Controllable Radiance Fields for Dynamic Face Synthesis

Peiye Zhuang[1*†], Liqian Ma[2†], Sanmi Koyejo[1,3,4], Alexander Schwing[1]

[1]University of Illinois Urbana-Champaign, [2]ZMO AI Inc., [3]Stanford University, [4]Google Inc.

{peiye, sanmi, aschwing}@illinois.edu, liqianma.scholar@outlook.com

Free-view dynamic head synthesis with motion control

Figure 1: **Controllable free-view dynamic head synthesis.** Each row presents an identity sampled from a prior distribution and two expressions guided by a reference image (*bottom left*), viewed from multiple directions (*column 1-3 and 4-6*).

Abstract

Recent work on 3D-aware image synthesis has achieved compelling results using advances in neural rendering. However, 3D-aware synthesis of face dynamics hasn't received much attention. Here, we study how to explicitly control generative model synthesis of face dynamics exhibiting non-rigid motion (e.g., facial expression change), while simultaneously ensuring 3D-awareness. For this we propose a Controllable Radiance Field (CoRF): 1) Motion control is achieved by embedding motion features within the layered latent motion space of a style-based generator; 2) To ensure consistency of background, motion features and subject-specific attributes such as lighting, texture, shapes, albedo, and identity, a face parsing net, a head regressor and an identity encoder are incorporated. On head image/video data we show that CoRFs are 3D-aware while enabling editing of identity, viewing directions, and motion.

1. Introduction

Face synthesis is an important task with applications in digital content creation, film-making and Virtual Reality (VR). Generative Adversarial Nets (GANs), as a powerful generative model, have demonstrated remarkable success in generating high-quality faces. However, despite the impressive performance of GANs in both 2D [15, 22, 2, 24, 25, 23] and 3D image synthesis [35, 46, 7, 16], its extension to *both 3D-aware and motion-controllable* image synthesis has not been fully explored. This is largely due to the complexity of physical motion and the challenging yet consistent appearance changes of an identity. For example, dynamic head synthesis requires maintaining some 3D consistency over time to preserve facial identity while permitting other 3D deformations due to expression changes. It is even more challenging to enable interactive manipulation of identity, motion, and viewing directions.

Addressing this task of *generalizable*, *3D-aware*, and *motion-controllable* synthesis of face dynamics enables to generate never-before-seen faces including their dynamic motion, while controlling the viewpoint, as shown in Fig. 1.

Most related to this task are motion transfer methods such as face reenactment [69, 64, 47, 61, 70, 67, 49, 48].

Specifically, prior work animates a subject shown in a source image given target motion from a driving video. For this, prior methods commonly use 2D GANs to produce dynamic subject images with additional guidance such as reference keypoints [69, 64, 47, 61, 70, 67] and self-learned motion representations [48, 49]. However, due to intrinsic limitations of 2D convolutions, these approaches lack 3D-awareness. This leads to two main failure modes: 1) spurious identity changes – the subject, head pose or facial expression of the source image is distinct from the target when rotating a source head by a large angle; 2) serious distortions when transferring motion across identity based on keypoints which contain identity-specific information such as face shapes. To address these failure modes, recent face reenactment works [40, 13] estimate parameters of 3D morphable face models, e.g., poses and expression, as additional guidance. Notably, Guy et al. [13] explicitly represent heads in a 3D radiance field showing impressive 3D-consistency. However, one model per face identity needs to be trained [13].

Different from this direction, we develop a generalizable method of motion-controllable synthesis for novel, never-before-seen identity generation. This differs from prior works which either consider motion control but drop 3D-awareness [43, 56, 44, 63, 69, 64, 48, 47, 61, 70, 67], or are 3D-aware but don't consider dynamics [46, 7, 16] or can't generate never-before-seen identities [40, 13].

To achieve our goal, we propose controllable radiance fields (CoRFs). We learn CoRFs from RGB images/videos with unknown camera poses. This requires to address two main challenges: 1) how to effectively represent and control identity and motion in 3D; and 2) how to ensure spatio-temporal consistency across views and time.

To control identity and motion, CoRFs use a style-based radiance field generator which takes low-dimensional identity and motion representations as input, as shown in Fig. 1. Unlike prior head reconstruction work [40, 13] that uses ground-truth images for self-supervision, there is no ground-truth target image for a generated image of a never-before-seen person. Thus, we propose an additional motion reconstruction loss to supervise motion control.

To ensure spatio-temporal consistency, we propose three consistency constraints on face attributes, identities, and background. Specifically, at each training step, we generate two images given the same identity representation yet different motion representations. We encourage that the two images share identical environment and subject-specific attributes. For this, we apply a head regressor and an identity encoder. A head regressor decomposes the images into representations of a statistic face model, including lighting, shape, texture and albedo. Then, we compute a consistency loss which compares the predicted attribute parameters of the paired synthetic images. Moreover, we use an identity encoder to ensure the paired synthetic images share the same identity. Further, to encourage a consistent background across time, we recognize background regions using a face parsing net and employ a background consistency loss between paired synthetic images that share the same identity representation.

As there is no direct baseline, we compare CoRFs to two types of work that are most related: 1) face reenactment work [48, 49] that can control facial expression transfer; and 2) video synthesis work [56, 44, 63, 55] that can produce videos for novel identities. We evaluate the improvements of our method using the Fréchet Inception Distance (FID) [19] for videos, motion control scores, and 3D-aware consistency metrics on FFHQ [22] and two face video benchmarks [42, 9] at a resolution of 256×256. Moreover, we show additional applications such as novel view synthesis and motion interpolation in the latent motion space.

