Expressive Talking Head Video Encoding in StyleGAN2 Latent Space

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

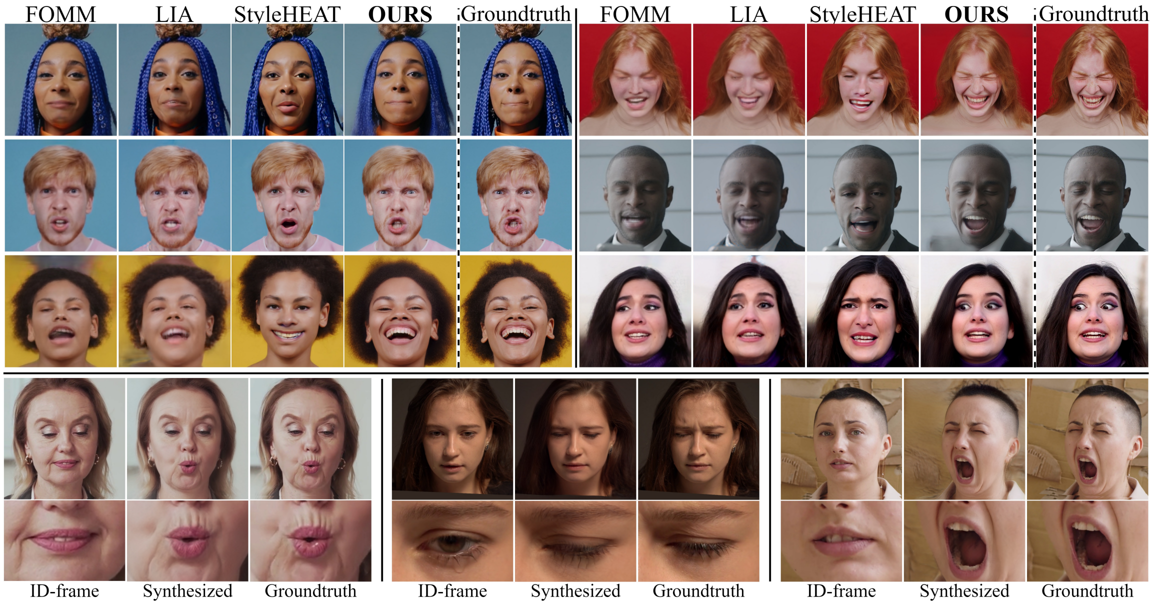

Figure 1. The proposed framework is capable of capturing fine, detailed, and highly expressive facial features (*e.g.*, lip-pressing, mouth puckering, mouth gaping, gaze, wrinkles). *Top:* Demonstrates how our re-synthesis results compare with a few state-of-the-art models: FOMM [25], LIA [31], and StyleHEAT[34]. *Bottom:* Depicts zoomed-in images of synthesized frames generated through our approach using the encoding of the ID-frame (ID-latent) and 35 parameters per frame capturing the facial deformations of the Groundtruth.

Abstract

*While the recent advances in research on video re-enactment have yielded promising results, the existing approaches fall short in capturing the fine, detailed, and expressive facial features (*e.g.*, lip-pressing, mouth puckering, mouth gaping, and wrinkles), which are crucial in generating realistic animated face videos. To this end, we propose an end-to-end expressive face video encoding approach that facilitates data-efficient high-quality video re-synthesis by optimizing low-dimensional edits of a single Identity-latent. The approach builds on StyleGAN2 image inversion and multi-stage non-linear latent space editing to generate videos that are nearly comparable to input videos. While existing StyleGAN latent-based editing techniques focus on simply generating plausible edits of static images, we automate the latent space editing to capture the fine expressive facial deformations in a sequence of frames using an encoding that resides in the Style-latent space (StyleSpace) of StyleGAN2. The encoding thus obtained could be super-imposed on a single Identity-latent to facilitate re-enactment of high-resolution face videos at 1024^2. The proposed framework economically captures face identity, head-pose, and complex expressive facial motions at fine levels, and thereby bypasses training, person modeling, dependence on landmarks/keypoints, and low-resolution synthesis which tend to hamper most re-enactment approaches. The proposed method is designed with maximum data efficiency, where a single $W+$ latent and 35 parameters per frame enable high-fidelity video rendering. This pipeline can also be used for puppeteering (*i.e.*, motion transfer). Project page:* `https://trevineoorloff.github.io/ExpressiveFaceVideoEncoding.io/`.

1. Introduction

Talking-head re-enactment, which involves animating a static portrait image to mimic the changes in head-pose and other facial attribute deformations of a driving video while maximally preserving the identity across the frames, has a wide range of applications such as AR/VR, telepresence, and movie production. Intuitively to facilitate re-enactment, one has to decompose the motion from the identity of the driving sequence of frames, and to this end, most contemporary methods utilize facial landmarks/keypoints-based [25, 29, 30], 3D facial representation-based [11, 22, 38], and latent-based [31, 32, 37] approaches to encode the facial deformations. While these methods generate promising results and each of them has its own pros and cons (Sec. 2), the most common drawbacks of the existing approaches include limitation to low resolution (commonly 256^2 and 512^2 at most), the requirement of extensive training data and person modeling, and especially the inability to capture extreme poses and intricate expressive facial details (see Fig. 1) which detracts from the realism of re-enacted videos.

On the other hand, the recent advances in StyleGAN2-based inversion techniques [1, 5, 6, 23] enable manipulation of high-resolution (1024^2) real-world images [2, 3, 4, 16, 24, 33] due to the highly disentangled property of their latent spaces. However, such latent-based manipulation techniques are mostly limited to static images and focus on simply generating plausible edits (*e.g.*, changes to smile, age, hair color, pose). While recent research [9, 17, 34] have employed StyleGAN2 to generate high-resolution reenactment video, they utilize 3D parametric models to capture facial deformations. While such priors are able to capture global facial attributes, they are not capable of capturing the fine and intensely expressive facial deformations.

