Enhancing Identity-Deformation Disentanglement in StyleGAN for One-Shot Face Video Re-Enactment

Qing Chang, Yao-Xiang Ding*, Kun Zhou

State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China
12321115@zju.edu.cn, dingyx.gm@gmail.com, kunzhou@acm.org

Abstract

The task of one-shot face video re-enactment aims at generating target video of faces with the same identity of one source frame and facial deformation of the driving video. To achieve high quality generation, it is essential to precisely disentangle identity-related and identity-independent characteristics, meanwhile build expressive features keeping high-frequency facial details, which still remain unaddressed for existing approaches. To deal with these two challenges, we propose a two-stage generation model based on StyleGAN, whose key novel techniques lie in better disentangling identity and deformation codes in the latent space through an identity-based modeling and manipulating intermediate StyleGAN features at the second stage for augmenting facial details of the generating targets. To further improve identity consistency, a data augmentation method is introduced during training for enhancing the key features affecting identity such as hair and wrinkles. Extensive experimental results demonstrate the superiority of our approach compared to state-of-the-art methods.

Code — https://nickchang97.github.io/ViVFace.github.io/

Introduction

Face-related generation is essential in many applications, spanning from movie production, digital avatars, and games. Among them, one-shot face video re-enactment studies the following task: given a source face frame and a driving video for another face, the task studies generating a video, which maintains identity of the source frame and the facial deformations from the video. This requires to develop a strong generation model which accurately integrates source frame identity and driving video deformations with limited information, thus has been recognized as a significant challenge in facial generation.

The generation process of the re-enactment task is typically divided into two steps: transfer and rendering. The transfer step involves extracting deformation information from the driving video and integrating them with the identity features from the source frames. The rendering step focuses on generating target video frames based on these features, which is commonly done by using generative models like StyleGAN2 (Karras et al. 2020), Stable Diffusion (Rombach et al. 2022), and neural rendering (Mildenhall et al. 2021). It is evident that the transfer step is crucial to generation quality.

Among common deformation modeling approaches, the key points (Siarohin et al. 2019; Mallya, Wang, and Liu 2022; Tao et al. 2022) and 3D morphable face models (3DMM) (Feng et al. 2021) usually have limited expressive power, which can result in the generated faces lacking accurate expression and high-fidelity details. The NeRF-based model (Xu et al. 2023) usually requires training a specific model for each individual face with numerous images from different viewpoints. Although some methods (Tran et al. 2024; Deng et al. 2024; Deng, Wang, and Wang 2024) utilize large-scale datasets to realize a one-shot generalization, these models need a very long training time to converge and output resolution is limited. In comparison, models based on latent space (Oorloff and Yacoob 2023; Xu et al. 2023), which represent deformation with code embeddings in the learned latent space, can avoid these limitations. As an example, StyleGAN2 (Karras et al. 2020), which is a latent space model, has shown impressive performance in generating high-resolution and highly realistic human faces under many face-related generation tasks. Furthermore, its latent space exhibits good semantic properties and strong editing capabilities. This makes the latent space model a desirable choice for face re-enactment.

However, to fully exploit the potential of latent space models, there are two significant challenges to address. The first is to precisely disentangle identity-related and identity-independent characteristics. Identity is a relatively abstract concept, which is represented by complex and distributed visual characteristics of a face. This makes the extraction of the identity features a significantly challenging task. The second challenge is to build expressive features keeping high-frequency facial details, which strongly affects the generation quality. Although rapid development has been achieved for latent-based re-enactment approaches (Wu, Lischinski, and Shechtman 2021; Oorloff and Yacoob 2023), these two challenges remain unaddressed.

In this work, we propose , which builds the latent-based face re-enactment generation model based on latent space of StyleGAN2 (Karras et al. 2020) to deal with the above

*Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

challenges. Previous work (Abdal, Qin, and Wonka 2019; Tov et al. 2021; Wu, Lischinski, and Shechtman 2021) typically retrieve inverted latent codes in the $W+$ space,SS space or a combination of both. Because impressive results in (Oorloff and Yacoob 2023), we also adopt a hybrid space, specially getting identity latent in $W+$, and facial deformation (expression and pose) latent in SS space. Building on this, we further introduce a novel entanglement formulation by adopting conditional modeling. Concretely, we use an encoder to extract generic deformation latent codes of driving frame for transferring and further utilize it to generate more accurate deformation latent codes conditioned on the source frame for rendering.

After building a more expressive re-enactment framework, we discovered that it was hard to reconstruct the original frame with fine details in $W+$ space, like hair and wrinkles. Inspired by previous GAN inversion work (Wang et al. 2022a), we design a refinement network to refine intermediate feature of the StyleGAN generator instead of retraining it. In summary, our work primarily introduces the following contributions: (1) We build a framework based on StyleGAN2 which realize more accurate face re-enactment through identity-based disentanglement. (2) We design a refinement network that enhances the identity consistency of animated frame in the feature space of StyleGAN2, eliminating the necessity of retraining it. (3) We leverage StyleGAN's latent space to achieve data-efficiency image augmentation for refinement network training. (4) We use extensive experiments to validate the effectiveness and superiority of our framework.