Contributions. 1) We study the new task of *generalizable*, *3D-aware*, and *motion-controllable* face generation. 2) We develop CoRFs which enable editing of motion, and viewpoints of never-before-seen subjects during synthesis. For this, we study techniques that aid motion control and spatio-temporal consistency. 3) CoRFs improve visual quality and spatio-temporal consistency of free-view motion synthesis compared to multiple baselines [56, 44, 63, 48, 69, 49, 55] on popular face benchmarks [22, 42, 9] using multiple metrics for image quality, temporal consistency, identity preservation, and expression transfer.

2. Related work

3D-aware image synthesis. GANs [15] have significantly advanced 2D image synthesis capabilities in recent years, addressing early concerns regarding diversity, resolution, and photo-realism [22, 2, 24, 25, 23, 17]. More recent GANs aim to extend 2D image synthesis to 3D-aware image generation [37, 53, 34, 35, 11], while permitting explicit camera control. For this, methods [37, 53, 34, 35, 11] use an implicit neural representation.

Implicit neural representations [32, 8, 31, 66, 28, 51, 50, 38, 5] have been introduced to encode 3D objects (i.e., its shape and/or appearance) via a parameterized neural net. Compared to voxel-based [21, 41, 65] and mesh-based [36, 59] methods, implicit neural functions represent 3D objects in continuous space and are not restricted to a particular object topology. A recent variation, Neural Radiance Fields (NeRFs) [32], represents the appearance and geometry of a *static* real-world scene with a multi-layer perceptron (MLP), enabling impressive novel-view synthesis with multi-view consistency via volume rendering. Different from the studied method, those classical approaches can't generate novel scenes.

To address this, recent work [46, 7, 16, 72, 71] studies 3D-aware generative models using NeRFs as a gener-
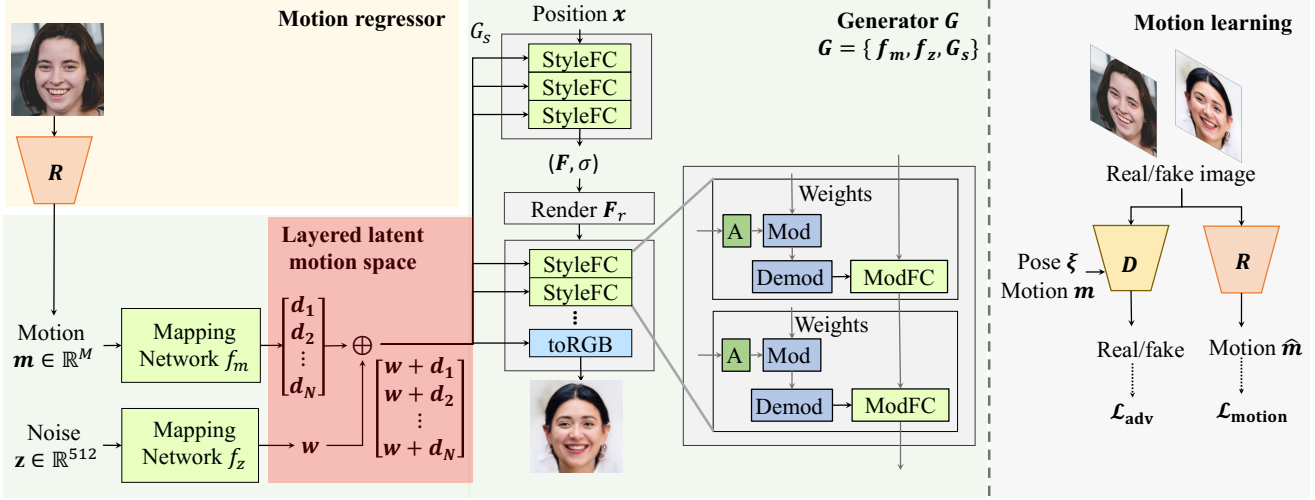
Figure 2: **Overview of Controllable Radiance Fields (CoRFs).** In the offline step, we collect motion representations m using a regressor R. At training time, the generator G renders an image given position input x, a noise z, and a motion representation m. The motion input m is embedded via a mapping network $f_m(m)$ and combined with a style vector w via a summation ($\oplus$) before being used to generate the face. The discriminator D compares real and generated images, conditioned on a camera pose ξ and a motion representation m. To ensure visual quality and motion control, we employ a discriminator loss $\mathcal{L}_{\text{adv}}$ and a motion reconstruction loss $\mathcal{L}_{\text{motion}}$.

ator. For instance, GRAF [46] and π-GAN [7] operate on a randomly sampled camera pose and latent vectors from prior distributions to produce a radiance field via an MLP. StyleNeRF and CIPS-3D [16, 72] combine a shallow NeRF network to provide low-resolution radiance fields and a 2D rendering network to produce high-resolution images with fine details. LolNeRF [39] learns 3D objects by optimizing foreground and background NeRFs together with a learnable per-image table of latent codes. Zhao *et al.* [71] develop a generative multi-plane image (GMPI) representation to ensure view-consistency. These works only generate multi-view images for *static* scenes. In contrast, the proposed CoRF targets generation of *face dynamics* while enabling to control motion. We also notice concurrent NeRF-based generative models [52, 20] aiming for 3D-aware semantic appearance or expression editing.

