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

Bowen Zhang[1*] Chenyang Qi[2*] Pan Zhang[1] Bo Zhang[3†] HsiangTao Wu[3]
Dong Chen[2] Qifeng Chen[2†] Yong Wang[1] Fang Wen[3]
[1]USTC [2]HKUST [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 [1, 6, 7, 24, 27, 31, 33, 39, 41, 47] 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

*Equal contribution, interns at Microsoft Research.
†Joint corresponding authors.

face model (3DMM) [2, 3]. 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 [43] trained on synthetic data [42], yielding 669 head landmarks that offer significantly richer information on facial geometry. In addition, we build upon the face-swapping approach [23] 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* [32] 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 a personal video. Such personalized training has been widely adopted in industry to ensure product-level quality, yet this process is computationally expensive, which greatly limits its use scenarios. Thus, speeding up this *personalized training*, a task previously under-explored, is of great significance to the application of talking head synthesis.

We propose to achieve fast personalization with meta-learning. The key idea is to find an initialization model that can be easily adapted to a given identity with limited training iterations. To this end, we resort to a meta-learning approach [9, 26] that finds success in quickly learning discriminative tasks, yet is rarely explored in generative tasks. Specifically, we optimize the model for specific personal data with a few iterations. In this way, we get a slightly fine-tuned personal model towards which we move the initialization model weight a little bit. Such meta-learned initialization allows us to train a personal model within 30 seconds, which is 3 times faster than a vanilla pretrained model while requiring less amount of personal data.

Moreover, we propose a novel temporal super-resolution network to enhance the resolution of the generated talking head video. To do this, we leverage the generative prior to boost the high-frequency details for portraits and meanwhile take into account adjacent frames that are helpful to reduce temporal flickering. Finally, we reach temporally coherent video results of 512×512 resolution with compelling facial details. In summary, this work innovates in the following aspects:

- We propose a carefully designed framework to significantly improve the identity-preserving capability when animating a one-shot in-the-wild portrait.
- To the best of our knowledge, we are the first to explore meta-learning to accelerate personalized training, thus obtaining ultra-high-quality results at affordable cost.
- Our novel video super-resolution model effectively enhances details without introducing temporal flickering.

2. Related Work

2D-based talking head synthesis. Methods along this line [24, 28, 33, 34, 39] predict explicit warping flow by interpolating the sparse flow defined by 2D landmarks. FOMM [33] assumes local affine transformation for flow interpolation. Recently, Mallya *et al*. [24] 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 Face-Vid2Vid [39]) to be interpolated into dense flows. Using predefined facial landmarks [12] is also a straightforward approach [13, 36, 48, 53] to represent the motion of driving images. For example, PFLD [12] uses 98 facial landmarks. However, they could not generate an accurate warping flow since the landmarks are not dense enough.

Talking head synthesis with 3D face prior. 3D Morphable Models [2, 3] represent a face image as PCA coefficients relating to identity, expression, and pose, which provides an easy tool to edit and render portrait images [10, 11, 19, 20, 35, 44]. Some attempts [19, 20] render the animated faces by combining the identity coefficients of the source image and the motion coefficients of the driving video. Recent works [6, 31] 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 [1, 31, 33, 41] is 256×256, bounded by the available video datasets [4, 25]. To enhance the visual quality of output videos, previous literature trains an additional single-image super-resolution network [7, 40, 45] or utilizes pretrained 2D StyleGAN [40, 47]. 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.

