Pattern Recognition Letters 185 (2024) 160–166

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Semantic-aware hyper-space deformable neural radiance fields for facial avatar reconstruction

Kaixin Jin, Xiaoling Gu *, Zimeng Wang, Zhenzhong Kuang, Zizhao Wu, Min Tan, Jun Yu

Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, No. 1158, No. 2 Baiyang Street, Hangzhou, 310018, Zhejiang, China

ARTICLE INFO

Editor: Gangyi Jiang

MSC:
41A05
41A10
65D05
65D17

Keywords:
Facial avatar reconstruction
Hyper-space deformation
Semantic guidance
Neural radiance fields

ABSTRACT

High-fidelity facial avatar reconstruction from monocular videos is a prominent research problem in computer graphics and computer vision. Recent advancements in the Neural Radiance Field (NeRF) have demonstrated remarkable proficiency in rendering novel views and garnered attention for its potential in facial avatar reconstruction. However, previous methodologies have overlooked the complex motion dynamics present across the head, torso, and intricate facial features. Additionally, a deficiency exists in a generalized NeRF-based framework for facial avatar reconstruction adaptable to either 3DMM coefficients or audio input. To tackle these challenges, we propose an innovative framework that leverages semantic-aware hyper-space deformable NeRF, facilitating the reconstruction of high-fidelity facial avatars from either 3DMM coefficients or audio features. Our framework effectively addresses both localized facial movements and broader head and torso motions through semantic guidance and a unified hyper-space deformation module. Specifically, we adopt a dynamic weighted ray sampling strategy to allocate varying degrees of attention to distinct semantic regions, enhancing the deformable NeRF framework with semantic guidance to capture fine-grained details across diverse facial regions. Moreover, we introduce a hyper-space deformation module that enables the transformation of observation space coordinates into canonical hyper-space coordinates, allowing for the learning of natural facial deformation and head-torso movements. Extensive experiments validate the superiority of our framework over existing state-of-the-art methods, demonstrating its effectiveness in producing realistic and expressive facial avatars. Our code is available at https://github.com/jematy/SAHS-Deformable-Nerf.

1. Introduction

Reconstructing high-fidelity facial avatars from monocular videos holds significance in computer graphics and computer vision, offering promising applications in digital human modeling, video conferencing, and virtual reality. Nevertheless, achieving superior fidelity in capturing the dynamics of human facial expressions remains a formidable challenge due to the inherent complexity of facial geometry and appearance variations [1–3]. The earliest attempts towards face reconstruction utilized parametric 3D Morphable Models (3DMMs) [4]. However, these models struggle to capture detailed facial features such as hair, skin texture, and accessories like glasses [5,6].

Recently, Neural Radiance Field (NeRF) [7] has revolutionized the field of three-dimensional reconstruction by offering photorealistic rendering of novel-view images. The impressive rendering quality of NeRF has inspired great interest in facial avatar reconstruction. For example, NerFACE [8] learns a dynamic NeRF to render 4D facial avatars by utilizing pose and expression coefficients estimated by 3DMMs [4]. However, NerFACE suffers from two limitations. Firstly, it models the torso and head with a single NeRF, resulting in unnatural torso movements. Secondly, it ignores variations in local facial dynamics, leading to poor generation of fine-grained facial details. HFA-GP [9] aims to mitigate these challenges by learning a local and low-dimensional subspace within the latent space of 3D-GAN as a generative prior for reconstructing facial avatars from monocular videos. Nevertheless, HFA-GP still exhibits minor visual artifacts that do not exist in the input image or alter the background of the input image. In contrast, AD-NeRF [10] is introduced for audio-driven talking head synthesis, utilizing two distinct NeRFs to model the head and torso separately. However, AD-NeRF often yields unnatural outcomes due to the head-torso separation during the rendering stage [11]. DFRF [12] further tackles the few-shot audio-driven talking head synthesis problem and achieves plausible results in the few-shot setting; nevertheless, it also

* Corresponding author.
E-mail addresses: jinkx@hdu.edu.cn (K. Jin), guxl@hdu.edu.cn (X. Gu), wangzimeng@hdu.edu.cn (Z. Wang), zzkuang@hdu.edu.cn (Z. Kuang), wuzizhao@hdu.edu.cn (Z. Wu), tanmin@hdu.edu.cn (M. Tan), yujun@hdu.edu.cn (J. Yu).