Dynamic object synthesis. Dynamic object synthesis has been studied using 2D- and 3D-based methods. Some 2D-based methods unconditionally generate dynamic instances via convolutional neural nets such as GANs [58, 43, 44, 56, 63], but struggle with motion control. Follow-up work incorporates additional reference information during synthesis [69, 64, 48, 47, 61, 70, 62, 40], e.g., for tasks like *face reenactment*. Our method is related but differs in that CoRF synthesizes images and controls motion for *never-before-seen* identities from *free viewpoints*. 3D-based methods recover face geometry and control the expression by changing geometry parameters [54, 26, 14, 33]. While they can control facial motion effectively, they struggle to produce real-

istic hair, teeth, and accessories [14, 33]. Recent work [13] adapts neural implicit methods, e.g., NeRFs to learn head reconstruction from images. However, one model is trained per identity, i.e., these methods don't generalize across different identities. Unlike prior work [13], CoRF generalizes across identities.

Motion conditioning. We draw inspiration from GAN-based image inversion and editing work [1, 73]. To be concrete, Abdal et al. [1] proposed to embed images back to an extended latent identity space of a StyleGAN [24], i.e., the W^+ space. This extended latent identity space was shown to be remarkably useful for image editing and expression transfer in follow-up work [73]. In our case, we embed a motion representation into a layered latent motion space, which we then broadcast to every synthesis layer in the generator. In Section 4, we illustrate that such a motion embedding strategy benefits motion control.

3D consistency preservation. To ensure 3D consistency over time, a 3D convolutional discriminator is commonly used to distinguish the source of input video clips [55]. Recent work [68] finds that when employing an implicit neural rendering net as a generator, a 2D discriminator for 2 synthetic frames in a video is sufficient. Inspired by this result, we propose consistency losses on paired synthetic frames without using a 3D convolutional discriminator.

3. Method

We aim for generalizable, 3D-aware and motion-controllable synthesis of face dynamics. Concretely, we

seek to generate never-before-seen faces while controlling dynamic motions and viewpoint arbitrarily. For this, we propose controllable radiance fields (CoRF), illustrated in Fig. 2, for which we provide an overview next.

Overview. CoRFs are composed of a generator G, a discriminator D, a regressor R and an identity encoder E. The CoRF generator G renders an RGB image I' based on a noise vector z and a motion representation m. Concretely, the generator G first applies two mapping modules f_z and f_m, before rendering the image via a style-based synthesis module G_s, i.e., $G = \{f_z, f_m, G_s\}$. Specifically, to ensure effective motion control, we embed motion representations m into a layered latent motion space using a mapping net f_m, which is then combined with a style vector $w := f_z(z)$ via a summation. The discriminator D is applied to ensure photo-realism. The regressor R and the encoder E ensure spatio-temporal consistency of synthetic faces over dynamic changes.

During training, different from prior work, we generate two images given the same identity yet different motion representations. To preserve consistency over views and poses, the regressor R and the identity encoder E extract features for the paired images which are encouraged to be similar via a consistency loss $\mathcal{L}_{\text{consist}}$ and an identity loss $\mathcal{L}_{\text{id}}$. To supervise motion control, we estimate the motion representation from images and incorporate a motion reconstruction loss $\mathcal{L}_{\text{motion}}$, to minimize the distance between the estimated motion and the conditioned one. Finally, we apply a background loss, $\mathcal{L}_{\text{bg}}$, to preserve background consistency over dynamic changes.

In the following, we briefly describe the generator G and the discriminator D (Section 3.1). We then describe our contributions: 1) how to control motion (Section 3.2) and 2) how to preserve spatio-temporal consistency (Section 3.3).

3.1. Generator and Discriminator

Generative radiance fields. We adopt the generative strategy of StyleNeRF [16] where a style-based radiance field generator G renders a synthetic image I'. Specifically, we render colors for each pixel from an integral feature F_r. To obtain an integral feature F_r, we integrate features F of 3D points along a camera ray r. We present the generator G and its workflow in Fig. 2.

Formally, we cast rays from the camera origin o through the pixels of the image plane, and accumulate the sampled density σ and feature values F along the rays with near and far planes t_n and t_f. The aggregated feature F_r is computed using the rendering equation [29] as follows:

$$F_r = \int_{t_n}^{t_f} T(t)\sigma(r(t)F(r(t), d, m, z))dt,$$

$$\text{where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(r(s), m, z)ds\right). \quad (1)$$

Here, d refers to the ray direction and $T(t)$ corresponds to the accumulated transmittance along the ray r from t_n to t. The functions $\sigma(r(t), m, z)$ and $F(r(t), d, m, z)$ are the density and the feature at a 3D location, respectively.

Different from StyleNeRF [16], the generator G in CoRF is conditioned on a motion representation $m \in \mathbb{R}^{50}$ for motion control. To this end, a ReLU mapping network f_m embeds the motion representation m into a latent space, which is then injected to the synthesis module G_s via weight modulation [24]. We discuss the details in Section 3.2. Also different from StyleNeRF [16], the generator G does not contain upsampling blocks. Instead, we increase the image resolution by sampling more rays.