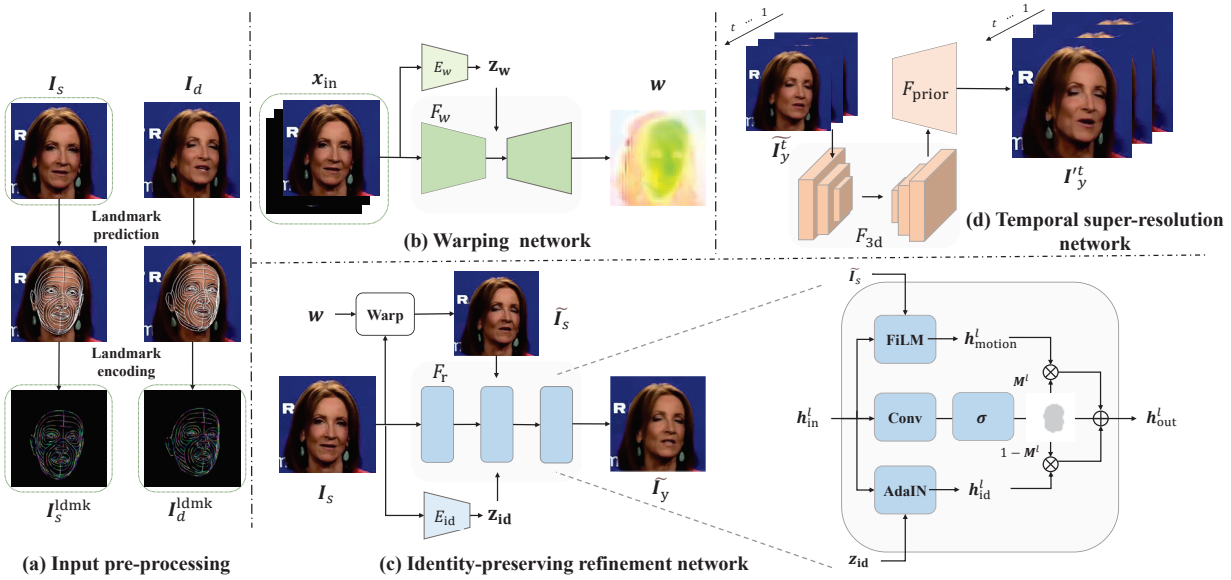

Figure 2. Overview of our one-shot framework. (a) Given a source image $\boldsymbol{I}_s$ and a driving video $\{\boldsymbol{I}_d^1, \boldsymbol{I}_d^2, \cdots, \boldsymbol{I}_d^t\}$, we first extract their dense landmarks $(\boldsymbol{I}_s^{\text{ldmk}}, \boldsymbol{I}_{\text{d}}^{\text{ldmk}})$ using a pretrained landmark detector. (b) Then, we estimate warping flows $\boldsymbol{w}$ between the source image and each driving frame according to concatenated input $\boldsymbol{x}_{\text{in}}$. (c) We further refine the warped source input $\tilde{\boldsymbol{I}}_s$ using an ID-preserving network. (d) Finally, we enhance and upsample the 256×256 results $\{\tilde{\boldsymbol{I}}_y^1, \tilde{\boldsymbol{I}}_y^2, \cdots, \tilde{\boldsymbol{I}}_y^t\}$ to high-fidelity output $\{\boldsymbol{I'}_y^1, \boldsymbol{I'}_y^2, \cdots, \boldsymbol{I'}_y^t\}$ in 512×512.

Meta learning and fast adaptation. The goal of meta-learning methods [9, 26, 38, 49] 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*. [49] propose to improve GAN inversion using a hyper-network. We leverage the idea of model-agnostic meta-learning (MAML) [9] 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 $\boldsymbol{I}_s$ of a person which we refer as source image and a sequence of t driving video frames $\{\boldsymbol{I}_d^1, \boldsymbol{I}_d^2, \cdots, \boldsymbol{I}_d^t\}$, we aim to generate an output video $\{\boldsymbol{I'}_y^1, \boldsymbol{I'}_y^2, \cdots, \boldsymbol{I'}_y^t\}$ 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 $\tilde{\boldsymbol{I}}_y$ at 256×256 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 512×512 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 [43] 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 [42], 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 [33, 34, 39]. 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*., $\boldsymbol{x}_{\text{in}} = \text{Concat}(\boldsymbol{I}_s, \boldsymbol{I}_s^{\text{ldmk}}, \boldsymbol{I}_d^{\text{ldmk}})$, for warping field prediction. To ensure a globally coherent prediction, we strengthen the warping capability using the condition of a latent motion code $\boldsymbol{z}_w$ which is derived from the input, *i.e*., $\boldsymbol{z}_w = E_w(\boldsymbol{x}_{\text{in}})$, where E_w is a CNN encoder. The motion code $\boldsymbol{z}_w$ is injected into the flow estimation network F_w through AdaIN [15]. 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 $\boldsymbol{w} = F_w(\boldsymbol{x}_{\text{in}}, \boldsymbol{z}_w)$.