In order to bridge the gap between high-fidelity static portrait image synthesis/manipulation and face reenactment of intense expressions and speech, we propose a novel end-to-end face video encoding approach that automates the latent-editing process to capture head-pose and fine and complex expressive facial deformations using merely 35 parameters per frame that reside in the Style-latent space (StyleSpace, SS) of StyleGAN2. We extend single image generation models, namely StyleGAN2 [18] and StyleFlow [3] to the temporal dimension. Quantitative evaluation of latent spaces: Z, W, $W+$, and SS, by [33], indicates that within the StyleGAN2's latent spaces, the proposed StyleSpace has the best disentanglement, completeness, and informativeness. Thus, we perform edits on SS as it enables control of individual facial attributes without re-training a network to enforce disentanglement [10]. Moreover, since the latent spaces are sparse (*i.e.*, only specific points in the space are visually valid and meaningful) we propose optimization frameworks that anchor the latent space attribute editing to the real images. The computed latent paths between frames are non-linear and therefore avoid the limitations of common linear latent editors [31].

In this research, we focus on both the re-synthesis and puppeteering of face videos using a compact encoding scheme while focusing on accurate reconstruction of expressive facial deformations. In re-synthesis, we encode a face video using a low-dimensional representation of small edits of a single Identity-latent (ID-latent). The proposed pipeline is capable of capturing and regenerating complex facial features as shown in Figs. 1 and 3 while achieving state-of-the-art performance at 1024^2. Further, since the encoding is independent of the subject in the video, we can substitute the ID-latent (*i.e.*, an inversion of a real face) of a different subject and apply the face deformation parameters to generate high-fidelity puppeteering videos. Our face video encoding is extremely compact: a single latent (18×512) corresponding to an ID-latent and only 35 parameters per frame that control the head-pose (3 parameters) and the facial features edits (32 parameters), which amounts to merely 70 bytes per frame.

In summary, the key contributions of the paper are:

- A novel algorithm for high-resolution (1024^2) face video encoding for re-synthesis and puppeteering with emphasis on precise reconstruction of both expressive and talking facial attributes in contrast to common models that do not focus on fine/complex expressive facial details,
- A novel approach that employs image inversion and sparse latent space editing to produce an extremely compact face video encoding scheme (35 parameters per frame), in contrast to most prevailing work on latent space editing that simply illustrate plausible semantic visual results,
- A novel method to find StyleSpace channels corresponding to facial attributes based on index sensitivity.

2. Related Work

2.1. Latent Space Based Editing

Understanding the latent space of a pre-trained GAN has led to better controllability over the generated output. Research such as [16, 24] explore the latent space of StyleGAN to identify the interpretable semantic directions that control attributes such as aging, smile, gender, pose, *etc.* within the latent space. However, the entangled nature of the latent space limits the manipulation, as it often leads to undesirable artifacts.

StyleSpace [33], StyleFlow [3], and StyleRig [26] are a few prominent algorithms based on the StyleGAN2 architecture that yield impressive control over latent-based manipulations. The authors of StyleSpace analyzed SS and

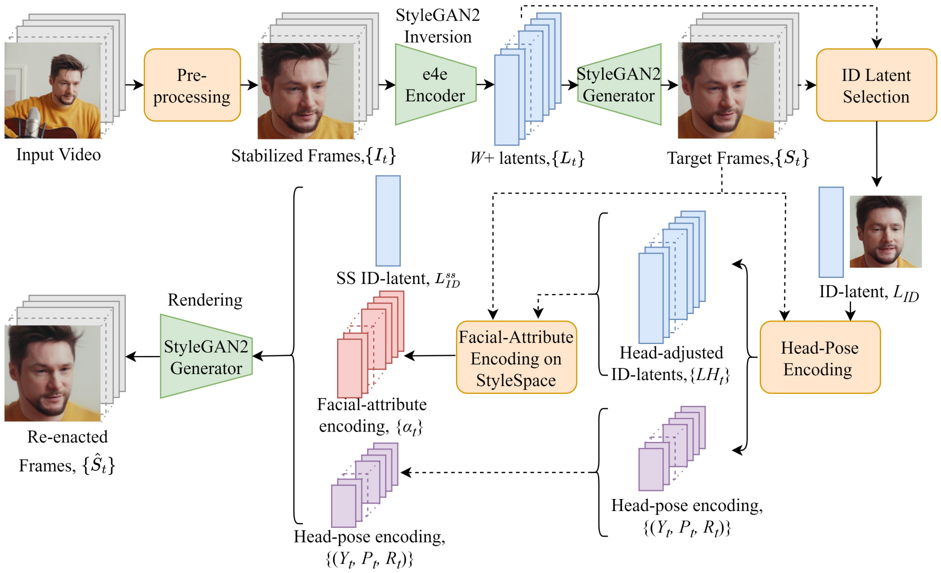
StyleGAN2 Inversion
Pre-processing
e4e Encoder
StyleGAN2 Generator
ID Latent Selection
Input Video
Stabilized Frames,$\{I_t\}$
W+ latents,$\{L_t\}$
Target Frames,$\{S_t\}$
SS ID-latent, L^{ss}_{ID}
ID-latent, L_{ID}
Rendering
StyleGAN2 Generator
Facial-Attribute Encoding on StyleSpace
Head-adjusted ID-latents,$\{LH_t\}$
Head-Pose Encoding
Facial-attribute encoding, $\{\alpha_t\}$
Re-enacted Frames, $\{\hat{S}_t\}$
Head-pose encoding, $\{(Y_t, P_t, R_t)\}$
Head-pose encoding, $\{(Y_t, P_t, R_t)\}$

