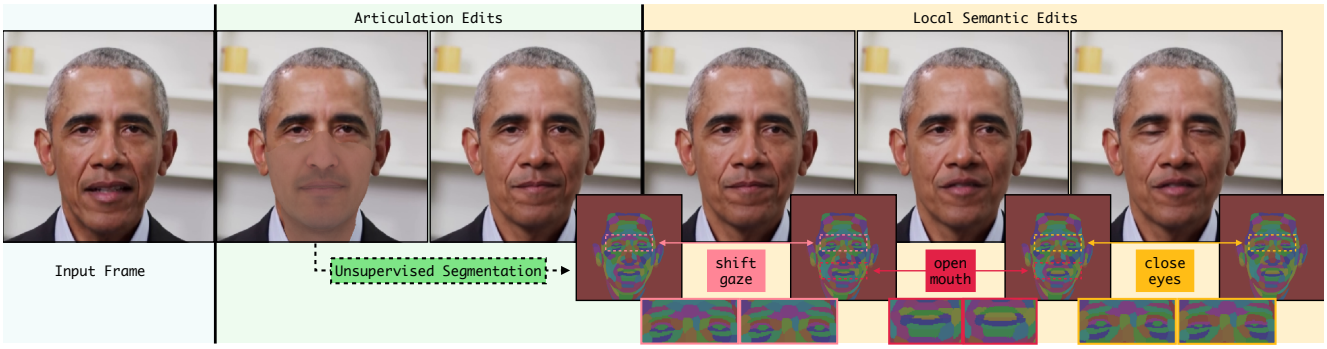
This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unsupervised Facial Performance Editing via Vector-Quantized StyleGAN Representations

Berkay Kicanaoglu Pablo Garrido Gaurav Bharaj

Flawless AI

{berkay.kicanaoglu, pablo.garrido, gaurav.bharaj}@flawlessai.com



Our method enables semantic video editing by converting StyleGAN [27] prior into unsupervised segmentation prior.

Abstract

High-fidelity virtual human avatar applications create a need for photorealistic video face synthesis with controllable semantic editing over facial features. While recent generative neural methods have shown significant progress in portrait video synthesis, intuitive facial control, e.g., of mouth interior and gaze at different levels of details, remains a challenge. In this work, we present a novel face editing framework that combines a 3D face model with StyleGAN vector-quantization to learn multi-level semantic facial control. We show that vector quantization of StyleGAN features unveils richer semantic facial representations, e.g., teeth and pupils, which are difficult to model with 3D tracking priors. Such representations along with 3D tracking can be used as self-supervision to train a generator with control over coarse expressions and finer facial attributes. Learned representations can be combined with user-defined masks to create semantic segmentations that act as custom detail handles for semantic-aware video editing. Our formulation allows video face manipulation with precise local control over facial attributes, such as eyes and teeth, opening up a number of face reenactment and visual expression articulation applications.

1. Introduction

High-resolution editing and rendering of photorealistic facial portrait videos are in high demand due to virtual human applications, such as actor performance editing, multi-language telepresence, and video conferencing.

In the area of high-resolution video face editing, state-of-the-art methods [41, 36, 56, 1, 42, 57] use pre-trained generative neural networks, such as StyleGAN [25, 27] that, when trained on large face datasets [25, 32], leads to learning a face prior that can be exploited for video editing. In the literature, various formulations leverage such a pre-trained GAN as input in an attempt to disentangle *style* spaces for semantic face image editing [44, 8]. Here, editing is limited by the training dataset distribution as well as attributes such as pose range, expression, ethnicity, and gender. The edits are often not explicit and can be unintuitive since it involves finding directions or discriminating features in a high-dimensional feature space to achieve the desired effects [44].

Several generative neural face synthesis methods have been proposed [51]. They take 2D videos frames and a tracked 3D face, often parameterized via a 3D morphable model (3DMM) [11], as supervised inputs and learn to *neurally* render the video. Once the neural network is learned, faces can be edited by modifying the tracked 3D face (or the 3DMM parameters) and rendering new subject videos. The tracked 3D face can be altered using audio input [53, 67], target reenactment videos [55], or artist-based modifications of the 3D face morphable model [49, 50]. Such *3DMM-aware*

methods allow global 3D head pose control and mouth shape articulation. Photorealism is handled implicitly by generative neural networks that act as a black-box renderer [31, 54]. However, these approaches lack local control, e.g., of mouth interior or gaze. The main reason is that such local details can not trivially be tracked via standard 3D tracking methods.

On the other hand, *3DMM-free* methods, where no 3D face tracking is given, learn to transfer the mouth and facial motion of a driving actor video to that of a target video via deep learning and 2D computer vision techniques [46, 39, 10]. Such methods warp the input (original) face frames into the target space while preserving the target face appearance and semantics. Unfortunately, their success is somewhat limited since topological changes, e.g., mouth opening, can not be explained by simple 2D image warping transformations [67].

Thus, we desire an approach that learns conditional (rig-like) control automatically for intuitive video editing applications. To this end, we use vector quantization (VQ) to discover meaningful segmentations in (pre-trained) StyleGAN feature-space. VQ has been used for generative modeling [40] and recently adopted in GAN frameworks and vision transformers [13]. Unlike existing methods, we use VQ to learn spatial segmentations automatically from a pre-trained StyleGAN for a given input video. It provides not only dense input to the generator, but also an effective mechanism to manipulate local facial details. Our approach is divided into two stages, as shown in Fig. 1. Stage-I establishes a map between an input (tracked) video sequence and a pretrained StyleGAN. Stage-II utilizes the learned VQ-representations, represented as a semantic segmentation mask, from Stage-I as input to a generator network that acts as the final renderer.

Our method provides a configurable interface for semantic edits thanks to the dense segmentations produced in Stage-I. Given the segmentation masks discovered in Stage-I, artists can label the spatial layout to add human interpretability and combine the segmented regions to create editing priors that act as artist handles for semantic-aware editing. As a result, several applications are feasible, such as semantic editing – gaze, nose, mouth interior, edits such as teeth removal, blinking, and nose and eyebrow shape change. To the best of our knowledge, we are the first to adopt VQ for unsupervised segmentation of a pre-trained StyleGAN given a video and use the learned segmentation features for localized semantic video editing. Our key novelties include:

1. A novel formulation that uses vector quantization of a pre-trained StyleGAN for automatic unsupervised video segmentation.

2. A VQ-aware decoder that converts automatically discovered VQ segmentation into decoder's SPADE feature blocks for high-resolution image synthesis.

3. User-guided semantic segmentation masks that act as editing handles for intuitive facial video editing.

4. Video editing applications that enable users to control semantic facial attributes not seen before (see Teaser).

2. Related Work

This section reviews neural portrait synthesis and StyleGAN inversion methods, applied to semantic image and video face editing. Please refer to Tewari *et al*. [51] for an extensive overview on neural synthesis and Bermano *et al*. [5] for in-depth review of GAN inversion.