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 $\boldsymbol{h}^l_{\text{in}}$ be the l-th layer feature map of the refinement network. We obtain the identity-aware feature output $\boldsymbol{h}^l_{\text{id}}$ by modulating $\boldsymbol{h}^l_{\text{in}}$ with the identity embedding $\boldsymbol{z}_{\text{id}}$ through AdaIN, where $\boldsymbol{z}_{\text{id}}$ is extracted using a pre-trained face recognition model E_{id} [5]. Meanwhile, we obtain a motion-aware feature $\boldsymbol{h}^l_{\text{motion}}$, which keeps the head motion and expression of the driving video. Specifically, $\boldsymbol{h}^l_{\text{motion}}$ is obtained via a Feature-wise Linear Modulate (FiLM) [8, 29, 30] according to the warped image $\tilde{\boldsymbol{I}}_s = \boldsymbol{w}(\boldsymbol{I}_s)$, *i.e.*, $\boldsymbol{h}^l_{\text{motion}} = \text{Conv}(\tilde{\boldsymbol{I}}_s) \times \boldsymbol{h}^l_{\text{id}} + \text{Conv}(\tilde{\boldsymbol{I}}_s)$.

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 [23], 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 $\boldsymbol{M}^l$ is used for the fusion of these two parts, which is predicted by,

$$\boldsymbol{M}^l = \sigma(\text{Conv}(\boldsymbol{h}^l_{\text{in}})), \tag{1}$$

where σ 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 $\boldsymbol{M}^l$, which is,

$$\boldsymbol{h}^l_{\text{out}} = \boldsymbol{M}^l \otimes \boldsymbol{h}^l_{\text{motion}} + (1 - \boldsymbol{M}^l) \otimes \boldsymbol{h}^l_{\text{id}}, \tag{2}$$

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

Training objective. Perceptual loss [16] is computed between the warped source image $\tilde{\boldsymbol{I}}_s$ and ground-truth driving image $\boldsymbol{I}_d$ for accurate warping prediction. The same loss is also applied to enforce the refinement output $\tilde{\boldsymbol{I}}_y$.

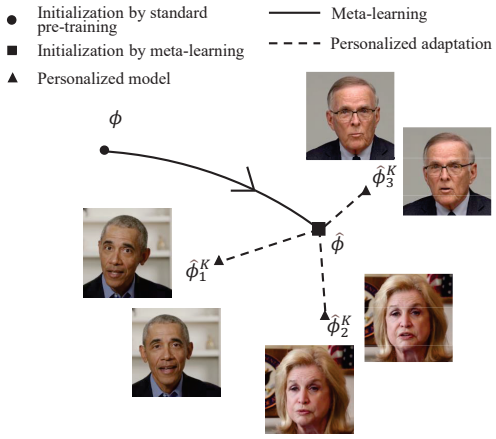

Figure 3. Visualization of meta-learning. Compared with the one-shot pre-trained model ϕ, the meta-learned model $\hat{\phi}$ can be rapidly adapted to a unique personal model $\hat{\phi}_j^K$ in K steps.

We also extract the feature using face recognition model E_{id} [5], and penalize the dissimilarity between the ID vectors of the output image $\tilde{\boldsymbol{I}}_y$ and source $\boldsymbol{I}_s$, using

$$\mathcal{L}_{\text{id}} = 1 - \cos\left(E_{\text{id}}(\tilde{\boldsymbol{I}}_y), E_{\text{id}}(\boldsymbol{I}_s)\right). \tag{3}$$

A multi-scale patch discriminator $\mathcal{L}_{\text{adv}}$ [29] 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.*, $\mathcal{L}_{\text{eye}}, \mathcal{L}_{\text{mouth}}$, for these parts.

