Robust One-Shot Face Video Re-enactment using Hybrid Latent Spaces of StyleGAN2

Trevine Oorloff Yaser Yacoob
University of Maryland, College Park, Maryland, USA
{trevine,yaser}@umd.edu

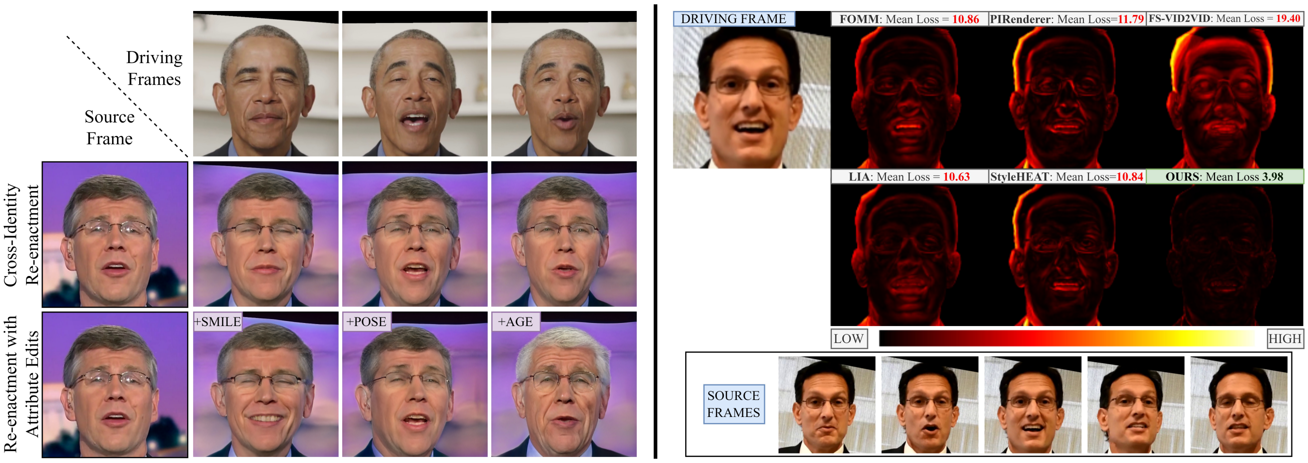
Driving Frames
Source Frame
Cross-Identity Re-enactment
Re-enactment with Attribute Edits
+SMILE
+POSE
+AGE
DRIVING FRAME
FOMM: Mean Loss = 10.86
PIRenderer: Mean Loss=11.79
FS-VID2VID: Mean Loss = 19.40
LIA: Mean Loss = 10.63
StyleHEAT: Mean Loss=10.84
OURS: Mean Loss 3.98
LOW
HIGH
SOURCE FRAMES

Figure 1: **The proposed end-to-end framework facilitates *robust* one-shot face video re-enactment at 1024^2, purely based on StyleGAN2's predefined latent spaces *without* explicit structural priors for guidance.** We show on the *Left*, *2nd row:* re-enactment examples successfully animating the source frame to mimic the motion of the driving sequence and *3rd row:* re-enactments with latent-based attribute edits facilitated by our approach. *Right:* Ideally, re-enactments need to be insensitive to the head pose and expressions of the source. The heatmap depicts the L1 loss (lower the better) averaged across 5 one-shot same-identity re-enactments of the driving frame, each initiated from a single source frame with a different head pose and expression. This proves the superior robustness of our approach in comparison to existing approaches (FOMM [30], PIRenderer [26], FS-Vid2Vid [35], LIA [37], StyleHEAT [42]).

Abstract

Recent research on one-shot face re-enactment has progressively overcome the low-resolution constraint with the help of StyleGAN's high-fidelity portrait generation. However, such approaches rely on explicit 2D/3D structural priors for guidance and/or use flow-based warping which constrain their performance. Moreover, existing methods are sensitive (not robust) to the source frame's facial expressions and head pose, even though ideally only the identity of the source frame should have an effect. Addressing these limitations, we propose a novel framework exploiting the implicit 3D prior and inherent latent properties of StyleGAN2 to facilitate one-shot face re-enactment at 1024^2 (1) with zero dependencies on explicit structural priors, (2) accommodating attribute edits, and (3) robust to diverse facial expressions and head poses of the source frame. We train an encoder using a self-supervised approach to decompose the identity and facial deformation of a portrait image within the pre-trained StyleGAN2's predefined latent spaces itself (automatically facilitating (1) and (2)). The decomposed identity latent of the source and the facial deformation latents of the driving sequence are used to generate re-enacted frames using the StyleGAN2 generator. Additionally, to improve the identity reconstruction and to enable seamless transfer of driving motion, we propose a novel approach, Cyclic Manifold Adjustment. We perform extensive qualitative and quantitative analyses which demonstrate the superiority of the proposed approach against state-of-the-art methods. Project page: `https://trevineoorloff.github.io/FaceVideoReenactment_HybridLatents.io/`.