Related Work

Latent Space of StyleGAN

A widely explored application for latent space of StyleGAN is its use for the editing of real images. Many efforts have been dedicated to leveraging StyleGAN (Tov et al. 2021) for this task, owing to its highly disentangled latent spaces. Many methods have been proposed for finding semantic latent directions using varying levels of supervision. Supervised (Shen et al. 2020; Jahanian, Chai, and Isola 2019; Goetschalckx et al. 2019) and unsupervised (Voynov and Babenko 2020; Peebles et al. 2020; Shen and Zhou 2021) approaches were proposed to edit semantics such as facial attributes, colors and basic visual transformations (e.g., smile and zooming) in generated or inverted real images (Zhu et al. 2020; Abdal, Qin, and Wonka 2020).

Face Video Re-enactment

Except for X-potrait (Xie et al. 2024) and Megactor (Yang et al. 2024) utilize paired images to realize one-step transfer, related works of face video re-enactment mainly contain two procedures, disentanglement and rendering. Others can generally be divided into two categories (warping-based and latent-based) based on how they transfer facial deformation.

Warping-based methods These approaches usually adopt supervised or unsupervised landmarks (Siarohin et al. 2019, 2021; Tao et al. 2022; Guo et al. 2024), 3D facial priors (Yin et al. 2022; Ding et al. 2023; Mensah et al. 2023) to guide the generation of warping field. Although 3D facial priors can alleviate the primary issue of unrealistic deformations caused by significant motion in 2D facial priors, they still have certain limitations. They fail to capture fine visual details such as wrinkles, hidden part (e.g., tongue and teeth), and non-rigid deformations. Additionally, they only focus on the inner face region, neglecting the broader context. The use of 2D/3D facial priors also introduces spatial-temporal incoherence due to inconsistencies in landmark parameters and struggles with generalization when faced with varying facial geometries between the driving and source identities. Concurently, LivePortrait (Guo et al. 2024) scales the training data to about 69 million high-quality frames demonstrating remarkable performance in animating images across various styles. However, this method has a limitation: it requires the first frame of the driving video to be a neutral frame, which restricts its use in more application scenarios. Recently, with the flourishing diffusion model (Song, Meng, and Ermon 2020; Rombach et al. 2022), many works (Zeng et al. 2023; Wei, Yang, and Wang 2024; Ding et al. 2023) opt to finetune on them for rendering process. In conclusion, these works (Yin et al. 2022; Ren et al. 2021) typically generate flow field to warp and obtain coarse target frame and further utilize a generator to refine and produce the final target frame.

Latent-based methods Warping-based methods can closely resemble original images in real-world conditions but often introduce unrealistic distortions and struggle with generating unseen facial structures, such as filling in teeth or adjusting eyes when the head rotates. In contrast, latent-based methods can leverage fruitful latent spaces of pre-trained models (Tov et al. 2021; Rombach et al. 2022) to address this issue. Previous works (Oorloff and Yacoob 2023; Bounareli et al. 2023) have employed StyleGAN2 for high-resolution one-shot face video re-enactment. And X-potrait (Xie et al. 2024) and Megactor (Yang et al. 2024) scale up the dataset to realize a one-step face re-enactment framework based on StableDiffusion (Rombach et al. 2022) but cost too many computing resources. In our proposed framework, we utilize latent spaces ($W+$ and SS) of StyleGAN2, leveraging its editability and disentanglement to encode both identity and facial deformation. Compared to most relative SOTA (Oorloff and Yacoob 2023), we adopt a different modeling for disentanglement and use StyleGAN's own editable latent space for image augmentation to train a refinement network which eliminates the need for retraining of StyleGAN during testing.

Framework Design

In this section, we first introduce the preliminaries on which our framework is based. There are three well-explored latent spaces (Z, $W+$ and SS) for StyleGAN2 (Karras et al. 2020). In Z space, a random noise latent vector $z \in Z$ is sampled from a normal distribution. After obtaining z, a multi-layer mapping network transforms it to $w \in W$ and $W \in \mathbb{R}^{512}$. In vanilla generation process, w is transformed by layer-wise affine transformations $A(\cdot)$ to produce ss la-

