Encode-in-Style: Latent-based Video Encoding using StyleGAN2

Trevine Oorloff and Yaser Yacoob

University of Maryland, College Park MD 20742, USA
{trevine,yaser}@umd.edu

Fig. 1: Re-synthesis examples of complex head and facial motions captured using our approach in comparison to fs-vid2vid [38] and FOMM [32]

Abstract. We propose an end-to-end facial 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. It economically captures face identity, head-pose, and complex facial motions at fine levels, and thereby bypasses training and person modeling which tend to hamper many re-synthesis approaches. The approach 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). The project page is located at https://trevineoorloff.github.io/Encode-in-Style.

Keywords: video resynthesis, video encoding, StyleGAN, image inversion, latent space editing
Fig. 2: The seven-stage pipeline for encoding a video in latent space: (1) pre-processing, (2) GAN inversion, (3) ID-latent selection, (4) head-pose encoding, (5) facial attribute encoding, (6) generator fine-tuning, and (7) rendering

1 Introduction

Recent advances in GAN inversion techniques \cite{1,5,6,30} enable manipulation of real-world images \cite{2,3,4,20,31,41}. Image manipulation through the latent space is dominated by the StyleGAN architectures since the space is less entangled. Quantitative evaluation of latent spaces: \(Z, W, W^+, \) and \(S\), by Wu et al. \cite{41}, indicates that within the StyleGAN’s latent spaces, the proposed style-space (denoted StyleSpace, \(SS\)) has the best disentanglement, completeness, and informativeness. Thus, \(SS\) enables control of individual facial attributes without re-training a network to enforce disentanglement \cite{13}.

Manipulation of videos using latent representations is a challenging problem since spatio-temporal and identity incoherence are highly noticeable. When viewing a video, after the identity of the person is established, humans mostly focus their attention on changes in the facial features (e.g., movements in mouth, eyes, jaw, head-pose, etc.) rather than the features that remain constant such as the facial geometry, hair, ears, etc. We build on this observation to extend StyleGAN2 architecture \cite{23} and high-quality head-pose \cite{3} and face editing \cite{41} approaches to create a novel latent-based facial video encoding that enables high-resolution and data-efficient facial video re-synthesis.

We focus on re-synthesis and re-enactment of videos using a compact encoding scheme. In re-synthesis, we encode a video using a low-dimensional representation of small edits of an Identity-latent (ID-latent). Our approach is capable
of capturing and regenerating complex facial features as shown in Fig. 1 while achieving state-of-the-art re-synthesis performance at $1024^2$. Further, since the encoding is independent of the subject in the video, we can substitute the ID-latent (e.g., an inversion of a real face) of a different subject and apply the face deformation parameters to generate high fidelity re-enactment video (i.e., puppeteering).

We extend single image generation models, namely StyleGAN2, StyleFlow, and StyleSpace to the temporal dimension. 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 limitations of common linear latent editors.

Video clip encoding is extremely compact, a single latent ($18 \times 512$) 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). The 70 bytes per frame are nearly half the state-of-the-art (see Table 5 [39]). The key contributions of the paper are:

- a novel algorithm for high-resolution ($1024 \times 1024$) facial video encoding for re-synthesis and re-enactment,
- an extremely compact encoding of head and face motions,
- a generative approach that employs optimization frameworks instead of person-centric data training,
- a framework employing image inversion to anchor the sparse latent-edits to image-based constraints, facilitating accurate spatio-temporal modeling,
- a novel method to find StyleSpace channels corresponding to facial attributes based on index sensitivity,
- a high-quality (4K) video dataset.

We focus on automating the editing of latent spaces in contrast to the prevailing work on latent-space editing that illustrates plausible semantic visual results (e.g., smiles, hair color, gaze) [3,27,41]. Specifically, we compute low-dimensional edits that closely replicate complex and visually-entangled facial motions that are consistent with real face deformations.

2 Related Work

2.1 GAN-based Video Synthesis

The algorithms proposed in [15,34,39,46] are the most related to our work. Wang et al. in [39] proposed a one-shot approach for facial video synthesis (both reconstruction and re-enactment with a maximum resolution of $512^2$) using a novel 3D keypoint representation of faces, where the information related to the identity and motion is decomposed. This concise representation enables transmitting the facial video using a low data rate. On the other hand, the works of [15,34] are focused on generating high resolution ($1024^2$) videos using a fixed pre-trained
image generator (StyleGAN2) and navigating the latent space. The MoCoGAN-HD \cite{34} trains a temporal generator that generates plausible motion trajectories in the latent space ($Z$) to generate realistic videos. The algorithm proposed in StyleVideoGAN \cite{15} trains a temporal architecture using a single video to learn the temporal correlations in the latent space ($W^+$) of a pre-trained StyleGAN, and then generate videos.

These approaches attempt to learn a model that decomposes a video to its identity-specific content and motion-related content and hence require a training phase. In contrast, our model extends the inherent disentangled nature of the StyleSpace (SS) \cite{41} and the versatility of StyleFlow \cite{3} in the StyleGAN2-based architectures to achieve this decomposition in our video encoding.

### 2.2 2D/3D Face Video Re-synthesis

Controlling the facial attributes and their motion through facial key-points/landmarks \cite{7,18,38,44,45} are used for video encoding. While landmark-based approaches are mostly subject-agnostic, they are challenged to capture fine facial details (e.g. hair, teeth, lip compression, etc.) and accessories (e.g., eye-glasses). Further, they are dependent on the accuracy of the landmarks and suffer in re-enactment video synthesis when the head and/or facial geometries of the source and target considerably differ \cite{36}. Wang et al. \cite{39} sought to address the latter through a novel 3D keypoint representation that decomposes the person-specific and motion-related information.

Approaches such as \cite{14,16,17,26,48} employ 3D facial models (e.g. 3DMM \cite{10}) to guide the synthesis, and are excellent at capturing facial movements. HeadGAN \cite{14} introduced a novel synthesis model that is conditioned on the 3D face representation to disentangle identity from facial expressions. This facilitates re-enactment as well as expression and pose editing in videos. 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 surrounding facial features (e.g. hair) or complex features such as teeth and hair, and require 3D training data that are resource and computation intensive.