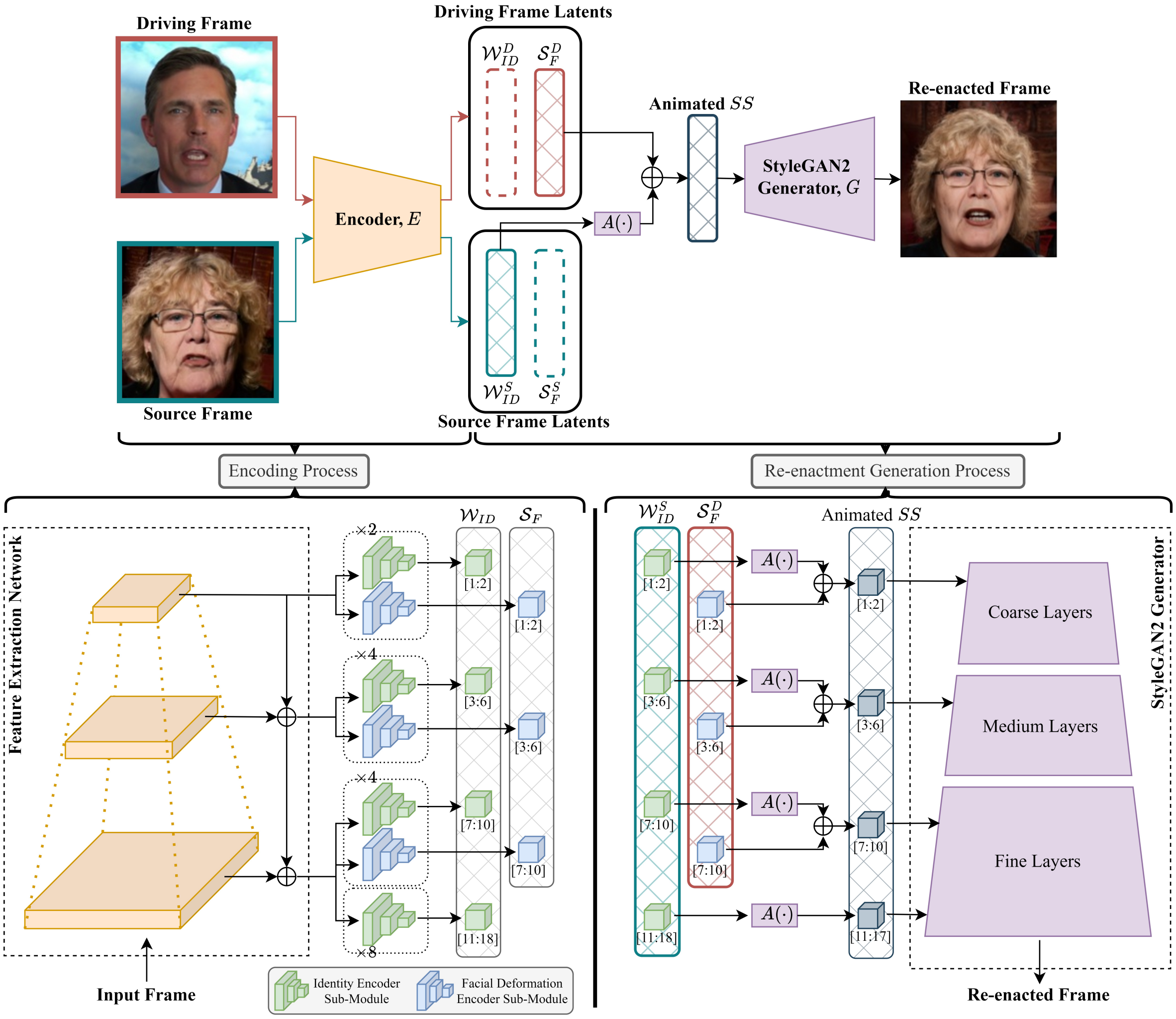

Figure 2: **The pipeline of the proposed framework.** The high-level re-enactment process (*Top*), the expanded architectures of the encoding (*Bottom-Left*) and re-enactment (*Bottom-Right*) processes are depicted. In encoding, given a frame, the Encoder, E, outputs a pair of latents: Identity latent, $\mathcal{W}_{ID}$, and Facial-deformation latent, $\mathcal{S}_F$, that reside within the predefined $W+$ and SS spaces of StyleGAN2. In re-enactment, $\mathcal{W}_{ID}^{S}$ (source) transformed using $A(\cdot)$ and $\mathcal{S}_F^{D}$ (driving frame) are combined to form the animated SS latent, which is used to obtain the re-enacted frame using the StyleGAN2 Generator, G.

1. Introduction

One-shot face video re-enactment refers to the process of generating a video by animating the identity of a portrait image (source frame) mimicking the facial deformations and head pose of a driving video. The increasing interest in virtual reality has stimulated video re-enactment due to its wide range of applications (*e.g*., digital avatars, animated movies, telepresence).

The task of one-shot face re-enactment is challenging since it requires extraction of (1) identity and 3D facial structure of the given 2D source frame and (2) motion information from the driving frames, to facilitate realistic animations, despite the unavailability of paired data. To this end, most research employ facial landmarks [18, 24, 32, 43] or 3D parametrized models [12, 19, 26, 42] to capture the underlying facial structure and/or motion. Even though explicit facial structure priors may support rigorous control, they suffer from lack of generalizability (for different face geometries), inability to capture fine/complex facial deformations (*e.g*., wrinkles, internal mouth details such as tongue and teeth), inability to handle accessories such as eyeglasses, and inconsistencies in predictions which hinder their performance. While latent-based models [9, 37, 39, 44] and predictive keypoint models [30, 36] alleviate the dependency on explicit structural priors they are limited to producing low resolution videos.

Moreover, since the source frame is only responsible for determining the identity of the animated frame in one-shot face re-enactment, ideally, the expression and head pose of the source frame should not have an effect on the animated frame. We define this property as *one-shot robustness*. Existing one-shot re-enactment methods do not address this issue and thus have poor robustness to diverse expressions and head poses of the source frame (see Fig. 1: *Right*).

StyleGAN2's [22] ability to generate high-resolution (1024^2) photo-realistic faces and semantic interpretability of its latent spaces [13, 29, 40] lead to improved re-enactment generations [19, 25, 42]. Considering StyleGAN2 latent space manipulations [1, 13, 25, 40], which facilitate edits on head pose, smile, gaze, blink, *etc*., it is evident that the predefined latent space of a pre-trained StyleGAN2 has implicit 3D information embedded within it. We conjecture that the StyleGAN2's latent spaces are not yet fully exploited for re-enactment and use of explicit structural representations as in [19, 42] (3DMM) is redundant and limits the performance of StyleGAN2 to the capacity-limits of such structural priors as discussed previously.

In this work, we address the following question: *Can we learn a general model to facilitate face identity, attributes, and motion edits leveraging the predefined latent spaces of StyleGAN2 without reliance on explicit 2D/3D facial structure priors while improving the performance of generating realistic, high-quality, and temporally consistent one-shot face re-enactment videos that are also robust to diverse expressions and head poses of the source frame?*

We train an encoder through a self-supervised approach to encode a given portrait image as an Identity latent, $\mathcal{W}_{ID}$, and a Facial deformation latent, $\mathcal{S}_F$, that reside in the *predefined* $W+$ and StyleSpace (SS) latent spaces [40] of StyleGAN2 respectively, thus forming a *hybrid latent space*. This novel approach of decomposing identity and facial deformation within the predefined latent spaces of StyleGAN2 itself obviates the need for explicit structural priors and accommodate latent-based attribute manipulation (*e.g*., smile, pose, age, *etc*.) proposed by previous research [13, 29]. We exploit the inherent latent properties of the $W+$ and SS latent spaces: best distortion-editability trade-off and best disentanglement [40] respectively, to design the decomposition framework as explained in Sec. 3. The decomposed identity latent of the source and the deformation latents of the diving sequence of frames are then used to generate high-fidelity one-shot face re-enactment video (with or without attribute edits) at 1024^2 using a pre-trained StyleGAN2 generator (Fig. 2). Further, the model was carefully designed such that the re-enacted video is robust to diverse head poses and expressions of the source frame.