Model-based Neural Portrait Synthesis. A major line of work attempts facial portraits synthesis using GAN-based approaches with semantic control, driven either via sparse keypoints [69, 62, 34], 3DMM priors [31, 30, 53, 35], or multi-modal input [64, 73]. Another research trend models view-consistent animatable 3D talking heads from 2D videos using neural implicit representations [4, 72, 48, 21, 19, 16, 17, 6]. These approaches synthesize detailed portraits, often controlled via 3DMM parameters to enable intuitive editing.

While these methods can achieve semantic control, they can only do global facial feature manipulations, thus limiting the range of editing applications. Besides, 3DMM-aware neural approaches often lack input conditioning inside the mouth interior, eyes, and regions beyond the inner face, resulting in poor or inconsistent synthesis thereof [49]. To overcome these issues, some approaches leverage GAN-based priors, such as StyleGAN [26, 28, 24], conditioned with 3DMM parameters to synthesize non-controllable regions in ways consistent with the underlying prior data distribution [49, 50, 7, 67] and to generate high-resolution face images. We also use a StyleGAN prior to render detailed facial features even for non 3DMM-driven regions. However, our approach also learns a semantic facial layout from StyleGAN features in an unsupervised fashion that, when combined with an artist prior, allows semantic video editing, e.g., of the mouth interior and eyes, with finer control not seen before.

StyleGAN Inversion for Image and Video Editing. There has been a large interest in designing StyleGAN-based inversion methods [5] for style-based manipulations in images [14, 42, 1, 41, 3, 36] and videos [2, 57, 15, 65, 66]. After inversion, semantic edits are conducted via style-space arithmetic and latent-space traversal with precomputed latent directions from projective subspaces [44, 22, 15] or classifiers [65, 2]. While inversion is a primer for style-based manipulations, it suffers a trade-off between editing capability and identity distortion, especially when edited features lie outside StyleGAN distributions. Tov *et al*. [56] address this problem using an encoder-for-editing (e4e) approach that restricts the distributions of inverted images to their original space. We use e4e for stable inversion as well.

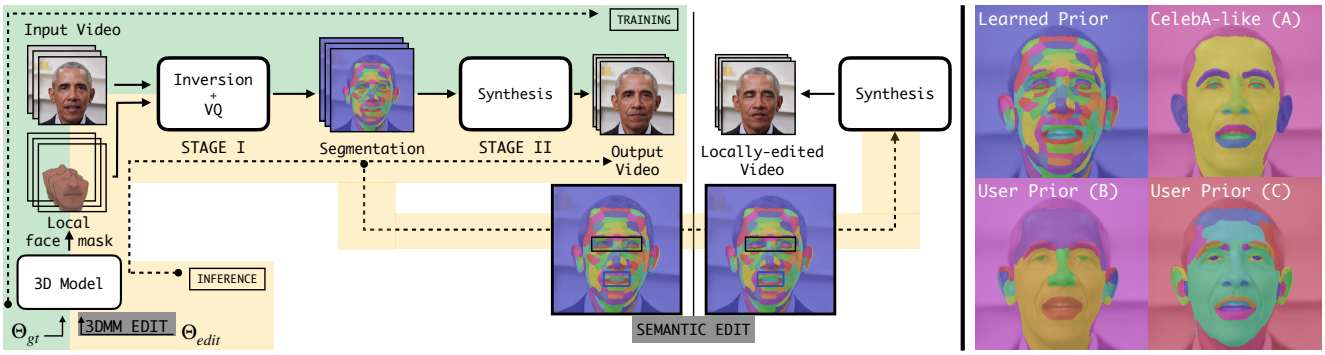


Figure 1. **(Left)** First, our method learns a video segmentation prior using vector-quantization (VQ) over StyleGAN features. Once trained, VQ generates segmentation maps without any extra effort. Then, we train a personalized generator conditioned on these maps, which acts as our final renderer. When combined with 3DMM inputs Θ, our full method achieves semantic control at different levels of details over the frames of a given video. **(Right)** The learned video prior allows artists to define a custom semantic segmentation mask based on their needs. In a clockwise fashion, starting from top-left, we show the learned video prior and three different user definitions.

From an inversion point of view, most approaches focus on finding optimal maps between images and style features that offer control over global attributes. Still, little effort has been invested into the spatial properties of StyleGAN latent-spaces. Yin *et al*. [67] show that there is an equivariant relationship between style-induced feature maps and the generator outputs under 2D image transformations, revealing that facial motion information, e.g., via 3DMM, can be used for detailed talking head editing. Based on these insights, we employ VQ to discover spatial semantics of style features at different levels of details in an unsupervised fashion, thus enabling semantic editing control.

StyleGAN Spatial Semantics. Recent studies of StyleGAN priors have shown an emergent semantic disentanglement property in spatial dimensions [12, 9, 37, 45, 33, 29]. Lee *et al*. [32] learn a style mapping between semantic masks and images via an annotated dataset. Endo *et al*. [12] improve upon it and learn a mapping between StyleGAN latent space and a few semantic user-defined masks to generate pseudo-semantic labels that can be used to train image generators with semantic editing control. In [9, 37], fixed semantic regions are discovered via k-means clustering of StyleGAN features. Such semantic regions can be used for transferring coarse facial attributes between images [9, 18] or text-driven global face manipulation with user-defined queries [37, 8]. Liu *et al*. [33] integrate a mask prediction branch into StyleGAN and model joint distribution of semantic masks and images, and learn to retain spatial context for semantic image translation. Unlike previous methods, we discover spatial semantics in an unsupervised manner, without semantic labels, spatial priors, or complex architecture modifications.

Since most StyleGAN-based edits affect global styles, there have been improvements to achieve more fine-grained control [29, 45]. These methods alter the StyleGAN architecture to include explicit semantic style injection via segmentation masks to factorize semantic regions [45] or modulate style feature transfer [29, 38]. Our approach also achieves fine-grained semantic control but needs no explicit spatial semantics. Instead, we discover them in a fully unsupervised fashion via VQ. Furthermore, we allow artists to create intuitive semantic handles for editing.

VQ Generators. Recently, vector quantization has shown great promise as discrete compact representations for modeling salient spatial features in generative image synthesis [60, 40, 13]. Oord *et al*. [60] introduce a VQ autoencoder to learn context-rich visual parts and model their distributions with an autoregressive convolutional network. Razavi *et al*. [40] extend it using a hierarchical multi-scale representation of codebooks. Esser *et al*. [13] show that a convolutional VQ-GAN combined with an autoregressive transformer can learn rich representations with global context to enable high-definition image synthesis. While these methods leverage VQ to encode efficient spatial representations for large-scale image synthesis, we utilize it as a means to reveal a detailed spatial layout of StyleGAN features in a video without supervision. The resulting learned semantic layout enables localized semantic video editing, e.g., eyes and mouth interior, with finer local control not seen before.