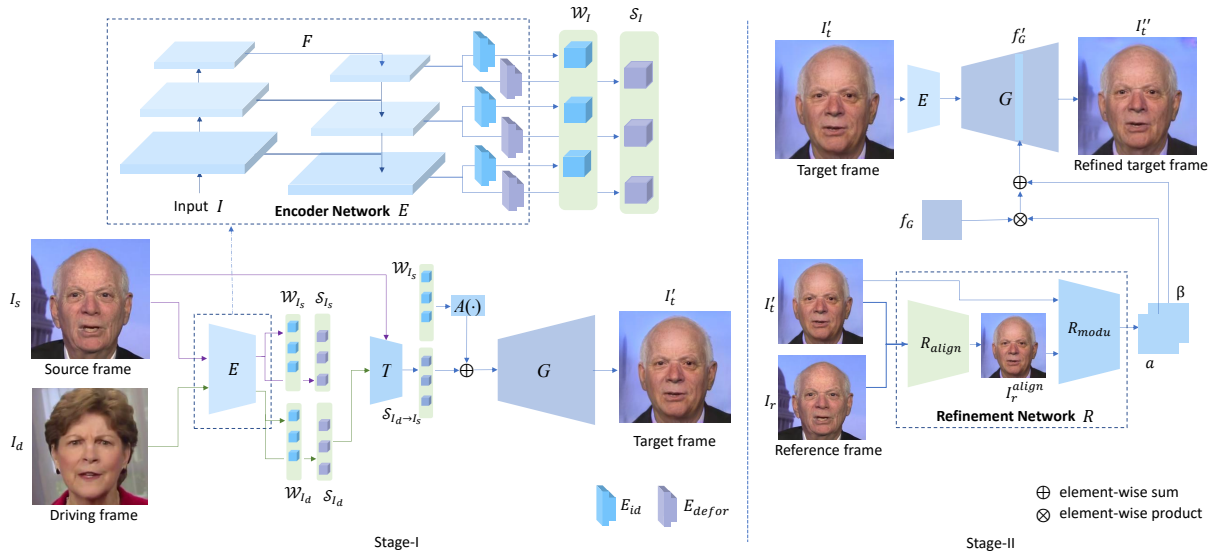

Figure 1: Pipeline of our method. It contains two stages. The stage-I presents disentanglement of the identity ($\mathcal{W}_I$) and facial deformation ($\mathcal{S}_I$) and how to realize face re-enactment with them. Stage-II shows how refinement network works.

tent which belongs to the SS space. Following definition of previous methods (Abdal, Qin, and Wonka 2019; Wu, Lischinski, and Shechtman 2021), an extended space, $W+$, is defined as $W+ \in \mathbb{R}^{18\times512}$, which allows the input to $A(\cdot)$ to vary across different layers.

Next, we introduce the re-enactment pipeline and model design of our approach, as illustrated in Fig. 1. We adopt a frame-wise strategy to address the frames of driving videos. This pipeline includes two stages. In Stage-I, a StyleGAN-based model is used to extract identity and deformation information from the source and driving frames and to generate the preliminary target face frame. The Stage-I model includes GAN-inverse encoding, conditional latent space fusion, and GAN-based decoding modules, in which the conditional latent space fusion module *serves as the key component for improving identity-deformation disentanglement*. In Stage-II, a refinement network is proposed to adjust the target frame, which *serves as the key part to enhance the generation of high-frequency facial details*. In the following, we introduce the details of the models.

Stage-I Model

As shown in Fig. 1 left, the Stage I consists of three parts: encoder network E, transformation network T, and pre-trained StyleGAN2 generator G.

GAN-inverse encoding. Firstly, both the source frame I_s and driving frame I_d are sent into E to obtain their corresponding inverted latent codes W_I and S_I. Specifically, the encoder network E starts with a backbone F, which consists of a ResNet50-SE (He et al. 2016; Hu, Shen, and Sun 2018) model with FPN (Lin et al. 2017) to extract basic information. Then, encoder E_{defor} is used to obtain the generic facial deformation latent code $\mathcal{S}_{I_d}$ in the SS space which is used for *transferring pose and expression*. In correspondence, encoder E_{id} is used for *obtaining identity latent codes* $\mathcal{W}_I$ in the $W+$ space. This GAN inversion process is formulated as

$$\mathcal{W}_I = E_{id}(f_I),\ \mathcal{S}_I = E_{defor}(f_I). \tag{1}$$

The architecture of E_{defor} and E_{id} are the same as map2style network in e4e (Tov et al. 2021).

Conditional latent space fusion and GAN decoding. The conditional latent space fusion module is the key for addressing the identity-deformation disentanglement challenge. Despite previous work (Karras et al. 2020; Tov et al. 2021; Oorloff and Yacoob 2023) has validated meaningful disentanglement for many facial attributes (eg., hair, beard, make-up) in SS space, we observed that the high-level semantic latent representations of facial deformation in SS space *are strongly entangled with the identity*. If we directly use $\mathcal{S}_{I_d}$ as the deformation code without further processing, it may introduce incorrect deformation information when further integrating with identity code of the source frame. We provide visual results to illustrate this phenomenon in Fig. 4. To address this issue in cross-identity re-enactment, we adopt an identity-based conditional disentanglement mechanism to obtain a more accurate deformation latent code for the target frame. Concretely, the generic deformation latent code from the driving frame $\mathcal{S}_{I_d}$ and source frame I_s are sent to a small transformation network $T(\cdot)$ to generate target facial deformation latent code $\mathcal{S}_{I_d\to I_s}$ conditioned on the latent space of source image:

$$\mathcal{S}_{I_d\to I_s} = T(\mathcal{S}_{I_d}, I_s). \tag{2}$$

This transformation network is implemented with four transformer decoder layers (Vaswani et al. 2017). This deformation code $\mathcal{S}_{I_d\to I_s}$ is then fused with the identity code $A(\mathcal{W}_{I_s})$ from the source frame to obtain the final latent code, which is sent to the decoder G for generation:

$$I'_t = G(A(\mathcal{W}_{I_s}) + \mathcal{S}_{I_d\to I_s}). \tag{3}$$

The process is detailedly illustrated in Fig. 1 left.