Additionally, we propose a novel algorithm, *Cyclic Manifold Adjustment (CMA)* inspired by PTI [28], to address (1) StyleGAN2's poor identity reconstruction of out-of-domain source frames and (2) the non-homogeneity of the local manifolds around the driving identity and the source identity latents (*i.e*., a facial deformation edit applied to two identity latents would have slight differences in the rendered animation due to the non-homogeneity of local manifolds around different identities). Thus improving the source identity reconstruction and enabling seamless transfer of facial deformations of the driving video. While research such as [5, 28, 34] addresses the former, to the best of our knowledge no research has been conducted to address the latter.

In summary, our key contributions include:

- A novel StyleGAN2-based hybrid latent space framework that enables high-fidelity one-shot face video re-enactment at 1024^2, robust to diverse head poses and expressions of the source yielding state-of-the-art results (quantitative improvement of upto 12% in cross-identity re-enactment and 50% in one-shot robustness),
- A novel hybrid latent space approach of decomposing the identity and facial deformation of a portrait image within the *predefined* latent spaces of StyleGAN2 itself obviating the need for explicit structural priors for guidance and facilitating re-enactments with latent-based attribute edits,
- A novel algorithm, Cyclic Manifold Adjustment, that locally adjusts the StyleGAN2's manifold to improve the reconstruction of an out-of-domain source and enable seamless transfer of facial deformations of the driving video to the source image.

To the best of our knowledge, we are the first to decompose identity and facial deformation within the pre-trained StyleGAN2's *predefined* latent spaces itself, the first to handle *robustness* to diverse head pose and expressions of the source frames, and the first to propose a manifold adjustment technique handling both *source identity reconstruction and non-homogeneity* of the latent space in the task of latent-based re-enactment.

2. Related Work

Face Video Re-enactment: Face video re-enactment approaches can be categorized based on the identity/motion representation or the approach used to generate the animated frames. Facial landmarks [18, 24, 32, 43], 3D facial priors [12, 19, 26], 2D/3D predictive keypoint methods [30, 36], and intermediate latent representations [9, 37, 39, 44] are commonly used for identity/motion representation. While 3D facial priors address the main issue of 2D facial priors, *i.e*., unrealistic deformations caused during significant motion, their performance is limited by the lack of fine visual-details surrounding face dynamics (wrinkles, teeth, non-rigid deformations), inability to represent accessories (*e.g.*, eyeglasses), and representation of only the inner face region. Further, the use of explicit 2D/3D facial structural priors leads to spatio-temporal incoherence stemming from inconsistencies of the landmarks/parameters and poor generalization in cases of varying facial geometries between the driving and source identities.

The work on re-enactment either follows an optical flow based warping strategy [11, 30, 36] or a generative approach [9, 16, 19] to animate frames. Even though warping methods yield results with high resemblance to the in-the-wild source images due to operating in image space, they are prone to unrealistic deformations in faces, have weaker generalization compared to generative approaches, and perform poorly in generating facial structures that were not visible in the source image (*e.g.*, filling in teeth in opening of the mouth, opening eyes, ears when rotating the head).

StyleGAN-based Face Re-enactment: Recent research [8, 19, 42] employ StyleGAN2 as a tool for high-resolution one-shot face re-enactment due to its ability to produce realistic portraits at 1024^2 and the rich semantic properties of its latent spaces [13, 29, 40]. MegaFR [19] encodes the residual deformation between the source image and a 3D rendering (combination of 3DMM parameters [7] of the source and driving frames) of the animated frame as an additive latent space offset. Bounareli *et al.* in [8] map the difference of 3D model parameters of the driving and source frames onto the latent space to obtain the re-enacted frame's latent. StyleHEAT [42] uses 3DMM parameters to capture the facial deformations to generate flow fields which are used to warp an intermediate feature space of the generator.

While the above methods yield promising results, they employ explicit 3D priors to capture the facial structure/motion, thus, suffer from the limitations of using explicit structural priors as explained previously (*i.e.*, inconsistencies, lack of fine-grained details, limited by the capacity of the 3D model, *etc.*). In contrast, our novel framework does not use *explicit* 2D/3D structural models, instead, we exploit the *implicit* 3D priors of StyleGAN2 and the inherent properties of its latent spaces to encode the decomposed identity and the motion within the StyleGAN2's *predefined* latent spaces itself, which also facilitates accommodation of latent-based attribute edits proposed by previous research. While StyleGAN2 inherently suffers from texture-sticking, StyleHEAT aggravates the issue as their model warps the intermediate feature space of the StyleGAN2 generator. Moreover, compared to [8, 19] that encode in the $W+$ space, we employ a hybrid latent approach using both $W+$ and SS latent spaces to exploit high inversion-editability and high disentanglement respectively. Further, our method is robust to diverse head poses and facial expressions of the source frame (*i.e.*, better one-shot robustness) as opposed to the existing one-shot re-enactment methods, which perform poorly.

3. Methodology

Preliminaries: Among the prominent predefined latent spaces of StyleGAN2, $W+$ space has the best trade-off between inversion quality and editability as demonstrated by StyleGAN2 inversion and latent manipulation research [3, 27, 31] and StyleSpace, SS, has the highest disentanglement [40]. Leveraging these signature properties, we encode a given portrait image to an Identity latent and a Facial deformation latent that resides within the predefined $W+$ and SS latent spaces respectively. The logical reasoning of this choice is explained involving Eqs. (2) and (4) in the Overview and is supported by ablations in Sec. 4.3.

Overview: As shown in Fig. 2, the proposed framework comprises of two main networks, an encoder, E, and a decoder, G: a pre-trained StyleGAN2 generator. The encoder consists of two heads: Identity Encoder, E_{ID}, and Facial Deformation Encoder, E_F, preceded by a common Feature Extraction Network, F. Given an input frame, the encoder outputs two latents, $\mathcal{W}_{ID}$ and $\mathcal{S}_F$, capturing identity and facial deformations respectively, such that,

$$\mathcal{W}_{ID} = E_{ID}(F(I)), \quad \mathcal{S}_F = E_F(F(I)), \tag{1}$$

$$I = G\left(A(\mathcal{W}_{ID}) + \mathcal{S}_F\right), \tag{2}$$

where, $A(\cdot)$ is the affine transformation from $W+$ to SS. While $\mathcal{W}_{ID}$ resides in the entire $W+$ space *i.e.*, $\mathcal{W}_{ID} \in W+ \subset \mathbb{R}^{18\times512}$, $\mathcal{S}_F$ resides in the space spanned by the first 10 SS layers *i.e.*, $\mathcal{S}_F \in SS_{1:10} \subset \mathbb{R}^{10\times512}$. This is based on (1) the observation in [25, 40] that SS latents corresponding to facial deformations of interest, (*i.e.*, pose, mouth, gaze, eyes, eyebrows, and chin) lie within the first 10 layers of SS and (2) to avoid the appearance jitters caused by edits on high-resolution feature layers [41].