3. Method

We propose a video editing system that unifies explicit model-based synthesis with StyleGAN's powerful prior. Our system synthesizes video frames with control over expressions and local facial details, such as eyes (pupils) and mouth interior. At its core, we employ VQ to convert StyleGAN into a video segmentation prior in an unsupervised fashion. The learned segmentation prior consists of a set of feature vectors, represented as a semantic spatial layout (semantic regions) over the face, e.g., upper teeth and pupils, see Fig. 3.

Unlike other StyleGAN-based editing methods, we defer

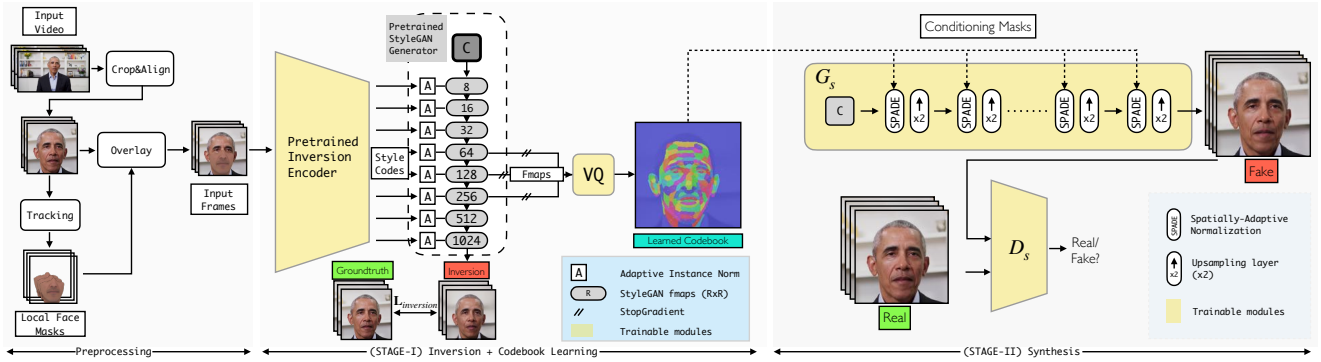


Figure 2. Our video editing system comprises two stages. **Stage-I** (middle): We learn to invert 3DMM-augmented images into pretrained StyleGAN's latent space. In this step, a vector-quantization (VQ) module learns a codebook of semantic segments over multi-scale StyleGAN feature maps with exponential moving average (EMA) updates in an unsupervised fashion. This stage achieves inversion and 3DMM expression control along with a set of codebook feature vectors corresponding to the centroids of learned segments. The learned codebook can then label the input easily. **Stage-II** (right): A GAN generator translates the automatically obtained semantic layouts into photorealistic video frames. Our generator uses SPADE [38] and upsampling layers, and it is directly driven by the semantic layout.

synthesis to another network. Thus, in a consecutive stage, we train a generator with learned segmentation information as a driving signal. This formulation not only allows for independent control at the rendering stage but also overcomes some limitations of inversion methods, such as identity drift, which is unacceptable for video editing applications.

3.1. Preprocessing

Given a source video with N frames $\mathcal{X} = \{x_i\}_{i=1}^N$, we crop and align frames following the procedure used in [57], and denote the resulting sequence (ground-truth) with $\mathcal{C} = \{c_i\}_{i=1}^N$. Next, we track the sequence $\mathcal{C}$ with our reimplementation of [52] adapted to video sequences. From the tracked face model, we render local face masks per frame, $\mathcal{H} = \{h_i\}_{i=1}^N$, to be overlayed over $\mathcal{C}$. We denote the overlaid frames (input) as $\hat{\mathcal{C}} = \{\hat{c}_i\}_{i=1}^N$. We employ the semantic face segmentation algorithm by Yu *et al.* [68] to generate a foreground mask $\mathcal{M}_{bg} = \{m_i^{bg}\}_{i=1}^N$ to remove background, hair, and clothing. We then apply $\mathcal{M}_{bg}$ to generate masked input frames $\hat{\mathcal{C}} = \{(1 - m_i^{bg}) \odot \hat{c}_i\}_{i=1}^N$ and ground truth frames $\mathcal{C} = \{(1 - m_i^{bg}) \odot c_i\}_{i=1}^N$[1]. From the semantic segmentation mask, we also extract the mouth, lips, and eyes regions and combine them into a single mask $\mathcal{M}_{roi} = \{m_i^{roi}\}_{i=1}^N$. We utilize $\mathcal{M}_{roi}$ to increase the influence of the aforementioned facial features during inversion (see 3.2).

3.2. Stage-I. Inversion & Vector-quantization (VQ)

Fig. 2 illustrates the preprocessing steps and inversion schemes, including vector quantization.

Inversion. To exploit StyleGAN prior, we first create an inversion mapping between the locally-masked frames and the style codes. We define inversion as

$$f : \hat{c}_i \mapsto w_i \in \mathcal{W}^+ \,, \tag{1}$$

where $\mathcal{W}^+$ is the extended style space. We denote the inverted frame as $r_i = G(f(\hat{c}_i); \Theta) = G(w_i; \Theta)$ where G is the pretrained StyleGAN generator parameterized by Θ. While our framework supports any inversion technique, we adopt an encoder-based model that is more robust to changing mask conditioning and provides faster inference time than an optimization-based method. Specifically, we opt for the encoder-based inversion scheme *e4e* [56]. Instead of training the encoder from scratch on our data distribution, we fine-tune an *e4e* encoder pretrained on FFHQ [25]. Although augmented frames impose an additional domain adaptation challenge to the pretrained *e4e*, we found that it is still a fast and effective way to establish f.

