not dense enough.

2. Related Work

2D-based talking head synthesis. Methods along this line [28, 32, 38, 39, 47] predict explicit warping flow by interpolating the sparse flow defined by 2D landmarks. FOMM [38] assumes local affine transformation for flow interpolation. Recently, Mallya et al. [28] computes landmarks from multiple source images using an attention mechanism. However, the landmarks learned in an unsupervised manner are too sparse (e.g., 20 in FaceVid2Vid [47]) to be interpolated into dense flows. Using predefined facial landmarks [14] is also a straightforward approach [15, 43, 57, 62] to represent the motion of driving images. Recent works [8, 35] predict a warping flow field for talking head synthesis using 3DMM. Although 3DMM-based methods allow explicit face control, using these coefficients to represent detailed face geometry and expression remains challenging.

Towards higher-resolution talking head. The video resolution of most talking head methods [2, 35, 38, 49] is $256 \times 256$, bounded by the available video datasets [6, 29]. To enhance the visual quality of output videos, previous literature trains an additional single-image super-resolution network [9, 48, 54] or utilizes pretrained 2D StyleGAN [48, 56]. However, these methods upsample the original video in a frame-by-frame manner and ignore the temporal consistency of adjacent video frames. In this work, we propose a novel temporal super-resolution module to boost temporal consistency while preserving per-frame quality.
Meta learning and fast adaptation. The goal of meta-learning methods [11, 30, 46, 58] is to achieve good performance with a limited amount of training data. In this paper, we aim to quickly build a personalized model given a new identity, which correlates with the goal of meta-learning. Zakharov et al. [58] propose to improve GAN inversion using a hyper-network. We leverage the idea of model-agnostic meta-learning (MAML) [11] to obtain the best initialization that can be easily adapted to different persons.

3. Method

Figure 2 illustrates an overview of our synthesis framework. Given an image $I_s$ of a person which we refer as source image and a sequence of $t$ driving video frames $\{I_d^1, I_d^2, \ldots, I_d^t\}$, we aim to generate an output video $\{I_y^1, I_y^2, \ldots, I_y^t\}$ with the motions derived from the driving video while maintaining the identity of the source image. Section 3.1 introduces our one-shot model (Figure 2(a, b, c)) for identity-preserving talking head generation $I_y$ at $256 \times 256$ resolution. We describe our meta-learning scheme in Section 3.2, which allows fast adaption using a few images. In Section 3.3, we propose a spatial-temporal enhancement network that generally improves the perceptual quality for both the one-shot and personalized models, yielding video frames with $512 \times 512$ resolution, as shown in Figure 2(d).

3.1. ID-preserving One-shot Base Model

In this section, we will introduce our warping network using dense facial landmarks and an identity-aware refinement network for one-shot talking head synthesis.

Warping prediction with dense landmarks. To predict an accurate geometry-aware warping field, we claim that dense landmark prediction [51] is the key to a geometry-aware warping field estimation. While dense facial landmarks are tedious to annotate, our dense landmark prediction is trained on synthetic faces [50], which reliably produces 669 points covering the entire head, including the ears, eyeballs, and teeth, for in-the-wild faces. These dense landmarks capture rich information about the person’s facial geometry and considerably ease the flow field prediction. However, it is non-trivial to fully make use of these dense landmarks. A naive approach is to channel-wise concatenate the landmarks before feeding into the network, as previous works [38, 39, 47]. However, processing such input is computationally demanding due to the inordinate number of input channels. Hence, we propose an efficient way to digest these landmarks. Specifically, We draw the neighboring connected landmark points, with each connection encoded in different colors as shown in Figure 2(a).

One can thus take the landmark images of the source and driving, along with the source image, i.e., $x_{in} = \text{Concat}(I_s, I_s^{ldmk}, I_d^{ldmk})$, for warping field prediction. To ensure a globally coherent prediction, we strengthen the warping capability using the condition of a latent motion code $z_w$ which is derived from the input, i.e., $z_w = E_w(x_{in})$, where $E_w$ is a CNN encoder. The motion code $z_w$ is injected into the flow estimation network $F_w$ through AdaIN [18]. By modulating the mean and variance of the normalized feature map, the network could be effectively...
guided by the motion vector which induces a globally coherent flow prediction. Formally, we obtain the prediction of a dense flow field through \( \mathbf{w} = F_w(\mathbf{x}_w, z_w) \).