In re-enactment, we follow a frame-wise approach, where a single source frame, I^S, and each frame, I_t^D, of the driving sequence are projected to $\{\mathcal{W}_{ID}^S, \mathcal{S}_F^S\}$ and $\{\mathcal{W}_{ID_t}^D, \mathcal{S}_{F_t}^D\}$ respectively (Eq. (1)). Thereafter, the animated frame, $I_t^{S\to D}$ is generated using G, sourcing $\mathcal{W}_{ID}^S$ and $\mathcal{S}_{F_t}^D$ latents, comprising of the source identity and the driving frame's facial deformations respectively.

$$I_t^{S\to D} = G\left(A(\mathcal{W}_{ID}^S) + \mathcal{S}_{F_t}^D\right) \tag{3}$$

As seen in Eq. (3), the additive latent $\mathcal{S}_{F_t}^D$ constitutes a latent edit performed on $\mathcal{W}_{ID}^S$. Thus, it is important for $\mathcal{W}_{ID}^S$ to reside in the latent space with the best inversion-editability trade-off (*i.e.*, $W+$) and accommodate a wide range of facial deformation latent edits imposed by the driving sequence ($\{\mathcal{S}_{F_t}^D\}$). It is equally important for $\mathcal{S}_{F_t}^D$ to reside in a highly disentangled latent space (*i.e.*, SS), so that it minimizes identity leakage, which would alter the source identity across frames. This choice of design also accommodates the latent-based attribute manipulations (*e.g.*, pose, age, smile, *etc.*) proposed by previous research [13, 29].

$$I^{S\to D}_{edit,t} = G\left(A(\mathcal{W}^S_{ID} + \mathcal{W}_{edit,t}) + \mathcal{S}^D_{F_t}\right) \tag{4}$$

3.1. Architecture

Feature Extraction Network, F: We use a ResNet50-SE backbone [14, 17] extended with a feature pyramid [23] to extract the coarse, medium, and fine features of each frame. These levels correspond to the levels of features addressed by each latent layer as in [21].

Identity Encoder, E_{ID}: The E_{ID} consists of network blocks similar to *map2style* in [27] where the feature maps of the corresponding granularity are gradually reduced to 512 dimensions using a fully convolutional network. The encoder consists of 18 such blocks each predicting a single layer (dimension) of $\mathcal{W}_{ID} \in \mathbb{R}^{18\times 512}$.

Facial Deformation Encoder, E_F: While E_F has a similar architecture to E_{ID}, it consists of only 10 *map2style* blocks as we limit the SS latent edits to only the first 10 layers of SS as explained in the Overview.

Decoder, G: We use the pre-trained StyleGAN2 generator, which facilitates the input of SS latents [40], as the decoder to generate re-enacted frames from the latents.

3.2. Implementation

Due to the unavailability of paired re-enactment datasets, we follow a self-supervised training approach to learn the weights of the encoder, E. During training, we randomly sample a single source frame, I^S, and two driving frames, I^{D1} and I^{D2}, one belonging to the same identity as I^S and the other from a randomly selected different identity respectively. The three frames, I^S, I^{D1}, and I^{D2} are encoded to the corresponding latents, $\{\mathcal{W}^S_{ID}, \mathcal{S}^S_F\}$, $\{\mathcal{W}^{D1}_{ID}, \mathcal{S}^{D1}_F\}$, and $\{\mathcal{W}^{D2}_{ID}, \mathcal{S}^{D2}_F\}$ respectively (Eq. (1)). We learn the weights of E by optimizing over the following loss functions.

Reconstruction Losses: Reconstruction losses are twofold comprising of a self-reconstruction loss, Eq. (5), and a re-enactment loss, Eq. (6), which measure the reconstruction of the source frame, $I^{S\to S}$, and the same-identity driving frame, $I^{S\to D1}$, using the source identity latent, $\mathcal{W}^S_{ID}$, and the corresponding facial deformation latents.

$$\mathcal{L}_{self} = \mathcal{L}_{rec}\left(I^S, G\{A(\mathcal{W}^S_{ID}) + \mathcal{S}^S_F\}\right) \tag{5}$$

$$\mathcal{L}_{reenact} = \mathcal{L}_{rec}\left(I^{D1}, G\left(A(\mathcal{W}^S_{ID}) + \mathcal{S}^{D1}_F\right)\right) \tag{6}$$

$$\mathcal{L}_{rec} = \lambda_{L2}\mathcal{L}_{L2} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_{GV}\mathcal{L}_{GV} \tag{7}$$

where, $\mathcal{L}_{rec}$ is a weighted sum of the MSE loss ($\mathcal{L}_{L2}$), LPIPS loss [45] ($\mathcal{L}_{LPIPS}$), and Gradient Variance loss [2] ($\mathcal{L}_{GV}$), weighed by λ_{L2}, λ_{LPIPS}, and λ_{GV} respectively.

Identity Loss: The identity loss is computed using,

$$\mathcal{L}_{id} = \{1 - \langle\phi(I^S), \phi(I^{S_{ID}})\rangle\} \tag{8}$$

$$+ \{1 - \langle\phi(I^S), \phi(I^{S\to D2})\rangle\} \tag{9}$$

where, cosine similarity ($\langle\cdot,\cdot\rangle$) of the ArcFace [10] feature space, is measured between the pair of images. $I^{S_{ID}} = G\left(A(\mathcal{W}^S_{ID})\right)$ is the source identity image and $I^{S\to D2} = G\left(A(\mathcal{W}^S_{ID}) + \mathcal{S}^{D2}_F\right)$ denotes the animation of the source representing the facial deformations of I^{D2}. Eq. (8) ensures the identity of the source is captured within $\mathcal{W}^S_{ID}$ and also prevents the optimization from converging to the trivial solution of Eqs. (5) and (6): *i.e.*, $\mathcal{W}^S_{ID} = 0$. Further, Eq. (9) makes sure that cross-identity re-enactment preserves identity *i.e.*, minimizes the identity information leakage to $\mathcal{S}_F$.

