

Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation

Xian Liu¹, Yinghao Xu¹, Qianyi Wu², Hang Zhou¹, Wayne Wu³, Bolei Zhou¹

¹The Chinese University of Hong Kong ²Monash University ³SenseTime Research

{alvinliu@ie, xy119@ie, zhouhang@link, bzhou@ie}.cuhk.edu.hk,
qianyi.wu@monash.edu, wuwenyan@sensetime.com

Abstract

Animating high-fidelity video portrait with speech audio is crucial for virtual reality and digital entertainment. While most previous studies rely on accurate explicit structural information, recent works explore the implicit scene representation of Neural Radiance Fields (NeRF) for realistic generation. In order to capture the inconsistent motions as well as the semantic difference between human head and torso, some work models them via two individual sets of NeRF, leading to unnatural results. In this work, we propose Semantic-aware Speaking Portrait NeRF (SSP-NeRF), which creates delicate audio-driven portraits using one unified set of NeRF. The proposed model can handle the detailed local facial semantics and the global head-torso relationship through two semantic-aware modules. Specifically, we first propose a Semantic-Aware Dynamic Ray Sampling module with an additional parsing branch that facilitates audio-driven volume rendering. Moreover, to enable portrait rendering in one unified neural radiance field, a Torso Deformation module is designed to stabilize the large-scale non-rigid torso motions. Extensive evaluations demonstrate that our proposed approach renders more realistic video portraits compared to previous methods. Project page: <https://alvinliu0.github.io/projects/SSP-NeRF>.

1. Introduction

Generating high-fidelity video portraits based on speech audio is of great importance to various applications like digital human, film-making and video dubbing. Many researchers tackle the task of audio-driven talking face or video portrait generation by using deep generative models. Several works rely solely on learning-based image reconstruction, which typically synthesize static results of low-resolution [8, 11, 47, 63, 69, 82, 83]. Other methods utilize explicit structural intermediate representations such as 2D landmarks [9, 15, 59] or 3D facial models [7, 51, 61, 67, 70, 76, 84]. Though some of them can generate high-fidelity images [59, 61], the errors in structured representation predic-

tion (e.g., expression parameters of a 3D Morphable Model (3DMM) [3]) lead to inaccurate face deformation [83].

Recently, the implicit 3D scene representation of Neural Radiance Fields (NeRF) [34] provides a new perspective for realistic generation. It enables free-view control with higher image quality compared to explicit methods, which is suitable for the video portrait generation task. Gafni *et al.* [18] first involve NeRF in the dynamic human head modeling from single-view data in a video-driven manner. However, an accurate explicit 3D model is still required in their settings. Moreover, they model torso consistently with the head, which leads to unstable results. Guo *et al.* [21] further propose AD-NeRF for audio-driven talking head synthesis. In particular, they build two individual sets of NeRF for head and torso modeling conditioned on audio input. Such a straightforward pipeline suffers from head-torso separation during the render stage, making generated results unnatural.

Based on previous studies, we identify two key challenges for incorporating NeRF into video portrait generation: 1) Each facial part’s appearance and moving patterns are intrinsically connected but substantially different, especially when associated with audios. Thus weighing all rendering areas equally without semantic guidance would lead to blurry details and difficulties in training. 2) While it is easy to bind head pose with camera pose, the global movements of the head and torso are in significant divergence. As the human head and torso are non-rigidly connected, modeling them with one set of NeRF is an ill-posed problem.

In this work, we develop a method called Semantic-aware Speaking Portrait NeRF (**SSP-NeRF**), which generates stable audio-driven video portraits of high-fidelity. We show that *semantic awareness is the key to handle both local facial dynamics and global head-torso relationship*. Our intuition lies in the fact that different parts of a speaking portrait have different associations with speech audio. While other organs like ears move along with the head, the high-frequency mouth motion that is strongly correlated with audios requires additional attention. To this end, we devise an *Semantic-Aware Dynamic Ray Sampling* module, which consists of an *Implicit Portrait Parsing* branch and a *Dy-*

namic Sampling Strategy. Specifically, the parsing branch supervises the modeling with facial semantics in 2D plane. Then the number of rays sampled at each semantic region could be adjusted dynamically according to the parsing difficulty. Thus more attention can be paid to the small but important areas like lip and teeth for better lip-synced results. Besides, we also enhance the semantic information by anchoring a set of latent codes to the vertices of a roughly predicted 3DMM [3] without expression parameters.

On the other hand, since the head and torso motions are rigidly bound together in the current NeRF, a correctly positioned torso cannot be rendered even with the portrait parsing results. We further observe the relationship between head and torso: while they share the same translational movements, the orientation of torso seldom changes with head pose under the speaking portrait setting. Thus we model non-rigid deformation through a *Torso Deformation* module. Concretely, for each point (x, y, z) in the 3D scene, we predict a displacement $(\Delta x, \Delta y, \Delta z)$ based on the head-canonical view information and time flows. Interestingly, although there are local deformations on the face, the deformation module implicitly learns to focus on the global parts. This design facilitates portrait stabilization in one unified set of NeRF. Experiments demonstrate that our method generates high-fidelity video portraits with better lip-synchronization and better image quality efficiently.

To summarize, our work has three main contributions: **(1)** We propose the *Semantic-Aware Dynamic Ray Sampling* module to grasp the detailed appearance and local dynamics of each portrait part without using accurate structural information. **(2)** We propose the *Torso Deformation* module that implicitly learns the global torso motion to prevent unnatural head-torso separated results. **(3)** Extensive experiments show that the proposed **SSP-NeRF** renders high-fidelity audio-driven video portraits with one unified NeRF in an efficient manner, which outperforms state-of-the-art methods on both objective evaluations and human studies.