Figure 2. **The multi-stage pipeline for encoding a video in latent space**: The (1) pre-processing stage aligns the input sequence of frames, which are fed to the (2) GAN inversion to obtain the corresponding sequence of $W+$ latents. Out of which the best inversion which also has near frontal head-pose is chosen to be the ID-latent in the (3) ID-latent selection stage. The (4) Head-pose encoding stage, encodes the yaw and pitch of the target frames, in reference to the ID-latent while generating a series of head-pose adjusted ID-latents. Subsequently, the (5) facial-attribute encoding stage, encodes the facial deformations using 32 parameters anchoring onto the head-pose adjusted ID-latents. Finally, the encoded parameters (35/frame) and the ID-latent are used to synthesize the re-enacted frames at the (6) Rendering stage.

formulated an algorithm to identify the style channels that control specific attributes by backtracking gradients. StyleFlow, on the other hand, uses a flow-based model conditioned on the attributes to enable non-linear and conditioned latent space edits. Even though the StyleRig algorithm enables a rig-like control over the 3D semantic parameters of faces generated through StyleGAN, it has limited manipulative directions [26]. In contrast to these latent editing approaches, which simply generate plausible edits to static images, our algorithm attempts to automate the latent editing to quantify facial deformations in the form of StyleSpace edits.

2.2. Face Video Re-enactment

Controlling the facial attributes and their motion through facial keypoints/landmarks are popularly used in video re-enactment [15, 30, 36, 35, 25]. While these approaches provide a strict guidance over the facial attributes, they are challenged to capture fine expressive facial details (*e.g.*, teeth, lip compression, wrinkle dynamics, *etc.*) and accessories (*e.g.*, eyeglasses). Further, they are dependent on the accuracy of the landmarks and suffer in re-enactment video synthesis when the head and/or face geometries of the source and target considerably differ [29].

Approaches such as [11, 13, 14, 20, 40] employ 3D facial structural models (*e.g.*, 3DMM) to guide the synthesis, and are excellent at capturing facial movements. Despite the potential of 3D model-based approaches to generate high-quality videos, they represent only the inner-face region; thus are comparatively poor at constructing complex features such as teeth, wrinkles, complex mouth motion and require 3D training data that are resource and computation intensive.

2.3. StyleGAN-based Video Synthesis

The ability to synthesize high-resolution photo-realistic images and the rich latent space of StyleGAN are stimulating video synthesis research. MoCoGAN-HD [27] and StyleVideoGAN [12], each train a temporal architecture that is used to navigate the latent space of a pre-trained StyleGAN2 to search for temporally coherent directions for synthesizing videos at 1024^2. While the former is limited to generating random video clips, StyleVideoGAN facilitates re-enactment using a PCA-based approach to transform the learned motion trajectories to the source image. While authors of [9] propose a method to find controllable directions of the $W+$ space of StyleGAN2 with the help of a 3D model synthesizing videos at 256^2, the research of [17, 34] utilize 3D models to capture the facial motion, hence share the drawbacks of 3D structural prior based models discussed above, despite their ability to generate 1024^2 videos.

In addition to the inability of capturing the highly expressive facial attributes precisely, all these approaches attempt to learn a model that decomposes the motion-related content and hence requires a training phase. In contrast, our model extends the inherent disentangled nature of the StyleSpace (SS) of a pre-trained StyleGAN2 to achieve this decomposition in our pipeline. Further, in contrast to the above StyleGAN-based approaches which require the entire latent (18×512) per frame, the proposed framework provides an extremely compact encoding scheme comprising of 0.38% of parameters per frame (35 vs.18×512) while generating videos at 1024^2.

3. Methodology

3.1. Overview

Our approach consists of six stages: video pre-processing, GAN inversion, ID-latent selection, head-pose encoding, facial attribute encoding, and rendering. The entire flow is represented in Fig. 2.

We use the following notation to describe the pipeline. Notations beginning with L and L^{ss} denote $W+$ latents and the corresponding SS latents, respectively. L^{ss} is obtained using the affine transform $\mathcal{A}(\cdot)$, *i.e.*, $L^{ss} = \mathcal{A}(L)$. I denotes a real image and S denotes a synthesized image from a latent. For example, $S_t = G(L_t)$ describes the generation of an image from a latent, and the subscript refers to the frame at time t. G is the original StyleGAN2 generator, but it is supplemented by two style generators, G_{sf} for StyleFlow and G_{ss} for StyleSpace. Both G_{sf} and G_{ss} are derivations of G, with the difference being in the input, where G_{sf} take attribute edits such as yaw, pitch, *etc*. as input operating on the $W+$ space and G_{ss} takes StyleSpace latents as input (*i.e.*, $G(L) = G_{ss}(A(L))$). E is the e4e encoder used for real image inversion into $W+$ space. Y_t and P_t are the optimal Yaw and Pitch used by G_{sf} at time t. Finally, α_t is a 32-dimensional vector that controls the facial deformations of the generator G_{ss}, given a latent L_t.

The pre-processing stage generates a set of face images that are stabilized and aligned so that their inversion to latent space achieves maximal identity preservation and continuity of spatio-temporal head and face motions. The inversion employs the e4e encoder [28] to generate a sequence of latents, $L_1, \ldots, L_t$ in the $W+$ space corresponding to the sequence of frames. The images generated from these latents serve as the basis for rigid and non-rigid optimizations, replacing the raw image input. They enable controlled editability in conjunction with image loss metrics (see Sec. 4.2). Further details on video pre-processing and GAN inversion stages are respectively in Sec. B and C of the supplementary.