https://doi.org/10.1016/j.patrec.2024.08.004
Received 11 January 2024; Received in revised form 10 May 2024; Accepted 9 August 2024
Available online 10 August 2024

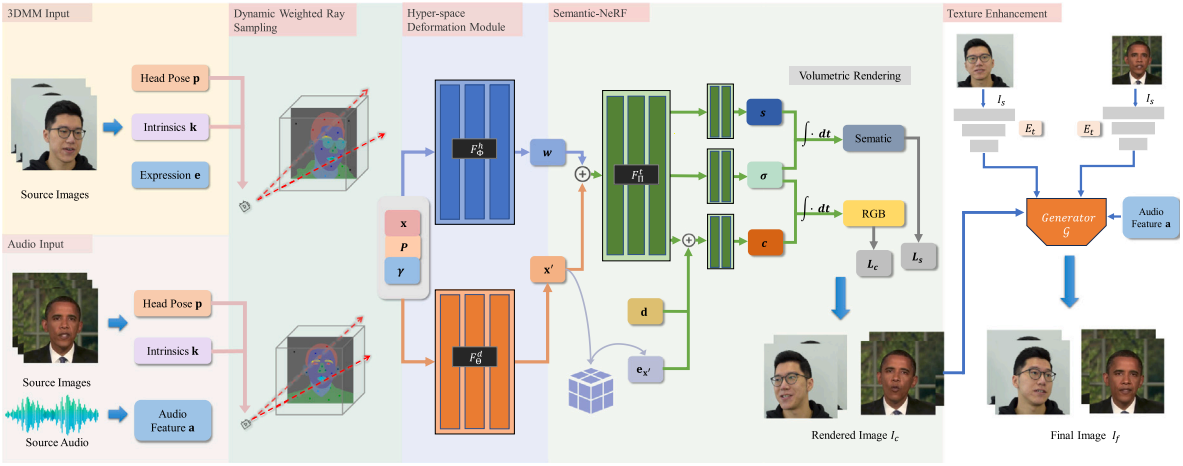

Fig. 1. Overview of our proposed semantic-aware hyper-space deformable NeRF-based framework for reconstructing high-fidelity face avatars from monocular videos. It consists of four main components: (1) a dynamic weighted ray sampling strategy, (2) a hyper-space deformation module, (3) a semantic guidance module that enhances the NeRF with semantic information, and (4) a texture enhancement technique.

encounters difficulties with head-torso separation by treating the torso as a static background. DIRFA [13] leverages a transformer-based probabilistic mapping network to generate talking faces driven by audio signals. While DIRFA excels in producing talking faces with diverse realistic facial animations from the same driving audio, it focuses solely on the movement of the lips and neglects the internal structure of the mouth, particularly the representation of the teeth and oral cavity.

The aforementioned challenges can be distilled into three primary issues when integrating NeRFs into facial avatar reconstruction. Firstly, the diverse motions exhibited by the head, torso, and facial regions present a significant hurdle in effectively modeling these movements with a single deformable module. Secondly, the varying motion patterns across different facial regions require semantic guidance to accurately capture local dynamics and fine details of facial appearance. Notably, the mouth region exhibits high-frequency motion, demanding heightened attention compared to other facial regions. Thirdly, the absence of a generalized NeRF-based framework adaptable to either 3DMM coefficients or audio input presents a notable limitation.

In this work, we propose a generalized semantic-aware hyper-space deformable NeRF-based framework for reconstructing high-fidelity facial avatars from monocular videos. Our framework can be driven by either 3DMM coefficients or audio features, effectively handling both localized facial movements and broader head and torso motions. To achieve this, we introduce four main components: (1) a dynamic weighted ray sampling strategy tailored to the semantic regions of the face, (2) a hyper-space deformation module designed to map observation space coordinates to the canonical hyper-space, facilitating the learning of more natural facial deformation and head-torso movements, (3) a semantic guidance module that enhances the NeRF with semantic information to capture the fine-grained appearances of different facial regions, and (4) a texture enhancement technique that improves the visual quality of the rendered images.

For this paper, the main contributions are as follows: (1) We propose a generalized semantic-aware hyper-space deformable NeRF-based framework for reconstructing high-fidelity facial avatars from monocular videos, which can be driven by either 3DMM coefficients or audio input. (2) We introduce a hyper-space deformation module that transforms the observation space coordinates to the canonical hyper-space coordinates, which can effectively handle both localized facial movements and broader head and torso motions. (3) Extensive experiments demonstrate that the proposed framework outperforms the existing state-of-the-art methods.