2. Related Work

Audio-Driven Talking Head Synthesis. Driving talking head with speech audio has a bunch of applications, which is of great research interest to computer vision and graphics. Conventional works mostly resort to stitching techniques [5, 6, 17], where a predefined set of phoneme-mouth correspondence rules is used to modify mouth shapes. With the rapid growth of deep neural networks, end-to-end frameworks are proposed. One category of methods, namely image reconstruction-based methods, generate talking face by latent feature learning and image reconstruction [11, 14, 22, 46, 47, 57, 60, 63, 66, 80, 82, 83, 85]. For example, Chung *et al.* [11] propose the first end-to-end method with an encoder-decoder pipeline. Zhou *et al.* [82] explicitly disentangle identity and word information for

better feature extraction. Prajwal *et al.* [47] achieve synchronous lip movements with a pretrained lip-sync expert. However, these methods can only generate fix-sized images with low resolution. Another line of approaches named model-based methods utilize structural intermediate representations like 2D facial landmarks or 3D representations to bridge the mappings from audio to complicated facial images [7, 9, 15, 33, 56, 59, 61, 64, 70, 76, 84]. Typically, Chen *et al.* [9] and Das *et al.* [15] first predict 2D landmarks then generate faces. Thies *et al.* [61] and Song *et al.* [56] infer facial expression parameters from audio in the first stage, then generate 3D mesh for final image synthesis. But errors in intermediate prediction often hinder accurate results. In contrast to these two lines of works, our method can render more realistic speaking portraits of high-fidelity without any accurate structural information.

Implicit Representation Methods. Recent works leverage implicit functions for learning scene representations [26, 28, 34, 37, 55, 79], where multi-layer perceptron (MLP) weights are used to represent the mapping from spatial coordinates to a signal in continuous space like occupancy [32, 44, 50, 54], signed distance function [20, 65, 75], color and volume density [2, 16, 29, 34], semantic label [24, 81] and neural feature map [10, 36]. A recent popular work named Neural Radiance Fields (NeRF) [34] optimizes an underlying continuous volumetric scene mapping from 5D coordinate of spatial location and view direction to implicit fields of color and density for photo-realistic view results. Naturally, naive NeRF is confined to static scenes, which triggers a branch of studies to extend NeRF for dynamic scenes [4, 18, 27, 31, 38–41, 43, 45, 48, 49, 58, 62]. However, few works focus on complicated dynamic scenes like speaking portraits. The main difficulty lies in the learning of cross-modal associations between different portrait parts and speech audio. One relevant work [21] synthesizes talking head with two individual sets of NeRF for head and torso, making generated results fall apart. In this work, we take semantics as guidance to grasp each portrait part’s local dynamics and appearances for fine-grained results efficiently. A deformation module further enables us to synthesize stable video portraits using one unified set of NeRF.

3. Our Approach

We present **Semantic-aware Speaking Portrait NeRF (SSP-NeRF)** that generates delicate audio-driven portraits with one unified set of NeRF. The whole pipeline is depicted in Fig. 1. In this section, we first review the preliminaries and the problem setting of video portrait synthesis with neural radiance fields (Sec. 3.1). We then introduce the *Semantic-Aware Dynamic Ray Sampling* module, which facilitates fine-grained appearance and dynamics modeling for each portrait part with semantic information (Sec. 3.2). Furthermore, we elaborate the *Torso Deformation* module

that handles non-rigid torso motion by learning location displacements (Sec. 3.3). Finally, the volume rendering process and network training details are described (Sec. 3.4).

3.1. Preliminaries and Problem Setting

Given images with calibrated camera intrinsics and extrinsics, NeRF [34] represents a scene using a continuous volumetric radiance field F . Specifically, F is modeled by an MLP, which takes 3D spatial coordinates $\mathbf{x} = (x, y, z)$ and 2D view directions $\mathbf{d} = (\theta, \phi)$ as input, then outputs the implicit fields of color $\mathbf{c} = (r, g, b)$ and density σ . In this way, the MLP weights store scene information by the mapping of $F : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$. To compute the color of a single pixel, NeRF [34] approximates the volume rendering integral using numerical quadrature [30]. Consider the ray $\mathbf{r}(v) = \mathbf{o} + v\mathbf{d}$ from camera center \mathbf{o} , its expected color $\hat{C}(\mathbf{r})$ with near and far bounds v_n and v_f is calculated as:

$$\hat{C}(\mathbf{r}) = \int_{v_n}^{v_f} T(v)\sigma(\mathbf{r}(v))\mathbf{c}(\mathbf{r}(v), \mathbf{d})dv, \quad (1)$$

where $T(v) = \exp(-\int_{v_n}^v \sigma(\mathbf{r}(u))du)$ is the accumulated transmittance along the ray from v_n to v . With the hierarchical volume sampling, both coarse and fine MLPs are optimized by minimizing the photometric discrepancy.

In this work, we focus on audio-driven video portrait generation in a basic setting: 1) The camera pose $\{R, \tau\}$ is given by the estimated rigid head pose, where the rotation matrix $R \in \mathbb{R}^{3 \times 3}$ and the translation vector $\tau \in \mathbb{R}^{3 \times 1}$ are estimated by 3DMM [3] on the face; 2) The audio feature $\mathbf{a} \in \mathbb{R}^{64}$ is extracted by a pretrained DeepSpeech [1] model and further processed with a light-weight audio encoder to get more compact representation. Therefore, the implicit function of audio-driven portrait **basic** setting is:

$$F^{\text{basic}} : (\mathbf{x}, \mathbf{d}, \mathbf{a}) \rightarrow (\mathbf{c}, \sigma). \quad (2)$$

Guo *et al.* [21] use an off-the-shelf parsing method [25] to divide training images into head and torso for individual NeRF modeling. Following their settings, we assume that the semantic parsing maps are also available in our method.

3.2. Semantic-Aware Dynamic Ray Sampling

To avoid the unnatural head-torso separation problem described in Sec. 1, we render the whole portrait with one unified set of NeRF. However, two problems remain: 1) The associations between different portrait parts and audio are different. For example, audio is more related to lip movements than torso motions. How to grasp the fine-grained appearance and dynamics of each portrait part remains unsolved; 2) Since the rays are uniformly sampled over the whole image, how to make the model pay more attention to small but important regions like mouth is challenging.