### 2.3 Latent Space Based Editing

Understanding the latent space of a pre-trained GAN has led to better controllability over the generated output. Härkönen et al. \cite{20} suggest a PCA-based approach applied onto the latent space of StyleGAN \cite{22} to identify key control directions for semantic edits such as aging, lighting, etc. InterFaceGAN \cite{31} maps the latent space with the semantics of the output images of StyleGAN using classifiers and identified a few disentangled controls for facial attributes such as age, gender, pose, and smile. However, the entangled nature of the latent space limits the manipulation, as it often leads to undesirable artifacts.

StyleSpace \cite{41}, StyleFlow \cite{3}, and StyleRig \cite{33} are few prominent algorithms based on the StyleGAN2 architecture that yield impressive control over
latent-based manipulations. The authors of StyleSpace analyzed the intermediate latent spaces for disentanglement, completeness, and informativeness and then formulated an algorithm to identify the style channels that control specific attributes. StyleFlow, on the other hand, uses a flow-based model conditioned on the attributes to enable non-linear and conditioned latent space edits. StyleRig argues that their algorithm enables a rig-like control over the 3D semantic parameters (obtained through 3DMM) of faces generated through StyleGAN.

3 Methodology

Our approach consists of seven stages: video pre-processing, GAN inversion, ID-latent selection, head-pose encoding, facial attribute encoding, generator fine-tuning, and rendering. The entire flow is represented in Fig. 2 and utilizes the e4e encoder [35], StyleFlow [3], StyleSpace [41], and PTI [30] (with significant changes to these components to achieve our objectives).

We use the following notation to describe the pipeline. Notations beginning with $L$ and $L_{ss}$ denote $W+$ latents and the corresponding StyleSpace latents, respectively. $L_{ss}$ is obtained using the affine transform $A(\cdot)$, i.e., $L_{ss} = A(L)$. $I$ denotes a real image and $S$ denotes a synthesized image. $G$ is a generator of an 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 three style generators, $G_{sf}$ for StyleFlow, $G_{ss}$ for StyleSpace, and $G_{ss}^\dagger$ for fine-tuned $G_{ss}$. $E$ 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, $\alpha_t$ 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 to generate a sequence of latents in $W+$ space that capture identity and head and face motions. These latents, $L_1, \ldots, L_t$, 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. It is important to note that optimization constraints are applied in the image space and not in the latent space, so in every iteration, the respective generator is employed in image synthesis.

In the third stage, a single latent from the sequence is selected as an ID-latent, $L_{ID}$, for generating the various head-poses of the person in the video.

$$L_{ID} = \arg\max_{L_t} (ID_{similarity}(I_t, G(L_t)))$$

(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. The fourth stage finds, for each frame, the head transformation (i.e., $Y_t$ and $P_t$) in StyleFlow latent style-space needed to render
\(L_{ID}\) as close as possible to the synthesized image \(G(L_t)\) (i.e., with respect to the e4e encoded latent), by minimizing,

\[
\min_{Y_t, P_t} \mathcal{L}\{ G_{sf}(L_{ID}, Y_t, P_t) \ , \ G(L_t) \}. \tag{2}
\]

\(G_{sf}(L_{ID}, Y_t, P_t)\) results in a new latent, \(LH_t\), that captures the correct head-pose at time \(t\) starting from \(L_{ID}\). The fifth stage solves, for each frame, the set of facial deformations \(\alpha_t\) in StyleSpace, that when applied to \(LH_t^{ss}\) matches as close as possible to \(G(L_t)\). The result is a set of 32 parameters that achieve \(G(L_t) \approx G_{ss}(LH_t^{ss}, \alpha_t)\) through minimizing,

\[
\min_{\alpha_t} \mathcal{L}\{ G_{ss}(LH_t^{ss}, \alpha_t) \ , \ G(L_t) \}. \tag{3}
\]

Inspired by the concept proposed in PTI \[30\], the sixth stage fine-tunes \(G_{ss}\) producing \(G_{ss}^f\) with the objective of improving photo-realism of StyleGAN’s out-of-domain subjects by minimizing,

\[
\min_{\theta} \sum_t \mathcal{L}\{ G_{ss}(LH_t^{ss}, \alpha_t ; \theta) \ , \ I_t \} \tag{4}
\]

Finally, an image is synthesized by applying,

\[
S_t = G_{ss}^f(LH_t^{ss}, \alpha_t) = G_{ss}^f(G_{sf}(L_{ID}, Y_t, P_t), \alpha_t). \tag{5}
\]

Thus, re-synthesizing an image at time \(t\) using a fixed \(L_{ID}\) and 34 style controlling parameters (plus the initial Roll angle, \(R_t\) used in pre-processing which could be re-applied since it is in 2D image space).

### 3.1 Video Pre-Processing

Face alignment is an important step in face image inversion regardless of whether an encoder or optimization approach is employed since StyleGAN2 is a fixed resolution architecture (1024\(^2\) pixels). Temporal consistency of the alignment is critical due to the role each frame plays in our optimizations. Slight misalignments may change identity, head-pose, or misinterpret facial feature attributes (shape and dynamics). The alignment used in StyleGAN2 depends on the commonly used 68 facial landmarks, including mouth and eye coordinates for warping. However, these undergo dynamic changes in a video clip which generate jitters and rescaling in face alignment. To avoid the impact of dynamic coordinates, Fox et al. \[15\] 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) when the pose of the head rotates out-of-plane it is better to rely on inversion to capture the relative head alignment.
We employ [8] 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. Dense optical-flow captures a combination of rigid and non-rigid facial motions. However, since our objective is to only align the rigid head motion between frames, we employ a parametric optical-flow model [9] to register a frame at time $t$ to a Key Frame $k_i$ at time $i$ ($i < t$). 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 (see Sec. 3 in Appendix for further details).