The overall training loss can be formulated as

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{w}}^{\text{VGG}} + \lambda_{\text{r}}\mathcal{L}_{\text{r}}^{\text{VGG}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \\ & \lambda_{\text{eye}}\mathcal{L}_{\text{eye}} + \lambda_{\text{mouth}}\mathcal{L}_{\text{mouth}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}},\end{aligned} \tag{4}$$

where $\lambda_{\text{r}}, \lambda_{\text{id}}, \lambda_{\text{eye}}, \lambda_{\text{mouth}}$, and λ_{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 ϕ and aims to reach the optimal personal model weight $\hat{\phi}_j$ by minimiz-

ing the error against the personal images $\hat{\boldsymbol{X}}_j$:

$$\hat{\boldsymbol{\phi}}_j = \min_{\boldsymbol{\phi}_j} \mathcal{L}(G_{\boldsymbol{\phi}_j}(\hat{\boldsymbol{X}}_j)), \tag{5}$$

where $G_{\boldsymbol{\phi}_j}$ denotes the whole generator with weight ϕ_j. Usually we perform K steps of stochastic gradient descent (SGD) from initialization $\boldsymbol{\phi}$ to approach $\hat{\boldsymbol{\phi}}_j$, so the weight updating process can be formulated as:

$$\boldsymbol{\phi}_j^k = \text{SGD}_K(\boldsymbol{\phi}, \hat{\boldsymbol{X}}_j). \tag{6}$$

Our goal is to find an optimal initialization $\hat{\boldsymbol{\phi}}^K$ which could approach any personal model after K steps of SGD update, even for a small K, *i.e.*,

$$\hat{\boldsymbol{\phi}}^K = \min_{\boldsymbol{\phi}} \sum_{j=1}^{M} \|\hat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j^k\|. \tag{7}$$

Indeed, the general one-shot pretrained model $\boldsymbol{\phi}$ 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 $\boldsymbol{\phi}$ and desired $\hat{\boldsymbol{\phi}}^K$, 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 [26], 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 $\boldsymbol{\phi}_j^0 = \boldsymbol{\phi}$, which is the initialization to be optimized. Formally, we sample a batch of personal training data $\hat{\boldsymbol{X}}_j$, the K steps of personalized training yield the finetuned model weights as:

$$\boldsymbol{\phi}_j^k = \text{SGD}\left(\boldsymbol{\phi}_j^{k-1}, \hat{\boldsymbol{X}}_j\right), \ k = 1, \cdots, K. \tag{8}$$

Finally, the personal update, *i.e.*, the difference of $\boldsymbol{\phi}_j^K$ and $\boldsymbol{\phi}_j^0$, is used as the gradient to update our initialization $\boldsymbol{\phi}$:

$$\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \beta\left(\boldsymbol{\phi}_j^K - \boldsymbol{\phi}\right), \tag{9}$$

where β 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.

Algorithm 1 Optimization of Initial Weights with Reptile

Input: weights $\boldsymbol{\phi}$ of generation network G, inner loop learning rate α, meta-learning rate β, number of training iterations N, number of training persons M, number of inner loop iterations K

for $i = 1, \cdots, N$ **do**
  for $j = 1, \cdots, M$ **do**
    $\boldsymbol{\phi}_j^0 = \textbf{Clone}(\boldsymbol{\phi})$
    Sample a training batch $\hat{\boldsymbol{X}}_j$ of person j
    for $k = 1, \cdots, K$ **do**
      $\boldsymbol{\phi}_j^k = \boldsymbol{\phi}_j^{k-1} - \alpha \nabla_{\boldsymbol{\phi}} \mathcal{L}(G_{\boldsymbol{\phi}}(\hat{\boldsymbol{X}}_j))$
    end for
  end for
  $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \frac{\beta}{M} \sum_{j=1}^{M} (\boldsymbol{\phi}_j^K - \boldsymbol{\phi})$
end for

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 [7,47] 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 [40, 46], the pre-trained generative models [17, 18, 50] like StyleGAN contain rich face prior and could significantly help to enhance the high-frequency details. Moreover, the disentangled $\mathcal{W}$ space in StyleGAN provides desirable temporal consistency during manipulation [37], 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 $\{\tilde{\boldsymbol{I}}_y^1, \tilde{\boldsymbol{I}}_y^2, ..., \tilde{\boldsymbol{I}}_y^t\}$ 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 [40].