Discriminator. We use the estimated camera pose ξ and the motion feature m of a training image $I \sim p_{\text{real}}$ to control generation of new images $I' \sim p_{\text{syn}}$. The discriminator D is conditioned on a camera pose ξ and a motion feature m, following the conditional strategy of StyleGAN2-ADA [23, 6]. We use a non-saturating adversarial loss with a gradient regularizer [30] applied on the discriminator D. We refer to this loss as $\mathcal{L}_{\text{adv}}$, i.e.,

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{I' \sim p_{\text{syn}}, m \sim p_m, \xi \sim p_\xi}[f(D(I', \xi, m))]$$
$$+ \mathbb{E}_{I \sim p_{\text{real}}}[f(-D(I)) + \|\nabla D(I)\|^2],$$

where $f(u) \triangleq -\log(1 + \exp(-u))$. Further, p_m and p_ξ refer to distributions over motion vectors and pose respectively. In practice, we extract motion features and pose from training data samples using [12]. During training, we randomly sample a training image with its corresponding estimated motion feature and pose to control image synthesis.

3.2. Motion control

A key goal of CoRF is to control synthesis of 3D-aware face dynamics. For this, we condition the generator G on low-dimensional motion representations m. In this section, we discuss our conditioning method, motion extraction from images, and the supervision which we use during training to encourage motion control.

Motion conditioning. We embed the motion representation in a layered latent motion space using the mapping module f_m. The layered latent motion space is a concatenation of N latent motion vectors, denoted as $[d_1, \ldots, d_N]^T := f_m(m)$, where $d_i \in \mathbb{R}^{512}$ for $i \in \{1, \ldots, N\}$. Here, N corresponds to the number of layers in the synthesis module G_s, which we find to lead to feature embeddings with superior performance. The latent motion vectors are then combined with the latent style vector w. Formally, we combine style vector and motion representation via a summation, i.e.,

$$[w + d_1, \ldots, w + d_N]^T.$$

The conditioned latent vectors are then given to each layer of the synthesis module G_s via modulation [24].
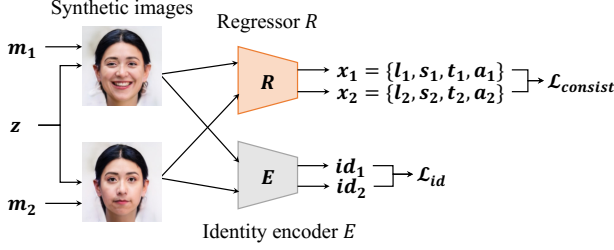
Figure 3: **Consistency learning.** We propose two losses, $\mathcal{L}_{\text{consist}}$ and $\mathcal{L}_{\text{id}}$, to ensure spatio-temporal consistency of paired generated images over dynamics, which includes lighting (l), texture (t), shape (s), albedo (a), and identity (id).

Instead of summation, other conditioning strategies for control exist. However, we find that not every strategy works equally well in our task. For example, recent head reconstruction work [13] concatenates a motion representation with a noise vector z which is then provided as input to the first layer of the model. We compare our conditioning strategy with this prior method [13] in Section 4 and find that our motion conditioning strategy accelerates convergence of the motion reconstruction loss $\mathcal{L}_{\text{motion}}$.

Efficient motion extraction. To obtain a low-dimensional motion representation m for face dynamics, we extract coefficients of a blendshape basis of a statistical head model using a pre-trained regressor R with an accurate statistical head model [27]. For this we use a publicly available package.*

Note that our method is also compatible with other pre-trained head models. Compared with existing work [13] that uses an *optimization-based* approach for expression estimation, use of a pre-trained regressor R enjoys two advantages. First, using the regressor R is computationally more efficient as coefficient prediction requires only a single forward-pass. Second, we can use the regressor R as a differentiable plug-in to regress and supervise the expression representation of synthetically generated frames.

Motion learning. A motion reconstruction loss $\mathcal{L}_{\text{motion}}$ is applied to supervise motion control. It assesses whether the synthetically generated image exhibits the motion for which the representation was provided as input. For this we use a simple quadratic loss, i.e.,

$$\mathcal{L}_{\text{motion}} = \mathbb{E}_{I' \sim p_{\text{syn}}} \| m - \hat{m} \|^2, \qquad (2)$$

where $\hat{m}$ is estimated by the regressor R, i.e., $\hat{m} = R(I')$.

3.3. Spatio-temporal consistency

The aforementioned losses are insufficient to ensure spatio-temporal consistency. To address this, we propose

---
*Code available from https://github.com/YadiraF/DECA

two losses which focus on the generated face and one loss which focuses on background regions. All three losses are simple yet effective at preserving spatio-temporal consistency. In summary, to optimize CoRF we minimize the weighted objective

$$\mathcal{L} \triangleq \lambda_1 \mathcal{L}_{\text{adv}} + \lambda_2 \mathcal{L}_{\text{motion}} + \lambda_3 \mathcal{L}_{\text{consist}} + \lambda_4 \mathcal{L}_{\text{id}} + \lambda_5 \mathcal{L}_{\text{bg}}, \quad (3)$$

where the hyper-parameters $\lambda_1, \lambda_2, \ldots, \lambda_5$ adjust the strength of the different loss terms. In practice, the regressor R and the identity encoder E are pre-trained offline on a large-scale dataset [12, 4]. We describe each of the three additional losses $\mathcal{L}_{\text{consist}}$, $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{bg}}$ next.