2. Method

Fig. 1 illustrates our semantic-aware hyper-space deformable NeRF-based framework, comprising four main components: (1) a dynamic weighted ray sampling strategy, (2) a hyper-space deformation module, (3) a semantic guidance module that enhances the NeRF with semantic information, and (4) a texture enhancement technique.

2.1. Dynamic weighted ray sampling

Building on SSP-NeRF [14], we introduce a dynamic weighted ray sampling strategy tailored to the motion patterns of distinct facial regions. Unlike the original NeRF [7], which samples rays uniformly on the image plane, our strategy assigns different sampling probabilities to different semantic categories during the training stage. SSP-NeRF defines the sampling probability of the ith semantic category during the training stage as follows:

$$p_i = \frac{\mathcal{L}_i}{\sum_{i=1}^{K} \mathcal{L}_i} \tag{1}$$

where K is the number of semantic categories in the parsing map, $\mathcal{L}_i$ is the sum of semantic loss and RGB loss of the ith semantic category for the previous epoch. Thus, the rays across K categories are sampled as:

$$N_i = p_i \cdot N_s \tag{2}$$

where N_i is the number of rays distributed to the ith category and N_s is the number of sampled rays.

We observed that the dynamic ray sampling in SSP-NeRF relies solely on the training loss to allocate the sampling ratio. To enhance the flexibility and controllability of dynamic sampling, we introduce sampling weights denoted as $\omega \in \mathbb{R}^K$ during the ray sampling process. We define the sampling probability of the ith semantic category during the training stage as follows:

$$p_i = \frac{\omega_i \mathcal{L}_i}{\sum_{j=1}^{K} \omega_j \mathcal{L}_j} \tag{3}$$

The enhanced dynamic weighted ray sampling strategy allows for manual adjustment of sampling weights, enabling more attention on important facial regions.

2.2. Hyper-space deformation module

As previously discussed, the torso exhibits more flexible movement compared to the head. Modeling the head and torso movement with a single NeRF may result in unnatural torso oscillations. To address the inconsistency in head-torso motion, AD-NeRF [10] employed two separate NeRFs, which led to head-torso separation issues. To mitigate these issues, we propose a unified hyper-space deformation module capable of modeling the motions of the head, torso, and face. This module comprises a global deformation field for capturing global non-rigid motion and a local deformation field for capturing detailed local motion.

Global Deformation Field. We define a global deformation field that maps observation coordinates to canonical coordinates, which are subsequently utilized to query a template NeRF. In order to facilitate the modeling of global non-rigid motion in the torso and facial regions, we incorporate head pose and facial features as inputs into the function. Formally, given the spatial coordinate $\mathbf{x} \in \mathbb{R}^3$, facial feature $\gamma \in \mathbb{R}^{dim}$ and head pose $\mathbf{p} \in \mathbb{R}^6$, the global deformation function F_{Θ}^{d} is trained to output the displacement $\Delta\mathbf{x} \in \mathbb{R}^3$ for converting the given point to its position in the canonical space as $\mathbf{x} + \Delta\mathbf{x}$.

$$F_{\Theta}^{d} : (\mathbf{x}, \mathbf{p}, \gamma) \in \mathbb{R}^{dim+9} \rightarrow \triangle\mathbf{x} \in \mathbb{R}^3 \tag{4}$$

where the function F_{Θ}^{d} is parameterized by a learned MLP. Note that the motion of the head can be parameterized via the head pose by binding the head pose with the camera pose.

Local Deformation Field. While the global deformation field effectively models the movement of the torso and facial regions in general, it falls short in capturing facial topological changes, such as eye blinking and mouth openings, and may even introduce undesirable artifacts. Inspired by Hypernerf [15], we leverage a lifting function to define the 5D radiance field of each input image as a slice in hyper-space. Similarly, given the spatial coordinate $\mathbf{x} \in \mathbb{R}^3$, facial feature $\gamma \in \mathbb{R}^{dim}$, and head pose $\mathbf{p} \in \mathbb{R}^6$, the lifting function F_{Φ}^{h} is trained to output a point $\mathbf{w}$ in ambient coordinate space, defining the cross-sectional subspace of the coordinate in hyper-space.

$$F_{\Phi}^{h} : (\mathbf{x}, \mathbf{p}, \gamma) \in \mathbb{R}^{dim+9} \rightarrow \mathbf{w} \in \mathbb{R}^{W} \tag{5}$$

where the function F_{Φ}^{h} is parameterized by a learned MLP.