Identity Latent Consistency Loss: To enforce one-shot robustness we obtain the $\mathcal{W}_{ID}$ of $I^{S\to D2}$ by passing it through E and compute the following loss over the latent space to encourage consistent identity latents irrespective of head pose and facial expressions of the source.

$$\mathcal{L}_{w\_id} = \|W^S_{ID} - W^{S\to D2}_{ID}\|_2 \tag{10}$$

Regularization Loss: Additional regulatory losses are used to reduce the variance within $\mathcal{W}^S_{ID}$ [31] and to control the facial-deformation edits, $\mathcal{S}^{D1}_F$, to be within the proximity of $A(\mathcal{W}^S_{ID})$ combined in a ratio of $1 : \lambda_S$. Indices $[i]$ and $[j]$ denote the i^{th} and j^{th} dimension of the latent.

$$\Delta_i = \mathcal{W}^S_{ID}[i] - \mathcal{W}^S_{ID}[1] \tag{11}$$

$$\mathcal{L}_{reg} = \sum_{i=1:18} \|\Delta_i\|_2 + \lambda_S \sum_{j=1:10} \|\mathcal{S}^{D1}_F[j]\|_2 \tag{12}$$

Feature Reconstruction Loss: Complementary to the reconstruction losses, feature reconstruction losses are computed using the same loss functions with the exception of the losses being computed on a dilated masked region consisting of the mouth, eyes, and eyebrows to increase the emphasis on capturing fine facial deformations accurately.

$$\begin{aligned}\mathcal{L}_f = {} & \mathcal{L}_{rec}\left(M^S \odot I^S, M^S \odot I^{S\to S}\right) \\ & + \mathcal{L}_{rec}\left(M^{D1} \odot I^{D1}, M^{D1} \odot I^{S\to D1}\right)\end{aligned} \tag{13}$$

Additionally, similar to [31] we train a **Latent Discriminator**, with adversarial loss, $\mathcal{L}_d$, to encourage the W^S_{ID} to be in the well-editable regions of StyleGAN2 latent space.

Total Loss: The total loss is as follows, where λ_* represents the corresponding weights.

$$\begin{aligned}\mathcal{L} = {} & \mathcal{L}_{self} + \mathcal{L}_{reenact} + \lambda_{id}\cdot\mathcal{L}_{id} + \lambda_{w\_id}\cdot\mathcal{L}_{w\_id} \\ & + \lambda_d\cdot\mathcal{L}_d + \lambda_{reg}\cdot\mathcal{L}_{reg} + \lambda_f\cdot\mathcal{L}_f\end{aligned} \tag{14}$$

3.3. Cyclic Manifold Adjustment (CMA)

As explained in Sec. 1, we propose a novel approach, *Cyclic Manifold Adjustment*, inspired by PTI[28], with the following objectives: (1) improving the identity reconstruction quality of out-of-domain subjects and (2) addressing the non-homogeneity of the local manifolds around the

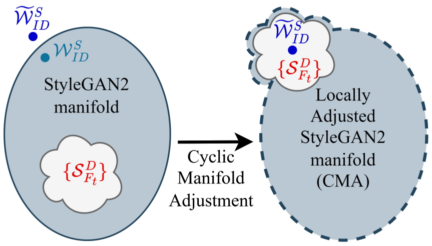
$\widetilde{\mathcal{W}}^S_{ID}$
$\mathcal{W}^S_{ID}$
StyleGAN2 manifold
$\{\mathcal{S}^D_{F_t}\}$
Cyclic Manifold Adjustment
$\widetilde{\mathcal{W}}^S_{ID}$
$\{\mathcal{S}^D_{F_t}\}$
Locally Adjusted StyleGAN2 manifold (CMA)

Figure 3: **Cyclic Manifold Adjustment (CMA).** For an out-of-domain subject, $\widetilde{\mathcal{W}}^S_{ID}$, $\mathcal{W}^S_{ID}$, and $\{\mathcal{S}^D_{F_t}\}$, represent the true identity, identity latent estimate obtained using E, and the sequence of facial deformation latents obtained from the driving sequence. We locally tweak the StyleGAN2's manifold around $\mathcal{W}^S_{ID}$, to include the latent space spanned by $\{\mathcal{S}^D_{F_t}\}$ centered around $\widetilde{\mathcal{W}}^S_{ID}$, thus improving the source identity reconstruction and enabling seamless transfer of facial deformations of the driving video.

driving and source identity latents. We achieve this by fine-tuning the StyleGAN2 generator, G, at inference (with encoder weights frozen), optimizing the carefully designed cyclic objective (Eqs. (15) to (18)).

Suppose the true source identity latent is $\widetilde{\mathcal{W}}^S_{ID}$ which is out of StyleGAN2's domain. Fine-tuning the StyleGAN2 generator, G, using PTI on the source image, would constrain the local latent space around the source identity, $\widetilde{\mathcal{W}}^S_{ID}$, thus limiting the editability through facial deformation latents of the driving frames, $\{\mathcal{S}^D_{F_t}\}$ (as PTI focuses only on objective (1)). In contrast, CMA tweaks the latent space manifold around the source identity latent estimate, $\mathcal{W}^S_{ID}$, obtained through E, to include the latent space spanned by the sequence $\{\mathcal{S}^D_{F_t}\}$ centered around $\widetilde{\mathcal{W}}^S_{ID}$ as depicted in Fig. 3. In other words, the cyclic reconstructions of the source and driving frames are obtained from the decomposed latents $\mathcal{W}^S_{ID\_cyc}, \mathcal{S}^D_{F\_cyc}$ of the re-enacted frame, $I^{S\rightarrow D}$ (Eqs. (17) and (18)). Thus, considering Eq. (16), $\mathcal{W}^S_{ID\_cyc}, \mathcal{S}^D_{F\_cyc} = E(I^{S\rightarrow D}) = E(G(A(W^S_{ID}) + S^D_F))$, minimizing the cost function, Eq. (15), finetunes G, such that the source identity, W^S_{ID}, and the driving face deformations, S^D_F, are maximally visible in the re-enacted frame, $I^{S\rightarrow D}$. This locally adjusts the StyleGAN2 manifold around W^S_{ID} to include $\widetilde{\mathcal{W}}^S_{ID}$ with S^D_F in its neighborhood. This novel approach improves the identity reconstruction of the out-of-domain source and enables seamless transfer of facial deformations of the driving video. We achieve this by optimizing the cost function,

$$\mathcal{L}\left(I^S_{cyc}, I^S\right) + \mathcal{L}\left(I^D_{cyc}, I^D\right) \tag{15}$$

where, $\mathcal{L}$ denotes a combination of LPIPS and L2 losses and I^S_{cyc} and I^D_{cyc} are cyclic reconstructions of the source and driving frames respectively generated as follows.

$$\mathcal{W}^S_{ID\_cyc}, \mathcal{S}^D_{F\_cyc} = E\left(I^{S\rightarrow D}\right) \tag{16}$$

$$I^S_{cyc} = G\left(A(\mathcal{W}^S_{ID\_cyc}) + \mathcal{S}^S_F\right) \tag{17}$$

$$I^D_{cyc} = G\left(A(\mathcal{W}^D_{ID}) + \mathcal{S}^D_{F\_cyc}\right) \tag{18}$$

4. Experiments and Results

Datasets: We pre-trained the encoder on the CelebV-HQ dataset [47], which includes diverse high resolutions (min. 512^2) with over 35K videos involving 15K+ identities. The HDTF dataset [46], which consists of 362 videos (720p/1080p) with 300+ identities, was used for the fine-tuning stage with an 80-20 non-overlapping train-test split. All the sampled frames were preprocessed according to [30]. For training 50 samples/video were used and for evaluation the first 500 samples/video of 75 unseen videos (total 37.5K frames) were chosen.