Consistency of attributes. The motion conditioning occasionally changes lighting and person-specific attributes such as albedo and shape, resulting in flickering textures in rendered images. To alleviate this issue, we propose a consistency loss to supervise attributes independent from motion (see Fig. 3). Specifically, we extract lighting, texture, shape, and albedo representations for paired artificially generated images which have an identical noise vector z, yet a different motion representation m. These extracted attributes are independent from motion and thus should be identical for paired images. To encourage their consistency via a loss, we subsume representations of these attributes (lighting, albedo, texture, shape) for both paired images in the set $\mathcal{A} = \{(l_1, l_2), (a_1, a_2), (t_1, t_2), (s_1, s_2)\}$. We then compute the consistency loss

$$\mathcal{L}_{\text{consist}} = \mathbb{E}_{I' \sim p_{\text{syn}}} \left[\sum_{(x_1, x_2) \in \mathcal{A}} \beta_x \| x_1 - x_2 \|^2 \right], \qquad (4)$$

by comparing pairs of estimated attributes x_1 and x_2. Here, l, a, t, and s are the attributes of lighting, albedo, texture, and shape, respectively. β_x is the loss weight. In practice, these attributes are predicted by the regressor R.

Consistency of identity. Beyond attribute consistency, we use an identity loss on paired images, following prior face reconstruction work [12] (see Fig. 3). To be concrete, an encoder E outputs identity embeddings of paired generated images which have an identical noise vector z yet a different motion representation. Let I_1' and I_2' refer to the two generated images. The identity loss is proportional to the negative cosine similarity between embeddings of both images, i.e., it is computed via

$$\mathcal{L}_{\text{id}} = \mathbb{E}_{I_1' \sim p_{\text{syn}}, I_2' \sim p_{\text{syn}}} \left[1 - \frac{E(I_1') \cdot E(I_2')}{\| E(I_1') \|_2 \cdot \| E(I_2') \|_2} \right]. \quad (5)$$

Consistency of background. Beyond the constraints on faces, we also want to encourage that the background remains mostly static over time. For this, we recognize background regions of paired synthetic images I_1' and I_2', using

Source   Driving   FOMM [48] MRAA [49] Ours   Source   Driving   FOMM [48] MRAA [49] Ours
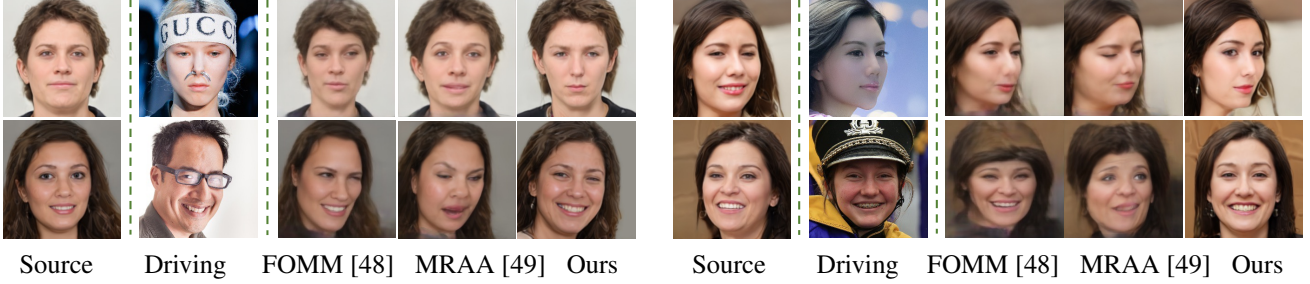
Figure 4: **Visual comparison of face reenactment on FFHQ.** We transfer expression and pose from driving images to the source images. We compare our results with results of FOMM [48] and MRAA [49] at a resolution of 256×256 in challenging cases. We show face reenactment across identity, gender, pose, or with a partially occluded driving image.

a pre-trained face parsing net.[†] The corresponding binary background masks, which we refer to via s_1 and s_2, use a value of 1 to indicate the background regions and a value of 0 to indicate the face regions. Using both masks, we define a background consistency loss on the overlapping background regions of paired images via

$$\mathcal{L}_{\text{bg}} = \mathbb{E}_{I_1' \sim p_{\text{syn}}, I_2' \sim p_{\text{syn}}} \left[\frac{\|(s_1 \cdot s_2) \cdot (I_1' - I_2')\|_2}{\|s_1 \cdot s_2\|} \right], \quad (6)$$

where $\cdot$ denotes an element-wise product. Due to the limited space, we illustrate details of the background loss $\mathcal{L}_{\text{bg}}$ in the supplementary material.

4. Experiments

4.1. Experimental settings

Datasets and preprocessing. We evaluate the proposed approach on FFHQ [22] and two face video datasets, including Faceforensics++ [42] and VoxCeleb2 [9]. (i) **FFHQ** [22]: We show CoRF learning motion dynamics on FFHQ, which is a single image dataset. (ii) **The Faceforensics++ dataset** [42] contains 1,000 raw video sequences. (iii) **The VoxCeleb2 dataset** [9] contains over 1 million videos for 6,112 identities. As videos in these two face datasets [42, 9] contain low-quality frames with small or occluded faces, a pre-processing step is applied to clean the data. Specifically, we detect the faces in the videos using a pretrained landmark detection model [3], sharpen faces, filter out frames with small, blurry, or occluded faces, and align the remaining faces. We finally use 990 and 4,804 videos in each dataset [42, 9] for CoRF training.