3.2 GAN Inversion

Two factors were considered in choosing an appropriate GAN inversion method: (1) faithful representation of the given image (i.e. minimal reconstruction loss), (2) ability to facilitate latent space edits. Tov et al. in [35] suggested that there exists a trade-off between these two factors, i.e. distortion and editability. Generally, inversion is done using a trained encoder and/or an optimization framework. While the former has better editability, it has a comparatively high reconstruction loss and vice versa. Recently, [30] proposed to bridge the gap between the two trade-offs by fine-tuning the generator but this adds computational and information transmission costs. We chose the e4e encoder [35], which was designed to facilitate the inversion of real images in proximity to the regions StyleGAN was trained on, thus mitigating the trade-off.

3.3 Identity-Latent Selection

The per-frame inversion creates a series of latents. Depending on the extent of head motion, deformation in StyleGAN2 space is likely. Therefore, we select the latent that matches best the identity of the source frame. We use ArcFace [12] to compute the similarity between the source and the reconstructed images of the face (Eq. 1). The closest of the face-matches that is also near frontal views of the person is chosen as the representative $L_{ID}$, the basis for re-synthesis (see Sec. 3 in Appendix for further details).

3.4 Head-Pose Encoding

Temporally consistent head-pose is challenging to recover and re-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. There are numerous landmark and mesh-based approaches for estimating head-pose. Instead, 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 pre-processing 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 the state-of-the-art that relies on linear and fixed directions that apply to all latents.

We re-formulate the head-motion as a head-pose matching problem between a rendered image of the encoded latent of the real frame $I_t$, i.e., $G(L_t)$, 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 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. Fig. 3 illustrates the optimization of Yaw and Pitch to minimize the loss between the encoded image at time $t$ and a rotated $L_{ID}$ with estimated $Y_t$ and $P_t$ rendered using StyleFlow.

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.

### 3.5 Facial Attribute Encoding

The facial attribute encoding extends [41], where the authors demonstrate the highly disentangled nature of the StyleSpace and provide a few StyleSpace indices that have mostly disentangled control over facial attributes. The facial-attribute encoding, $\alpha_t$, (32 parameters) of each frame is applied to the latent $LH_t$ to StyleSpace via $LH_t^{ss} = A(LH_t)$.

#### Choice of StyleSpace Indices

The StyleSpace indices are analyzed to make sure that maximally disentangled indices that capture complex and detailed
Table 1: StyleSpace indices corresponding to the deformation of facial attributes. The indices take the form of \{l; c_1, c_2, \ldots \}, where \(l\) and \(c\) denote the layer index and channel index of the StyleSpace.

| Facial Attribute, \(F\) | StyleSpace Indices, \(\mathcal{V}\) |
|-------------------------|----------------------------------|
| Mouth                   | \{6: 113, 202, 214, 259, 378, 501\}, \{8: 17, 387\}, \{14: 12\}, \{11: 6, 41, 78, 86, 313, 361, 365\}, \{15: 45\} |
| Chin/Jaw                | \{5: 50, 505\}, \{6: 131\}, \{8: 390\} |
| Eyes                    | \{9: 63\}, \{11: 257\}, \{12: 82, 414\}, \{14: 239\}, \{17: 28\} |
| Eyebrows                | \{8: 6, 28\}, \{9: 30\}, \{11: 320\} |
| Gaze                    | \{9: 409\} |

Facial deformation attributes as shown in Fig. [4] are selected. For a specific facial feature \(f \in F\), 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,

\[
\Gamma_{f,i} = \frac{1}{|\{\delta_k\}|} \sum_k \left\{ \frac{L_{LPIPS}(S_k \ast M, S_{k-1} \ast M)}{\|\delta_k - \delta_{k-1}\|} \right\},
\]

where \(S_k = G_{ss}(L_{ID}^{st} + \delta_k \mathbb{1}_i)\) is the synthesized image generated using \(L_{ID}\) perturbed by \(\delta_k\) at StyleSpace index \(i\), \(M\) is the binary mask over the facial attribute considered, and \(1_i = \{1 \text{ when } (l, c) = i; 0 \text{ elsewhere}\}\). 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 \(\Gamma_i\). Subsequently, we rank the indices based on \(\Gamma_{f,i}\) and \(\Gamma_i\) values and choose the indices that have a higher \(\Gamma_{f,i}\) and a negligible \(\Gamma_i\) 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 \(F\) and the set StyleSpace indices \(\mathcal{V}\), thus chosen are given in Tab. [1].

**Facial Deformation Attribute Encoding** We compute the optimal encoded latent values, \(\alpha_t\), that edit facial attributes to capture the facial deformations. \(\alpha_t\) represents the offset values from \(LH_t^{st}\) and is obtained through a per-frame optimization (Eq. (3)) over the StyleSpace indices and is presented in Algorithm [1] in the Appendix. The reconstruction of the latent \(L_t\) obtained from the e4e encoder (Sec. 3.2) is used as the target image in the optimization and denoted by \(S_t\), while the rendered frame during the optimization is denoted by \(\hat{S}_t\).

**Initialization of indices (\(LH_t^{st}\))**: 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 to 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
the re-synthesis. Therefore, to bias the algorithm to solve for $\alpha_t$ in the vicinity of the previous frame’s optimum, we initialize the StyleSpace indices we optimize, $i = (l, c) \in V$ of $LH^{ss}_t$ as,

$$LH^{ss}_t(l, c) = LH^{ss}_{t-1}(l, c), \forall (l, c) \in V.$$  \hspace{1cm} (7)

**Index specific learning rate, $\eta_{f,i}$:** We observed that different subjects and indices have different sensitivities to a unit change in the StyleSpace ($\Gamma_{f,i}$) (see Sec. E.2 in Appendix). This observation corroborates the non-linear nature of latent editing discussed in StyleFlow. 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 an index-specific learning rate using,

$$\eta_{f,i} = \exp \left\{ -\frac{3}{2} \frac{\Gamma_{f,i}}{\max_{i \in V_f}(\Gamma_{f,i})} \right\},$$ \hspace{1cm} (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,