Stage-II Model

Refinement network. Despite the Stage-I model achieves improved disentanglement between identity and deformation, we find it remains difficult to find a precise inverted code in the $W+$ space that accurately captures the high-fidelity details of the source frame. Motivated by HFGI (Wang et al. 2022a), low-rate latent codes are insufficient for representing high-fidelity details. To address this issue, we propose a refinement network R, as illustrated on the right side of Fig. 1, which enhances missing details by manipulating the intermediate feature map of the generator based on the reference frame I_r. Here, the source frame is taken as reference frame and R captures the details of I_r which are lost in I'_t. If their structures are pixel-aligned, it's more easy for the network to find the corresponding missed information. Hence, we implement R with two subnetworks R_{align} and R_{modu}. I_r and I'_t are concatenated and input into R_{align}, which aims at deforming I_r according to I'_t to generate I_r^{align}. We then concatenate them and feed them into the R_{modu} that is a convolutional network and its outputs α and β are parameters of a gated fusion operator. Specifically, we have $f'_G = \alpha \circ f_G + \beta$ and their shape are the same as f_G where f_G means the intermediate feature map of StyleGAN generator when generating I'_t. After this, we obtain the refined target frame through $I''_t = G(f'_G)$.

Training Process

The training involves two stages: the Stage-I and Stage-II models are trained separately and in sequence.

Data sampling. In each training iteration l, we randomly sample three different frames—I_{l1}, I_{l2}, and I_{l3}—from the dataset, where I_{l1} and I_{l2} are taken from the same video, while I_{l3} is taken from a different video. Only one identity acts in every video. To simplify the subsequent descriptions, we omit the symbol l and instead use I_1, I_2, I_3 to represent the frames sampled in each iteration. Further we use $I'_{n\to k}$ to denote generated target frame when source frame is I_k and driving frame is I_n where $k, n \in \{1, 2, 3\}$.

Stage-I Model Training

In the first stage, encoder network E and transformation network T are trainable. Their weights are optimized with following loss functions.

GAN inversion. The reconstruction loss L_r is introduced:

$$\begin{aligned} L_r &= L_I(I_1, G(A(\mathcal{W}_{I_1}) + \mathcal{S}_{I_1\to I_1})), \\ L_I(I_i, I_j) &= L_{mse}(I_i, I_j) + L_{lpips}(I_i, I_j), \end{aligned} \tag{4}$$

where L_{mse} represents mean square error in the RGB space and L_{lpips} aims to improve perception similarity through intermediate feature map of deep convolution network (Krizhevsky, Sutskever, and Hinton 2012).

Disentanglement. Secondly, we employ a re-enactment loss, which aims at disentangling the latent codes of facial deformation and identity:

$$L_{reen} = L_I(I'_{2\to 1}, I_1). \tag{5}$$

On the other hand, solely using this loss is insufficient to realize well disentanglement. A collapsed solution occurs when part or all of the identity information of face is embedded in the deformation latent code, which can also lead to the loss function being minimized to zero. Hence identity constraint is needed when transferring $\mathcal{S}'_I$ between different identities to avoid this situation. We adopt the identity loss to ensure this:

$$\begin{aligned} L_{id} =& 1 - \cos\big[R(I_1), R(G(A(\mathcal{W}_{I_1}) + 0))\big] \\ &+ 1 - \cos\big[R(I_1), R(I'_{3\to 1})\big], \end{aligned} \tag{6}$$

where R represents the ArcFace (Deng et al. 2019) face recognition network. The first-line term avoids useless identity code and the second-line term helps prevent identity information leakage to deformation latent S_I.

Besides the constraints on image and feature level, we also employ the following two losses

$$L_{I-c} = ||\mathcal{W}_{I_1} - \mathcal{W}_{I'_{3\to 1}}||_2, \; L_{S-c} = ||\mathcal{S}_{I_3} - \mathcal{S}_{I'_{3\to 1}}||_2, \tag{7}$$

targeting at encourage cycle latent consistency for cross-identity transferring.

Latent regularization Similar to previous studies exploiting latent space of styleGAN (Tov et al. 2021; Richardson et al. 2021; Oorloff and Yacoob 2023), we use a latent discriminator with an adversarial loss L_d to encourage the inverted latent $\mathcal{W}_I$ to be not far away from the original W latent space and latent regularization loss L_{reg} to limit the range of $\mathcal{S}_I$. Their details are introduced in the appendix.