Training: Training was performed in two stages: (1) a *Pre-training Stage*: the entire encoder, E is trained for 200K iterations, followed by (2) a *Fine-tuning Stage*: only the two latent-prediction heads: E_{ID} and E_F are fine-tuned (weights of the Feature Exctraction Network, F frozen) at a reduced learning rate for 20K additional iterations to avoid over-fitting. While the former focuses mainly on learning the features of images, improving generalization, and learning the implicit prior of the StyleGAN2's latent space, the latter stage focuses on capturing the detailed facial deformations and face details. See supplementary for further details.

Inference: During inference, a driving video and a single source frame were obtained from the evaluation samples of the HDTF dataset. While the source frame and the driving video are of the same identity in same-identity re-enactment, the identities differ in the case of cross-identity re-enactment. The re-enacted frames were generated as explained in Eqs. (1) to (4). Cyclic manifold adjustment (Sec. 3.3) was used to improve realism and visual quality. Since a generative architecture is used, our model is capable of rendering re-enactment videos in real-time ($\sim$ 30fps).

4.1. Baselines and Metrics

Baselines: We compare our results against a diverse range of state-of-the-art approaches that are based on: predictive keypoints (FOMM [30]); 3D models (PIRenderer [26], StyleHEAT[42]), facial landmarks (FS-Vid2Vid[35]); intermediate latents (LIA [37]); flow-based warping (FOMM, FS-Vid2Vid, PIRender, StyleHEAT); and StyleGAN2-based (StyleHEAT).

Metrics: We extensively evaluate the results of the proposed framework against the baselines using (1) *Reconstruction fidelity*: L1-norm pixel loss, Peak Signal-to-Noise Ratio (PSNR), (2) *Identity preservation:* Identity loss ($\mathcal{L}_{ID}$ - computed using [10]), (3) *Perceptual quality:* LPIPS [45],

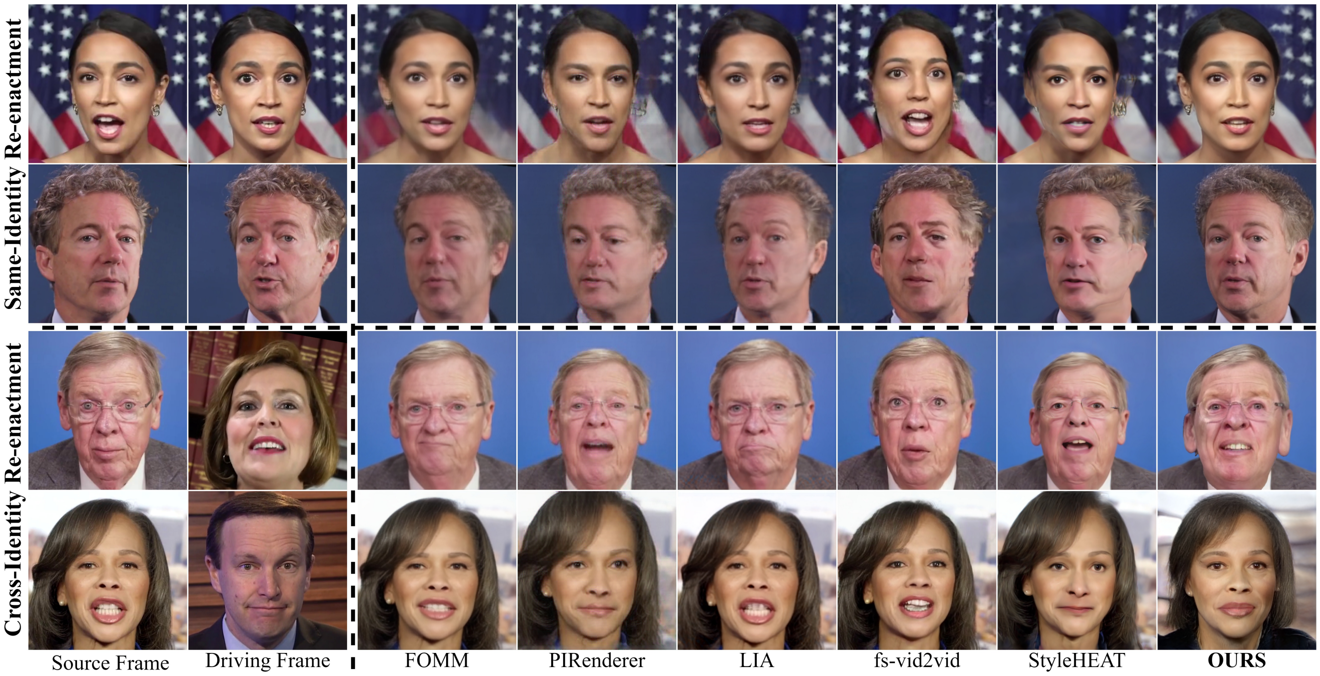

Figure 4: **Qualitative evaluation of same-identity *(Top)* and cross-identity *(Bottom)* re-enactment**. *Same-Identity Re-enactment:* Observe the lack of sharpness in facial features (*e.g.*, teeth, wrinkles, eyes), visual artifacts around eyes, ears, and mouth, and incorrect/missing facial features in baseline methods in comparison to our approach. *Cross-Identity Re-enactment:* Observe in comparison to the baselines: *3rd row:* teeth, mouth formation, and head pose; *4th row:* preservation of source identity and lip structure, and the expression of driving. Refer supplementary video for better visualization.

SSIM [38], FID [15], (4) *Spatio-temporal perceptual quality:* FVD [33], FVD_M (FVD over the mouth), (5) *Temporal coherence in facial attributes:* ρ_{AU}, ρ_{GZ}, and ρ_{pose}, the temporal correlation of Action Units (expressions), gaze, and pose respectively measured using [6].