Implicit Portrait Parsing Branch. Our solution to the first problem is to add a parsing branch. Since the portrait parts of the same semantic category share similar motion patterns and texture information, it will be beneficial for the appearance and geometry learning in NeRF, which is also proven in recent implicit representation studies [73–75, 81]. As shown in Fig. 1, we extend the original NeRF with an additional parsing branch that predicts the semantic information. Note that since a certain 3D coordinate’s semantic label is view-invariant, the parsing branch does not condition on view direction \mathbf{d} . Specifically, suppose there are totally K semantic categories, the parsing branch maps the 3D spatial coordinate \mathbf{x} to semantic logits $\mathbf{s}(\mathbf{x})$ over K classes, which is further conditioned on audio \mathbf{a} . Hence the expected semantic logits $\hat{S}(\mathbf{r})$ along the ray $\mathbf{r}(v)$ with near and far bounds v_n and v_f can be calculated as:

$$\hat{S}(\mathbf{r}) = \int_{v_n}^{v_f} T(v)\sigma(\mathbf{r}(v), \mathbf{a})\mathbf{s}(\mathbf{r}(v), \mathbf{a})dv, \quad (3)$$

$$\text{where } T(v) = \exp(-\int_{v_n}^v \sigma(\mathbf{r}(u), \mathbf{a})du). \quad (4)$$

Such semantic awareness can naturally distinguish each part over the whole image, thus figuring out different associations between audio and different portrait regions.

Dynamic Ray Sampling Strategy. To generate delicate facial images with lip-synced results, we have to care for each portrait part, especially those small but crucial regions. Original NeRF **uniformly** samples rays on the image plane [34]. Such an unconstrained ray sampling process focuses on big regions (*e.g.*, background and cheek) yet ignores small regions (*e.g.*, lip and teeth) that are important for fine-grained results. Therefore, we use semantic information to guide the ray sampling process dynamically. In particular, we denote all the points that are sampled on the image as $\Omega = \bigcup_{i=1}^K \Omega_i$, where K is the total number of semantic categories in parsing map and Ω_i is the set of points that are sampled on the i -th semantic class. During the training stage, we calculate the average loss of each category \mathcal{L}_i for the previous epoch (the sum of semantic loss and RGB loss, which will be introduced in Sec. 3.4), and then dynamically sample rays across K categories by:

$$N_{\Omega_i} = \frac{\mathcal{L}_i}{\sum_{i=1}^K \mathcal{L}_i} \cdot N_s, \quad (5)$$

where N_{Ω_i} denotes the number of rays distributed to the i -th category and N_s is the total number of sampled rays. We identify two benefits for such design: 1) The average loss of a semantic category is area-agnostic. Thus the learning process will equally sample those small-area regions; 2) Some image parts are comparatively easier to learn. For example, the texture of eye is more complicated than that of background. This leads to lower loss of background category

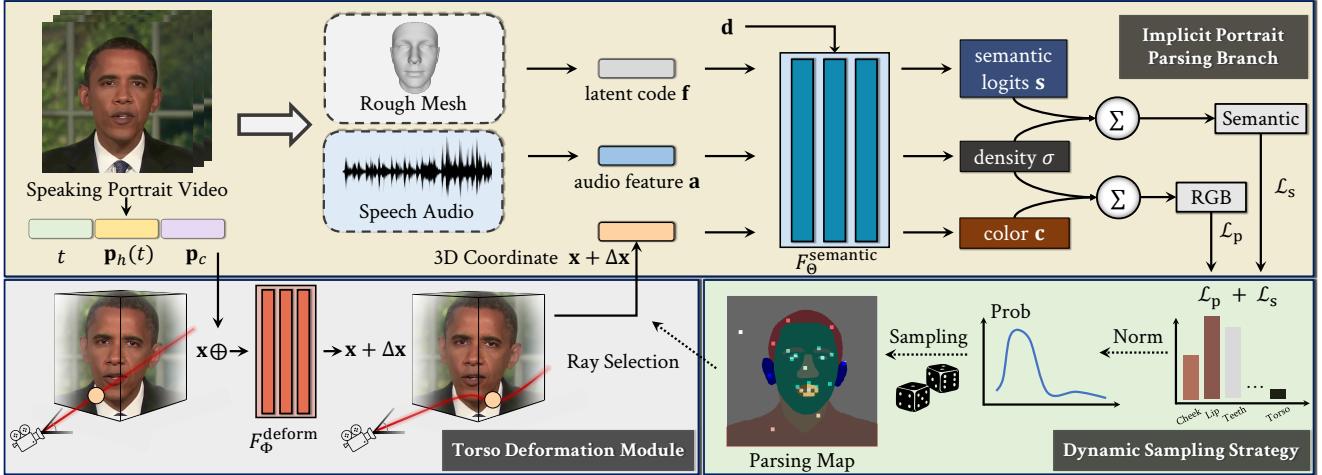


Figure 1. **Overview of Semantic-aware Speaking Portrait NeRF (SSP-NeRF) framework.** In Implicit Portrait Parsing Branch (yellow), the semantic-aware implicit function $F_{\Theta}^{\text{semantic}}$ takes latent code f , audio feature a , 3D coordinate x and view direction d as input, then outputs the semantic logits s , density σ and color c of the scene. In Dynamic Sampling Strategy (green), the RGB loss \mathcal{L}_p and semantic loss \mathcal{L}_s are utilized to guide the distribution of rays sampled at each semantic region. In particular, the Torso Deformation module (grey) uses an implicit function F_{Φ}^{deform} to map from the time t , head pose $p_h(t)$, canonical pose p_c and 3D coordinate x into the displacement Δx , which generates the deformed 3D coordinate $x + \Delta x$ to model non-rigid torso motions.

and dynamically drives the implicit function to pay more attention to hard-to-learn regions. Our experiment further shows that this design can accelerate training as well.