In the ID-latent selection stage, a single latent from the sequence, $L_1, \ldots, L_t$, is selected as the ID-latent, L_{ID}, which serves as the base identity for the face and head-pose deformations across the entire sequence of frames. L_{ID} is obtained using,

$$L_{ID} = \underset{L_t}{\operatorname{argmax}} \left(ID_{similarity}(I_t, G(L_t) \right). \tag{1}$$

Using a single L_{ID} as the anchor to perform head-pose and facial motion edits, not only reduces the data requirement of rendering but also minimizes the identity variation across frames. In a re-enactment setting, the image corresponding to L_{ID} functions as the source image and the sequence of frames $\{I_t\}$ function as the driving frames. Please refer to Sec. D of the supplementary for further details.

The fourth stage: head-pose encoding, finds, for each frame, the head transformation (*i.e.*, Y_t and P_t) in StyleFlow latent space needed to render L_{ID} as close as possible to the synthesized image $G(L_t)$ by minimizing,

$$\min_{Y_t, P_t} \mathcal{L} \left(G_{sf}(L_{ID}, Y_t, P_t) \ , \ G(L_t) \right) \tag{2}$$

$G_{sf}(L_{ID}, Y_t, P_t)$ results in a new latent, $LH_t \in W+$, that captures the correct head-pose at time t starting from L_{ID}.

The fifth stage: facial attribute encoding, solves for each frame, the set of facial deformations α_t in SS, that when applied to LH_t^{ss} matches as close as possible to $G(L_t)$ (where, LH_t^{ss} denotes the corresponding SS latent of LH_t obtained using $LH_t^{ss} = \mathcal{A}(LH_t)$). The result is a set of 32 parameters, α_t, that achieve $G(L_t) \approx G_{ss}(LH_t^{ss}, \alpha_t)$ through minimizing,

$$\min_{\alpha_t} \mathcal{L} \left(G_{ss}(LH_t^{ss}, \alpha_t) \ , \ G(L_t) \right). \tag{3}$$

Finally at the rendering stage, the re-enacted frame at time t is synthesized using a fixed L_{ID} and 34 style controlling parameters (plus the initial Roll angle, R_t used in pre-processing) as follows,

$$S_t = G_{ss}(LH_t^{ss}, \alpha_t) = G_{ss} \left(G_{sf}(L_{ID}, Y_t, P_t), \alpha_t \right). \tag{4}$$

3.2. Video Pre-Processing

Face alignment is an important step in StyleGAN2-based face image inversion regardless of whether an encoder or optimization approach is employed since a pre-trained generator is used. Moreover, temporal consistency of the alignment is critical due to the role each frame plays in our optimizations. Slight misalignments may alter identity, head-pose, or misinterpret facial feature attributes (shape and dynamics). The alignment used in StyleGAN2 depends on the commonly used 68 facial landmarks [19], including mouth and eye coordinates for warping. However, the eyes and mouth undergo dynamic changes in a video clip which generate jitters and rescaling in face alignment. To avoid the impact of dynamic coordinates, [12] cropped the full face excluding the eyes and mouth coordinates. We consider this insufficient to alleviate the combined effects of head-pose and facial motions. Instead, our alignment aims to:

(1) completely stabilize the head when head-pose does not change between consecutive frames, so that non-rigid face motions are captured in a maximally aligned form, (2) rely on inversion to capture the relative head alignment when the head pose rotates out-of-plane.

We employ [7] for detecting faces and tracking features in a video clip. However, the landmarks are not sufficiently accurate for face alignment over a sequence of frames. Since our objective is to only align the rigid head motion between frames, we employ a parametric optical-flow model [8] to register a frame at time t to a key frame k_i at time i ($< t$). When the rigid head motion is small or limited to the 2D plane, the registration is accurate for the duration (occasionally, several tens of frames), but upon out-of-plane head rotation, the registration requires adjusting the key frame to a new k_{i+1}. When the head out-of-plane rotation is rapid, consecutive frames may become key frames. Kindly refer to Sec. B in supplementary for further details.

3.3. Head-Pose Encoding

Temporally consistent head-pose is challenging to recover and synthesize. Head-pose is represented by three degrees of rotation, Yaw, Pitch, and Roll, computed with respect to a virtual point at the center of the head. While there are numerous landmark and mesh-based approaches for estimating head-pose, the estimate of angles from a single image is fragile and insufficient for accurate re-synthesis. Thus, in this research, we choose an analysis-by-synthesis approach to estimate the closest rendering of a latent to the target image (Eq. (2)). StyleFlow proposed an effective system for a single latent-based edit of head-pose by controlling the Yaw and Pitch angles. The Roll angle is a 2D image-based transformation and is relegated to a preprocessing step necessary for face-alignment as required by StyleGAN2.

An important feature of StyleFlow is that the attribute editing direction is dependent and conditioned on the given latent (*i.e.*, it is specific to a person and relevant attributes captured by the generator). This conditional architecture leads to improved disentangling and it also allows continuous parameter editing. Critically, the edit path is non-linear in the latent space in contrast to previous latent manipulation algorithms that rely on linear and fixed directions that apply to all latents [3].

We re-formulate the head-motion as a head-pose matching problem between a rendered image of the real-frame's encoded latent, L_t, and the rendered image of a rotated L_{ID} which is solved as a minimization problem (Eq. (2)). The minimization employs two losses, L2 and LPIPS [39] to search the Yaw-Pitch space using gradient descent. These losses are computed over a masked area of the face that is based on an 81-landmark model (an extension of the 68 landmarks model to include the forehead). However, the eyes, mouth, and eyebrows are excluded in the L2 loss, since these non-rigid areas are not relevant to 3D head rotations. The outcome of this stage is an alignment of the L_{ID} to match the head-pose at time t, and it is represented by a new latent LH_t (in $W+$) that will be further edited to capture the non-rigid motions of the eyebrows, eyes, mouth, and chin.