The complete loss. The complete Stage-I loss function is

$$\begin{aligned} L_{stage1} = \lambda_r \cdot L_r + \lambda_{reen} \cdot L_{reen} + \lambda_{id} \cdot L_{id} + \\ \lambda_{I-c} \cdot L_{I-c} + \lambda_{S-c} \cdot L_{S-c} + \lambda_d \cdot L_d + \lambda \cdot L_{reg}, \end{aligned} \tag{8}$$

which is the integration of the sub-loss terms introduced above. The detailed setups of the loss weights are introduced in the appendix.

The total loss of second stage L_{stage2} is as follows:

$$L_{stage2} = \lambda_{refine} \cdot L_{refine} + \lambda_{align} \cdot L_{align}. \tag{9}$$

Stage-II Model Training

In the second stage, except for the refine network, all other modules are frozen. In each iteration, I_2 is considered as the reference frame, which is concatenated with I'_1 and sent to R_{align} to generate an aligned image I_r^{align}. We use L_2 loss

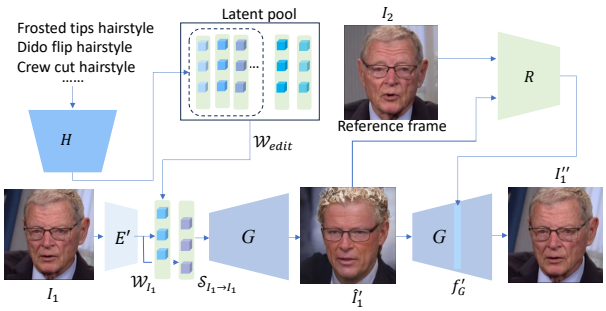

Figure 2: Training augmentation in second stage. To simplify, we combine the E and the T here and represent them together as E'.

Method	Same-identity Re-enactment								Cross-identity Re-enactment			
	L1↓	LPIPS↓	PSNR↑	SSIM↑	CSIM↑	FID ↓	AED ↓	APD↓	CSIM↑	FID↓	AED↓	APD↓
PASL	0.41	0.52	12.01	0.32	0.33	192.4	**1.76**	0.19	0.43	202.7	13.2	0.21
HyperReenact	0.24	0.34	15.72	0.43	0.25	156.8	4.1	**0.06**	0.21	162.5	9.1	**0.07**
Ours	**0.17**	**0.27**	**17.43**	**0.57**	**0.57**	**143.2**	3.5	**0.06**	**0.53**	**149.7**	**8.8**	**0.07**
VOODOO3D	0.22	0.23	20.37	0.62	0.64	94.7	4.7	**0.08**	0.58	101.2	9.6	**0.12**
Potrait4Dv2	0.19	0.20	21.2	0.63	0.70	**73.5**	5.0	0.11	0.66	**92.7**	9.8	0.18
Ours	**0.17**	**0.18**	**21.7**	**0.65**	**0.73**	90.6	**4.2**	**0.08**	**0.69**	118.6	**9.3**	0.13
LIA	0.14	0.15	20.0	0.62	0.80	39.5	5.9	0.19	**0.89**	**84.9**	11.9	0.38
PIRender	0.11	0.12	21.1	0.65	**0.85**	**38.2**	6.2	0.09	0.82	97.6	10.7	0.16
StyleHeat	0.15	0.19	19.36	0.60	0.68	106.9	5.6	0.09	0.67	104.9	9.8	0.16
Robust *	0.11	0.12	21.81	0.66	0.71	45.2	5.1	0.11	0.72	123.7	9.7	0.17
Ours	**0.10**	**0.11**	**22.32**	**0.68**	0.80	42.6	**4.7**	**0.08**	0.79	107.6	**9.1**	**0.15**

Table 1: Quantitative results of the same-identity re-enactment and the cross-identity re-enactment. * means we reproduce it ourselves. Bold indicates the best result. For fair competition with previous works, we adopt three training settings. The results in the first three rows of the table are obtained from models trained on the Celeb (Nagrani, Chung, and Zisserman 2017) dataset. The middle section reports results from models trained on the CelebV-HQ dataset, while the bottom section presents results from models trained on both the CelebV-HQ and HDTF datasets. Additional comparisons with models (Yang et al. 2024; Xie et al. 2024; Guo et al. 2024) trained on very large-scale datasets are included in the appendix.

$L_{align} = ||I_1 - I_r^{align}||_2$ to supervise it. Afterwards, I_2^{align} and I_1' are concatenated and send them into R_{modu} to generate α and β to adjust intermediate feature map of StyleGAN. We adopt the reconstruction loss $L_{refine} = L_I(I_1, I_1'')$ to supervise the refined target frame I_1''.