Structured 3D Information. 3D cues are crucial for NeRF to grasp better spatial geometry information as proved in [16, 65, 71, 78]. In our framework, we identify that the awareness of *rough* 3D facial information can serve as guidance for face semantic and geometry learning. Concretely, a 3D facial model is built with *mean* expression parameters. We take inspiration from [44, 45, 72] to anchor a set of latent codes to the vertices of 3DMM model and diffuse to 3D space with SparseConvNet [19] to extract latent code volume. We query the latent code $f \in \mathbb{R}^{88}$ at each point by trilinear interpolation¹ similar to Peng *et al.* [45]. Such structured 3D information could enhance semantic learning by giving similar features to the same semantic class while discriminative features among different semantic categories. Till now, we can update the **basic** setting in Eq. 2 to **semantic-aware** implicit function with parameters Θ :

$$F_{\Theta}^{\text{semantic}} : (\mathbf{x}, \mathbf{d}, \mathbf{a}, \mathbf{f}) \rightarrow (\mathbf{c}, \sigma, \mathbf{s}). \quad (6)$$

3.3. Torso Deformation Module

As mentioned in Sec. 3.1, the estimated head pose serves as camera pose. However, such straightforward treatment ignores the fact that head and torso motions are inconsistent. To tackle this problem, we design a Torso Deformation module to stabilize the large-scale non-rigid torso motions. **Torso Deformation Implicit Function.** Concretely, an implicit function is optimized to estimate the deformation field

of $\Delta \mathbf{x} = (\Delta x, \Delta y, \Delta z)$ at a specific time instant t . Based on the observation that torso pose changes slightly and is weakly related to speech audio, the head pose $p_h(t)$ at time t and a canonical pose p_c are further given as references to learn the displacement Δx , while audio feature a does not serve as input. Note that for convenience, the canonical pose p_c is set as the head pose of the first frame, thus the displacement $\Delta x = 0$ when $t = 0$. The implicit function for torso deformation with parameters Φ is formulated as:

$$F_{\Phi}^{\text{deform}} : (\mathbf{x}, t, p_h(t), p_c) \rightarrow \Delta \mathbf{x}. \quad (7)$$

Notably, although such deformation is added to the *whole image*, we empirically find that *only the torso part* tends to be deformed, while the facial dynamics are naturally modeled by semantic-aware implicit function in Eq. 6. Such disentanglement will be further analyzed in Sec. 4.5.

Overall Implicit Function. Combine the semantic-aware implicit function with our proposed Torso Deformation module, we can model the overall implicit function as:

$$F_{\Theta}^{\text{overall}} : (\mathbf{x} + \Delta \mathbf{x}, \mathbf{d}, \mathbf{a}, \mathbf{f}) \rightarrow (\mathbf{c}, \sigma, \mathbf{s}), \\ \text{where } \Delta \mathbf{x} = F_{\Phi}^{\text{deform}}(\mathbf{x}, t, p_h(t), p_c). \quad (8)$$

3.4. Volume Rendering and Network Training

Volume Rendering with Deformation. Since the Torso Deformation module is proposed to compensate for non-rigid torso motions, we accordingly adapt the NeRF’s original volume rendering formulas for color and semantic distribution in Eq. 1, Eq. 3 and Eq. 4. Consider a certain 3D point $\mathbf{x}(v) = \mathbf{o} + v\mathbf{d}$ located on the ray emitted from center

¹Please refer to the supplementary material for more details.

\mathbf{o} on view direction \mathbf{d} , its warped coordinate at time t with head pose $\mathbf{p}_h(t)$ and canonical pose \mathbf{p}_c is computed as:

$$\mathbf{x}'(v, t) = \mathbf{x}(v) + F_\Phi^{\text{deform}}(\mathbf{x}(v), t, \mathbf{p}_h(t), \mathbf{p}_c). \quad (9)$$

With the deformed 3D coordinate $\mathbf{x}'(v, t)$ along the modified ray path $\mathbf{r}'(v, t)$, we can calculate the expected color $\hat{C}(\mathbf{r}'(v, t))$ and semantic logits $\hat{S}(\mathbf{r}'(v, t))$ with near and far bounds v_n and v_f under **semantic-aware** setting as:

$$\begin{aligned}\hat{C}(\mathbf{r}') &= \int_{v_n}^{v_f} T'(v, t) \sigma(\mathbf{r}'(v, t), \mathbf{a}, \mathbf{f}) \mathbf{c}(\mathbf{r}'(v, t), \mathbf{d}, \mathbf{a}, \mathbf{f}) dv, \\ \hat{S}(\mathbf{r}') &= \int_{v_n}^{v_f} T'(v, t) \sigma(\mathbf{r}'(v, t), \mathbf{a}, \mathbf{f}) \mathbf{s}(\mathbf{r}'(v, t), \mathbf{a}, \mathbf{f}) dv,\end{aligned}$$

and $T'(v, t) = \exp(-\int_{v_n}^v \sigma(\mathbf{r}'(u, t), \mathbf{a}, \mathbf{f}) du)$, (10)

where $T'(v, t)$ is the accumulated transmittance along the ray path $\mathbf{r}'(v, t)$ from v_n to v . Note that the estimated semantic logits $\hat{S}(\mathbf{r}')$ are subsequently transformed into multi-class distribution $p(\mathbf{r}')$ through softmax operation.

Network Training. Similar to NeRF [34] that simultaneously optimizes coarse and fine models with hierarchical volume rendering, we train the network with following photometric loss \mathcal{L}_p and semantic loss \mathcal{L}_s :

$$\begin{aligned}\mathcal{L}_p &= \sum_{\mathbf{r}' \in \mathcal{R}'} \left[\left\| \hat{C}_c(\mathbf{r}') - C(\mathbf{r}') \right\|_2^2 + \left\| \hat{C}_f(\mathbf{r}') - C(\mathbf{r}') \right\|_2^2 \right], \\ \mathcal{L}_s &= - \sum_{\mathbf{r}' \in \mathcal{R}'} \left[\sum_{k=1}^K p^k(\mathbf{r}') \log \hat{p}_c^k(\mathbf{r}') + \sum_{k=1}^K p^k(\mathbf{r}') \log \hat{p}_f^k(\mathbf{r}') \right],\end{aligned} \quad (11)$$

where \mathcal{R}' is the set of *deformed* camera rays passing through image pixels; $C(\mathbf{r}')$, $\hat{C}_c(\mathbf{r}')$ and $\hat{C}_f(\mathbf{r}')$ denote the ground-truth, coarse volume predicted and fine volume predicted pixel color for the deformed ray \mathbf{r}' , respectively; and $p^k(\mathbf{r}')$, $\hat{p}_c^k(\mathbf{r}')$ and $\hat{p}_f^k(\mathbf{r}')$ denote the ground-truth, coarse volume predicted and fine volume predicted multi-class semantic distribution for the deformed ray \mathbf{r}' , respectively. The overall learning objective for the framework is:

$$\mathcal{L} = \mathcal{L}_p + \lambda \mathcal{L}_s, \quad (12)$$

where λ is the weight balancing coefficient. At the training stage, the network parameters Θ and Φ of the implicit functions in Eq. 8 are updated based on above loss function.