**ID-preserving refinement network.** Directly warping the source image with the predicted flow field inevitably introduces artifacts and the loss of subtle perceived identity. Therefore, an ID-preserving refinement network is needed to produce a photo-realistic result while maintaining the identity of the source image. Prior works primarily focus on the geometry-aware flow field prediction, whereas such deformation may not well characterize the fine-grained facial geometry. In this work, we resolve the identity loss via a well-designed identity-preserving refinement network.

We propose to attentionally incorporate the semantic identity vector with the intermediate warping results. Let \( h^l_{in} \) be the \( l \)-th layer feature map of the refinement network. We obtain the identity-aware feature output \( h^l_{id} \) by modulating \( h^l_{in} \) with the identity embedding \( z_{id} \) through AdaIN, where \( z_{id} \) is extracted using a pre-trained face recognition model \( E_{id} \) [7]. Meanwhile, we obtain a motion-aware feature \( h^l_{motion} \) which keeps the head motion and expression of the driving video. Specifically, \( h^l_{motion} \) is obtained via a Feature-wise Linear Modulate (FiLM) [10, 33, 34] according to the warped image \( \tilde{I}_s = w(I_s) \), i.e., \( h^l_{motion} = \text{Conv}(\tilde{I}_s) \times h^l_{id} + \text{Conv}(I_s) \).

With both the identity-aware and motion-aware features, we adaptively fuse the features through an attention-based fusion block. Inspired by recent face-swapping approaches [27], we suppose that the driving motions and source identity should be fused in a spatially-adaptive manner, in which the facial parts that mostly characterize the key facial features express the identity should rely more on the identity-aware feature, whereas other parts (e.g. hair and clothes) should make greater use of the motion-aware feature. A learnable fusion mask \( M^l \) is used for the fusion of these two parts, which is predicted by,

\[
M^l = \sigma(\text{Conv}(h^l_{id})),
\]

where \( \sigma \) indicates the sigmoid activation function. In this way, the model learns to properly inject the identity-aware features into identity-related regions. The output of layer \( l \) can be derived by fusing features according to the mask \( M^l \), which is,

\[
h^l_{out} = M^l \otimes h^l_{motion} + (1 - M^l) \otimes h^l_{id},
\]

where \( \otimes \) denotes the Hadamard product. Through a cascade of such blocks, we obtain the final output image \( \tilde{I}_y \), which well preserves the source identity while accurately following the head motion and expression as the driving person.

**Training objective.** Perceptual loss [19] is computed between the warped source image \( \tilde{I}_s \) and ground-truth driving image \( I_d \) for accurate warping prediction. The same loss is also applied to enforce the refinement output \( \tilde{I}_y \).

\[
\mathcal{L} = \mathcal{L}_w^{VGG} + \lambda_r \mathcal{L}_r^{VGG} + \lambda_{id} \mathcal{L}_{id} + \lambda_{eye} \mathcal{L}_{eye} + \lambda_{mouth} \mathcal{L}_{mouth} + \lambda_{adv} \mathcal{L}_{adv},
\]

where \( \lambda_r, \lambda_{id}, \lambda_{eye}, \lambda_{mouth}, \) and \( \lambda_{adv} \) are the loss weights.

**3.2. Meta-learning based Faster Personalization**

While achieving state-of-the-art generation quality using our one-shot model, there always exists challenging cases that are intractable since the one-shot setting is inherently ill-posed. Person-specific characteristics and occlusion would never be faithfully recovered using a one-shot general model. Thus, personalized fine-tuning is necessary to achieve robust and authentic results that are truly usable. Nonetheless, fine-tuning the one-shot pre-trained model on a long video is computationally prohibitive for common users. To solve this, we propose a meta-learned model whose initialization weights can be easily adapted according to low-shot personal data within a few training steps, as illustrated in Figure 3.

Formally, given a person \( j \), the personalized adaptation starts from the pre-trained model weight \( \phi \) and aims to reach the optimal personal model weight \( \phi^* \) in \( K \) steps.