Baselines and evaluation metrics. We evaluate CoRF from two perspectives. First, we quantitatively compare our CoRF with face reenactment work [48, 55]. To this end, we use the Fréchet Inception Distance (FID) [19] and 3 cosine similarity scores for identity preservation and expression transfer, namely ID, EXP[†] and EXP[‡]. Regarding identity

---

[†]Code available from https://github.com/zllrunning/face-parsing.PyTorch
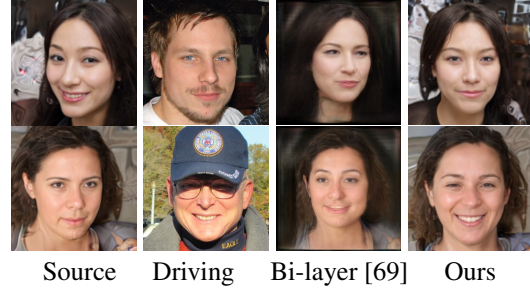

Source   Driving   Bi-layer [69]   Ours

Figure 5: **Visual comparison to Bi-layer [69].**

preservation, we extract face features using a popular image identity recognition mode pretrained on VGGface2 [4]. Regarding expression transfer, for a fair evaluation, we use another pretrained facial expression estimation network [10] (EXP[‡]), which differs from the encoder [12] (EXP[†]) used in our model. Second, as CoRF generates never-before-seen dynamic faces from a prior noise distribution, similar to unconditional video synthesis approaches [56, 44, 63, 55], we compare CoRF with these baselines using the Fréchet Video Distance (FVD) [57] score. The FVD score is a video-level FID score commonly used to examine both visual quality and temporal coherence of synthesized dynamics [63, 18].

4.2. Results

Comparison to face reenactment methods. We present a visual comparison to FOMM [48], MRAA [49], and Bi-layer [69] in Fig. 4 and Fig. 5. For this, we transfer expressions and poses from real 'driving' images to synthetic source images in four challenging cases, including cross-identity, cross-gender, cross-pose, and driving image occlusion. We observe that the baseline methods struggle in these cases, while CoRF better preserves face geometry and identity. Additionally, we show targeted video generation on the VoxCeleb2 dataset [9] at a 256×256 resolution in Fig. 6. Different from the face reenactment baselines [48, 49], CoRF is capable of controlling more factors such as novel views and field-of-view (FOV) without the requirement of a reference image, e.g., in Fig. 7. We show additional results in the supplementary material.

Figure 6: **Targeted video generation on VoxCeleb2 data.** We generate head videos (*row 2 and 4*) conditioned on facial expression information of the driving frames (*row 1 and 3*) at a 256×256 resolution. Note, the synthetic portraits have a similar expression as the driving frame (see highlighted eyes in row 2, and the mouth in row 4).

In Tab. 1, we quantitatively compare face reenactment methods on 10,000 pairs of source and driving images (variances in the supplementary material). We find CoRF to significantly improve upon the baselines [48, 49] with regards to identity preservation and expression transformation.

Complex expression animation. We show complex expression animation in Fig. 8. Among all randomly generated images, we observe the results of the proposed method to follow the expressions in the driving image, highlighting the method's capability.

Figure 7: **Rendered results under camera zoom-in to zoom-out.**

Motion interpolation in the latent space. To analyze the latent motion space, we render from multiple views a trajectory of synthetic images conditioned on linear interpolations between two motion representations, as shown in Fig. 9. The smooth interpolation between two expressions over viewpoint changes demonstrates that the latent motion space captures semantics and CoRF enables to generate dynamic faces from multiple views with 3D-awareness.

Comparison to unconditional video synthesis methods. We present a quantitative temporal coherence comparison in Tab. 2 where we use 10,000 images/videos for numerical evaluation. Tab. 2 shows that CoRF significantly improves video-level (FVD) quality compared to the baselines [63,

	FaceF [42]			Vox2 [9]		
	[48]	[47]	Ours	[48]	[47]	Ours
FID $\downarrow$	24.44	20.62	**1.53**	32.01	26.26	**3.64**
ID $\uparrow$	0.934	0.965	**0.973**	0.942	0.968	**0.988**
EXP† $\uparrow$	0.592	0.763	**0.815**	0.748	0.837	**0.938**
EXP‡ $\uparrow$	0.543	0.595	**0.601**	0.459	0.481	**0.731**

Table 1: **Quantitative evaluation for face reenactment.** We compare our method with FOMM [48] and MRAA [47] at a 256×256 resolution. We compute the FID score [19], the face identity preservation score (ID), and expression similarities (two expression extractors [12] and [10] are used, denoted as EXP† and EXP‡, respectively).

Driving Result Driving Result Driving Result

Figure 8: **Rendered results with complex driving expressions.**

Figure 9: **Multi-view motion interpolation in the latent space.** We linearly interpolate two motion representations during synthesis (*column 1-6*), viewed from multiple directions (*row 1-3*).

55]. In Fig. 10, we also show a comparison of unconditional video synthesis methods [63, 55] at a 64×64 resolution (size is chosen to be compatible with the baselines [63, 55]). We observe that CoRF produces sharper faces and persistent head geometry over time and space (*row 3*).