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_c + \mathcal{L}_p + \mathcal{L}_{ID} + \mathcal{L}_{FP},$$ \hspace{1cm} (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 $m$, $c$, and $p$ correspond to the losses computed over extracted regions of the \{mouth + chin/ jaw\}, \{eyes + eyebrows\}, and \{pupil\}, respectively.

$$\mathcal{L}_m = \mathcal{L}_{LPIPS_m} + \mathcal{L}_{L2_m} + \mathcal{L}_{IF_m},$$ \hspace{1cm} (10)
$$\mathcal{L}_c = \mathcal{L}_{LPIPS_c} + \mathcal{L}_{L2_c} + \mathcal{L}_{IF_c},$$ \hspace{1cm} (11)
$$\mathcal{L}_p = \mathcal{L}_{L2_p} + \mathcal{L}_{IF_{L2_p}},$$ \hspace{1cm} (12)

where $\mathcal{L}_{LPIPS}, \mathcal{L}_{L2},$ and $\mathcal{L}_{IF}$ represent the LPIPS loss, L2 loss, and Inter-frame loss, respectively. (Refer to Sec. E.3 in Appendix for a detailed explanation).

### 3.6 Generator Fine-Tuning, PTIs

There exist an inherent quality loss in the initial encoding as the real-world subjects would mostly be out-of-domain of StyleGAN resulting in notable deviations between the encoded frames and real frames. Roich et al. in PTI [30] propose fine-tuning the StyleGAN’s generator around “pivots” to improve the photo-realism of images while maintaining editability. Adapting from this concept, we fine-tune only the layers post-StyleSpace, $G^{ss}$ by solving Eq. (4) using  \{\$LH^{ss}_t, \alpha_t\} as pivots (in contrast to $W+$ latents in [30]) with real frames as reference to produce $G^{ss}_t$. The optimization is performed over the entire sequence of frames simultaneously compared to the single image tuning used in PTI.
3.7 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 re-synthesize the video, first $LH_t$ is obtained from $L_{ID}$ to adjust for the head-pose using StyleFlow for each frame. Then the $LH_t$ is transformed to the StyleSpace, $LH_{ss}^t$ using $A(\cdot)$, and the 32 indices responsible for the facial attributes, $\alpha_t$ are applied to reconstruct the image using the generator, $G_{ss}^t$:

$$\hat{S}_t = G_{ss}^t(LH_{ss}^t + \alpha_t 1)$$

4 Experiments and Results

4.1 Dataset and Evaluation

We selected 150 video clips (4K videos) from the video-sharing site [www.pexels.com](http://www.pexels.com). We thank the community for free licensing and allowing editing of human data. Each video contains a single face performing significant face deformations, head motion, and speech. The code for StyleVideoGAN [15], the SOTA approach to re-synthesize $1024^2$ face videos using the latent space is not publicly available. The authors of [15,39] kindly processed six videos for comparison. Additionally, we evaluated the publicly available models of FOMM [32] and fs-vid2vid [38] over our dataset at the algorithms’ native resolution. We evaluate video re-synthesis on these measures: spatial quality, spatio-temporal quality and appearance, and temporal consistency of identity (see [11] and refer to Section F.2 in Appendix for additional details on metrics).

Fig. 4: Qualitative evaluation of the results yielded through our approach. 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.
Table 2: Quantitative comparison of video re-synthesis against baselines. The first section (*) consists of metrics computed for 6 videos received upon requests made to the authors. The second section comprises metrics evaluated against the dataset of 150 videos. The final section consists of the ablations performed, where “Ours (ReStyle)” refers to the e4e encoder replaced with ReStyle and “Ours – PTI” refers to the stage prior to generator fine-tuning. We outperform the SOTA \[15\] at 1024^2 on all metrics while using only 0.38% of latent space parameters used by them.

| Method            | res. | L1   | LPIPS | L_{ID} | PSNR | SSIM | FID | FVD | FVD_M | ρ_{AU} | ρ_{GZ} | ρ_{pose} |
|-------------------|------|------|-------|--------|------|------|-----|-----|-------|--------|--------|----------|
| Wang et al. \[39\] | 512  | 1.95 | 0.024 | 0.088 | 33.85| 0.966|11.00| 61.2| 9.19  | 0.884  | 0.972  | 0.992    |
| StyleVidGAN \[15\] | 1024 | 4.04 | 0.109 | 0.164 | 28.76| 0.928|28.77|223.3|24.07  | 0.739  | 0.884  | 0.979    |
| Ours*             | 1024 | 1.97 | 0.029 | 0.067 | 34.03| 0.961|13.63| 85.7|14.66  | 0.884  | 0.972  | 0.992    |
| FOMM \[32\]      | 512  | 5.75 | 0.093 | 0.158 | 25.21| 0.900|42.36|359.6|74.40  | 0.768  | 0.596  | 0.418    |
| fs-vid2vid \[38\] | 512  | 4.01 | 0.066 | 0.174 | 31.03| 0.932|28.73|130.4|38.94  | 0.745  | 0.596  | 0.418    |
| Ours              | 1024 | 1.99 | 0.030 | 0.096 | 34.27| 0.961|15.66| 86.0|21.88  | 0.768  | 0.838  | 0.841    |
| Ours (ReStyle)    | 1024 | 2.01 | 0.050 | 0.101 | 34.16| 0.959|17.36| 98.7|24.52  | 0.764  | 0.829  | 0.839    |
| Ours – PTI        | 1024 | 2.95 | 0.062 | 0.124 | 31.01| 0.956|23.86|140.3|31.27  | 0.719  | 0.807  | 0.814    |