We also extract the feature using face recognition model \( E_{id} \) [7], and penalize the dissimilarity between the ID vectors of the output image \( \tilde{I}_y \) and source \( I_s \), using

\[
\mathcal{L}_{id} = 1 - \cos(E_{id}(\tilde{I}_y), E_{id}(I_s)).
\]

A multi-scale patch discriminator \( \mathcal{L}_{adv} \) [33] is adopted to enhance the photo-realism of the outputs. To further improve the generation quality on the hard eye and mouth areas, we add additional \( \mathcal{L}_1 \) reconstruction losses, i.e., \( \mathcal{L}_{eye}, \mathcal{L}_{mouth} \), for these parts.

The overall training loss can be formulated as

\[
\mathcal{L} = \mathcal{L}_w^{VGG} + \lambda_r \mathcal{L}_r^{VGG} + \lambda_{id} \mathcal{L}_{id} + \lambda_{eye} \mathcal{L}_{eye} + \lambda_{mouth} \mathcal{L}_{mouth} + \lambda_{adv} \mathcal{L}_{adv},
\]
ing the error against the personal images \( \hat{X}_j \):
\[
\hat{\phi}_j = \min_{\phi_j} \mathcal{L}(G_{\phi_j}(\hat{X}_j)),
\]
(5)

where \( G_{\phi_j} \) denotes the whole generator with weight \( \phi_j \). Usually we perform \( K \) steps of stochastic gradient descent (SGD) from initialization \( \phi \) to approach \( \phi_j \), so the weight updating process can be formulated as:
\[
\phi^k_j = \text{SGD}_K(\phi, \hat{X}_j).
\]
(6)

Our goal is to find an optimal initialization \( \hat{\phi}^K \) which could approach any personal model after \( K \) steps of SGD update, even for a small \( K \), i.e.,
\[
\hat{\phi}^K = \min_{\phi} \sum_{j=1}^{M} \| \phi_j - \phi^k_j \|.
\]
(7)

Indeed, the general one-shot pretrained model \( \phi \) is a special case in the above formulation, which essentially learns the model weight in Equation 7 when \( K = 0 \). When we are allowed to perform a few adaption steps \( K > 0 \), there is a gap between \( \phi \) and desired \( \phi^K \), since the optimization target of the standard pre-training is to minimize the overall error across all training data, it does not necessarily find the best weight suitable for personalization.

Compared with general one-shot models, we leverage the idea of Model-Agnostic Meta-Learning (MAML) to bridge this gap and enable surprisingly fast personalized training. The goal of MAML-based methods is to optimize the initialized weights such that they could be fast adapted to a new identity within a few steps of gradient descent, which directly matches our goal. Directly optimizing the initialization weight using Equation (7) involves the computation of second-order derivatives, which is computationally expensive on large-scale training. Therefore, we utilize Reptile [30], a first-order MAML-based approach to obtain suitable initialization for fast personalization.

To be more specific, our meta-learning model explicitly considers the personalized adaptation during training. For each person \( j \), we start from the \( \phi_0^j = \phi \), which is the initialization to be optimized. Formally, we sample a batch of personal training data \( \hat{X}_j \), the \( K \) steps of personalized training yield the finetuned model weights as:
\[
\phi^k_j = \text{SGD} \left( \phi^{k-1}_j, \hat{X}_j \right), \quad k = 1, \cdots, K.
\]
(8)

Finally, the personal update, i.e., the difference of \( \phi^K_j \) and \( \phi_0^j \), is used as the gradient to update our initialization \( \phi \):
\[
\phi \leftarrow \phi - \beta \left( \phi^K_j - \phi \right),
\]
(9)

where \( \beta \) is the meta-learning rate. The full algorithm is shown in Algorithm 1. Our model progressively learns a more suitable initialization through meta-training as visualized in Figure 3, which could fast adapt to personalized models after limited steps of adaptation.

### 3.3. Temporal-consistent Super-resolution Network

To further enhance the generation resolution and improve the high-fidelity details of our output, a video super-resolution network is needed as the final stage of the generation framework. Previous talking head synthesis works [9, 56] utilize single-frame super-resolution as the last stage and ignore the quality of temporal consistency and stability. Performing super-resolution in frame-by-frame manner tends to produce texture flickering which severely hampers the visual quality. In this work, we consider multiple adjacent frames to ensure temporal coherency.