Ablation study. In Fig. 11 we show the motion reconstruction loss curves during training with different motion conditioning strategies. The baseline method concatenates a motion representation m with a noise vector z, as used in prior work [13] (*blue curve*). Compared to the baseline method, our motion embedding strategy (*green and red curves*) accelerates convergence of the motion reconstruction loss $\mathcal{L}_{motion}$. Moreover, we evaluate the effectiveness

	64×64				256×256	
	[56]	[63]	[44]	Ours	[55]	Ours
FaceF [42]	70.71	65.58	51.66	**34.33**	53.26	**35.67**
Vox2 [9]	40.06	36.18	31.79	**20.12**	35.81	**21.91**

Table 2: **Quantitative evaluation for video synthesis.** We evaluate our method using FVD [57] scores ($\downarrow$) and compare to MoCoGAN [56], G^{3}AN [63], TGANv2 [44] at the 64×64 resolution, and with MOCOGAN-HD [55] at the 256×256 resolution.
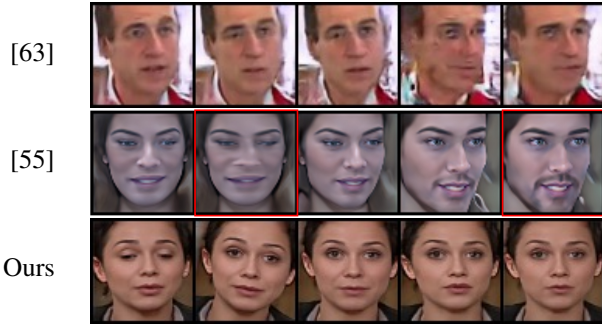

Figure 10: **Visual comparison of video synthesis.** We compare our method with popular unconditional video synthesis approaches G^{3}AN [63] and MoCoGAN-HD [55] on VoxCeleb2 [9]. The appearance of our method's results remains consistent over view and expression changes, while baselines struggle (see frames highlighted with red border).

of conditioning the discriminator on motion. We find that motion conditioning of the discriminator (*green curve*) is slightly better than omission of motion conditioning (*red curve*). Thus, we use a conditional discriminator in all the experiments.

Fig. 12 provides a visual comparison for the ablation settings of the proposed consistency losses, illustrated below the figure (*column 1-3*). For each setting, we render two synthetic images applying an identical identity representation yet a different motion representation (*top* and *bottom*). We find that the identity loss $\mathcal{L}_{id}$, the background loss $\mathcal{L}_{bg}$, and the consistency loss $\mathcal{L}_{consist}$ help to preserve the face identity and a stable background over expression and pose changes. The table on the right shows the averaged loss values with or without optimizing the corresponding loss terms, which supports the visual observation.

5. Conclusion

We propose controllable radiance fields (CoRFs) to address the task of generalizable, 3D-aware, and motion-controllable synthesis of face dynamics. CoRFs allow not only explicit 3D-aware motion control and arbitrary view-
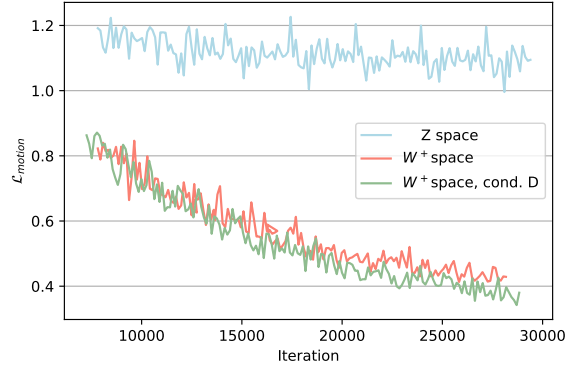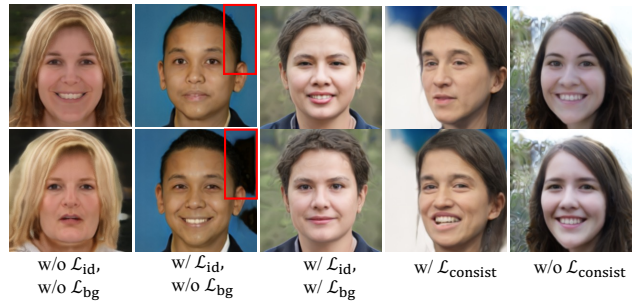

Figure 11: **Analysis of motion conditioning strategies.**

	$\mathcal{L}_{id}$	$\mathcal{L}_{bg}$
w/o $\mathcal{L}_{id}$, w/o $\mathcal{L}_{bg}$	0.07	0.26
w/ $\mathcal{L}_{id}$, w/ $\mathcal{L}_{bg}$	0.04	0.10

Figure 12: **Ablation study for losses.** We render paired synthetic images conditioned on the same identity representation but different motion representations (*row 1-2*) using different ablation settings for identity, background, and consistency losses. In the table, we show averaged values of the losses $\mathcal{L}_{id}$ and $\mathcal{L}_{bg}$ (*column 2-3*), without (or with) optimizing them.

points but also diverse identity generation in photo-realistic quality. Extensive experiments highlight that CoRFs improve performance w.r.t. 3D consistency and identity preservation over pose and motion changes, compared to state-of-the-art methods on benchmark video datasets. We hope this work inspires future work on the challenging task of 3D-aware dynamic face synthesis.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018.

[5] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *ECCV*, 2020.

[6] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, et al. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022.

[7] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021.

[8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019.

[9] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[10] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR-Workshops*, 2019.

[11] Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. Learning to generate chairs, tables and cars with convolutional networks. *TPAMI*, 2017.

[12] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *SIGGRAPH*, 2021.

[13] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, 2021.

[14] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. In *TOG*, 2018.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[16] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *ICLR*, 2022.

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.

[18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018.

[19] 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*, 2017.

[20] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *CVPR*, 2022.