3.4. Facial Attribute Encoding

The facial attribute encoding extends [33], where the authors demonstrate the highly disentangled nature of the SS. The facial-attribute encoding, α_t, (32 parameters) of each frame is applied to the latent LH_t^{ss}, which is a transformation of LH_t to SS via $LH_t^{ss} = \mathcal{A}(LH_t)$.

Choice of StyleSpace Indices: The StyleSpace indices are analyzed to make sure that maximally disentangled indices that capture complex and detailed expressive facial attributes as shown in Figs. 1 and 3 are selected. For a specific facial feature $f \in \mathcal{F}$, we score each index $i \in \{l, c\}$ using index sensitivity, $\Gamma_{f,i}$, which measures the change in image space for a unit change in the StyleSpace index. $\Gamma_{f,i}$ is defined as,

$$\Gamma_{f,i} = \frac{1}{|\{\delta_k\}|} \sum_k \left(\frac{\mathcal{L}_{LPIPS}(S_k * M, S_{k-1} * M)}{|\delta_k - \delta_{k-1}|} \right), \quad (5)$$

where $S_k = G_{ss}(L_{ID}^{ss} + \delta_k \mathbb{1}_i)$ is the synthesized image generated using L_{ID} perturbed by δ_k at SS index i, M is the binary mask over the facial attribute considered, and $\mathbb{1}_i = \{1$ when $(l, c) = i$; 0 elsewhere$\}$. We choose $\{\delta_k\}$ to be a sequence of successive values with $|\{\delta_k\}|$ elements, and the subscript k indicates the iterating index. Additionally, we calculate the index sensitivity over the whole face (*i.e.*, M is a matrix of ones that covers the whole face) and is denoted by Γ_i. Subsequently, we rank the indices based on $\Gamma_{f,i}$ and Γ_i values and choose the indices that have a higher $\Gamma_{f,i}$ and a negligible Γ_i based on simple thresholding. We repeat the scoring on multiple subjects and frames sampled from the dataset and obtain the prominent indices across the sampled data. This novel approach enables the selection of maximally disentangled StyleSpace indices corresponding to the specific facial attribute chosen. The list of facial attributes $\mathcal{F}$ and the set StyleSpace indices, thus chosen (denoted as $\mathcal{V}$), are tabulated in Tab. 1 of the supplementary.

The significance of our SS indices selection process as opposed to the algorithm proposed in [33] is as follows. We observed that the StyleSpace, SS representation is not unique. *i.e.*, optimizing

$$\min_{\alpha_{inv_t}} \mathcal{L}\left(G_{ss}(LH_t^{ss} + \alpha_t + \alpha_{inv_t}), G_{ss}(LH_t^{ss})\right) \quad (6)$$

does not necessarily yield $\alpha_t + \alpha_{inv_t} \approx 0$. Therefore, as [33] back propagates to compute the gradient with respect to an SS index, the gradients are less accurate, as the

SS indices contributing to an identical facial deformation of two frames would differ (as not unique). Instead, we use a forward approach, perturbing each index separately and computing the corresponding deformation loss, thus directly computing the true gradient (sensitivity in the image space for a unit change of each SS index) which is more accurate.

Facial Deformation Attribute Encoding: We compute the optimal encoded latent values, α_t, that edit facial attributes to capture the facial deformations. α_t represents the offset values from LH_t^{ss} and is obtained through a per-frame optimization (Eq. (3)) over the SS indices and is presented in Algorithm 1. The reconstruction of the latent L_t obtained from the e4e encoder is used as the driving frame in the optimization and denoted by S_t, while the rendered re-enacted frame during the optimization is denoted by $\hat{S}_t$.

Initialization of indices ($\boldsymbol{LH_t^{ss}}$): Due to the sparsity of the latent space and as the optimization is over a multi-dimensional space, it is highly probable for the optimization algorithm to converge consecutive frames, which are nearby in image-space, onto local-minima that are distant in the latent space. The slight differences in the optimum point of consecutive frames could introduce jitter in re-enactment. Therefore, to bias the algorithm to solve for α_t in the vicinity of the previous frame's optimum, we initialize the SS indices that we optimize, $i = (l, c) \in \mathcal{V}$ of LH_t^{ss} as,

$$LH_t^{ss}(l, c) = LH_{t-1}^{ss}(l, c), \forall (l, c) \in \mathcal{V}. \tag{7}$$

Index-specific learning rate, $\boldsymbol{\eta_{f,i}}$: We observed that different subjects and indices have different sensitivities to a unit change in the StyleSpace ($\Gamma_{f,i}$) (see Sec. E.3 in supplement). This observation corroborates the non-linear nature of latent editing discussed in StyleFlow and the non-homogentiy of latent spaces discussed in [21]. Hence, using the same learning rate across all indices would result in an undue dominance of high-sensitivity indices, thus generating non-optimal results. Therefore, for each input video and each facial attribute, we compute the index-specific learning rate,

$$\eta_{f,i} = \exp\left(-1.5\,\Gamma_{f,i}\,/\max_{i\in\mathcal{V}_f}\left(\Gamma_{f,i}\right)\right), \tag{8}$$

that was obtained empirically. For each epoch, optimization is done in parallel for all the attributes and the optimization over indices corresponding to the gaze is skipped for frames where blinking is detected.