Inspired by previous 2D face restoration works [48, 55], the pre-trained generative models [20, 21, 59] like StyleGAN contain rich face prior and could significantly help to enhance the high-frequency details. Moreover, the disentangled \( W \) space in StyleGAN provides desirable temporal consistency during manipulation [44], which also benefits our framework. Therefore, we propose a temporally consistent super-resolution module by leveraging pretrained StyleGAN and 3D convolution, where the latter brings quality enhancement in spatio-temporal domain. As shown in Fig. 2, we feed the concatenated sequence of \( t \) video frames \( \{\tilde{I}_{y}^1, \tilde{I}_{y}^2, \cdots, \tilde{I}_{y}^t\} \) into a U-Net composed of 3D convolution with reflection padding on the temporal dimension. To ensure pre-trained per-frame quality while improving temporal consistency, we initialize the 3D convolution weight in U-Net with a pretrained 2D face restoration network [48].

These spatio-temporally enhanced features from the U-Net decoder further modulate the pretrained StyleGAN features through FiLM. Thus, the super-resolution frames \( \{I_{y}^1, I_{y}^2, \cdots, I_{y}^t\} \) are obtained as:
\[
\{I_{y}^1, I_{y}^2, \cdots, I_{y}^t\} = F_{3D}(F_{3D}(\{\tilde{I}_{y}^1, \tilde{I}_{y}^2, \cdots, \tilde{I}_{y}^t\})).
\]
(10)

During training, we optimize the output \( I_y' \) towards the 512 \times 512 ground truth \( I_d \) using \( \ell_1 \) and perceptual loss.
4. Experiments

4.1. Experiment Setup

**Dataset.** Following [9], we train our warping and refinement networks on cropped VoxCeleb2 dataset [6] at 256² resolution. We randomly select 500 videos from the test set for evaluation. For our meta-learning-based fast personalization, we finetune our base model on HDTF dataset [61], which is composed of 410 videos from 300 different identities. We downsample cropped faces to 256² resolution and split the original HDTF dataset into 400 training videos and 10 test videos. After the convergence of our meta-training, we further evaluate the personalization speed of our model on the HDTF test set. Our temporal super-resolution module $F_{3d}$ is trained on the HDTF dataset [61] which has 300 frames per video with 512² resolution. We feed downsampled 256² frames into fixed warping and refinement network and use the outputs of the refinement network as inputs for the next temporal super-resolution module. More training details are provided in the supplementary.

**Metrics** We evaluate the fidelity of self-reconstruction using FID [16] and LPIPS [60]. Our motion transfer qual-
Methods | Self Reconstruction (256 × 256) | Cross Reenactment (256 × 256)
--- | --- | ---
X2Face [49] | 45.2908 | 0.6806 | 0.9632 | 0.2147 | 0.1007 | 91.1485 | 0.6496 | 0.3112 | 0.1210
Bi-layer [57] | 100.9196 | 0.5881 | 0.5280 | 0.1258 | 0.0139 | 127.7823 | 0.6336 | 0.2330 | 0.0208
FOMM [38] | 12.1979 | 0.2338 | 0.2096 | 0.0964 | 0.0100 | 80.1637 | 0.5760 | 0.2340 | 0.0239
PIRender [35] | 14.4065 | 0.2639 | 0.3024 | 0.1080 | 0.0162 | 78.8430 | 0.5440 | 0.2330 | 0.0214
Ours | **11.9528** | **0.2262** | **0.1296** | **0.0942** | **0.0124** | **77.5048** | **0.2944** | **0.2524** | **0.0258**

Methods | Self Reconstruction (512 × 512) | Cross Reenactment (512 × 512)
--- | --- | ---
StyleHEAT [56] | 44.5207 | 0.2840 | 0.4112 | 0.1155 | 0.0131 | 111.3450 | 0.4720 | 0.2505 | 0.0218
Ours | **21.4974** | **0.2079** | **0.0832** | **0.0904** | **0.0121** | **49.6020** | **0.1952** | **0.2737** | **0.0242**

Table 1. Quantitative results for self-reconstruction and cross-reenactment. We evaluate both 256 × 256 results and 512 × 512 results. Our method outperforms all baselines on both resolutions across image fidelity metrics with comparable motion transfer results.