4. Experiments

4.1. Dataset and Preprocessing

Dataset Collection. Our method targets to synthesize audio-driven facial images. Hence a certain person’s speaking portrait video with audio track is needed. Unlike

previous studies that demand large-corpus data or hours-long videos, we can achieve high-fidelity results with short videos of merely a few minutes. In particular, we extend the *publicly-released* video set of Guo *et al.* [21] and obtain videos of average length 6,750 frames in 25 fps.²

Training Data Preprocessing. We follow the basic setting [21] of audio-driven video portrait generation to preprocess training data: (1) For the speech audio, it is first processed by a pretrained DeepSpeech [1] model. Then a 1D convolutional network with self-attention mechanism is adopted [21, 61] for smooth feature learning. The extracted audio feature $\mathbf{a} \in \mathbb{R}^{64}$ is fed into the implicit function in Eq. 8. (2) For the video frames, they are cropped and resized to 450×450 to make talking portrait in the center. An off-the-shelf method [25] is leveraged to obtain parsing maps of total 11 semantic classes. The background image and head pose are estimated in a similar way to Guo *et al.* [21]. Note that the estimated head pose $\mathbf{p}_h(t)$ at time t is treated as camera pose and the canonical pose \mathbf{p}_c is set as the starting frame’s head pose, *i.e.*, $\mathbf{p}_c = \mathbf{p}_h(0)$ in Eq. 8.

4.2. Experimental Settings

Comparison Baselines. We compare our method with recent representative works: (1) **ATVG** [9], which uses 2D landmark to guide facial image synthesis; (2) **Wav2Lip** [47] that achieves state-of-the-art lip-sync performance by pre-training a lip-sync expert; (3) **MakeitTalk** [84], a representative 3D landmark-based approach; (4) **PC-AVS** [83] which generates pose-controllable talking face by modularized audio-visual representation; (5) **NVP** [61] that first infers expression parameters from audio, then generates images with a neural renderer; (6) **SynObama** [59] which learns mouth shape changes for facial image warping; (7) **AD-NeRF** [21], which is the first work that uses implicit representation of NeRF to achieve arbitrary-size talking head synthesis. In particular, we also show the evaluations directly on the **Ground Truth** for a clearer comparison.

Implementation Details.² The $F_\Theta^{\text{semantic}}$ and F_Φ^{deform} together with their associated fine models all consist of simple 8-layers MLPs with hidden size of 128 and ReLU activations. Following NeRF [34], positional encoding is applied to each 3D coordinate \mathbf{x} , view direction \mathbf{d} and time instant t to map the input into higher dimensional space for better learning. The positional encoder is formulated as: $\gamma(q) = \langle (\sin(2^l \pi q), \cos(2^l \pi q)) \rangle_0^L$, where we use $L = 10$ for \mathbf{x} , and $L = 4$ for \mathbf{d} and t . For the parsing maps, we use $K = 11$ categories for semantic guidance, including cheek, eye, eyebrow, ear, nose, teeth, lip, neck, torso, hair and background. The structured 3D feature extractor is borrowed from [45] that processes feature volume with 3D sparse convolutions and outputs latent code with $2 \times, 4 \times, 8 \times, 16 \times$ downsampled sizes. The semantic weight λ is em-

²Please refer to supplementary material for more details.

Methods	Testset A				Testset B [61]			Testset C [59]	
	PSNR ↑ N/A	SSIM ↑ 1.000	LMD ↓ 0	Sync ↑ 6.632	LMD ↓ 0	Sync ↑ 5.973	LMD ↓ 0	Sync ↑ 6.204	
ATVG [9]	24.125	0.725	5.261	4.708	5.074	6.208	5.869	4.419	
Wav2Lip [47]	26.667	0.793	5.811	6.952	4.893	6.980	5.740	6.806	
MakeitTalk [84]	25.522	0.704	7.238	3.873	6.704	4.105	6.512	3.925	
PC-AVS [83]	25.712	0.756	5.406	5.834	5.247	6.113	5.771	5.983	
NVP [61]	-	-	-	-	5.072	5.689	-	-	
SynObama [59]	-	-	-	-	-	-	5.485	5.938	
AD-NeRF [21]	29.814	0.844	5.183	6.092	5.119	5.613	5.392	6.012	
SSP-NeRF (Ours)	32.649	0.868	4.934	6.438	4.892	5.886	5.208	6.186	

Table 1. **The quantitative results of *cropped setting* on Testset A, B [61] and C [59].** We compare the proposed Semantic-aware Speaking Portrait NeRF (SSP-NeRF) against recent SOTA methods [9,21,47,59,61,83,84] and ground truth under four metrics. For LMD the lower the better, and the higher the better for other metrics. Note that the detailed comparison settings are elaborated in Sec. 4.3.

Testset A (450×450)					
Methods	PSNR	SSIM	LMD	Sync	# of params
GT	N/A	1.000	0	5.291	-
AD-NeRF	29.186	0.827	4.892	4.237	2.69M
Ours	32.785	0.876	4.495	4.993	1.10M

Table 2. **The quantitative results of *full resolution setting* on Testset A.** We compare our method with AD-NeRF [21] that also generates whole portrait with full resolution of 450×450 . We evaluate image quality and lip-sync accuracy of synthesized results. The number of parameters for each model is shown in table.

pirically set to 0.04. The model is trained with 450×450 images during $400k$ iterations with a batch size of $N_s = 1024$ rays. The framework is implemented in PyTorch [42] and trained with Adam optimizer [23] of learning rate $5e - 4$ on a single Tesla V100 GPU for 36 hours.