Loss Functions: The algorithm is optimized by minimizing over multiple losses. The total loss is defined as,

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_e + \mathcal{L}_p + \mathcal{L}_{ID} + \mathcal{L}_{FP}, \tag{9}$$

where the loss terms $\mathcal{L}_{ID}$ and $\mathcal{L}_{FP}$ represent the identity loss and the Face-Parsing loss respectively and the subscripts m, e, and p correspond to the losses computed over extracted regions of the {mouth + chin/ jaw}, {eyes + eyebrows}, and {pupil}, respectively.

$$\mathcal{L}_m = \mathcal{L}_{LPIPS_m} + \mathcal{L}_{L2_m} + \mathcal{L}_{IF_m}, \tag{10}$$

$$\mathcal{L}_e = \mathcal{L}_{LPIPS_e} + \mathcal{L}_{L2_e} + \mathcal{L}_{IF_e}, \tag{11}$$

$$\mathcal{L}_p = \mathcal{L}_{L2_p} + \mathcal{L}_{IF\_L2_p}, \tag{12}$$

where $\mathcal{L}_{LPIPS}$, $\mathcal{L}_{L2}$, and $\mathcal{L}_{IF}$ represent the LPIPS loss, L2 loss, and Inter-frame loss, respectively. Please refer Sec. E.4 in supplementary for further details.

Algorithm 1: Optimization Flow for frame t

```
Inputs:
  • Head-pose adjusted $W+$ latents: $LH_t$ and $LH_{t-1}$,
  • Target frames: $S_{t-1}$ and $S_t$,
  • Rendered frames: $\hat{S}_1$ and $\hat{S}_{t-1}$,
  • StyleSpace of $t-1$: $LH_{t-1}^{ss}$ and $\alpha_{t-1}$.
  • Optimizer $F'$, $N$ number of epochs, and $G_{ss}$
Initialization:
  • Obtain the StyleSpace latent, $LH_t^{ss} = \mathcal{A}(LH_t)$
  • Initialize $LH_t^{ss}(l, c)$, $\forall i = (l, c) \in \mathcal{V}$
  • $\alpha_t = [0, \ldots, 0]$
  • Compute the index-specific learning rates, $\eta_{f,i}$
        $\eta = \{\eta_{f,i};\ \forall f \in \mathcal{F}, i \in \mathcal{V}\}$
Optimization:
for n = [1:N] do
    $\hat{S}_t = G_{ss}\{LH_t^{ss} + \alpha_t \mathbb{1}_i\}$
      where $\mathbb{1}_i = \{1$ when $(l, c) \in \mathcal{V}$; $0$ elsewhere$\}$
    $\mathcal{L} = \mathcal{L}\{\hat{S}_1, \hat{S}_{t-1}, \hat{S}_t, S_{t-1}, S_t\}$
    $\alpha_t \leftarrow \alpha_t - \eta F'(\nabla_{\alpha_t}\mathcal{L}, \alpha_t)$
end
Output:
  • 32-dimensional $\alpha_t$
```

3.5. Rendering

Once the encoding is complete, the L_{ID} and the time-series of the 35 parameters, $\{\alpha_t, Y_t, P_t, R_t\}$ are transmitted to the renderer. To synthesize the re-enactment video, first LH_t is obtained from L_{ID} to adjust for the head-pose using StyleFlow for each frame. Then LH_t is transformed to $LH_t^{ss} \in SS$, on to which the 32 indices responsible for the facial attributes, α_t are applied to synthesize the image using the generator, G_{ss}.

$$\hat{S}_t = G_{ss}(LH_t^{ss} + \alpha_t \mathbb{1}) \tag{13}$$

4. Experiments and Results

4.1. Dataset and Evaluation

We selected 150 video clips (4K videos) from the video-sharing site `www.pexels.com` that combine high visual quality with expressive head and facial motions that are

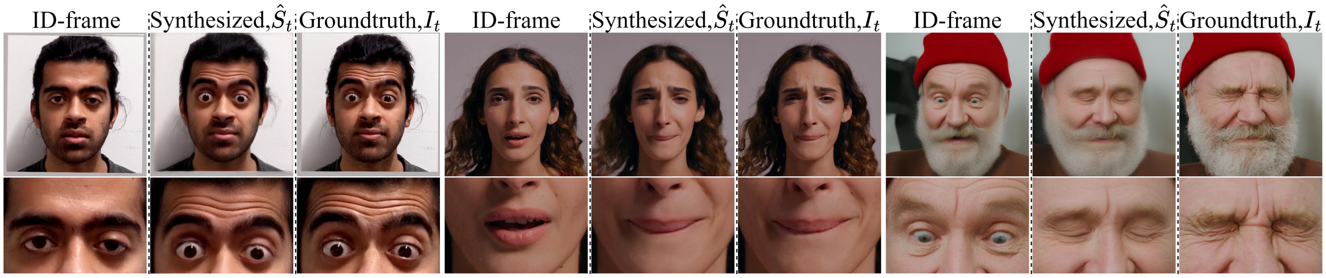

Figure 3. **Qualitative examples yielded through our approach** (in addition to Fig. 1). The StyleSpace indices and the optimization procedure were carefully designed such that complex and fine facial details such as lip-pressing, mouth puckering, mouth gaping, and wrinkles around the eyes, mouth, nasal-bridge, and forehead are well-captured.