Additionally, we design an experiment to evaluate the robustness to diverse head poses and facial expressions of the source frame (*one-shot robustness*). The performance of same-identity one-shot re-enactment of 5 driving videos with 5 diverse source frames per driving video (25 source image-driving video combinations) are evaluated based on the mean loss and standard deviation across re-enactments.

4.2. Analysis and Discussion

Quantitative Analysis: The performance of our approach in comparison to the baselines in the tasks of same-identity and cross-identity re-enactments, are tabulated in Tabs. 1 and 2 respectively. Our approach yields the best performance across all metrics except $\mathcal{L}_{ID}$ where it achieves comparable values to the best. A higher $\mathcal{L}_{ID}$ for FOMM is expected as it operates in the image space as opposed to our generative approach. However, the performance of our model generating results at 1024^2, in comparison to the baselines is not fully reflected through the metrics, as (1) most of the metrics ($\mathcal{L}_{ID}$, FID, FVD, *etc.*) are computed on downsampled images at a low-resolution which do not reflect the fine-grained details and (2) the lack of high-resolution ($\geq 1024^2$) data samples. Further, we yield the lowest mean and standard deviation in one-shot robustness experiment (Tab. 3) (upto 50% improvement) proving the superior robustness of our framework to diverse head poses and expressions of the source frames compared to baselines.

Qualitative Analysis: The qualitative examples in Fig. 4 and supplementary video demonstrate that our approach yields much improved visual results compared to the baselines that are prone to visual artifacts, incorrect facial attributes, and lack of sharpness in scenarios where the source and driving frames have significant difference in pose and/or expression, subject is wearing eyeglasses, filling in unseen details (*e.g.*, teeth, ear). Our approach successfully handles these cases while producing more realistic re-enactment at high-resolution (1024^2). The versatility of our model is more pronounced in the cross-identity examples, where facial deformations of the driving frames are accurately mimicked in the re-enacted frame while preserving the source identity. Fig. 1 demonstrates the ability of the model to generate realistic re-enactments with latent-based attribute edits.

Further, all approaches except StyleHEAT are only capable of low-resolution generation. While StyleHEAT uses

Method	res.	L1 ↓	LPIPS↓	$\mathcal{L}_{ID}$↓	PSNR↑	SSIM↑	FID↓	FVD↓	FVD_M↓	ρ_{AU}↑	ρ_{GZ}↑	ρ_{pose}↑
FOMM [30]	256^2	2.37	0.042	**0.095**	32.8	0.959	22.3	102.1	19.9	0.808	0.737	0.948
PIRenderer [26]	256^2	3.52	0.053	0.118	29.1	0.932	28.0	145.1	26.8	0.748	0.770	0.906
LIA [37]	256^2	2.72	0.049	0.102	31.5	0.951	26.4	105.0	19.6	0.822	0.732	0.944
FS-Vid2Vid [35]	512^2	4.38	0.065	0.151	27.4	0.919	31.6	255.0	45.0	0.572	0.635	0.822
StyleHEAT [42]	$\mathbf{1024^2}$	3.59	0.059	0.133	28.8	0.934	38.8	171.0	31.4	0.745	0.850	0.921
Ours	$\mathbf{1024^2}$	**2.28**	**0.027**	0.097	**33.1**	**0.967**	**19.3**	**101.0**	**16.6**	**0.829**	**0.872**	**0.952**
Ours w/o ID reg	1024^2	2.34	0.028	0.111	32.8	0.953	20.3	115.4	20.7	0.810	0.857	0.948
Ours w/o Hybrid	1024^2	2.52	0.031	0.114	31.7	0.950	20.6	116.9	23.6	0.746	0.798	0.923

Table 1: **Quantitative comparison of one-shot same-identity re-enactment against baselines.** *Top*: Evaluation results computed over 75 unseen videos (37.5K total frames) of the HDTF dataset [46]. Our approach yields the best performance across all metrics except $\mathcal{L}_{ID}$ (comparable with best) while generating high-resolution re-enactment. *Bottom*: Ablations performed for Ours w/o ID reg.: Framework without the Identity regularization and Ours w/o Hybrid: Framework without the Hybrid latent spaces *i.e*., both latents in $W+$ space.

Method	FID↓	FVD↓	FVD_{M}↓	ρ_{AU+GZ}↑	ρ_{pose}↑
FOMM [30]	94.0	529.4	78.4	0.450	0.782
PIRenderer [26]	84.8	417.3	54.2	0.668	0.880
LIA [37]	94.8	536.2	76.7	0.404	0.788
FS-Vid2Vid [35]	90.6	532.7	86.7	0.493	0.745
StyleHEAT [42]	97.2	408.8	58.9	0.645	0.875
Ours	**74.3**	**375.4**	**50.2**	**0.718**	**0.915**
Ours w/o ID reg	85.6	399.8	53.1	0.685	0.894
Ours w/o Hybrid	93.0	427.3	60.8	0.649	0.878
Ours w/o CMA	86.9	388.1	52.4	0.646	0.896
Ours – CMA + PTI	84.8	421.2	56.8	0.627	0.890

Table 2: **Quantitative evaluation of cross-identity one-shot re-enactment.** *Top*: Evaluation computed over 75 unseen videos (37.5K frames in total) of the HDTF dataset [46] and random source frames of different identities. Our approach yields the best performance across all metrics which is reflective of the visual results. *Bottom*: Ablations for Ours w/o ID reg.: Framework without the Identity regularization; Ours w/o Hybrid: Framework without the Hybrid latent spaces *i.e*., both latents in $W+$ space; Ours w/o CMA: Framework without Cyclic Manfold Adjustment (CMA) and Ours – CMA + PTI: Framework replacing Cyclic Manfold Adjustment (CMA) with PTI[28].

Method	LPIPS↓ $\times 10^{-2}$	$\mathcal{L}_{ID}$ ↓ $\times 10^{-1}$	FID↓ $\times 10^{1}$	FVD↓ $\times 10^{2}$
FOMM	7.5 ± 2.1	2.3 ± 1.1	3.2 ± 1.0	2.0 ± 0.7
PIRenderer	5.7 ± 0.4	1.2 ± 0.2	2.2 ± 0.2	1.5 ± 0.3
LIA	7.8 ± 2.0	2.3 ± 1.0	3.4 ± 1.0	2.0 ± 0.8
FS-Vid2Vid	7.5 ± 1.1	1.8 ± 0.5	3.3 ± 0.6	2.2 ± 0.5
StyleHEAT	6.0 ± 0.4	1.4 ± 0.2	3.3 ± 0.5	1.5 ± 0.2
Ours	$\mathbf{2.9 \pm 0.2}$	$\mathbf{1.1 \pm 0.1}$	$\mathbf{1.5 \pm 0.2}$	$\mathbf{0.8 \pm 0.1}$

Table 3: **Evaluation of One-Shot Robustness.** We evaluate the robustness of each model in the task of same-identity re-enactment using 5 driving videos (500 samples/video) and 5 diverse source frames per driving video (25 source image-driving video combinations). Our approach yields the least mean and standard deviation proving its robustness to diverse head pose and expressions of source frames.