2.3. Semantic-NeRF and network training

Semantic-NeRF. Considering that each semantic region possesses distinct appearances and movement patterns, we extend the deformable NeRF by incorporating semantic guidance for capturing local dynamics and fine-grained appearances of different facial regions. As illustrated in Fig. 1, we introduce a segmentation renderer into the deformable NeRF. Formally, we map a query coordinate $\mathbf{x}'$ and ambient coordinate $\mathbf{w}$ to a distribution over K semantic labels through pre-softmax semantic logits $\mathbf{s}(\mathbf{x}', \mathbf{w})$. Subsequently, we adapt NeRF volume rendering equations to compute the semantic label and the color of a single pixel. Let $\mathbf{x}(t)$ be a point along the camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ emitted from the center of projection $\mathbf{o}$ to a pixel v. Considering near and far bounds t_n and t_f in that ray, the expected color C and semantic logits S of the pixel v is defined by:

$$\hat{S}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{x}'(t), \mathbf{w}(t))\mathbf{s}(\mathbf{x}'(t), \mathbf{w}(t))dt \tag{6}$$

$$\hat{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{x}'(t), \mathbf{w}(t))\mathbf{c}(\mathbf{x}'(t), \mathbf{w}(t), \mathbf{d}, \mathbf{e}_{\mathbf{x}'})dt \tag{7}$$

$$\text{where} \quad \mathbf{x}'(t) = \mathbf{x}(t) + F_{\Theta}^{d}(\mathbf{x}(t), \mathbf{p}, \gamma), \tag{8}$$

$$\mathbf{w}(t) = F_{\Phi}^{h}(\mathbf{x}(t), \mathbf{p}, \gamma), \tag{9}$$

$$F_{II}^{t}(\mathbf{x}'(t), \mathbf{w}(t), \mathbf{d}, \mathbf{e}_{\mathbf{x}'}) = [\mathbf{c}(\mathbf{x}'(t), \mathbf{w}(t), \mathbf{d}, \mathbf{e}_{\mathbf{x}'}), \sigma(\mathbf{x}'(t), \mathbf{w}(t)), \mathbf{s}(\mathbf{x}'(t), \mathbf{w}(t))], \tag{10}$$

$$\text{and} \quad T(t) = exp(-\int_{t_n}^{t} \sigma(\mathbf{x}'(\mathbf{t}), \mathbf{w}(t))ds). \tag{11}$$

where F_{II}^{t} is the semantic-NeRF which takes canonical space coordinate, ambient space coordinate, and view direction as input, and outputs RGB value, semantic logits, and volume density. Subsequently, the semantic logits can be subjected to a softmax activation function to transform them into multi-class probabilities. To enhance high-frequency image details, we employ a learnable 3D feature grid [16] that utilizes bicubic interpolation to extract local feature vector $\mathbf{e}_{\mathbf{x}'}$ from it. The volume rendering integrals in Eq. (6), Eq. (7) and (11) can be approximated through numerical integration.

Network Training. Following NeRF [7], we adopt a hierarchical volume sampling strategy to simultaneously optimize coarse and fine networks, which are trained using a combination of photometric loss $\mathcal{L}_p$ and semantic loss $\mathcal{L}_s$:

$$\mathcal{L}_p = \sum_{\mathbf{r}\in\mathcal{R}} \left[\left\| \hat{C}_c(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 + \left\| \hat{C}_f(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 \right] \tag{12}$$

$$\mathcal{L}_s = -\sum_{\mathbf{r}\in\mathcal{R}} \left[\sum_{l=1}^{L} p^l(\mathbf{r}) \log \hat{p}_c^l(\mathbf{r}) + \sum_{l=1}^{L} p^l(\mathbf{r}) \log \hat{p}_f^l(\mathbf{r}) \right] \tag{13}$$

where $\mathcal{R}$ is the set of rays in each batch. $C(\mathbf{r})$, $\hat{C}_c(\mathbf{r})$, and $\hat{C}_f(\mathbf{r})$ are the ground truth, coarse volume predicted, and fine volume predicted RGB colors for ray $\mathbf{r}$ respectively. $p^l(\mathbf{r})$, $\hat{p}_c^l(\mathbf{r})$ and $\hat{p}_f^l(\mathbf{r})$ denote the ground truth, coarse volume predicted, and fine volume predicted multi-class semantic probability at class l for ray $\mathbf{r}$ respectively. Therefore, the overall training loss for our framework is:

$$\mathcal{L} = \mathcal{L}_p + \lambda\mathcal{L}_s \tag{14}$$

where λ is the weight of the semantic loss and to balance the magnitude of both losses.