To fine-tune the pretrained *e4e* encoder, we adopt the loss functions and hyperparameters from [56] *but* with LPIPS [71] loss, factorized to weigh more the mouth and eye regions. Below we show the overall (modified) objective function for completeness. Please refer to [56] for details:

$$\begin{aligned}\mathcal{L}_{inversion}^{e4e} = \lambda_{l2} \cdot \mathcal{L}_2(c_i, r_i) + \\ \lambda_{lpips}^{roi} \cdot \mathcal{L}_{lpips}(c_i \odot m_i^{roi}, r_i \odot m_i^{roi}) + \\ \lambda_{lpips}^{nonroi} \cdot \mathcal{L}_{lpips}(c_i \odot (1 - m_i^{roi}), r_i \odot (1 - m_i^{roi})) + \\ \lambda_{sim} \cdot \mathcal{L}_{sim}(c_i, r_i) + \lambda_{d-reg} \cdot \mathcal{L}_{d-reg}(w_i) + \lambda_{adv} \cdot \mathcal{L}_{adv}\end{aligned}$$

Here, $\mathcal{L}_2$, $\mathcal{L}_{lpips}$ and $\mathcal{L}_{sim}$ are reconstruction and identity losses, whereas $\mathcal{L}_{d-reg}$ and $\mathcal{L}_{adv}$ are losses to ensure w_i stays close to the original $\mathcal{W}$ space. As we defer synthesis to Stage-II, our goal is to achieve geometric alignment, especially in the eye and mouth region, between ground truth c_i and StyleGAN inversion r_i (and in turn the learned video segmentation) during Stage-I. Thus, we use $\lambda_{lpips}^{roi} > \lambda_{lpips}^{nonroi}$ to minimize the alignment errors in regions of interest.

[1] Unless otherwise stated, $\mathcal{C}$ and $\hat{\mathcal{C}}$ refer to foreground-segmented sequences in the rest of the paper.

Vector Quantization (VQ). During (inversion) finetuning, we simultaneously deploy a VQ module that serves as a *passive* (online) clustering mechanism to learn a video segmentation prior. VQ module consists of a learnable codebook and a codebook update mechanism. In practice, the codebook $\mathbf{Z}_K^d \in \mathbb{R}^{K\times d}$ is implemented as a tensor, where d and K are the feature dimensionality and the number of entries, respectively. Once the entries $z_j = \mathbf{Z}[j,:] \in \mathbb{R}^d$ have converged to representative vectors, i.e., centroids, for the classes discovered over the face, we can use them to generate semantic video face segmentation at inference time.

We denote a StyleGAN feature map of resolution q and dimensionality d as a set of vectors $G_d^q = \{g_i \in \mathbb{R}^d : i = 1,\dots,q^2\}$. VQ assigns each feature vector $g \in G_d^q$ the nearest codebook entry z_j measured with *cosine* distance. For codebook learning, we adopt the exponential moving average (EMA) update mechanism from [40] as follows:

$$
\begin{aligned}
N_j^{(t)} &:= N_j^{(t-1)} * \gamma + n_j^{(t)}(1-\gamma) \\
m_j^{(t)} &:= m_j^{(t-1)} * \gamma + \Sigma i^{n_j^{(t)}} g_i^{(t)}(1-\gamma) \\
z_j^{(t)} &:= \frac{m_j^{(t)}}{N_j^{(t)}} ,
\end{aligned} \tag{2}
$$

where $n_j^{(t)}$ is the number of features that are assigned to cluster j at time step t.

StyleGAN representations, from coarse to fine, encode different levels of semantic information. We discover *empirically* that coarser layers, i.e., 16^2 and 32^2, embed contours and rough pose, whereas feature maps in larger scales carry finer semantics, such as texture and facial parts, as shown in 3. Therefore, we combine activations from multiple scales to learn a segmentation prior, which benefits from both lower and higher semantic levels. Specifically, in every forward call during inversion training, we gather layer activations G_{512}^{64}, G_{256}^{128} and G_{128}^{256}, i.e., the output of the last style-convolution for each layer. Before feeding in the activations to VQ module, we pass each activation map through instance normalization [58] layers and upsample to 256^2 using bilinear interpolation. Finally, we concatenate each activation along the channel dimension, resulting in a tensor $X_{VQ} \in \mathbb{R}^{996\times256\times256}$. Although we segment out the background, hair and clothing in the input, we find it beneficial to apply the same segmentation mask over this tensor, i.e., $X_{VQ} := X_{VQ} \odot (1 - m^{bg})$. This way, we make sure the codebook capacity is primarily focused on the face region. This mipmap-style input is then fed to VQ to update.

In learning the codebook, we ensure quantization does not interfere with the pretrained StyleGAN generator; otherwise, the training stability is degraded. As such, we neither use the quantized features in the generator nor allow the inversion objective to influence quantization. By the end of Stage-I training, we accomplish a mapping from augmented input frames to one-hot encoding of K unsupervised semantic classes $\hat{c}_i \mapsto x_{seg} \in \mathbb{R}^{K\times256\times256}$ for driving the synthesis network in Stage-II.

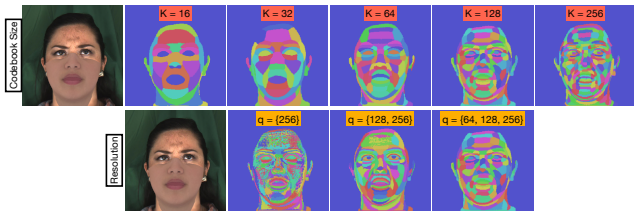


Figure 3. We show ablation results w.r.t **(top)** codebook size K and **(bottom)** input resolutions of the stacked StyleGAN layers. Increasing the codebook size yields finer partitioning of the subject's face into temporally consistent regions given hierarchical features as input (64x64, 128x128, 256x256).

3.3. Stage-II. VQ-driven Frame Synthesis

To guarantee identity fidelity and enable local semantic edits, our approach defers synthesis to a dedicated (or personalized) generator driven by video segmentations obtained from Stage-I. This way, we can also overwrite subject-specific videos without retraining, which is an important aspect for applications where repeated edits with varying conditions are needed, e.g., changing the scripts.

Unlike image-to-image translation models [23] mostly using variants of U-Net [43] architectures, we opt for an encoder-free generator design to reduce model size and to render frames $c_s = G_s(x_{seg})$ at 1024x1024. Thus, our generator G_s is a fully-convolutional decoder admitting semantic masks via spatial normalization layers. As normalization layers, we use SPADE [38]. To achieve a better personalization prior for a given subject and remove stochasticity at inference time, our design also utilizes a constant tensor, akin to unconditional generative models [26].