- **Frame-wise reconstruction fidelity** using Structured Similarity Index (SSIM) \[40\], Peak Signal-to-noise Ration (PSNR), and mean L1-distance
- **Frame-wise perceptual quality** using Fréchet Inception Distance (FID) \[21\] and LPIPS loss \[47\]
- **Spatio-temporal perceptual quality** using Fréchet Video Distance (FVD) \[37\] and FVD_M over the mouth
- **Spatio-temporal face deformation** time-series correlation between Action Unit (AU) activations \(ρ_{AU}\), gaze \(ρ_{GZ}\), and pose \(ρ_{pose}\) measured by \[8\]
- **Facial identity preservation** using Identity loss (ID-loss) \[29\] which is derived from ArcFace \[12\]

Referring to the top of Tab. 2, we achieve state-of-the-art performance at 1024^2 with significantly improved re-synthesis results compared to StyleVideoGAN while utilizing only 0.38% of the latent-space parameters used by them (35 vs. 18 × 512 per frame). Further, our encoding scheme outperforms \[39\] in \(L_{ID}\), PSNR, and \(ρ_{AU}\) while yielding comparable results elsewhere despite metrics of \[39\] being computed in its native resolution of 512^2. Evaluation across the full dataset, (middle of Tab. 2), shows that our approach outperforms both fs-vid2vid and FOMM in all scores by large margins. It is critical to note that lower native resolutions (e.g. \[32,38,39\]) significantly favor several metrics since there is no penalty to loss of details (e.g. L1, SSIM, FVD, etc.) with respect to 1024^2 metrics.

Fig. 1 and Fig. 4 illustrate, qualitatively, the capturing of fine facial details such as lip pressing, mouth puckering and gaping, wrinkles around the eyes, mouth, nasal-bridge, and forehead, etc. enhancing photo-realism of the re-synthesis videos which are not necessarily captured by the metrics (refer Sec. F.3 for more examples). To the best of our knowledge, such fine details were not explicitly addressed by previous research.
4.2 Ablation Study

As ablations, we studied several design choices in our pipeline, namely, the use of a different encoder, using the real frames as the reference in the facial attribute optimization, and the effect of the generator fine-tuning stage.

The e4e encoder responsible for the initial inversion (Sec. 3.2) was substituted by the ReStyle encoder [5], which uses an iterative residual-based algorithm to generate the $W+$ latent corresponding to a real image. As we observed a deterioration of reconstruction quality in the ReStyle inversions as passes accumulate, we chose the inversion at the end of the 4th pass for the evaluation. Based on the results in Tab. 2, the proposed pipeline with the e4e encoder yields marginally better scores compared to using ReStyle.

We further investigated using the real frames $\{I_t\}$ as reference for the facial attribute encoding optimization (Sec. 3.5) instead of the synthesized frames $\{S_t\}$. This resulted in visually sub-optimal results requiring us to abandon tighter pixel-level metrics as $L_{1,2}$, which are essential in capturing fine facial details such as wrinkles, gaze, etc. Hence, we opted to use synthesized frames 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.

The impact of the generator fine-tuning stage (Sec. 3.6) was studied by evaluating the re-synthesis results rendered using the default StyleGAN2 generator $G_{ss}$ (refer to “Ours – PTI_{ss}” in Tab. 2). It could be observed that while the re-synthesis results without the fine-tuning stage yet outperform fs-vid2vid and FOMM in almost all scores, the fine-tuning stage improves the photo-realism which is well reflected in the performance improvement with the addition of the fine-tuning stage. This is justifiable due to the tendency of real-world subjects being out of the domain of StyleGAN and the inherent loss of the encoder.

![Fig. 5: 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 reasonably well across subjects](image-url)
Table 3: Quantitative comparison of puppeteering against baselines evaluated across 50 puppet-puppeteer pairs. Our approach achieves the best performance across all metrics except FVD while rendering high-resolution re-enactment videos.

| Method       | res. | $L_1$↓ | FID↓ | FVD↓ | $FVD_{AG}$↓ | $\rho_{AU}$↑ |
|--------------|------|--------|------|------|-------------|-------------|
| FOMM [32]    | 256  | 0.159  | 77.00| **396.78** | 103.04      | 0.501       |
| fs-vid2vid [38] | 512  | 0.202  | 73.57| 445.05| 112.65      | 0.640       |
| Ours         | 1024 | **0.094** | **63.49** | **405.52** | **82.30** | **0.708** |

4.3 Puppeteering

One-shot puppeteering, which is another use-case of our algorithm, generates a video using a single frame of a puppet to mimic a puppeteer. The 35 parameters are to a large degree independent of the subject (given the disentangled nature of StyleFlow and StyleSpace). We circumvent automatic alignment of puppet’s single-shot face deformations to the puppeteer’s initial face appearance by choosing candidates based on $\rho_{AU}$, the correlation of Action Unit scores generated using [8] between the puppet and puppeteer. Fig. 5 shows qualitative samples from puppeteering and quantitatively in Tab. 3, where our approach produces the best scores except for FVD. However, puppeteering could be improved further by computing an optimal mapping of latent edits if multi-shot puppet data is available (see Sec. 3.5).

4.4 Failure Cases

There are multiple scenarios where latent-based video encoding may fail: (1) during pre-processing if the face is misaligned with respect to StyleGAN2 expectations, (2) extreme facial deformations and profile views, stemming from the low representation in the FFHQ dataset used in training StyleGAN2, (3) identity drift in editing StyleFlow or StyleSpace, (4) wearables such as eyeglasses can be challenging in some cases due to remaining latent space entanglement (see Sec. C in Appendix for discussion).

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 facial video encoding, which enables high-fidelity ($1024^2$) video re-synthesis and re-enactment using a single W+ latent and 35 parameters per frame. Our algorithm achieves SOTA performance for video re-synthesis at $1024^2$ while using a fraction (0.38%) of parameters compared to StyleVideoGAN. 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 facial video encoding scheme.
Appendix

A Overview

The outline of the appendix is as follows.

- Sec. B: Detailed steps on alignment in the pre-processing stage
- Sec. C: Brief discussion on the limitations of the encoder
- Sec. D: Additional details on identity-latent selection
- Sec. E: Illustrated explanations of noteworthy sections of facial attribute encoding
- Sec. F: Further details and examples of experiments and results

B Video Pre-Processing: Alignment

The alignment carried out in the pre-processing stage could be elaborated further using the three steps below.