These spatio-temporally enhanced features from the U-Net decoder further modulate the pretrained StyleGAN features through FiLM. Thus, the super-resolution frames $\{\boldsymbol{I'}_y^1, \boldsymbol{I'}_y^2, ..., \boldsymbol{I'}_y^t\}$ are obtained as:

$$\{\boldsymbol{I'}_y^1, \boldsymbol{I'}_y^2, ..., \boldsymbol{I'}_y^t\} = F_{\text{StyleGAN}}(F_{\text{3D}}(\{\tilde{\boldsymbol{I}}_y^1, \tilde{\boldsymbol{I}}_y^2, ..., \tilde{\boldsymbol{I}}_y^t\})). \tag{10}$$

During training, we optimize the output $\boldsymbol{I'}_y$ towards the 512×512 ground truth $\boldsymbol{I}_d$ using ℓ_1 and perceptual loss.

Figure 4. Qualitative results of our self-reconstruction (top row) and cross-identity reenactment (bottom two rows) at 256×256 resolution. Our method synthesizes more faithful expression and motion, while better preserving the identity of the source portrait.

Figure 5. Qualitative results of self reconstruction and cross-identity reenactment at 512×512 resolution. Our spatial-temporal super-resolution module further enhances more high-frequency details on teeth, eyes, and hair.

4. Experiments

4.1. Experiment Setup

Dataset. Following [7], we train our warping and refinement networks on cropped VoxCeleb2 dataset [4] at 256^2 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 [52], which is composed of 410 videos from 300 different identities. We downsample cropped faces to 256^2 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 [52] which has 300 frames per video with 512^2 resolution. We feed downsampled 256^2 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 [14] and LPIPS [51]. Our motion transfer qual-

Methods	**Self Reconstruction** (256×256)					**Cross Reenactment** (256×256)			
	FID↓	LPIPS↓	ID Loss↓	AED↓	APD↓	FID↓	ID Loss↓	AED↓	APD↓
X2Face [41]	45.2908	0.6806	0.9632	0.2147	0.1007	91.1485	0.6496	0.3112	0.1210
Bi-layer [48]	100.9196	0.5881	0.5280	0.1258	0.0139	127.7823	0.6336	0.2330	**0.0208**
FOMM [33]	12.1979	0.2338	0.2096	0.0964	**0.0100**	80.1637	0.5760	0.2340	0.0239
PIRender [31]	14.4065	0.2639	0.3024	0.1080	0.0162	78.8430	0.5440	**0.2113**	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 [47]	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.

Source | Driving | FOMM | Ours w/o Reptile | Ours w/ Reptile

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.

ity is measured using average expression distance (AED) and average pose distance (APD) with driving videos. For our meta-learning-based fast personalization, we illustrate the LPIPS at different personalization epochs for each fine-tuning approach. Following previous works [21, 22], we evaluate our temporal consistency using warping error E_{warp}. For each frame $\tilde{\boldsymbol{I}}_y^t$, we calculate the warping error with the previous frame $\tilde{\boldsymbol{I}}_y^{t-1}$ warped by the estimated optical flow in the occlusion map.

4.2. Comparison with State-of-the-art Methods

We compare our warping and refinement models at 256×256 resolution against several state-of-the-art face reenactment works: X2Face [41], Bi-Layer [1], First-Order Motion Model [33], and PIRender [31]. The top row of Figure 4 presents our qualitative results of self-reconstruction. Different from sparse landmarks in unsupervised learning [33] or 1D 3DMM [31], our dense landmarks provide strong guidance to accurate and fine-detailed synthesis on gaze, mouth, and expressions. The last two rows show the results of our cross-identity reenactment. Since our landmarks have a better decomposition of identity and motion and our refinement network is identity-aware, our method is the only one that well preserves the identity of source image. In contrast, previous methods suffer from appearance data leakage directly from the driver and generate faces with a similar identity to the driving image. Note that the 3DMM coefficients of AED and APD evaluation are not fully decomposed. Thus, other baselines with identity leakage may have better quantitative metrics. We compare our full framework with the proposed temporal super-resolution module at 512×512 resolution against StyleHEAT [47], which is the only open-sourced method that generates high-resolution talking heads. StyleHEAT [47] fails to synthesize

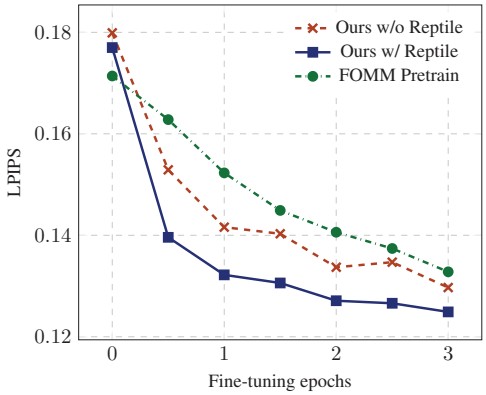

Figure 7. Comparison of test LPIPS at fine-tuning epochs for different approaches. Our reptile-based model achieves more than 3× speedup compared with the base model and previous baseline.