2.4. Texture enhancement

To rectify texture defects in the rendered image I_c, such as the teeth and skin, we employ a texture enhancement network $\mathcal{G}$ with an encoder–decoder architecture to produce the final image I_f. Inspired by these methods [17–19], we utilize a texture extractor E_t to extract multi-scale texture features from the source image I_s, which are then fused with the extracted features of the rendered image I_c. These fused features are subsequently fed into $\mathcal{G}$ to enhance the image quality of I_f. Furthermore, to ensure audio–visual consistency in mouth and lip movements during audio-driven facial avatar synthesis, we integrate audio features into $\mathcal{G}$. Overall, the training objective of $\mathcal{G}$ is to minimize the L_2 loss between I_{gt} and I_f.

3. Experiment

3.1. Training data

To train 3DMM-driven facial avatar reconstruction, we utilized three monocular RGB videos produced by NerFACE [8], each with a resolution of 512 × 512 and a duration of about 2 min at 50fps (6000 frames). The last 1000 frames of each video were reserved as a test set. We extracted head pose, camera intrinsic, and facial expression from each frame using a state-of-the-art face tracking method [20]. To train audio-driven facial avatar reconstruction, we utilized three monocular RGB videos provided by AD-NeRF [10], each with a resolution of 512 × 512 and a duration of 3 to 5 min. We split the last 1/8 of each video as a test set. The audio features were obtained from the speech audio using a pre-trained DeepSpeech [21] model. Furthermore, we employed BiseNet [22] to generate semantic maps consisting of 12 classes for semantic guidance, including background, hair, cheeks, eyebrows, eyes, nose, ears, lips, mouth, glasses, neck, and torso.

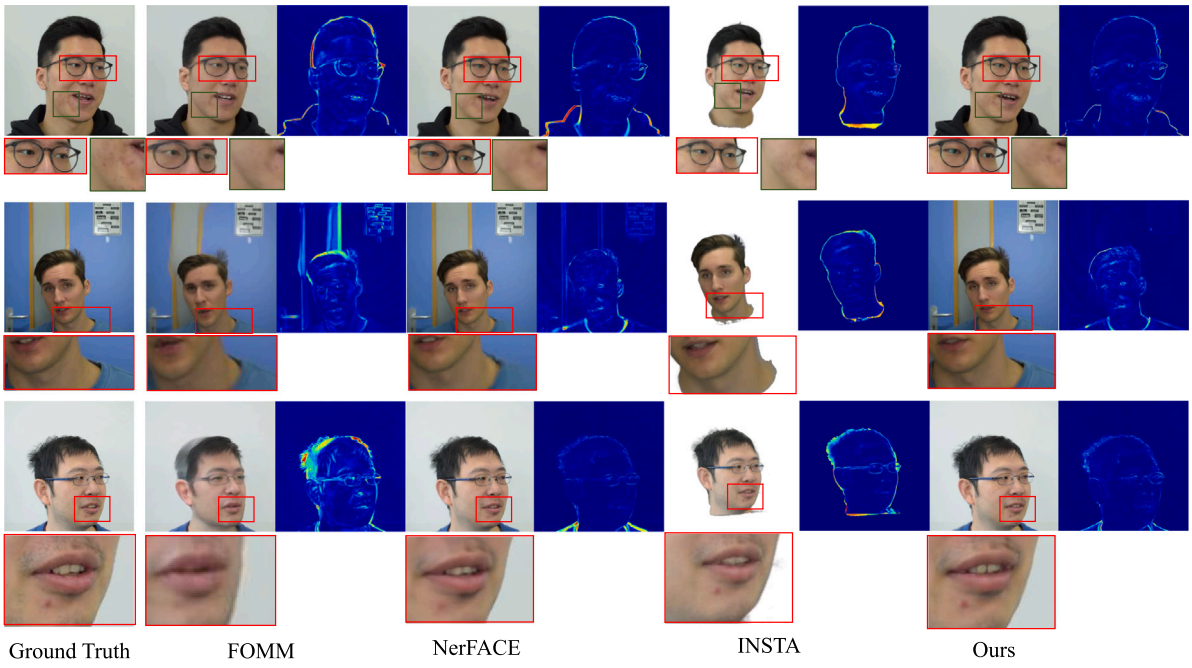

Fig. 2. Qualitative evaluation of 3DMM-driven face reenactment. From left to right: the ground truth image, FOMM [25], NerFACE [8], INSTA [26] and our method. The results show that our method can capture fine-grained details of various facial parts such as glasses, eyes, mouth, teeth, wrinkles, and even acne.