Figure 6. A comparison of different personalized models at the same epochs. Our personalized base model without reptile adaptation reconstructs more accurate skin color and source identity than the personalized FOMM method. For the model personalized from a reptile learning stage, finer detail on teeth and eye colors could be further generated.
sharp and accurate teeth in our experiment, and the identity of the output image is quite different from the source portrait due to the limitations of GAN inversion. In contrast, our refinement network is identity-aware and we leverage pretrained StyleGAN in our temporal super-resolution module to fully exploit its face prior knowledge. Figure 5 shows that our method produces sharp teeth and hairs, while bicubic-upsampled results are blurry with artifacts.

Table 1 demonstrates that our method achieves the best quantitative fidelity and comparable motion transfer quality on both self-reconstruction and cross-identity reenactment.

### 4.3. Evaluation of Fast Personalization

Although our base model achieves state-of-the-art quality as a general model, there still exists ill-posedness in some cases. For example, it is very difficult to synthesize teeth according to a closed source mouth. Thus, industry products typically conduct personalization using fine-tuning strategy, which can be computationally expensive. To achieve faster convergence, we finetune our base model using a meta-learning strategy to provide a better weight initialization for the following personalization. In Fig 7, we evaluate the personalization speed of our meta-learned model against our base model and previous baseline FOMM [38]. It takes our meta-learned model 0.5 epoch to decrease LPIPS to 0.14, which is 3× speedup against our base model, and 4× against FOMM. Figure 6 compares the personalization of our method and FOMM [38] at the same epoch, which illustrates our fast adaptation speed on ambiguous areas (e.g., teeth, eyes, and wrinkle details).

### 4.4. Evaluation of Temporal Super-resolution

In Table 4, we also evaluate the performance of our temporal super-resolution using 2D image fidelity and warping error $E_{warp}$. We train a 2D super-resolution baseline using GFPGAN [48]. The quantitative result in Table 4 shows that although naive 2D super-resolution improves per-frame fidelity, it also brings more flickering and larger warping error (0.0242) than simple bicubic upsampling (0.0184). To achieve temporally coherent results, we combine a U-Net composed of 3D convolution with facial prior [21], which significantly reduces $E_{warp}$ of our final videos from 0.0242 to 0.0213, and preserves compelling 2D facial details.

### 4.5. Ablation Study of Base Model

<table>
<thead>
<tr>
<th>Methods</th>
<th>FID↓</th>
<th>LPIPS↓</th>
<th>ID Loss↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours Sparse Landmark</td>
<td>14.3190</td>
<td>0.2485</td>
<td>0.1424</td>
</tr>
<tr>
<td>Ours w/o ID</td>
<td>12.2736</td>
<td>0.2256</td>
<td>0.2144</td>
</tr>
<tr>
<td>Ours</td>
<td>11.9528</td>
<td>0.2262</td>
<td>0.1296</td>
</tr>
</tbody>
</table>

Table 3. Quantitative ablation study of landmarks and ID coefficients in our base model.

We conduct ablation studies to validate the effectiveness of our driving motion and source identity representation in our base model. If we replace our 669 dense landmarks with sparse landmarks, the LPIPS of warped source images degrades by 0.2. To evaluate our identity-aware refinement, the removal of the identity input causes significant increase of identity loss from 0.1296 to 0.2144.

### 5. Conclusion

We present a novel framework for identity-preserving one-shot talking head generation. To faithfully maintain the source ID, we propose to leverage accurate dense landmarks in the warping network and explicit source identity during refinement. Further, we significantly advance the applicability of personalized model by reducing its training to 30 seconds with meta-learning. Last but not least, we enhance the final resolution and temporal consistency with 3D convolution and generative prior. Comprehensive experiments demonstrate the state-of-the-art performance of our system.

### References

[54] Lingbo Yang, Chang Liu, Pan Wang, Shanshe Wang, Peiran Ren, Siwei Ma, and Wen Gao. HiFaceGAN: Face renovation via collaborative suppression and replenishment. ACM Multimedia, 2020. 2
[56] Fei Yin, Yong Zhang, Xiaodong Cun, Yanbo Fan, Xuan Wang, Qinyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. In ECCV, 2022. 1, 2, 5, 7, 12, 13
[61] Zhiheng Zhang, Lincheng Li, and Yu Ding. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In CVPR, 2021. 6
Appendix A. Implementation Details