We train G_s in an adversarial fashion with a patch discriminator D_s. As adversarial loss, we adopt binary cross-entropy with gradient penalty regularization on discriminator weights. To ensure framewise quality, we use a weighted combination of smooth L1 reconstruction and factorized, i.e., roi, perceptual (LPIPS) losses. Our final objective then reads as follows:

$$
\begin{aligned}
\mathcal{L}_D^{synth} = &-\log D_s(c) - \log\left(1 - D_s(c_s)\right) + \\
&\lambda_{gp}\|\nabla_{x_{seg}} D_s(c_s)\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L}_G^{synth} = &-\log\left(D_s(c_s)\right) + \lambda_{l1}\mathcal{L}_{l1}(c, c_s) \\
&+ \lambda_{lpips}^{roi}\mathcal{L}_{lpips}(c_s \odot m^{roi}, c \odot m^{roi}) + \\
&\lambda_{lpips}^{nonroi}\mathcal{L}_{lpips}(c_s \odot (1 - m^{roi}), c \odot (1 - m^{roi}))
\end{aligned}
$$

3.4. Implementation Details

We use the PyTorch implementation[1] of StyleGAN2 [27] with the original network parameters from the official implementation. In Stage-I, we set $\lambda_{l2} = 5.0$, $\lambda_{lpips}^{nonroi} = 1.0$, $\lambda_{lpips}^{roi} = 5.0$, $\lambda_{sims} = 0.1$, $\lambda_{d-reg} = 2e{-}4$ and $\lambda_{adv} = 0.1$. During inversion, we set the learning rates to $5e{-}6$ and $1e{-}6$ for *e4e* encoder and discriminator, respectively. In Stage-II, we set $\lambda_{l1} = 0.5$, $\lambda_{adv} = 0.01$, $\lambda_{lpips}^{nonroi} = 0.2$, $\lambda_{lpips}^{roi} = 1.0$ and $\lambda_{gp} = 10.0$. We utilize Adam optimizer and a learning rate of $5e-4$ to train the generator and discriminator.

As local face masks, we choose albedo maps to make vector-quantization illumination invariant. Note that our approach is agnostic to the kind of masks used, e.g., diffuse texture, provided that the desired mouth articulation is represented clearly.

Temporal Stability Measure. When training G_s, we enforce temporal consistency via an extended temporal window scheme. Specifically, for a given batch, a target frame is padded with the leading and subsequent frame's semantic masks. We find that this simple extension suffices in preventing temporal flickering artifacts in the synthesized videos.

4. Experiments

Datasets & Training Details. In our experiments, we use MEAD dataset [61] and an Obama video downloaded from the internet[2]. We choose three different subjects from MEAD and use nearly 210 videos with neutral, happy and angry emotions in frontal, left/right 30-degree head orientation per subject. Each subject has approximately $20k$ frames in total. For Obama video, we use a 6-minute sequence, totaling around $10k$ frames. In Stage-I, we train the encoder and vector-quantizer for $6k$ iterations with a batch size of 4. In Stage-II, we train the generator for $60k$ iterations with a batch size of 2. We use the same training scheme for all datasets and train on a single A10 GPU.

As for quantization, we set $K = 128$ to reveal smaller details, such as pupils and teeth, in the learned segmentations. We warm-start the codebook $\mathbf{Z}_{128}^{996}$ by running k-means algorithm for 100 iterations on the first batch. During training, we set decay factor $\gamma = 10^{-3}$. To maximize the codebook's usage, we apply a dead-code replacement routine with a threshold of 2, akin to [70].
In the following, we provide quantitative and qualitative analysis and show different applications of our method.

4.1. Quantitative Results.

Metrics. We adopt both perceptual and pixel accuracy metrics to measure image quality and a video-based metric to judge temporal quality. As for image-based metrics, we compute MSE, SSIM [63], LPIPS [71], and FID scores [20]. To measure temporal consistency and quality of rendered frames, we compute Fréchet Video Distance (FVD) [59]. FVD is a sampling-based metric that computes a score using random temporal windows of varying lengths. As such, every run varies significantly depending on what frames are used as anchors, thus requiring multiple runs. To generate statistically meaningful results, we perform 50 runs on each video sequence and report the mean and standard deviations. We use two variants of FVD that use 16 and 128 frame windows and denote them as FVD_{16} and FVD_{128}, respectively. We also provide a third variant, $\text{FVD}_{128}^{\text{sub}}$, which uses a 128 frame window but samples every 8th frame when computing fake and real data statistics, as suggested in [59]. We remark that we run all three FVD metric variants on Obama sequence, whereas we only employ FVD_{16} on MEAD videos as they are only a few seconds long. Please refer to [47] for an in-depth explanation of FVD.

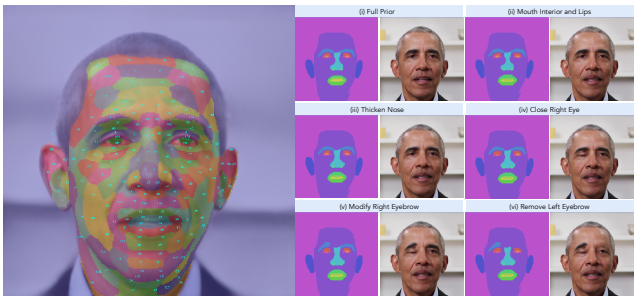


Figure 4. **Examplar visualization & user-defined masks.** (Left) Exemplar frame visualization. (Right) Consecutive local semantic edits via user-defined (CelebA-like) masks.

Analysis. We compare our approach against baselines PTI [42], FOMM [46] and DVP [31] in Tab. 1. Please refer to the supplementary document for details about the different baselines and evaluation settings. Our method consistently outperforms FOMM and PTI on almost all image-based and video-based metrics, except for FID metric where our method has a worse score on *M013*. Such a decrease in performance stems from mild artifacts around face boundaries and eyebrows, especially for non-frontal views. See supplementary video for details. Tab. 1 also shows that our method achieves the best scores on MSE, SSIM, LPIPS metrics. DVP attains better FVD scores even though our method produces results of similar visual quality (see supplementary video). We attribute the relatively large FVD values to a global color flicker in the temporal domain, which is barely perceivable to the human eye. We believe these residual temporal artifacts can be resolved with more advanced spatio-temporal architectures, which is an interesting avenue for future advances in the field.

[1] github.com/rosinality/stylegan2-pytorch
[2] www.youtube.com/watch?v=NGEvASSaPyg&t=471s

	Video-based Metrics	Image-based Metrics			
(MEAD)	**FVD**$_{16}$ ↓ (*M013* | *M027* | *W009*)	**FID** ↓ (*M013* | *M027* | *W009*)	**LPIPS** ↓ (*M013* | *M027* | *W009*)	**MSE** ↓ (*M013* | *M027* | *W009*)	**SSIM** ↑ (*M013* | *M027* | *W009*)
FOMM	(148.0; 9.1) | (147.7; 14.7) | (189.2; 12.9)	43.4 | 39.8 | 53.5	0.08 | 0.13 | 0.13	0.0024 | 0.0095 | 0.0072	0.97 | 0.86 | 0.88
PTI	(162.4; 10.3) | (177.4; 27.8) | (277.1; 23.3)	**12.3** | 11.7 | 16.1	0.05 | **0.05** | 0.06	0.0039 | 0.0052 | 0.0102	0.97 | 0.89 | 0.87
Ours	(**142.4**; 14.0) | (**136.7**; 23.2) | (**168.5**; 13.9)	22.5 | **11.3** | **11.1**	**0.04** | **0.05** | **0.05**	**0.0009** | **0.0011** | **0.0018**	**0.98** | **0.94** | **0.93**
(Obama)	**FVD**$_{16}$ ↓ | **FVD**$_{128}$ ↓ | **FVD**$^{sub}_{128}$ ↓	**FID** ↓	**LPIPS** ↓	**MSE** ↓	**SSIM** ↑
DVP	(**371.3**; 143.6) | (**160.0**; 65.4) | (**229.3**; 62.5)	**3.84**	0.025	0.0026	0.95
Ours	(382.1; 182.2) | (215.9; 87.6) | (260.1; 104.8)	4.99	**0.019**	**0.0009**	**0.96**

Table 1. Our method performs favorably against FOMM and PTI on MEAD dataset when evaluating both image-based and video-based metrics. It also outperforms DVP for most image-based metrics on a held-out set of Obama dataset. DVP achieves better temporal scores. Note that FVD$_\star$ scores are the mean and standard deviation computed over 50 runs. **Bold** text highlights the best result.