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 [33]. 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 [33] 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 2, 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 [40]. The quantitative result in Table 2 shows

Methods	FID↓	LPIPS↓	E_{warp} ↓
Ground Truth	-	-	**0.0182**
Base w/ bicubic	25.5762	0.2285	0.0184
Base w/ GFPGAN [40]	22.6351	0.2178	0.0242
Ours	**21.4974**	**0.2079**	0.0213

Table 2. Quantitative evaluation of our temporal super-resolution on self-reconstruction at 512 × 512 resolution.

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 [18], 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

Methods	FID↓	LPIPS↓	ID Loss↓
Ours Sparse Landmark	14.3190	0.2485	0.1424
Ours w/o ID	12.2736	**0.2256**	0.2144
Ours	**11.9528**	0.2262	**0.1296**

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

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1, 2, 7

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. *international conference on computer graphics and interactive techniques*, 1999. 2

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

[4] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *conference of the international speech communication association*, 2018. 2, 6

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 4

[6] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *ICCV*, 2021. 1, 2

[7] 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, 2, 5, 6

[8] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 2018. https://distill.pub/2018/feature-wise-transformations. 4

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 2, 3

[10] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 2019. 2

[11] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)*, 2018. 2

[12] Xiaojie Guo, Siyuan Li, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. Pfld: A practical facial landmark detector. *arXiv: Computer Vision and Pattern Recognition*, 2019. 2

[13] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, 2020. 2

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 5

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 5, 8

[19] Hyeongwoo Kim, Mohamed Elgharib, Hans-Peter Zollöfer, Michael Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 2019. 2

[20] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *Association for Computing Machinery*, 2018. 2

[21] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 7

[22] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In *NeurIPS*, 2020. 7

[23] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv: Computer Vision and Pattern Recognition*, 2019. 2, 4

[24] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit Warping for Animation with Image Sets. In *NeurIPS*, 2022. 1, 2

[25] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. *conference of the international speech communication association*, 2017. 2

[26] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv: Learning*, 2018. 2, 3, 5

[27] Hao Ouyang, Bo Zhang, Pan Zhang, Hao Yang, Jiaolong Yang, Dong Chen, Qifeng Chen, and Fang Wen. Real-time neural character rendering with pose-guided multiplane images. *ECCV*, 2022. 1

[28] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. *arXiv preprint arXiv:2301.06281*, 2023. 2

[29] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 4

[30] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 4

[31] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, 2021. 1, 2, 7

[32] Mincheol Shin, Se Jung Kim, and Frank Biocca. The uncanny valley: No need for any further judgments when an avatar looks eerie. *Computers in Human Behavior*, 94:100–109, 2019. 2

[33] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 1, 2, 3, 7, 8

[34] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019. 2, 3

[35] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. Explicitly controllable 3d-aware portrait generation. *arXiv preprint arXiv:2209.05434*, 2022. 2

[36] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Facegan: Facial attribute controllable reenactment gan. In *CVPR*, 2021. 2

[37] Rotem Tzaban, Ron Mokady, Rinon Gal, Amit H. Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos, 2022. 5

[38] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *NeurIPS*, 2021. 3

[39] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2020. 1, 2, 3

[40] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 2, 5, 8

[41] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018. 1, 2, 7

[42] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *ICCV*, 2021. 2, 3

[43] 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. 2, 3

[44] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. *arXiv preprint arXiv:2301.02379*, 2023. 2

[45] 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

[46] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 2021. 5

[47] 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. 1, 2, 5, 7

[48] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020. 2, 7

[49] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 3

[50] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *CVPR*, 2022. 5

[51] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[52] Zhimeng Zhang, Lincheng Li, and Yu Ding. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 6

[53] Ruiqi Zhao, Tianyi Wu, and Guodong Guo. Sparse to dense motion transfer for face image animation. In *ICCV*, 2021. 2