3.2. Implementation details

We employ dynamic weighted ray sampling, assigning a sampling weight ω_i of 1 for all classes except for the lips and mouth, which are given a weight of 1.1. The total number of semantic categories K is set to 12. For 3DMM-driven models, the feature dimension dim is set to 76, while for audio-driven models, it is set to 29. Within the hyper-space deformation module, F_{Θ}^{d} comprises 6-layer MLPs with a hidden size of 128 and ReLU activations, while F_{Φ}^{h} comprises 4-layer MLPs with a hidden size of 64 and ReLU activations. We define the ambient dimensions as $W = 2$. The Semantic-NeRF F_{Π}^{t} consists of 8-layer MLPs with a hidden size of 256 and ReLU activations. It is followed by three branches: a semantic logits branch (4-layer MLPs), a color branch (4-layer MLPs), and a density branch (1-layer MLP). Following NeRF [7], we use positional encoding to map the input into a higher-dimensional space. Specifically, we set $L = 10$ for $\mathbf{x}$ and $\mathbf{x}'$, and $L = 4$ for $\mathbf{w}$ and $\mathbf{d}$. The hyper-space deformation module and the Semantic-NeRF are trained using 512×512 images for 500k iterations with a batch size of $N_s = 2048$. Additionally, we set $\lambda = 0.02$ in the training loss. The texture enhancement network architecture is based on HiDe-NeRF [19]. During audio-driven facial avatar synthesis, the last feature layer of the texture extractor E_t is replaced with audio features to maintain the audio–visual consistency of mouth and lip movements. The texture enhancement network is trained for 150k iterations. All experiments are conducted in PyTorch [23] and optimized using the Adam optimizer [24] with a learning rate of $5e{-}4$.

3.3. Evaluation

Evaluation Metrics. We adopt a range of metrics to quantitatively evaluate the quality of the generated results. These metrics include L_1 distance, PSNR, SSIM [28], and LPIPS [29], which are commonly used for 3DMM-driven face reenactment. Additionally, for audio-driven face reenactment, we incorporate the SyncNet Confidence [30] to assess the accuracy of mouth shapes and lip sync.

Evaluation of 3DMM-driven Face Reenactment. To comprehensively evaluate the performance of 3DMM-driven face reenactment, we select four representative methods: first-order motion models (FOMM) [25], NerFACE [8], INSTA [26], and PointAvatar [31]. Following prior research, we utilize three monocular videos provided by NerFACE and extract the 3DMM expression coefficients from each frame for both training and testing. To ensure fairness, considering that INSTA solely renders the face without incorporating the torso and background parts, we utilize semantic segmentation to isolate only the face from our rendered results (denoted as Ours_face) for evaluation. Table 1 presents the average quantitative results, demonstrating our method's superiority over the baselines across all metrics. Additionally, it is evident that the three NeRF-based methods exhibit a significant advantage over FOMM, underscoring the efficacy of employing neural radiance fields for face reenactment. Fig. 2 illustrates the qualitative comparison results in the self-reenactment scenario. The figure also depicts the squared error between the ground truth and the generated images, with brighter regions indicating larger errors. As can be observed from the results, FOMM introduces noticeable artifacts and distortions in facial photos and fails to preserve the identity of the original face. While NerFACE captures subject appearances, it struggles with unnatural torso movements and lacks fine-grained facial details. Fig. 2 highlights the larger squared error in the torso region of NerFACE. INSTA exhibits contour artifacts around the chin and hair areas. In contrast, our method faithfully renders various facial features, including glasses, eyes, mouth, teeth, wrinkles, and acne, while achieving realistic torso movements. These findings validate the high-quality 3DMM-driven face reenactment capabilities of our approach.