4.2. Qualitative Results.

Ablation Study. Vector-quantization involves a few hyper-parameters requiring experimental tuning. Fig. 3 provides ablation results w.r.t. varying codebook sizes. We specifically ablate over the feature map stacks used in VQ. The results demonstrate that the proposed StyleGAN activations provide more locally-consistent segments and overall reduce mask jitter. Please see our supplementary video for examples. While features at resolution 32^2 provide a good basis for this stability, Fig. 3 shows that higher resolution representations, e.g., 256^2 accommodate more semantic groups within the mouth. For instance, we can see individual teeth. This result implies a tradeoff between stability and details of the learned prior.

Qualitative Comparisons. Fig. 7 compares visually our method with FOMM [46] and PTI [42]. Overall, FOMM cannot reproduce results faithfully and our method fairs comparably to PTI. However, we capture better high-frequency details, such as freckles and moles. Please refer to the supplementary material for a better temporal appreciation of the synthesized results and for a comparison with DVP [31].

4.3. Applications.

Performance Preserving Reanimation. In many artistic applications, it is crucial to retain the original performance while performing local visual edits. Our approach makes possible to modify a subject-specific video in a localized fashion thanks to local face masks. Fig. 5 illustrates how we can reanimate a person's lower face via 3DMM-based expression transfer. Our method preserves the actor's eye gaze and facial details (unless instructed otherwise) via semantic maps. We further demonstrate cross-shot expression transfer using VoxCeleb2HQ subjects in Fig. 6. Please refer to the supplementary material for more details and results. As our approach, in essence, *inpaints* the mask region, we find that naive alpha-blending of the segmented foreground into the original footage suffices for compositing purposes. Therefore, it is not required to pass the final renders through refinement networks, e.g., as in Tzaban *et al*. [57].

Figure 5. **Performance editing via 3DMM.** Our method can overwrite input video frames (left) with given expressions via 3DMM (middle), allowing applications such as local mouth expression replacement. The overwritten video frames can be further enhanced with semantic editing touches.

Local Semantic Edits in Video Frames. Fig. 8 shows editing results for two subjects, where eye-gaze and mouth interior are seamlessly changed by modifying the segmentation masks. In particular, we demonstrate that fine touches in semantic layouts allow for subtle yet photorealistic edits in the final render. Since our video prior discovers mouth interior segments, e.g., tongue and upper/lower teeth, we can change the appearance of teeth or even remove them completely. In addition, our method can also perform topological modifications, such as opening and closing of eyes with the detailed segmentation prior, whereas model-based methods often fall short. As vector-quantization is unsupervised, the learned priors may vary semantically depending on input video's dynamics. For instance, the eyelid segment is better

Figure 6. **Cross-shot expression transfer on VoxCeleb2HQ subjects.** We use non-overlapping shots from the original video as a driver sequence and only transfer the driver's 3DMM expression parameters to that of the source. Note that our method robustly transfers expressions even when the driver's and source's head pose differ. Please see the supplementary video for more examples.

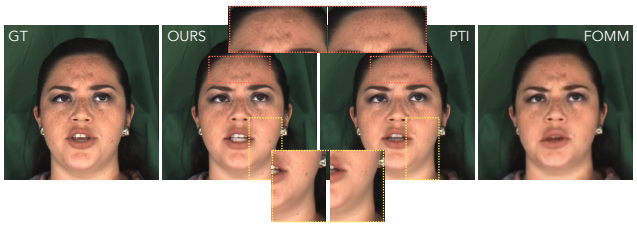


Figure 7. Our method and PTI achieve similar qualitative results [42, 57]. However, our formulation captures better some local details. In addition, our method allows explicit visual manipulation, which is not doable with PTI or FOMM.

recovered if the subject blinks often.

Local Editing w/ User Priors. In increasing K, it is possible to obtain finer, albeit over-segmented, semantic details. With this feature, artists can easily design personalized semantic priors in a bottom-up manner for differing purposes, as shown in Fig. 1. This opens up doors for novel interfacing applications since the defined priors propagate in a temporally-consistent manner at no labeling cost. Please refer to the supplementary video. An artist prior can simply be defined by grouping indices k from codebook $\mathbf{Z}_K$ into new segments after Stage-I. In defining artist-specific priors, we can train Stage-II with more intuitive local editing control. Fig. 4 (right) illustrates a prior inspired by CelebA-HQ's label definition derived from the learned segmentations. In Stage-II, we can train the generator using this semantic prior. Fig. 4 shows that our method can generate plausible visual edits, e.g., eyebrow removal and nose thickening, using a more human-friendly interface. Fig. 9 shows that, with user-defined priors, our method can deliver localized edits as illustrated in the activated heatmaps.

Examplar-frame Visualization. As the video segmentation prior is learned in an unsupervised manner, the correspondence between codebook entries and visual segments is not known beforehand. For local editing and user-defined segmentation masks, we rely on a *single* exemplar frame visualization, as shown in Fig. 4. It thus acts as a visual guide to defining user-priors since over-fragmentation can be overwhelming for users without a reference at this stage.

5. Limitations and Future Work

As with various StyleGAN-based editing approaches, the representation power of our method is limited by the training data distribution of StyleGAN. Although our personalized generator can learn to cope with deficiencies in texture and identity representation induced by inversion, it might still obtain suboptimal reconstructions, especially in presence of larger head poses and harsh lighting conditions. In this work, we focus on unveiling a pretrained StyleGAN's semantic prior with unprecedented detail using vector quantization. Further research directions will investigate architectural spatio-temporal modeling for improved performance.