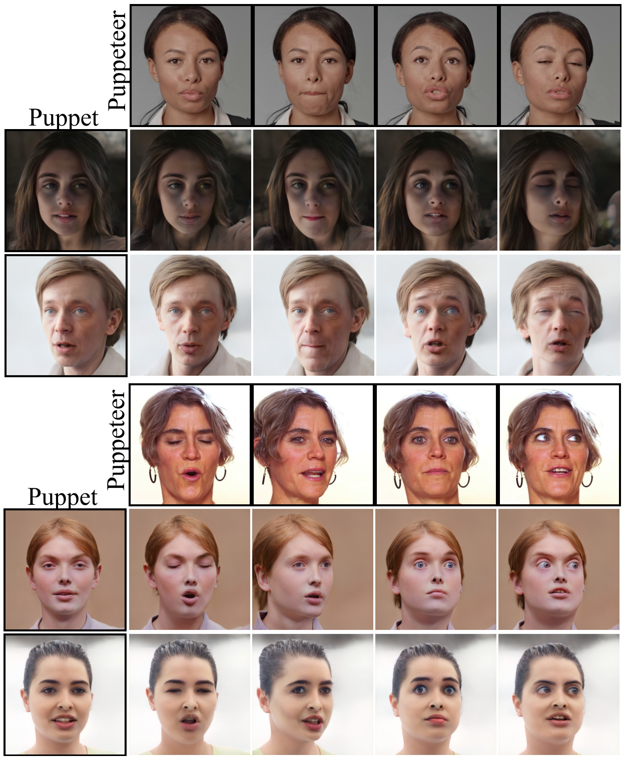

Figure 4. **Qualitative evaluation of puppeteering**, where the encoded parameters of the puppeteer are applied to the ID-latent of the puppet. It could be observed that even complex facial deformations are transferred well across different identities.

present in common low-resolution datasets. Each video contains a single face performing significant face deformations, head motion, and speech. Additional details of the dataset are included in Sec. F.1 in supplementary. There exists an inherent quality loss in the GAN inversion stage as the real-world subjects would mostly be out-of-domain of StyleGAN resulting in notable deviations between the e4e encoded frames and real frames. Thus, to improve the photo-realism of the initial GAN inversions while maintaining editability, PTI [23] was used. Kindly refer Sec. C of the supplementary.

We compare our results against two SOTA StyleGAN2-based models (most relevant): StyleHEAT and StyleVideoGAN, a latent-based model: LIA, and two other SOTA models (keypoint/landmark-based) that facilitate re-enactment: fs-vid2vid and FOMM. Publicly available models were used for all algorithms except StyleVideoGAN for which the authors kindly processed six videos. Note: All algorithms were evaluated at their native resolution using multiple metrics scoring: spatial quality, spatio-temporal quality and appearance, and temporal consistency of identity (details of metrics in Sec. F.2 in supplementary).

Referring to the top of Tab. 1, we achieve state-of-the-art performance at 1024^2 with significantly improved re-synthesis results compared to StyleGAN2-based models, StyleVideoGAN and StyleHEAT while utilizing only 0.38% of the latent space parameters used by them (35 vs. 18×512 per frame). Moreover, our approach also outperforms fs-vid2vid, FOMM, and LIA in all scores by large margins. It is critical to note that lower native resolutions [25, 30, 31] significantly favor several metrics since there is no penalty for loss of details (*e.g.*, L1, SSIM, FID, FVD, *etc.*) with respect to 1024^2 metrics. Hence it is essential to emphasize on the qualitative analysis which more accurately reflects the potential of our framework.

Figs. 1 and 3 illustrate, qualitatively, the capturing of fine facial details such as lip pressing, mouth puckering and gaping, dynamic wrinkles around the eyes, mouth, nasal-bridge, and forehead, *etc.* enhancing photo-realism of the re-enacted videos which are not necessarily captured by the metrics (see supplementary figures and videos for more examples). To the best of our knowledge, such fine expressive details were not explicitly addressed by previous research.

Similarly, as shown in Tab. 2, we achieve the best puppeteering results across all metrics. Further, Fig. 4 demonstrates the versatility of our method as even complex facial attribute deformations (*e.g.*, lip pressing, puckering, *etc.*) of the driving frames are transferred successfully to the puppet frame through the proposed framework.

4.2. Ablation Study

As ablations, we study several design choices in our pipeline, namely: the use of a different GAN inversion en-

Method	res.	L1 ↓	LPIPS↓	$\mathcal{L}_{ID}$↓	PSNR↑	SSIM↑	FID↓	FVD↓	ρ_{AU}↑	ρ_{GZ}↑	ρ_{pose}↑
FOMM	256^2	3.07	0.036	0.174	31.0	0.932	28.7	140.3	0.710	0.755	0.648
LIA	256^2	3.24	0.042	0.164	30.0	0.929	30.2	162.9	0.546	0.693	0.619
fs-vid2vid	512^2	5.75	0.093	0.158	25.2	0.900	42.4	359.6	0.571	0.784	0.629
StyleHEAT	**1024^2**	4.13	0.097	0.134	27.6	0.933	25.1	281.9	0.673	0.701	0.763
Ours	**1024^2**	**1.99**	**0.030**	**0.097**	**34.2**	**0.963**	**15.9**	**85.2**	**0.771**	**0.834**	**0.880**
StyleVid.GAN *	**1024^2**	4.04	0.109	0.104	28.8	0.926	28.8	223.3	0.739	0.884	0.979
Ours*	**1024^2**	**1.96**	**0.026**	**0.067**	**34.1**	**0.960**	**13.6**	**79.8**	**0.899**	**0.971**	**0.987**
Ours (ReStyle)	1024^2	2.01	0.031	0.099	34.0	0.959	16.9	93.9	0.767	0.831	0.843
Ours – PTI	1024^2	2.71	0.048	0.127	32.0	0.956	23.2	125.7	0.726	0.819	0.833

Table 1. **Quantitative comparison of video re-synthesis against baselines.** *Top* consist of metrics evaluated against the dataset of 150 videos. *Middle* includes scores computed over 6 videos received upon requests to authors. *Bottom* consists of ablation results evaluating the effect of using a different encoder and the generator fine-tuning stage. We yield SOTA results at 1024^2 on all metrics while using only 0.38% of latent space parameters of StyleGAN2.