4.3. Quantitative Evaluation

Evaluation Metrics. We employ evaluation metrics that have been previously used in talking face generation. We adopt **PSNR** and **SSIM** [68] to evaluate the image quality of generated results; Landmark Distance (**LMD**) [8] and **SyncNet Confidence** [12, 13] to account for the accuracy of mouth shapes and lip sync. Note that the landmarks are detected from synthesized images for the computation of LMD metric. Other metrics such as **CSIM** [7, 77] for measuring identity preserving and **CPBD** [35] for measuring result sharpness are shown in supplementary material.

Comparison Settings. The reconstruction/model-based methods require large-corpus training data or long videos, hence we directly inference with their publicly-released best models. Note that all baseline methods except for [21] fail to generate the whole portrait with full resolution, we divide our comparisons into two settings: 1) The *cropped setting*

in Table 1, where we crop the generated facial image with same region and resize into same size for fair evaluation metric comparison. 2) The *full resolution setting* in Table 2, where we compare with AD-NeRF [21] that could also synthesize the whole portrait with full resolution of 450×450 .

In the first setting, since NVP [61] and SynObama [59] do not provide pretrained models, we conduct comparisons on three datasets: (1) **Testset A**, the collected dataset mentioned in Sec. 4.1; (2) **Testset B**, where we extract speech audio from the demo of NVP to drive other baselines; (3) **Testset C**, where the audio from SynObama’s demo is used for animation. Note that the metrics for measuring image quality (PSNR and SSIM) are not evaluated on Testset B and C due to the low image quality of original videos. In the second setting, the experiment is only conducted on Testset A for high-resolution comparison. We further compare the number of model parameters against AD-NeRF [21] to show the efficiency of our proposed approach.

Evaluation Results. The results of the *cropped setting* and *full resolution setting* are shown in Table 1 and Table 2, respectively. It can be seen that the proposed **SSP-NeRF** achieves the best evaluation results in most metrics: (1) In the cropped setting, we synthesize fine-grained facial images with detailed local appearance and dynamics of each portrait part. Note that Wav2Lip [47] uses SyncNet [12, 13] for pretraining, which makes their results on SyncNet Confidence even better than the ground truth. Our performance on the LMD metric is the best, and the SyncNet Confidence of our model is close to the ground truth on all three datasets, showing that we can generate accurate lip-sync video portraits. (2) In the full resolution setting, the human face as well as torso part is evaluated. Different from AD-NeRF’s separated rendering pipeline, our design of Torso Deformation module facilitates steady results. The statistics on both model’s parameter number are shown in Table 2. Notably, our method is trained with $400k \times 1024$

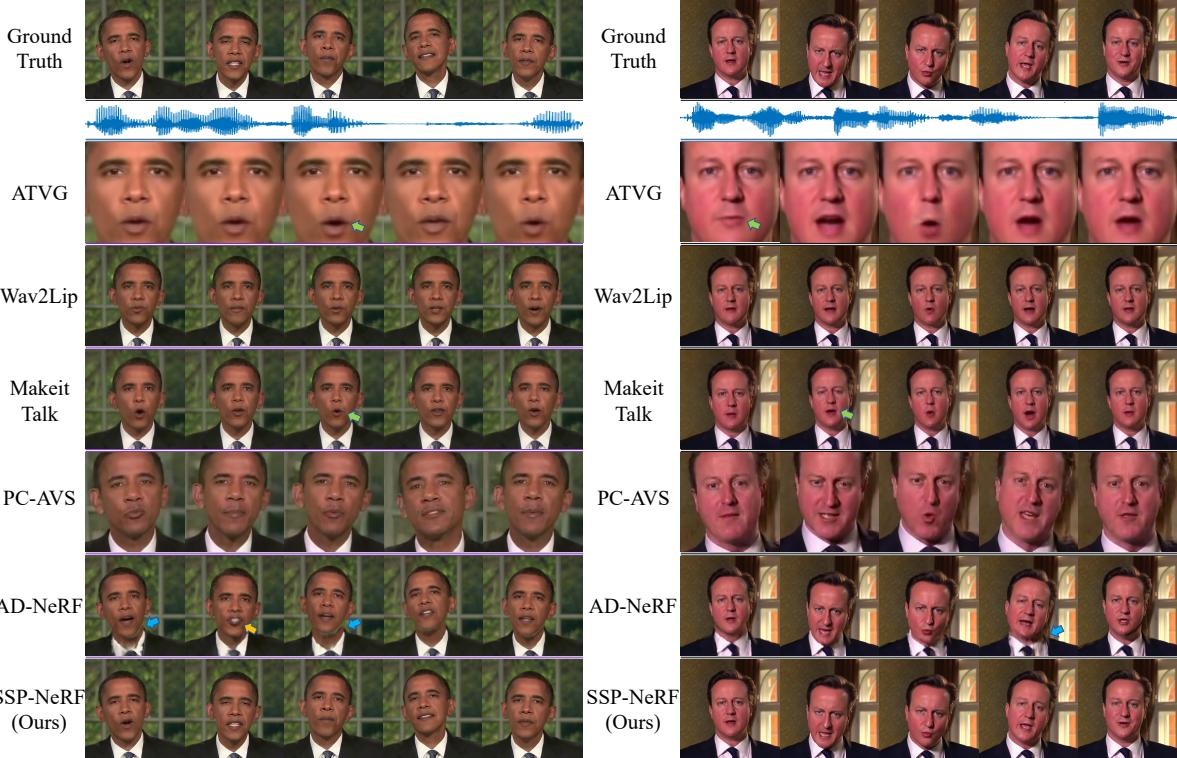


Figure 2. **The comparison of generated key frame results on Testset A.** We show the synthesized talking heads of ground truth, baseline methods [9, 21, 47, 83, 84] and ours. Please zoom in for better visualization. More qualitative comparisons can be found in demo video.