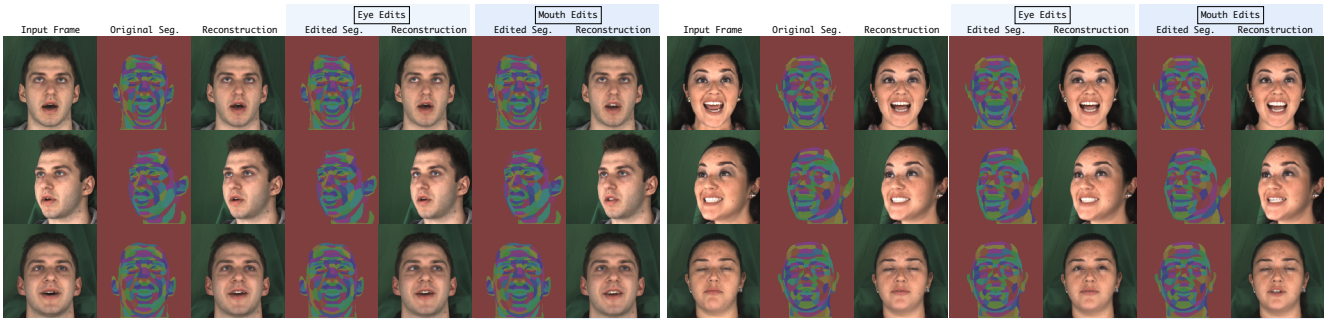


Figure 8. **Local editing.** Our method allows for fine-grained local semantic edits for mouth interior and gaze. The learned semantics further allow mouth opening/closing and eye opening.

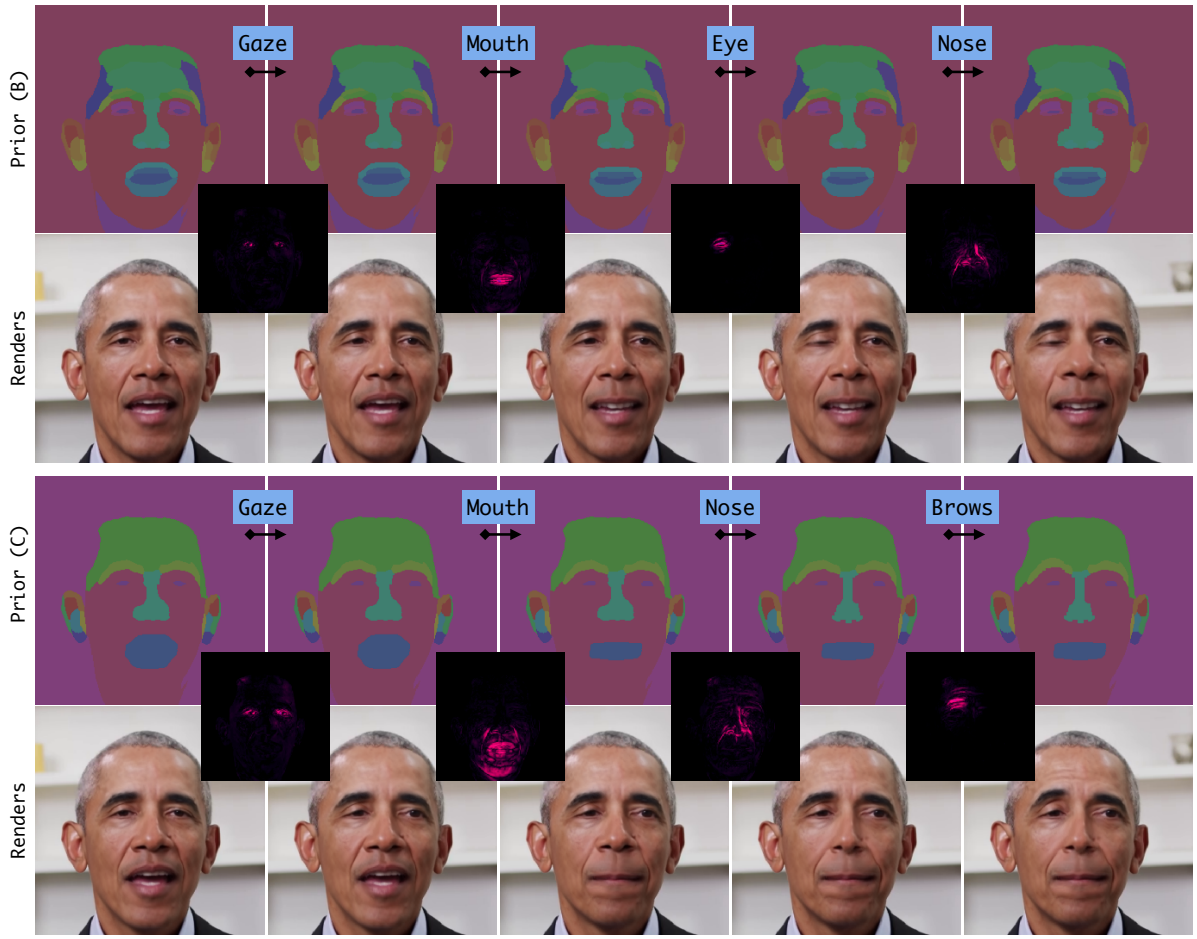


Figure 9. User-defined priors can be tailored to the artists' needs and provide intuitive semantics for laborious editing tasks. We show two different priors with which localized modifications can be applied. We highlight the affected image regions in the middle rows through photometric error maps. Note that local edits leave untouched areas unchanged, as shown in the maps.

6. Conclusion

We present a face editing framework that allows intuitive facial attribute control in portrait videos. Our system stands on the base of StyleGAN, and unifies explicit model-based generative synthesis with a novel StyleGAN based spatial prior. Expression editing control is achieved via 3DMM, while semantic editing control is realized through an induced StyleGAN spatial layout learned via vector quantization in an unsupervised fashion. Our system easily integrates user input to define semantic spatial regions that act as custom handles for intuitive local editing. We show the flexibility and effectiveness of our framework on several face editing tasks: reenactment, attribute manipulation (eye, gaze, mouth interior), and enhancements.