Method	res.	$\mathcal{L}_{ID}$↓	FID↓	FVD↓	FVD_M↓	ρ_{AU+GZ}↑
FOMM	256^2	0.153	77.0	396.8	103.0	0.501
LIA	256^2	0.174	82.3	406.0	112.4	0.527
fs-vid2vid	512^2	0.202	73.6	445.1	112.7	0.640
StyleHEAT	**1024^2**	0.181	81.0	437.5	109.8	0.667
Ours	**1024^2**	**0.095**	**63.9**	**386.5**	**82.3**	**0.708**

Table 2. **Quantitative comparison of puppeteering against baselines** evaluated across 50 puppet-puppeteer pairs. Our approach achieves the best performance across all metrics.

Method	Vid.1	Vid.2	Vid.3	Vid.4	Vid.5
StyleFlow	46.7	41.3	33.7	17.0	38.5
Ours	**16.0**	**19.9**	**16.3**	**11.5**	**21.9**

Table 3. **Quantitative comparison of our approach vs. straightforward head-pose adjustment using StyleFlow.** The mean head-pose loss (lower ↓ the better) of a few videos are tabulated.

coder, the significance of the head-pose encoding approach, using real frames as reference in facial attribute optimization, and the effect of PTI.

Using ReStyle encoder [5] replacing e4e generates comparable results (Tab. 1) implying that the proposed scheme is functional irrespective of the encoder provided that the inversion is within the editable domain of the latent space.

Further using real frames $\{I_t\}$ as reference for the facial attribute encoding optimization (Sec. 3.4) instead of the synthesized frames $\{S_t\}$ resulted in visually sub-optimal results requiring us to abandon tighter pixel-level metrics as $\mathcal{L}_{L2}$, which are essential in capturing fine facial details such as wrinkles, gaze, *etc*. Hence, we opted to use $\{S_t\}$ for the optimization stage. We suspect this behavior to be caused due to the natural noise present in real images to which the StyleSpace optimization might be sensitive to.

Even though StyleFlow is capable of directly generating a head-pose adjusted latent, provided $\{Y_t, P_t\}$, the quantified estimates of head-pose (using OpenFace) for a video stream are not sufficiently accurate to render using StyleFlow, resulting in inaccurate poses and significant jitter. Our synthesis-based optimization approach based on losses in image-space generates more accurate head-pose images consistent with reference frames (Tab. 3 and supplementary video).

It could be observed that the re-synthesis results without PTI (**Ours**-PTI in Tab. 1) yet outperform all baselines in almost all scores. The performance improvement seen with PTI is due to the tendency of real-world subjects to be out of the domain of StyleGAN and the inherent loss of the encoder used during the GAN inversion stage.

4.3. Limitations

Despite the promising results, the proposed approach has a few limitations. As the pipeline is based on the StyleGAN2 architecture, it inherits the limitations from StyleGAN2 and its inversion methods (*e.g*., fixed resolution, alignment requirements, *etc*.). Further, the encoding pipeline is sensitive to occlusions resulting in visual artifacts in the synthesized images. Additionally, certain scenarios with extreme facial deformations and profile views could yet be challenging, which stems from the low representation of the FFHQ dataset used in training StyleGAN2.

5. Conclusion

We extend the StyleGAN2's photo-realism and disentanglement of its StyleSpace spatio-temporally, to propose a novel end-to-end pipeline for latent-based expressive face video encoding, which enables high-fidelity (1024^2) video re-enactment using a single $W+$ latent and 35 parameters per frame. Our algorithm achieves state-of-the-art performance while using a fraction (0.38%) of parameters compared to StyleGAN2 latent-based approaches. To the best of our knowledge we are the first to (1) automate latent space editing (that was previously used to merely generate plausible facial edits) to capture extremely fine, rich, and complex facial deformations, and (2) to propose an extremely compact latent-based face video encoding scheme based on StyleGAN2 enabling re-enactment. The negative societal impact is discussed in Sec. G 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, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020.

[3] 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.

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *arXiv preprint arXiv:2102.02754*, 2021.

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

[6] 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.

[7] 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.

[8] M.J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision 25, 23–48*, 1997.

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

[10] Chris Donahue, Zachary C Lipton, Akshay Balsubramani, and Julian McAuley. Semantically decomposing the latent spaces of generative adversarial networks. *arXiv preprint arXiv:1705.07904*, 2017.

[11] 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.

[12] Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan. *arXiv preprint arXiv:2107.07224v1*, 2021.

[13] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[14] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018.

[15] Kuangxiao Gu, Yuqian Zhou, and Thomas Huang. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10861–10868, 2020.

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

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

[18] 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.

[19] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[20] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. *ACM Transactions on Graphics (TOG)*, 37(6):1–12, 2018.

[21] Trevine Oorloff and Yaser Yacoob. One-shot face video reenactment using hybrid latent spaces of stylegan2. *arXiv preprint arXiv:2302.07848*, 2023.

[22] 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.

[23] 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.

[24] 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*, pages 9243–9252, 2020.

[25] 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.

[26] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6142–6151, 2020.

[27] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*, 2021.

[28] 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.

[29] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Single source one shot reenactment using weighted motion from paired feature points. *arXiv preprint arXiv:2104.03117*, 2021.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468, 2019.

[37] 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.

[38] lei Zhang and Chris Pollett. Facial expression video synthesis from the StyleGAN latent space. In Xudong Jiang and Hiroshi Fujita, editors, *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*, page 7, Singapore, Singapore, June 2021. SPIE.

[39] 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.

[40] 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.