Evaluation of Audio-driven Face Reenactment. We then evaluate the performance of audio-driven face reenactment against three state-of-the-art methods: AD-NeRF [10], DFRF [12] and SadTalker [27]. Table 2 presents the quantitative results, showcasing our method's superior performance over AD-NeRF across all metrics. Nevertheless, DFRF achieves higher quantitative results by treating the torso as a static background. To make a fair comparison, we also evaluate the performance of DFRF without considering the torso and background (denoted as DFRF_face) and our method without the torso and background (denoted as Ours_face) using semantic segmentation. Table 2 displays that Ours_face outperforms DFRF_face, particularly in terms of SyncNet Confidence. SadTalker, trained on a large video

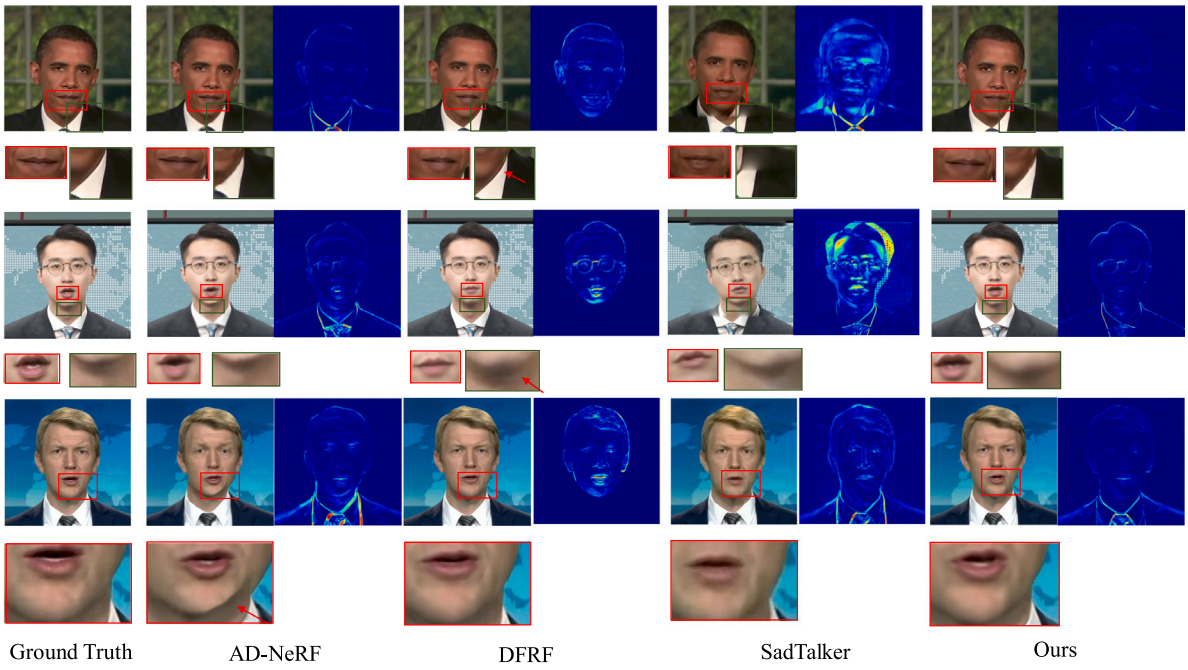

Fig. 3. Qualitative evaluation of audio-driven face reenactment. From left to right: the ground truth image, AD-NerF [10], DFRF [12], SadTalker [27] and our method. Here, we mark the head-torso separation issue with red arrows. The results show that our method can not only alleviate the head-torso separation issue but also capture better mouth and teeth shapes.

Table 1
Quantitative evaluation results of 3DMM-driven face reenactment.

Methods	Metrics			
	$L_1 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓
FOMM	0.034	23.55	0.892	0.079
NerFACE	0.017	27.84	0.956	0.058
PointAvatar	0.023	23.88	0.888	0.125
Ours	**0.013**	**30.00**	**0.964**	**0.037**
INSTA	0.013	24.87	0.939	0.100
Ours_face	**0.009**	**27.96**	**0.962**	**0.033**

* Trained on RTX 3090 graphics card.

Table 2
Quantitative evaluation results of audio-driven face reenactment.