Methods	ATVG	Wav2Lip	MakeitTalk	PC-AVS	NVP	SynObama	AD-NeRF	SSP-NeRF (Ours)
Lip-sync Accuracy	3.02	4.23	2.89	4.05	4.26	4.21	4.16	4.26
Video Realness	1.63	2.86	2.45	3.83	3.89	3.64	4.09	4.28
Image Quality	1.72	2.42	2.78	2.36	4.02	3.49	4.18	4.43

Table 3. **User study results on the generation quality of audio-driven portrait.** The rating is of scale 1-5, with the larger the better. We compare the lip-sync accuracy, video realness and image quality of baseline methods [9, 21, 47, 59, 61, 83, 84] and our proposed **SSP-NeRF**.

sampled rays, while AD-NeRF [21] uses $400k \times 2048$ rays for each model. Hence we generate portraits of *better* image quality and *better* lip-synchronization in a *more compact* model with *fewer* iterations, proving the effectiveness and efficiency of **SSP-NeRF**.²

4.4. Qualitative Evaluation

To compare the generated results of each method, we show the key frames of two clips in Fig. 2. The figure shows that our method synthesizes more lip-synced video portraits of higher image quality. In particular, ATVG [9] and MakeitTalk [84] rely on precise facial landmarks, which leads to inaccurate mouth shapes (green arrow); Wav2Lip [47] creates static talking heads; PC-AVS [83] fails to preserve the speaker’s identity, making generated results unrealistic. Moreover, all the image reconstruction-based methods [47, 83] or model-based methods [9, 84] fail to synthesize the whole portrait of high-fidelity simultane-

ously. Although AD-NeRF [21] manages to create full-resolution results, the separated rendering pipeline with uniform ray sampling leads to head-torso separation (as highlighted by blue arrows) and blurry results (orange arrow).

User Study.² Since subjective evaluation can reflect the quality of audio-driven portrait, a user study is further conducted. Specifically, we sample 30 audio clips from Testset A, B and C for all methods to generate results, and then involve 18 participants for user study. The Mean Opinion Scores rating protocol is adopted for evaluation, which requires the participants to rate three aspects of generated speaking portraits: (1) *Lip-sync Accuracy*; (2) *Video Realness*; (3) *Image Quality*. The rating is based on a scale of 1 to 5, with 5 being the maximum and 1 being the minimum.

The results are shown in Table 3. Since NeRF enables full-resolution whole portrait generation with expressive pose, both AD-NeRF [21] and our method score comparatively high on *Image Quality* and *Video Realness*. Be-

Methods	PSNR \uparrow	SSIM \uparrow	LMD \downarrow	Sync \uparrow
w/o F_{Φ}^{deform}	27.472	0.791	4.635	4.744
deform by \mathbf{a}	28.013	0.802	5.329	3.871
SSP-NeRF	32.785	0.876	4.495	4.993

Table 4. Ablation study results of Torso Deformation module.



Figure 3. The visualized deformation heatmap. From left to right, we show the predicted displacements over the whole portrait image under three cases of small, medium and large pose. We can observe that: 1) The deformations are mostly on the torso region; 2) The larger pose is, the more displacements it will learn.

sides, the users prefer our generated speaking portraits to AD-NeRF’s [21] due to the fine-grained local rendering and stable torso motions provided by our framework design. Although PC-AVS [83] also creates pose-controllable talking faces, the inaccuracy of implicit pose code extraction weakens their realness. Note that Wav2Lip [47], NVP [61] and SynObama [59] achieve competitive scores on *Lip-sync Accuracy*. However, they rely on large corpus or long training videos, while we merely take a short video as input, showing the efficacy of our method. To further measure the disagreement on scoring among the participants, the Fleiss’s-Kappa³ statistic is calculated on 18 participants’ ratings. The Fleiss-Kappa value is 0.816, which can be interpreted as “almost perfect agreement”.

4.5. Ablation Study

In this section, we present ablation study on the Testset A in terms of two key modules proposed in our framework. **Torso Deformation Module.** We conduct ablation experiments under two settings: (1) w/o F_{Φ}^{deform} , where we directly synthesize the whole portrait without deforming 3D coordinates. The results are shown in Table 4 (line1), where the ill-posed rendering leads to blurry torso with low image quality. To further investigate the efficacy of Torso Deformation module, we visualize the heatmap of learned displacements in Fig. 3. Since audio feature is not input to the deformation implicit function, it tends to warp the weakly audio-related torso part, while the strongly audio-related mouth movements are mostly modeled by $F_{\Theta}^{\text{semantic}}$. The marginal drop in lip-sync metrics also suggests that the deformation module majorly takes effect on the torso part.

Another ablation setting is: (2) deform by \mathbf{a} , where the audio input \mathbf{a} is fed to F_{Φ}^{deform} rather than $F_{\Theta}^{\text{semantic}}$, i.e., the audio feature \mathbf{a} , head pose $\mathbf{p}_h(t)$ and canonical pose \mathbf{p}_c are leveraged to deform both the human face and torso part si-

Methods	PSNR \uparrow	SSIM \uparrow	LMD \downarrow	Sync \uparrow
w/o semantic branch	29.479	0.832	4.886	4.562
w/o dynamic sample	29.514	0.826	4.916	4.490
w/o 3D information	31.059	0.845	4.683	4.739
SSP-NeRF	32.785	0.876	4.495	4.993

Table 5. Ablation study of Semantic-Aware Dynamic Ray Sampling Module. The ablation settings are elaborated in Sec. 4.5.

multaneously. The lip-sync performance drops dramatically as shown in Table 4 (line2). We guess the reason lies in distinct correlations between audio and different portrait parts. It is hard for deformation module to handle audio synchronization and portrait parts deformation at the same time.

Semantic-aware Dynamic Ray Sampling Module. The ablative experiments contain: (1) w/o semantic branch, which means the semantic supervision \mathcal{L}_s is not used; (2) w/o dynamic sample, where the rays are uniformly sampled over image plane; (3) w/o 3D information, which means the 3D feature \mathbf{f} is eliminated. The results in Table 5 verify that the semantic awareness enables the model to better grasp each part’s appearance and geometry. The dynamic ray sampling further facilitates fine-grained results.