1. Detect eye blinking and compensate for its effect on landmark location of the eyes. This improves StyleGAN2-based alignment by removing the sensitivity to eye shape change during blinking.

2. Registration of the face between a frame and a Key Frame uses a parameterized affine optical-flow model of the head [9], excluding the non-rigid face features (eyebrows, eyes, and mouth). The over-constrained optical-flow model is very effective at stabilizing the face between consecutive frames unless there are changes in the Yaw/Pitch of the head. We employ the mean L2 distance to automatically determine the quality of the inter-frame alignment over the non-rigid parts of the face (i.e., compute the residual error in RGB values of face stabilization). A mean distance beyond a fixed threshold indicates that the affine motion model is not successful at stabilizing the rigid part of the face, triggering step (3).

3. Key Frame change that forces a new Key Frame to be the basis for future frames’ face stabilization (aligned according to (1)).

For optical flow head registration, the threshold of the mean RGB registration error over the face (excluding eyes, mouth, and eyebrow areas) had to exceed 45.0 (if the inter-frame Yaw and Pitch change is less than 2°), or 30.0 (if the inter-frame Yaw or Pitch change exceeds 2°). The objective is to avoid forcing face registration when the head is moving out-of-plane. Instead, a change in the Key Frame is triggered, allowing the StyleGAN2 encoder to capture the new head-pose. Fig. 1 shows the Key Frames from a short sequence when the head moves to near profile and then back.
C  GAN Inversion

The e4e encoder \cite{35} while producing state-of-the-art results in GAN inversion of real images, has a few failing instances. For certain subjects, (e.g., Fig. 2 (a)) the identity of the encoded image deviates considerably from the real frame. In such cases, as we perform the inversion per-frame, there is a tendency for the identity to change across the frames of a single video clip as well. The identity change across frames could be due to the poor convergence of the encoder resulting from the existence of a higher per-frame loss due to poor identity. Additionally, there exist cases where the e4e encoder failed to capture certain facial attributes successfully (e.g., Fig. 2 (b) and (c)) which could be due to the low representation of complex features in the StyleGAN2 training dataset (FFHQ). Further, certain visual artifacts and deformations tend to appear in certain cases similar to the examples shown in Fig. 2 (d), (e), and (f), which could be caused due to occlusions (d) and the noisiness in the neighborhood of the inverted $W^+$.

![Fig. 1: Key Frames in a video sequence with head out-of-plane rotation](image)

![Fig. 2: Few examples of isolated instances where the e4e encoder fails. The examples depict: (a) poor identity, (b) incorrect gaze, (c) inability to capture extreme mouth movements, (d) deformations caused due to occlusions, (e) visual artifacts, and (f) flaws in facial features captured (open eyes while closed in real)](image)
However, the impact of most of these issues on the re-synthesis is mitigated as (1) we anchor our deformations with respect to a single ID-frame that has the highest identity match with the real and (2) utilize a generator fine-tuning stage (PTT*) to minimize the disparity between the real and synthesized frames.

D Identity-Latent Selection

The choice of ID-frame generated from $L_{ID}$ is of great significance as it serves as the base identity for the face and head-pose deformations across the entire sequence of frames. Hence, as discussed in Sec. 3.3 in the main-paper, we choose the identity latent based on the maximum identity preservation between the source frame and the encoded frame under the constraints that the selected ID-frame has near frontal head-pose and has no blink. An example plot depicting the variation of the identity similarity (computed based on ArcFace [12]) is given in Fig. 3 (a) and the corresponding best and worst ID-frame candidates based on our criteria are shown in Fig. 3 (b).

![Identity Similarity Plot](image.png)

**Fig. 3:** Identifying $L_{ID}$ is based on identity matching using ArcFace [12]. (a) depicts the identity similarity scores computed between the encoded and real frames. In this case, (b) the best ID is at frame 21 while (c) the worst is at frame 172.
E Facial Attribute Encoding

E.1 StyleSpace Indices

We illustrate the facial deformations corresponding to the manipulation of each of the 32 StyleSpace indices tabulated in Tab. 1 of the main-paper in Fig. 5. A pair of images marked as \((l, c) : +/−\) is included for each StyleSpace index, \((l, c) \in \mathcal{V}\) denoting the sign of the perturbation added to the respective StyleSpace index.

E.2 Index Specific Learning Rate

The variation of index sensitivity computed over the indices corresponding to the \{mouth + chin/jaw\} is shown in Fig. 4 (a). The significant variations seen in the plot make it evident that the index sensitivities cannot be simply ignored and hence, the indices cannot be treated the same during optimization. In order to alleviate the dominance of indices with a higher index sensitivity, we compute an index specific learning rate, \(\eta_{f,i} \), \(\Gamma_{f,i} \) specified in Eq. (6) in the main-paper. The \(\Gamma_{f,i} \) corresponding to the indices in Fig. 4 (a) are depicted in Fig. 4 (b). It could be seen that the \(\eta_{f,i} \) of indices having a higher \(\Gamma_{f,i} \) is comparatively lower than the indices of lower \(\Gamma_{f,i} \), thus effectively alleviating the dominance.