Dataset. Following previous work [9], we train our warping and refinement networks on cropped VoxCeleb2 dataset [6], 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 × Tesla V100 GPUs. We set hyperparameters of losses as: λt = 10, λid = 20, λeye = 50, λmouth = 50 and λadv = 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 β = 2 × 10−5. In each iteration, we train an inner loop for K = 24 iteration with inner loop learning rate α = 2 × 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 [25, 26], we evaluate our temporal consistency using warping error EWarp. For each frame O_t, we calculate the warping error with previous frame O_t−1 as:

\[ E_{\text{pair}}(t) = \sum_{i=1}^{N} M_t(i) \| y_t(i) - W(y_{t-1}(i)) \|_1, \]

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

where M_t is the occlusion map [36] 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 [42]. The averages warping error EWarp(\{t\}_{t=1}^{T}) 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 [5] based on the dense facial landmarks [51], 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 8 and Figure 9. “3 × 3-Conv-k-1” indicates a convolutional layer with kernel size of 3, channel dimensions of k and stride of 1. “ReLU, ReLU” indicates ReLU function respectively. Figure 10 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 dimensions and stride of 1.

Appendix B. Additional Ablation

B.1. Visualization of ID, Landmark Ablation

Figure 13 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 14, 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.

B.2. Ablation of Temporal Super-Resolution

In Figure 11, 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
HEAT generates unrealistic over-smooth and flickering images, which demonstrates our high temporal fidelity. Since StyleHEAT [56] at the same-id case, we evaluate using 500 frames of each video to drive 50 source images, its FID, LPIPS, and FVD are much worse than ours. Note that L1 loss $E_{\text{warp}}$ is biased towards over-smoothed results. Thus, the $E_{\text{warp}}$ of StyleHEAT and bicubic upsampled video can be lower than ours.

### Appendix C. Additional Comparison Results

#### C.1. Additional Qualitative Comparison with Recent Methods on a Larger Scale Test Set

**Main results.** We perform our evaluation including recent methods [17, 22, 41] on a larger test set, which contains 20 test videos following the setting of StyleHEAT [56]. 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 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 5, 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...
### Table 5. Evaluation against more baselines on a larger scale test set

<table>
<thead>
<tr>
<th>Methods</th>
<th>Params</th>
<th>FPS</th>
<th>Same-ID $256^2$</th>
<th>Cross-ID $256^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>FID↓</td>
<td>FVD↓</td>
</tr>
<tr>
<td>FOMM [38]</td>
<td>59.80</td>
<td>51.57</td>
<td>22.7112</td>
<td>136.4454</td>
</tr>
<tr>
<td>PRender [35]</td>
<td>22.52</td>
<td>9.72</td>
<td>28.2376</td>
<td>367.3942</td>
</tr>
<tr>
<td>DaGAN [17]</td>
<td>60.36</td>
<td>29.04</td>
<td>21.0879</td>
<td><strong>108.5139</strong></td>
</tr>
<tr>
<td>DAM [41]</td>
<td>59.75</td>
<td>43.74</td>
<td>23.7192</td>
<td>140.8459</td>
</tr>
<tr>
<td>ROME [22]</td>
<td>123.85</td>
<td>2.63</td>
<td>119.9319</td>
<td>1204.52</td>
</tr>
<tr>
<td>Ours</td>
<td>130.28</td>
<td>16.78</td>
<td><strong>18.1581</strong></td>
<td>219.6183</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Params</th>
<th>FPS</th>
<th>Same-ID $512^2$</th>
<th>Cross-ID $512^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>FID↓</td>
<td>FVD↓</td>
</tr>
<tr>
<td>StyleHEAT [56]</td>
<td>367.70</td>
<td>0.03</td>
<td>41.3364</td>
<td>244.6287</td>
</tr>
<tr>
<td>Ours</td>
<td>284.97</td>
<td>1.22</td>
<td><strong>21.1314</strong></td>
<td><strong>131.8511</strong></td>
</tr>
</tbody>
</table>

Table 6. Average ranking score of user study. User prefer ours the best in both three aspects.

Appendix D. Limitation

Our one-shot model may not handle occlusions well. As shown in Figure 12, 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 [9], which we leave for future work.
Figure 13. Qualitative comparison of warping quality of sparse landmarks and dense landmark encoding.

Figure 14. Qualitative comparison of identity-preserving architecture.