5. Broader Impact

Ethical Consideration. Animating realistic talking portrait has extensive applications like digital human and filmmaking. On the other hand, it could be misused for malicious purposes such as identity theft, deepfake generation, and media manipulation. Recent studies have shown promising results in detecting deepfakes [52, 53]. However, the lack of realistic data limits their performance. As part of our responsibility, we feel obliged to share our generated results with the deepfake detection community to improve the model’s robustness. We believe that the proper use of this technique will enhance the healthy development of both machine learning research and digital entertainment.

Limitation and Future Work. Our proposed SSP-NeRF achieves audio-driven video portrait generation of high-fidelity. However, the method still has limitations. The speed of synthesizing images is slow due to the heavy computation of rendering high-quality images. We also observe that the language gap between training and driven audio makes the synthesized mouth look unnatural occasionally [21]. We will address these issues in future work.

6. Conclusion

In this paper, we propose a novel framework Semantic-aware Speaking Portrait NeRF (**SSP-NeRF**) for audio-driven portrait generation. We introduce Semantic-Aware Dynamic Ray Sampling module to grasp the detailed appearance and the local dynamics of each portrait part with-

³https://en.wikipedia.org/wiki/Fleiss%27_kappa

out using accurate structural information. We then propose a Torso Deformation module to learn global torso motion and prevent head-torso separated results. Extensive experiments show that our approach can synthesize more realistic video portraits compared to the previous methods.

References

- [1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016. [3](#), [5](#)
- [2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. [2](#)
- [3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. [1](#), [2](#), [3](#)
- [4] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1450–1459, 2021. [2](#)
- [5] Matthew Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28, 1999. [2](#)
- [6] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360, 1997. [2](#)
- [7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020. [1](#), [2](#), [6](#)
- [8] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018. [1](#), [6](#)
- [9] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019. [1](#), [2](#), [5](#), [6](#), [7](#)
- [10] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021. [2](#)
- [11] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. [1](#), [2](#)
- [12] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016. [6](#)
- [13] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. [6](#)
- [14] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. [2](#)
- [15] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision*, pages 408–424. Springer, 2020. [1](#), [2](#)
- [16] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021. [2](#), [4](#)
- [17] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4):796–804, 1968. [2](#)
- [18] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, June 2021. [1](#), [2](#)
- [19] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. [4](#)
- [20] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. [2](#)
- [21] Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [22] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. [2](#)
- [23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [6](#)
- [24] Amit Kohli, Vincent Sitzmann, and Gordon Wetzstein. Inferring semantic information with 3d neural scene representations. *arXiv e-prints*, pages arXiv–2003, 2020. [2](#)
- [25] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. [3](#), [5](#)
- [26] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *arXiv preprint arXiv:2007.11571*, 2020. [2](#)
- [27] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *arXiv preprint arXiv:2106.02019*, 2021. [2](#)

- [28] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 2
- [29] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2
- [30] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3
- [31] Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escalano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, et al. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–21, 2020. 2
- [32] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2
- [33] Moustafa Meshry, Saksham Suri, Larry S Davis, and Abhinav Shrivastava. Learned spatial representations for few-shot talking-head synthesis. *arXiv preprint arXiv:2104.14557*, 2021. 2
- [34] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3, 5
- [35] Niranjan D Narvekar and Lina J Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience*, pages 87–91. IEEE, 2009. 6
- [36] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2
- [37] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2
- [38] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. *arXiv preprint arXiv:2104.03110*, 2021. 2
- [39] Pablo Palafox, Aljaz Bozic, Justus Thies, Matthias Nießner, and Angela Dai. Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, volume 3, 2021. 2
- [40] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2
- [41] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2
- [42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 6
- [43] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. *arXiv preprint arXiv:2105.02872*, 2021. 2
- [44] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 2, 4
- [45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2, 4, 5
- [46] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: a deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 80–88, 2017. 2
- [47] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 1, 2, 5, 6, 7, 8
- [48] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2
- [49] Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pva: Pixel-aligned volumetric avatars. *arXiv preprint arXiv:2101.02697*, 2021. 2
- [50] Daxuan Ren, Jianmin Zheng, Jianfei Cai, Jiatong Li, Haiyong Jiang, Zhongang Cai, Junzhe Zhang, Liang Pan, Mingyuan Zhang, Haiyu Zhao, et al. Csg-stump: A learning friendly csg-like representation for interpretable shape parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12478–12487, 2021. 2
- [51] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando De la Torre, and Yaser Sheikh. Audio-and gaze-driven

- facial animation of codec avatars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 41–50, 2021. 1
- [52] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 8
- [53] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 8
- [54] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 2
- [55] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *arXiv preprint arXiv:1906.01618*, 2019. 2
- [56] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody’s talkin’: Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020. 2
- [57] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*, 2018. 2
- [58] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Neff: Neural light-transport field for portrait view synthesis and relighting. *arXiv preprint arXiv:2107.12351*, 2021. 2
- [59] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 1, 2, 5, 6, 7, 8
- [60] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017. 2
- [61] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pages 716–731. Springer, 2020. 1, 2, 5, 6, 7, 8
- [62] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhofer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2
- [63] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413, 2020. 1, 2
- [64] 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 *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 2
- [65] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2, 4
- [66] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 2
- [67] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021. 1
- [68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [69] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 1
- [70] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1478–1486, 2021. 1, 2
- [71] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. In *Advances in Neural Information Processing Systems(NeuroIPS)*, 2021. 4
- [72] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 4
- [73] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, October 2021. 3
- [74] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *arXiv preprint arXiv:2106.12052*, 2021. 3
- [75] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 3
- [76] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 1, 2
- [77] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 6
- [78] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural reflectance surfaces for

- sparse-view 3d reconstruction in the wild. In *Conference on Neural Information Processing Systems*, 2021. 4
- [79] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escalano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)*, 40(1):1–17, 2021. 2
- [80] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 2
- [81] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3
- [82] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1, 2
- [83] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 1, 2, 5, 6, 7, 8
- [84] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makettalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 1, 2, 5, 6, 7
- [85] Hao Zhu, Huaibo Huang, Yi Li, Aihua Zheng, and Ran He. Arbitrary talking face generation via attentional audio-visual coherence learning. *arXiv preprint arXiv:1812.06589*, 2018. 2