StyleGAN2 to generate re-enactment videos at 1024^2, the use of an explicit 3D prior to capture the facial attributes and warping of the intermediate feature spaces inhibits its performance. In contrast, our model exploits the implicit priors of the predefined latent spaces of StyleGAN2 to achieve state-of-the-art performance. Further, Fig. 1 (*Right*) and supplementary video depict heatmaps of L1 loss (lower the better) averaged across 5 one-shot re-enactment runs each initiated with a single source frame with a different head pose and expression, which clearly shows the superior *one-shot robustness* of our approach in comparison to baselines.

4.3. Ablation Study

Identity Regularization: We evaluate the impact of using the identity-loss based regularization (Eq. (9)), which is in place to minimize the identity leakage into the facial deformation latent, $\mathcal{S}_F$. The results are in Tabs. 1 and 2 (*Bottom*) as *Ours w/o ID reg*. Considerable quantitative improvement is seen in cross-identity re-enactment with the inclusion of identity regularization.

Hybrid Latent Spaces: We employ a hybrid latent approach, where Identity latent, $\mathcal{W}_{ID}$, and Facial deformation latent, $\mathcal{S}_F$, reside in the domains of $W+$ and SS respectively to make use of their inherent properties. We validate the use of a hybrid approach against *Ours w/o Hybrid*, where both the Identity and Facial deformation latents reside in the $W+$ space. Based on the results at the bottom of Tabs. 1 and 2, the scores deteriorate compared to the Hybrid approach, most likely due to the entanglements within the $W+$ space.

Cyclic Manifold Adjustment (CMA): We evaluate the performance improvement of using CMA in comparison to (1) our framework without CMA (Tab. 2: *Ours w/o CMA*) and (2) replacing CMA with PTI (Tab. 2: *Ours – CMA + PTI*). While using CMA achieves the best results, it could be observed that the use of PTI deteriorates the scores compared to *Ours w/o CMA*. This could be because using PTI on the source constrains the local manifold around the source Sec. 3.3 and it does not guarantee the local manifold around the source to accommodate the facial deformations of the driving sequence (due to non-homogeneity) leading to distorted/ suboptimal results.

4.4. Limitations

Since we base our model on StyleGAN2, we inherit its limitations of texture sticking and alignment requirements. Further, handling occlusions and reconstruction of changing backgrounds are challenging since StyleGAN2 generator is pre-trained for faces. While our model could be adapted to StyleGAN3[20] to mitigate the issue of texture sticking, the use of StyleGAN2 is preferred due to its latent space being more structured and expressive [4] and the lower computational requirement (lower parameters in the encoder for StyleGAN2 as opposed to StyleGAN3). See supplementary for further details.

5. Conclusion

We propose a novel StyleGAN2-based hybrid latent space framework to facilitate one-shot face re-enactment at 1024^2 (1) without relying on explicit structural priors for guidance, (2) accommodating latent-based attribute edits, and (3) robust to diverse facial expressions and head poses of the source frame, while achieving state-of-the-art results both quantitatively and qualitatively. The framework reveals the full potential of StyleGAN2 for face edits, specifically identity, attributes, and facial deformations in videos, and is centered around an intuitively designed novel decomposition of identity and facial information that reside within the predefined latent spaces itself of a pre-trained StyleGAN2. Additionally, we propose a novel algorithm, Cyclic Manifold Adjustment that improves reconstruction of the source and enables seamless transfer of facial deformations of the driving video. The negative societal impact of our model is similar to that of other DeepFake algorithms and is discussed in the supplementary.

Acknowledgements

This work is supported by the US Defense Advanced Research Projects Agency (DARPA) Semantic Forensics (SemaFor) Program under HR001120C0124. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021.

[2] Lusine Abrahamyan, Anh Minh Truong, Wilfried Philips, and Nikos Deligiannis. Gradient variance loss for structure-enhanced image super-resolution. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3219–3223. IEEE, 2022.

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. *arXiv preprint arXiv:2104.02699*, 2021.

[4] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *arXiv preprint arXiv:2201.13433*, 2022.

[5] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. *arXiv preprint arXiv:2111.15666*, 2021.

[6] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[7] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552, 2016.

[8] Stella Bounareli, Vasileios Argyriou, and Georgios Tzimiropoulos. Finding directions in gan's latent space for neural face reenactment. *arXiv preprint arXiv:2202.00046*, 2022.

[9] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020.

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *CVPR*, 2019.

[11] Michail Christos Doukas, Evangelos Ververas, Viktoriia Sharmanska, and Stefanos Zafeiriou. Free-headgan: Neural talking head synthesis with explicit gaze control. *arXiv preprint arXiv:2208.02210*, 2022.

[12] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14398–14407, 2021.

[13] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] 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. *Advances in neural information processing systems*, 30, 2017.

[16] Gee-Sern Hsu, Chun-Hung Tsai, and Hung-Yi Wu. Dual-generator face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 642–650, 2022.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[18] Po-Hsiang Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Learning identity-invariant motion representations for cross-id face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7084–7092, 2020.

[19] Wonjun Kang, Geonsu Lee, Hyung Il Koo, and Nam Ik Cho. One-shot face reenactment on megapixels. *arXiv preprint arXiv:2205.13368*, 2022.

[20] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *arXiv preprint arXiv:2106.12423*, 2021.

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[24] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.

[25] Trevine Oorloff and Yaser Yacoob. Expressive talking head video encoding in stylegan2 latent-space. *arXiv preprint arXiv:2203.14512*, 2022.

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

[27] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[28] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021.

[29] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[30] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32:7137–7147, 2019.

[31] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[32] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1329–1338, 2021.

[33] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[34] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11379–11388, 2022.

[35] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. In *NeurIPS*, 2019.

[36] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021.

[37] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[39] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018.

[40] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021.

[41] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. Feature-style encoder for style-based gan inversion. *arXiv e-prints*, pages arXiv–2202, 2022.

[42] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pretrained stylegan. *arXiv preprint arXiv:2203.04036*, 2022.

[43] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020.

[44] Xianfang Zeng, Yusu Pan, Mengmeng Wang, Jiangning Zhang, and Yong Liu. Realistic face reenactment via self-supervised disentangling of identity and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12757–12764, 2020.

[45] 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.

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

[47] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022.