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *ICCV*, pages 6691–6700. IEEE, 2021. 1, 2

[2] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *CoRR*, abs/2201.13433, 2022. 2

[3] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H. Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *CVPR*, pages 18490–18500. Computer Vision Foundation / IEEE Computer Society, 2022. 2

[4] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. *CoRR*, abs/2206.06481, 2022. 2

[5] Amit H. Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Or Patashnik, and Daniel Cohen-Or. State-of-the-art in the architecture, methods and applications of stylegan. *Comput. Graph. Forum*, 41(2):591–611, 2022. 2

[6] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhöfer, Shunsuke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason M. Saragih. Authentic volumetric avatars from a phone scan. *ACM TOG*, 41(4):163:1–163:19, 2022. 2

[7] Prashanth Chandran, Sebastian Winberg, Gaspard Zoss, Jérémy Riviere, Markus H. Gross, Paulo F. U. Gotardo, and Derek Bradley. Rendering with style: combining traditional and neural approaches for high-quality face rendering. *ACM TOG*, 40(6):223:1–223:14, 2021. 2

[8] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. Image-based clip-guided essence transfer. *CoRR*, abs/2110.12427, 2021. 1, 3

[9] Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. Editing in style: Uncovering the local semantics of gans. In *CVPR*, pages 5770–5779. Computer Vision Foundation / IEEE Computer Society, 2020. 3

[10] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 2

[11] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models - past, present, and future. *ACM TOG*, 39(5):157:1–157:38, 2020. 1

[12] Yuki Endo and Yoshihiro Kanamori. Few-shot semantic image synthesis using stylegan prior. *CoRR*, abs/2103.14877, 2021. 3

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883. Computer Vision Foundation / IEEE Computer Society, 2021. 2, 3

[14] Qianli Feng, Viraj Shah, Raghudeep Gadde, Pietro Perona, and Aleix Martinez. Near perfect GAN inversion. *CoRR*, abs/2202.11833, 2022. 2

[15] Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan. In *BMVC*, page 220. BMVA Press, 2021. 2

[16] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, pages 8649–8658, 2021. 2

[17] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM TOG*, 41(6), 2022. 2

[18] Markos Georgopoulos, James Oldfield, Grigorios G. Chrysos, and Yannis Panagakis. Cluster-guided image synthesis with unconditional models. In *CVPR*, pages 11533–11542, 2021. 3

[19] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. *CoRR*, abs/2112.01554, 2021. 2

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017. 6

[21] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. *CoRR*, abs/2112.05637, 2021. 2

[22] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *NeurIPS*, 2020. 2

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 5

[24] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, pages 852–863, 2021. 2

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE TPAMI*, 43(12):4217–4228, 2021. 1, 4

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE TPAMI*, 43(12):4217–4228, 2021. 2, 5

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116. Computer Vision Foundation / IEEE Computer Society, 2020. 1, 6

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116. Computer Vision Foundation / IEEE Computer Society, 2020. 2

[29] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in gan for real-time image editing. In *CVPR*, pages 852–861. Computer Vision Foundation / IEEE Computer Society, 2021. 3

[30] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM TOG*, 38(6):178:1–178:13, 2019. 2

[31] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM TOG*, 37(4):163, 2018. 2, 6, 7

[32] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pages 5548–5557. Computer Vision Foundation / IEEE Computer Society, 2020. 1, 3

[33] Junling Liu, Yuexian Zou, and Dongming Yang. Semanticgan: Generative adversarial networks for semantic image to photo-realistic image translation. In *ICASSP*, pages 2528–2532. IEEE, 2020. 3

[34] Moustafa Meshry, Saksham Suri, Larry S. Davis, and Abhinav Shrivastava. Learned spatial representations for few-shot talking-head synthesis. In *ICCV*, pages 13809–13818. IEEE, 2021. 2

[35] 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 TOG*, 37(6):258, 2018. 2

[36] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *CoRR*, abs/2203.17272, 2022. 1, 2

[37] Daniil Pakhomov, Sanchit Hira, Narayani Wagle, Kemar E. Green, and Nassir Navab. Segmentation in style: Unsupervised semantic image segmentation with stylegan and CLIP. *CoRR*, abs/2107.12518, 2021. 3

[38] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346. Computer Vision Foundation / IEEE Computer Society, 2019. 3, 4, 5

[39] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM ICM*, pages 484–492. ACM, 2020. 2

[40] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vqvae-2. In *NeurIPS*, pages 14837–14847, 2019. 2, 3, 5

[41] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296. Computer Vision Foundation / IEEE Computer Society, 2021. 1, 2

[42] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 42(1), 2022. 1, 2, 6, 7, 8

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 5

[44] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI*, 44(4):2004–2018, 2022. 1, 2

[45] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *CVPR*, pages 11244–11254. IEEE, 2022. 3

[46] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, pages 7135–7145, 2019. 2, 6, 7

[47] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022. 6

[48] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and Hongsheng Li. Controllable 3d face synthesis with conditional generative occupancy fields. *CoRR*, abs/2206.08361, 2022. 2

[49] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *CVPR*, pages 6141–6150. Computer Vision Foundation / IEEE Computer Society, 2020. 1, 2

[50] Ayush Tewari, Mohamed Elgharib, Mallikarjun B. R., Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. PIE: portrait image embedding for semantic control. *ACM TOG*, 39(6):223:1–223:14, 2020. 1, 2

[51] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason M. Saragih, Matthias Nießner, Rohit Pandey, Sean Ryan Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, and Michael Zollhöfer. State of the art on neural rendering. *Comput. Graph. Forum*, 39(2):701–727, 2020. 1, 2

[52] Ayush Tewari, Michael Zollhöfer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE TPAMI*, 42(2):357–370, 2020. 4

[53] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, volume 12361 of *Lecture Notes in Computer Science*, pages 716–731. Springer, 2020. 1, 2

[54] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: image synthesis using neural textures. *ACM TOG*, 38(4):66:1–66:12, 2019. 2

[55] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *CVPR*, pages 2387–2395. IEEE Computer Society, 2016. 1

[56] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM TOG*, 40(4):133:1–133:14, 2021. 1, 2, 4

[57] Rotem Tzaban, Ron Mokady, Rinon Gal, Amit H. Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. *CoRR*, abs/2201.08361, 2022. 1, 2, 4, 7, 8

[58] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016. 5

[59] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018. 6

[60] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, pages 6306–6315, 2017. 3

[61] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 6

[62] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, pages 10039–10049. Computer Vision Foundation / IEEE Computer Society, 2021. 2

[63] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 6

[64] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. Photorealistic audio-driven video portraits. *IEEE TVCG*, 26(12):3457–3466, 2020. 2

[65] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A latent transformer for disentangled face editing in images and videos. In *ICCV*, pages 13769–13778. IEEE, 2021. 2

[66] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. A style-based GAN encoder for high fidelity reconstruction of images and videos. In *ECCV*, volume 13675 of *Lecture Notes in Computer Science*, pages 581–597. Springer, 2022. 2

[67] 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 pre-trained stylegan. In *ECCV*, volume 13677 of *Lecture Notes in Computer Science*, pages 85–101. Springer, 2022. 1, 2, 3

[68] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, volume 11217 of *Lecture Notes in Computer Science*, pages 334–349. Springer, 2018. 4

[69] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, pages 9458–9467. IEEE, 2019. 2

[70] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 6

[71] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 4, 6

[72] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C. Bühler, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. *CoRR*, abs/2112.07471, 2021. 2

[73] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *CVPR*, pages 4176–4186. Computer Vision Foundation / IEEE Computer Society, 2021. 2