Latent-Based Augmentation Training. According to observation on the visual results generated in the first stage, hair and wrinkles have the most significant impact on the perception of identity consistency. Therefore, at this stage, we leverage editing ability of StyleGAN's $W+$ space to realize data augmentation that disrupts hair and wrinkles without affecting expressions. It is demonstrated in Fig 2. Specifically, we use HairCLIP (Wei et al. 2022) (denoted as H) to generate various hair-editing latent codes according to text descriptions and out-of-the-box latent code from (Härkönen et al. 2020) for age modification. These editing latent codes form a latent pool. In each training iteration, we randomly sample a latent as $\mathcal{W}_{edit}$ from it. Further, we add it with $\mathcal{W}_{I_1}$ for generating $\hat{I}_1'$. This process is formulated as $\hat{I}_1' = G(A(\mathcal{W}_{I_1} + \mathcal{W}_{edit}) + \mathcal{S}_{I_1 \to I_1})$. With this, we replace I_1' with $\hat{I}_1'$ as the input of R in training time and other procedures remain unchanged. This helps the refinement network focus more on key region. Besides this, since the second stage uses the same dataset as the first stage, most frames do not provide effective supervision, leading to slow convergence. Adopting such data augmentation can also improve the training efficiency in this stage.

Experiments

Experiments Setup

This section first reports the dataset setting of training and evaluation and metrics to evaluate our methods and related works. Next, we present and analysis quantitative and qualitative results. Training details are provided in the appendix.

Datasets. To compare with previous SOTA (Yin et al. 2022; Oorloff and Yacoob 2023), we adopt similar training settings. First, we pre-train the entire network on the CelebV-HQ dataset (Zhu et al. 2022), which consists of 35K diverse videos. After this pre-training stage, we fine-tune our network on HDTF dataset (Zhang et al. 2021). This dataset contains about 300 high-resolution videos with over 300 identities. Thirty videos of them are split as test set and the remaining videos are used as training set. For videos in the training set, we randomly sample 50 frames of each video for training. As for videos in the test set, we choose the first 500 frames of each video for evaluation. Since recent works PASL (Hsu et al. 2024) and HyperReenact (Bounareli et al. 2023) only release inference model and codes, we train our model on Celeb (Nagrani, Chung, and Zisserman 2017) that offers about 100k videos to compare with them.

Metrics. To compare with SOTA methods, we adopt the same metrics in (Yin et al. 2022). Concretely, we use L1 norm pixel loss, Peak Sigal-to-Noise Ratio (PSNR), identity loss (CSIM) which is computed by a face recognition network (Deng et al. 2019), LPIPS (Zhang et al. 2018), SSIM (Wang et al. 2004), FID (Heusel et al. 2017) , Average Expression Distance (AED) and Average Pose Distance (APD) (Yin et al. 2022) for same-identity re-enactment. For cross-identity re-enactment, identity loss, FID, AED and APD are adopted to evaluate results quantitatively.

Quantitative Results

Quantitative results for the same-identity re-enactment and cross-identity re-enactment are reported in Tab 1. Because some compared methods can only generate target frames with a resolution of 256x256 , we compare all of these methods in this resolution. To demonstrate the strength of using StyleGAN as generator, we show high-resolution results in qualitative results and appendix.

For same-identity re-enactment comparisons, our method outperforms previous methods according to the metrics L1,

Figure 3: Visualization of same-identity and cross-identity re-enactment results. A horizontal dashed line is used to separate.

LPIPS, PSNR, SSIM which evaluate image similarity. The CSIM and FID metric usually balance with AED and APD metric. Naturally, no editing means the target frame is identical to source frame completely which means terrible AED and best CSIM and FID. Among models trained on Celeb, PASL (Hsu et al. 2024) shows best AED but the worst CSIM. For models trained on HDTF, LIA (Wang et al. 2022b) and PIRender (Ren et al. 2021) which utilize flow fields to transfer deformations, their results show better FID scores but relative worse expression and pose accuracy. Compared to them, our method also keep the identity well and has a better balance between FID and AED.

For cross-identity re-enactment, which is more meaningful and has broader applications. As shown in Tab 1, results of PASL show a significant performance drop compared to its results under same-identity setting and all metrics are the worst. HyperReenact demonstrate good accuracy in expressions and poses but performs the worst in maintaining identity consistency. In comparison, our method performs the best across all metrics in this setting. Other methods, PIRenderer, LIA and StyleHeat all utilize warping to transfer deformations and their results performs better in FID metric than ours. The results of LIA and PIRenderer demonstrate better identity consistency, but they perform the worst according to the AED and APD metrics. Our method performs well in maintaining identity and achieves the best accuracy in expressions and poses.

Qualitative Results

The results of same-identity re-enactment are reported in upper part of Fig 3 . As shown in the first row, our method produces expressions that are much more similar to the ground truth compared to other methods (Wang et al. 2022b; Yin et al. 2022; Bounareli et al. 2023; Ren et al. 2021). The second and third rows highlight that the wrinkle details of our results are much more consistent as driving frames compared to these comparisons. Our approach generates target frames with more accurate expressions while maintaining identity with minimal distortion.