[21] Danilo Jimenez Rezende, SM Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. In *NeurIPS*, 2016.

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[23] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *NeurIPS*, 2020.

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020.

[26] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *TOG*, 2018.

[27] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 2017.

[28] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. In *NeurIPS*, 2019.

[29] Nelson Max. Optical models for direct volume rendering. *TVCG*, 1995.

[30] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018.

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019.

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[33] 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. In *TOG*, 2018.

[34] Thu Nguyen-Phuoc, Chuan Li, Stephen Balaban, and Yong-Liang Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. In *NeurIPS*, 2018.

[35] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019.

[36] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *ICCV*, 2019.

[37] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, 2017.

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.

[39] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *CVPR*, 2022.

[40] Yurui Ren, Ge Li, Yuanqi Chen, et al. Pirenderer: Controllable portrait image generation via semantic neural rendering. *ICCV*, 2021.

[41] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, 2017.

[42] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

[43] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017.

[44] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. In *IJCV*, 2020.

[45] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[46] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020.

[47] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019.

[48] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019.

[49] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, et al. Motion representations for articulated animation. In *CVPR*, 2021.

[50] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020.

[51] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019.

[52] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. *CVPR*, 2022.

[53] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *ECCV*, 2018.

[54] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[55] Yu Tian, Jian Ren, Menglei Chai, et al. A good image generator is what you need for high-resolution video synthesis. In *ICLR*, 2021.

[56] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.

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

[58] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016.

[59] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018.

[60] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

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

[62] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021.

[63] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *CVPR*, 2020.

[64] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, 2018.

[65] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NeurIPS*, 2016.

[66] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, 2019.

[67] Ceyuan Yang, Zhe Wang, Xinge Zhu, et al. Pose guided human video generation. In *ECCV*, 2018.

[68] Sihyun Yu, Jihoon Tack, Sangwoo Mo, et al. Generating videos with dynamics-aware implicit generative adversarial networks. *ICLR*, 2022.

[69] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *ECCV*, 2020.

[70] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019.

[71] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G. Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *ECCV*, 2022.

[72] Peng Zhou, Lingxi Xie, Bingbing Ni, et al. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. *arxiv*, 2021.

[73] Peiye Zhuang, Oluwasanmi Koyejo, and Alexander G Schwing. Enjoy your editing: Controllable gans for image editing via latent space navigation. *ICLR*, 2021.

A. Additional experiments

A.1. Comparison of video generations

We show synthetic videos from 2 baseline methods MoCoGAN [56] and TGANv2 [44], and our CoRF at 64×64 resolution in Figure 13.

Figure 13: **Video generation results of baseline methods.** We show videos generated by 2 baseline methods MoCo-GAN [56] (*row 1-2*) and TGANv2 [44] (*row 3-4*), and ours (*row 5-6*).

Figure 14: **Rendered results under camera zoom-in to zoom-out.**

In Tab. 3 we show the standard deviations of the evaluation metrics using MoCoGAN [56], G^{3}AN [63], TGANv2 [44] (corresponding mean values in Table 1 of the main manuscript), FOMM [48] (corresponding mean values in Table 2 of the main manuscript), and our CoRF.

A.2. 3D reconstruction

In Fig. 15 we show COLMAP [45] reconstruction results for an identity with 2 different expressions. Following [16], we sample 36 camera poses to generate images and obtain the reconstructed point clouds from COLMAP with default parameters and without known camera poses.

A.3. Camera zoom

In Fig. 14 we show that CoRF enables to generate reasonable results with camera zoom effects.

B. Ethics discussion

From an application perspective, we develop 3D-aware and motion-controllable image synthesis. We hope our method provides inspiration for generalization of NeRF-GAN models. Importantly, we want to highlight the dangers of automated image manipulation. Similar to deepfake tasks whose aim is to produce fabricated images and videos that appear to be real. Improper use of image manipulation approaches might raise issues with regard to information security, property, etc., and can be used for information fabrication or information surgery. Edited image detection techniques [60] have been proposed recently to avoid the aforementioned issues, which promotes growth in both domains.

Table 3: **Standard deviations.** We show the standard deviations of KID, ID, EXP †, EXP $\ddagger$, and AKD scores using MoCo-GAN [56], G^{3}AN [63], TGANv2 [44], FOMM [48]. The standard deviations are computed on the two datasets [42, 9] at 64×64 resolution (*row 1-4*) and 256×256 resolution (*row 5-6*).

No.	Method	FaceForensics++ [42]					Vox2 [9]				
		KID	ID	EXP †	EXP ‡	AKD	KID	ID	EXP †	EXP ‡	AKD
1	[56] (64)	1e-3	-	-	-	-	1e-3	-	-	-	-
2	[63] (64)	9e-4	-	-	-	-	1e-3	-	-	-	-
3	[44] (64)	1e-3	-	-	-	-	9e-4	-	-	-	-
4	Ours (64)	5e-4	-	-	-	-	5e-4	-	-	-	-
5	[48] (256)	-	4e-2	2e-1	2e-1	2.2	-	5e-2	2e-1	2e-1	2.5
6	Ours (256)	-	2e-2	1e-1	1e-2	2.5	-	4e-2	1e-1	1e-1	2.9

Expression 1

COLMAP

Expression 2

COLMAP

Figure 15: **COLMAP reconstructions for models trained on FFHQ.** We show COLMAP reconstruction results (*row 2 and 4*) for an identity with 2 different expressions (*row 1 and 3*).