E.3 Details on Optimization

The face deformation attribute encoding algorithm in Section 3.5.2 is presented in Algorithm 1. The AdamW \cite{19} optimizer with AMSGrad \cite{28} was utilized with an initial learning rate of \(\eta = \{\eta_{f,i} ; \forall f \in \mathcal{F}, i \in \mathcal{V}\} \), \((\beta_1, \beta_2) = (0.9, 0.999)\), and \(\epsilon = 1e^{-8}\). The optimization was over 100 epochs (\(N = 100\)) and the learning

Fig. 4: (a) Index sensitivity and the corresponding (b) index specific learning rate. This figure represents values computed for an example subject over the \{mouth+chin/jaw\} indices.
Fig. 5: Example face deformations resulting from manipulation of each StyleSpace index, $(l, c) \in V$ in the negative (-) and positive (+) directions.
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 $t-1$: $LH_{ss}^{t-1}$ and $\alpha_{t-1}$,
- Optimizer $F'$, $N$ number of epochs, and $G_{ss}$

Initialization:
- obtain the StyleSpace latent, $LH_{ss}^t = A(LH_t)$
- initialize $LH_{ss}^t(l, c), \forall (l, c) \in V$
- compute the index-specific learning rates, $\eta_{f,i}$
  \[
  \eta = \{\eta_{f,i}; \forall f \in F, i \in V\}
  \]

Optimization:
for $n = [1:N]$ do
  \[
  \hat{S}_t = G_{ss}\{LH_{ss}^t + \alpha_t \mathbb{1}_i\}
  \]
  \[
  \text{where } \mathbb{1}_i = \{1 \text{ when } (l, c) \in V; 0 \text{ elsewhere}\}
  \]
  \[
  \mathcal{L} = \mathcal{L}\{\hat{S}_1, \hat{S}_{t-1}, \hat{S}_t, \hat{S}_{t-1}, S_t\}
  \]
  \[
  \alpha_t \leftarrow \alpha_t - \eta F'(\nabla \alpha_t \mathcal{L}, \alpha_t)
  \]
end

Output:
- 32-dimensional $\alpha_t$

rate was decayed every 10 epochs with a decaying factor of 0.8 using a learning rate scheduler for improved convergence. Additional details on the loss terms defined in Eqs. (9) to (12) of the main-paper are given below.

$\mathcal{L}_{LPIPS}$: The LPIPS loss [47], which is known to learn perceptual similarities well [19,29], was used to capture the structural details of the facial attributes between $S_t$ and $\hat{S}_t$. Nevertheless, $\mathcal{L}_{LPIPS}$ was not used in solving for the gaze ($L_p$) as it is invariant to slight spatial changes and hence introduces a slight jitter when used.

$\mathcal{L}_{L2}$: This denotes the L2 norm between the $S_t$ and $\hat{S}_t$, and enables precise reconstruction (e.g., the case of gaze).

$\mathcal{L}_{ID}$: To mitigate the risk of changing the identity of the subject across frames, while optimizing over the latent space the identity loss [29] is in place as a regularization term. This is computed between $\hat{S}_1$ and $\hat{S}_t$.

$\mathcal{L}_{FP}$: As we optimize over 32 indices in parallel, we noted occasional nose, mouth, and chin/jaw deformations. To discourage unwarranted deformations, the Face-Parsing loss, which is the L2 norm of the difference between the masked face-parsing scores [43] of the rendered and target frames, is used instead of facial-landmark coordinates loss (e.g., [8]). Face-parsing scores facilitate the gradient flow through the optimization and are more precise and stable across the
Appendix: Encode-in-Style

frames.

\[ \mathcal{L}_{FP} = \| FP(\hat{S}_t) \ast M - FP(S_t) \ast M \|_2 \]  

(1)

where function \( FP(\cdot) \) yields face-parsing scores and \( M \) denotes the binary mask of the face.

\( \mathcal{L}_{IF} \): The inter-frame loss is a derivation of the Frame Difference-Based (FDB) loss proposed in [42], to enforce temporal coherence between frames. We minimize this loss along with the other spatial losses to avoid enforcing temporal continuity posteriori. Provided the target video is temporally coherent, this loss is based on the concept that the image space and feature space differences between consecutive frames embed the temporal coherence. We use LPIPS and L2 losses to compute differences in the feature and image spaces, respectively.

\[ \mathcal{L}_{IF} = \mathcal{L}_{IF, LPIPS} + \mathcal{L}_{IF, L2} \]  

(2)

\[ \mathcal{L}_{IF, \ast} = \mathcal{L}_\ast\{S_t, S_{t-1}\} - \mathcal{L}_\ast\{\hat{S}_t, \hat{S}_{t-1}\} \]  

(3)

where \( \ast \) denotes either LPIPS or L2.

F Experiments and Results

F.1 Dataset

As stated in Sec. 4.1 of the main-paper, we compose a dataset consisting of video clips of 4K resolution sourced from the site www.pexels.com. The videos were chosen such that diverse subjects belonging to various ethnicities, age groups, and having different facial geometries, performing significant head-pose movements and facial deformations (both expressions and speech) were included. The results were computed based on 150 videos chosen from the dataset, with a mean of 304 frames, a minimum of 100 frames, and a maximum of 1000 frames.

F.2 Evaluation Metrics

The below metrics were used for the quantitative evaluation of our results in comparison with baselines, which are tabulated in Tables 2 and 3 of the main-paper.

Mean L1-distance, \( L_1 \): The per-pixel L1-distance was averaged across pixels, channels, and frames to obtain the score. The pixel values of the input images were in the range of [0,255].

Learned Perceptual Image Patch Similarity Loss, LPIPS: The metric was computed per-frame using the original implementation of [47] computed using the feature space of AlexNet [24].

Identity Loss, \( L_{ID} \): The identity loss was computed using,

\[ L_{ID} = 1 - \langle \phi(S_t), \phi(\hat{S}_t) \rangle \]  

(4)
where $\phi$ represents the pretrained ArcFace network \cite{arcface} and $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. While in re-synthesis (Table 2 in the main-paper) the loss was computed between the synthesized frame and the real frame, for puppeteering (Table 3 in the main-paper) the loss was computed between each frame and the puppet’s source frame.

**Peak Signal to Noise Ratio, PSNR:** This was computed using the built-in function of python’s scikit-image package using images having pixel values in the range $[0, 255]$.

**Fréchet Inception Distance, FID:** This metric, which is used to measure the photo-realism between two datasets, was computed based on the original implementation of \cite{fid} with a batch size of 100. Note: The input images are rescaled to $299 \times 299$ at the input of the inception network.