Next we present visual results of cross-identity re-enactment in bottom part of Fig 3. As shown, results of PIRenderer (Ren et al. 2021) and LIA (Wang et al. 2022b) are in a low-resolution and often yield wrong amplitude of many expressions. HyperReenact (Bounareli et al. 2023) can generate target frames with accurate expressions but terrible identity consistency. Robust (Oorloff and Yacoob 2023)'s results achieve a good identity preservation but fail to show accurate deformations for fine expressions. In contrast, our approach yields better results compared to previous methods. Concretely, from the last three rows, we can observe more accurate lip sync from our results in various expressions without sacrificing identity consistency. Due to our condi-

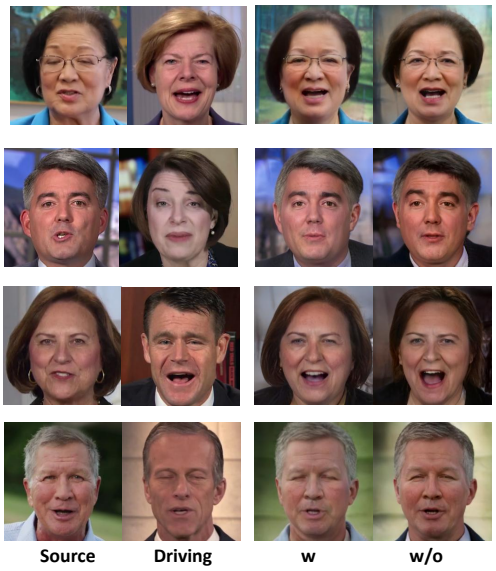

Figure 4: Ablation with identity-based disentanglement. The incorrect parts in target frames are highlighted with red bounding boxes.

tional latent space modeling, our method show more vivid expressions in cross-identity setting. More video comparison results are provided in the appendix. These demonstrate that our approach achieves better visual results compared to previous methods.

Ablation Study

In this section, we design experiments to verify the effectiveness of our designed transformation network, refinement network and model's robustness when different source frames are specified. Unless otherwise specified, the models in this section are trained solely on HDTF (Zhang et al. 2021) with 40K iterations.

T	CSIM↑	AED↓	APD↓
	0.65	10.6	0.19
✓	**0.66** (+1%)	**9.8**(+0.8)	**0.17**(+0.02)

Table 2: Ablation with identity-based disentanglement. This table reports CSIM, AED,APD.

Ablation with transformation network The key difference between our work and previous is the disentanglement of latent space of StyleGAN. Here we validate it quantitatively and qualitatively. We report numeric results in Tab 2. As this table shows, we have 0.8 AED improvement and 0.02 APD improvement without destroying identity consistency. Visual comparisons are presented in Fig 4. we can observe more accurate expression transfer in the first three rows. Without this, closed eyes and mouth in driving frame are not presented in target frames. From the results in the last two rows of this figure, it can be observed that our transformation network also serves to filter out identity information leaked into the deformation latent code.

R	CSIM↑	AED↓	APD↓
	0.66	9.8	0.17
✓	**0.72**(+6%)	**9.7**(+0.1)	0.17

Table 3: Ablation with refinement network. This table reports CSIM, AED, APD.

Ablation with refinement network Here we explore the difference with and without refinement network. We report numeric results in Tab 3. Through the refinement in feature space of StyleGAN, the metric CSIM has relative 9.1% improvement. Corresponding visual results are presented in appendix and it can be seen that refining the feature space of the target frame allows for better preservation of the wrinkles, hairstyle, and skin tone of the source frame.

Robustness with different source frames We further verify the robust performance of these methods. Concretely we use five different source frames of five videos to evaluate the stability of their same-identity re-enactment results. Results are reported in Tab 4 and we can observe the variance of our method and Robust (Oorloff and Yacoob 2023) are the least which are better than other comparisions (Wang et al. 2022b; Yin et al. 2022; Ren et al. 2021; Bounareli et al. 2023). It shows the stability of our method and we can find that methods based on latent space are more robust than warping-based methods when utilizing different source frames.

Method	L1↓	CSIM↑	AED↓	APD↓
LIA	0.16±0.06	0.85±0.10	5.3±1.2	0.16±0.03
HyperReenact	0.27±0.03	0.52±0.06	4.3±0.9	0.07±**0.02**
PIRenderer	0.15±0.06	0.84±0.10	4.8±1.7	0.11±0.04
StyleHeat	0.17±0.05	0.68±0.09	5.3±1.1	0.11±0.04
Robust	0.11±**0.01**	0.72± 0.07	5.1±**0.7**	0.12±0.03
Ours	0.09±**0.01**	0.77±**0.05**	4.6±**0.7**	0.09±**0.02**

Table 4: Ablation with one-shot robustness.

Discussions and Limitations

In this work, we design an end-to-end framework based on StyleGAN2 to generate high-fidelity one-shot facial video re-enactment results at 1024 resolution. our approach incorporates a conditional disentanglement to yield more accurate target frames. Besides this, we further design a refinement network that operates on the intermediate feature space of generator to augment high-frequency details missed in W space. By combining these designs, our framework achieve better results in both numeric results and visual comparisons. However, since our method is based on StyleGAN2, face alignment is necessary. Additionally, modeling backgrounds in the latent space of StyleGAN2 trained on faces remains a significant challenge. For potential negative societal impact, like other DeepFake algorithms (Koujan et al. 2020; Nirkin, Keller, and Hassner 2019; Thies et al. 2016), our method may be used to generate fake results for cheating by malicious users through just one image.

Acknowledgements

This work was supported by National Key R&D Program of China (2022ZD0114804) and National Natural Science Foundation of China (62206245). We thank Linzhou Li for the helpful discussions and the anonymous reviewers for their constructive suggestions that improved this paper.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, 4432–4441.

Abdal, R.; Qin, Y.; and Wonka, P. 2020. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8296–8305.

Bounareli, S.; Tzelepis, C.; Argyriou, V.; Patras, I.; and Tzimiropoulos, G. 2023. Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7149–7159.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Deng, Y.; Wang, D.; Ren, X.; Chen, X.; and Wang, B. 2024. Portrait4D: Learning One-Shot 4D Head Avatar Synthesis using Synthetic Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Deng, Y.; Wang, D.; and Wang, B. 2024. Portrait4D-v2: Pseudo Multi-View Data Creates Better 4D Head Synthesizer. *arXiv preprint arXiv:2403.13570*.

Ding, Z.; Zhang, X.; Xia, Z.; Jebe, L.; Tu, Z.; and Zhang, X. 2023. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4).

Goetschalckx, L.; Andonian, A.; Oliva, A.; and Isola, P. 2019. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*, 5744–5753.

Guo, J.; Zhang, D.; Liu, X.; Zhong, Z.; Zhang, Y.; Wan, P.; and Zhang, D. 2024. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arXiv:2407.03168*.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hsu, G.-S. J.; Zhang, J.-Y.; Hsiang, H. Y.; and Hong, W.-J. 2024. Pose Adapted Shape Learning for Large-Pose Face Reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7413–7422.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jahanian, A.; Chai, L.; and Isola, P. 2019. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Koujan, M. R.; Doukas, M. C.; Roussos, A.; and Zafeiriou, S. 2020. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 16–23. IEEE.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection.

Mallya, A.; Wang, T.-C.; and Liu, M.-Y. 2022. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35: 22438–22450.

Mensah, D.; Kim, N. H.; Aittala, M.; Laine, S.; and Lehtinen, J. 2023. A Hybrid Generator Architecture for Controllable Face Synthesis. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–10.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Oorloff, T.; and Yacoob, Y. 2023. Robust one-shot face video re-enactment using hybrid latent spaces of stylegan2. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20947–20957.

Peebles, W.; Peebles, J.; Zhu, J.-Y.; Efros, A.; and Torralba, A. 2020. The hessian penalty: A weak prior for unsupervised disentanglement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 581–597. Springer.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13759–13768.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2287–2296.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shen, Y.; Gu, J.; Tang, X.; and Zhou, B. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9243–9252.

Shen, Y.; and Zhou, B. 2021. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1532–1540.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in neural information processing systems*.

Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13653–13662.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tao, J.; Wang, B.; Xu, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3637–3646.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an Encoder for StyleGAN Image Manipulation. *arXiv preprint arXiv:2102.02766*.

Tran, P.; Zakharov, E.; Ho, L.-N.; Tran, A. T.; Hu, L.; and Li, H. 2024. VOODOO 3D: Volumetric Portrait Disentanglement for One-Shot 3D Head Reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Voynov, A.; and Babenko, A. 2020. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, 9786–9796. PMLR.

Wang, T.; Zhang, Y.; Fan, Y.; Wang, J.; and Chen, Q. 2022a. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11379–11388.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022b. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wei, H.; Yang, Z.; and Wang, Z. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*.

Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Tan, Z.; Yuan, L.; Zhang, W.; and Yu, N. 2022. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18072–18081.

Wu, Z.; Lischinski, D.; and Shechtman, E. 2021. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12863–12872.

Xie, Y.; Xu, H.; Song, G.; Wang, C.; Shi, Y.; and Luo, L. 2024. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Xu, Y.; Zhang, H.; Wang, L.; Zhao, X.; Huang, H.; Qi, G.; and Liu, Y. 2023. Latentavatar: Learning latent expression code for expressive neural head avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–10.

Yang, S.; Li, H.; Wu, J.; Jing, M.; Li, L.; Ji, R.; Liang, J.; and Fan, H. 2024. MegActor: Harness the Power of Raw Video for Vivid Portrait Animation. arXiv:2405.20851.

Yin, F.; Zhang, Y.; Cun, X.; Cao, M.; Fan, Y.; Wang, X.; Bai, Q.; Wu, B.; Wang, J.; and Yang, Y. 2022. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, 85–101. Springer.

Zeng, B.; Liu, X.; Gao, S.; Liu, B.; Li, H.; Liu, J.; and Zhang, B. 2023. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 628–637.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. CelebV-HQ: A large-scale video facial attributes dataset. In *European conference on computer vision*, 650–667. Springer.

Zhu, J.; Shen, Y.; Zhao, D.; and Zhou, B. 2020. In-domain gan inversion for real image editing. In *European conference on computer vision*, 592–608. Springer.