**Fréchet Video Distance, FVD:** The spatio-temporal perceptual score measured through FVD was computed using the original implementation of \cite{fvd}. Video fragments of length 120 frames were scored with a batch size of 8 and averaged to obtain the final FVD score due to resource limitations. Note: The frames are rescaled to $224 \times 224$ by the algorithm.

**Fréchet Video Distance - Mouth, FVD_M:** Similar to FVD, with the exception of the metric being scored over the masked area of the mouth region.

**Action Unit, Gaze, Pose Correlations, $\rho_{AU}$, $\rho_{GZ}$, $\rho_{pose}$:** These metrics measure the time-series correlation between the Action Unit activations, Gaze angles, and Yaw and Pitch angles respectively, which are computed using OpenFace 2.0 \cite{openface} of the synthesized and the reference sequences. These provide an insight into how well the facial deformations ($\rho_{AU}$), eye motion ($\rho_{GZ}$), and pose ($\rho_{pose}$) are captured by the algorithm in a spatio-temporal sense.

Note: All metrics except FVD, were computed per frame and averaged across all the frames. Further, unless noted otherwise all metrics were computed over a masked-out region of the reference face of each frame.

**F.3 Video Results**

The additional examples of video re-synthesis and puppeteering depicted in Fig. 6 and Fig. 7 respectively reaffirm the versatility of our approach. The corresponding videos could be viewed in the project page.
Fig. 6: Additional examples demonstrating the versatility of our algorithm in video re-synthesis

Fig. 7: Example puppeteering results generated by applying the encoded parameters computed for the puppeteer through our encoding algorithm onto the ID-latent of the puppet
References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4432–4441 (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8296–8305 (2020)
3. Abdal, R., Zhu, P., Mitra, N.J., Wonka, P.: Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics (TOG) 40(3), 1–21 (2021)
4. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Only a matter of style: Age transformation using a style-based regression model. arXiv preprint arXiv:2102.02754 (2021)
5. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. arXiv preprint arXiv:2104.02699 (2021)
6. Alaluf, Y., Tov, O., Mokady, R., Gal, R., Bermano, A.H.: Hyperstyle: Stylegan inversion with hypernetworks for real image editing. arXiv preprint arXiv:2111.15666 (2021)
7. Averbuch-Elor, H., Cohen-Or, D., Kopf, J., Cohen, M.F.: Bringing portraits to life. ACM Transactions on Graphics (TOG) 36(6), 1–13 (2017)
8. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018)
9. Black, M., Yacoob, Y.: Recognizing facial expressions in image sequences using local parameterized models of image motion. International Journal of Computer Vision 25, 23–48 (1997)
10. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
11. Chen, L., Cui, G., Kou, Z., Zheng, H., Xu, C.: What comprises a good talking-head video generation? A survey and benchmark. arXiv preprint arXiv:2005.03201 (2020)
12. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. CVPR (2019)
13. Donahue, C., Lipton, Z.C., Balsubramani, A., McAuley, J.: Semantically decomposing the latent spaces of generative adversarial networks. arXiv preprint arXiv:1705.07904 (2017)
14. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14398–14407 (2021)
15. Fox, G., Tewari, A., Elgharib, M., Theobalt, C.: Stylevideogan: A temporal generative model using a pretrained stylegan. arXiv preprint arXiv:2107.07224v1 (2021)
16. Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C., Agrawala, M.: Text-based editing of talking-head video. ACM Transactions on Graphics (TOG) 38(4), 1–14 (2019)
17. Geng, J., Shao, T., Zheng, Y., Weng, Y., Zhou, K.: Warp-guided gans for single-photo facial animation. ACM Transactions on Graphics (TOG) 37(6), 1–12 (2018)
18. Gu, K., Zhou, Y., Huang, T.: Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10861–10868 (2020)
19. Guan, S., Tai, Y., Ni, B., Zhu, F., Huang, F., Yang, X.: Collaborative learning for faster stylegan embedding. arXiv preprint arXiv:2007.01758 (2020)
20. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. arXiv preprint arXiv:2004.02546 (2020)
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25, 1097–1105 (2012)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
26. Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., Agarwal, A., Fursund, J., Li, H.: pagan: real-time avatars using dynamic textures. ACM Transactions on Graphics (TOG) 37(6), 1–12 (2018)
27. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
28. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. In: International Conference on Learning Representations (2018)
29. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296 (2021)
30. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. arXiv preprint arXiv:2106.05744 (2021)
31. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9243–9252 (2020)
32. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems 32, 7137–7147 (2019)
33. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zollhoefer, M., Theobalt, C.: Stylerig: Rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6142–6151 (2020)
34. Tian, Y., Ren, J., Chai, M., Obszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. arXiv preprint arXiv:2104.15069 (2021)
35. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) 40(4), 1–14 (2021)
36. Tripathy, S., Kannala, J., Rahtu, E.: Single source one shot reenactment using weighted motion from paired feature points. arXiv preprint arXiv:2104.03117 (2021)
37. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
38. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Catanzaro, B., Kautz, J.: Few-shot video-to-video synthesis. In: NeurIPS (2019)
39. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10039–10049 (2021)
40. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4), 600–612 (2004)
41. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12863–12872 (2021)
42. Xu, J., Xiong, Z., Hu, X.: Frame difference-based temporal loss for video stylization. arXiv preprint arXiv:2102.05822 (2021)
43. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018)
44. Zakharov, E., Ivakhnenko, A., Shysheya, A., Lempitsky, V.: Fast bi-layer neural synthesis of one-shot realistic head avatars. In: European Conference on Computer Vision. pp. 524–540. Springer (2020)
45. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9459–9468 (2019)
46. Zhang, L., Pollett, C.: Facial expression video synthesis from the StyleGAN latent space. In: Jiang, X., Fujita, H. (eds.) Thirteenth International Conference on Digital Image Processing (ICDIP 2021). p. 7. SPIE, Singapore, Singapore (Jun 2021)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
48. Zhang, Z., Li, L., Ding, Y., Fan, C.: 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. pp. 3661–3670 (2021)