Methods	Metrics				
	$L_1 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓	Sync ↑
Ground Truth	–	–	–	–	7.670
AD-NeRF	0.020	26.31	0.938	0.081	0.751
DFRF	**0.008**	**33.21**	**0.976**	**0.028**	3.79
SadTalker	0.060	19.41	0.829	0.191	**6.392**
Ours	0.015	27.90	0.956	0.042	3.984
DFRF_face	0.010	26.34	0.964	**0.028**	3.553
Ours_face	**0.009**	**26.55**	**0.965**	0.029	**3.621**

dataset, exhibits remarkable performance in audio-to-lip synchronization. However, it is restricted to processing videos at a resolution of 256 × 256. When dealing with higher-resolution videos, SadTalker can only handle facial regions, resulting in misalignments between the face and torso. Fig. 3 illustrates the qualitative comparison results in the self-reenactment scenario. Our method consistently generates more accurate mouth and teeth shapes compared to baseline methods. Additionally, the baseline methods exhibit noticeable head-torso separation issues. These experimental results demonstrate the superiority of our proposed method.

3.4. Ablation study

In the previous sections, we have shown the superior performance of our method over the baseline methods through extensive experiments. In this section, we analyze the effect of each key component of our method by conducting a stepwise ablation study where components are progressively removed. This allows us to evaluate the impact of each component individually and comprehend their collective effect on the model's performance. We consider the following variants of our method:

1. **w/o TE**: Our method without texture enhancement.
2. **w/o TE_DWS**: Our method without texture enhancement and dynamic weighted ray sampling.
3. **w/o TE_DWS_LDF**: Our method without texture enhancement, dynamic weighted ray sampling, and local deformation field.
4. **w/o TE_DWS_HDM**: Our method without texture enhancement, dynamic weighted ray sampling, and hyper-space deformation module.

Note that we perform the ablation studies on the first subject for 3DMM-driven face reenactment and the first subject (Obama) for audio-driven face reenactment, respectively. Tables 3 and 4 report the quantitative comparison results. The findings indicate that the full model outperforms the degraded models in both 3DMM-driven and audio-driven face reenactment. However, the quantitative comparison results may not be very distinct since the full model primarily enhances the degraded models in terms of fine-grained facial features. Moreover, the PSNR metric may not accurately reflect perceptual quality as it tends to favor blurry results over sharp images.

Figs. 4 and 5 illustrate the qualitative ablation results for 3DMM-driven face reenactment and audio-driven face reenactment, respectively. We notice that without the texture enhancement technique, the rendered images lack realistic details. We also observe that the absence of dynamic weighted ray sampling results in a blurry mouth region, especially noticeable in audio-driven face reenactment. Furthermore, by comparing the results of w/o TE_DWS_LDF and w/o TE_DWS_HDM, we find that both the local deformation field and global deformation field are crucial for achieving high-quality facial reenactment. In the

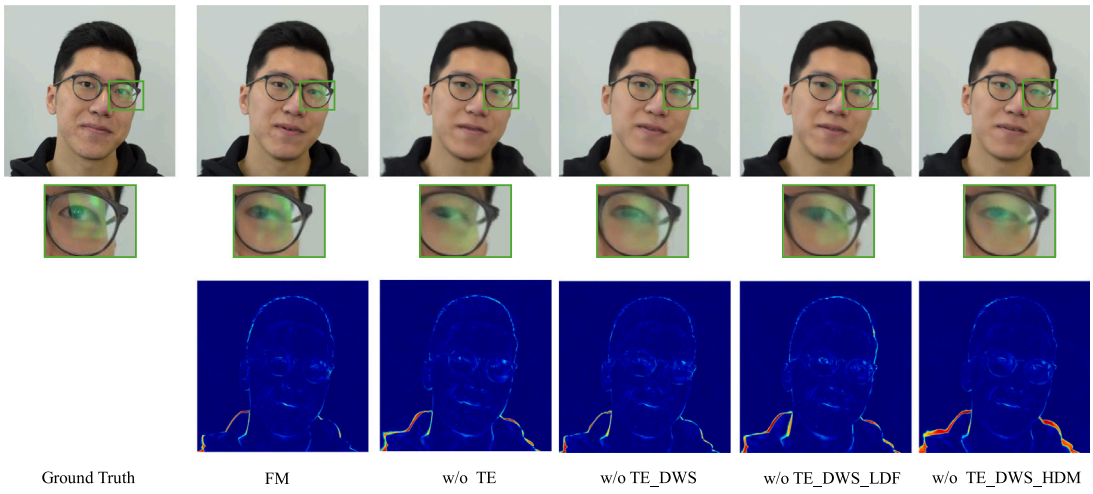

Fig. 4. The qualitative ablation results on 3DMM-driven face reenactment.

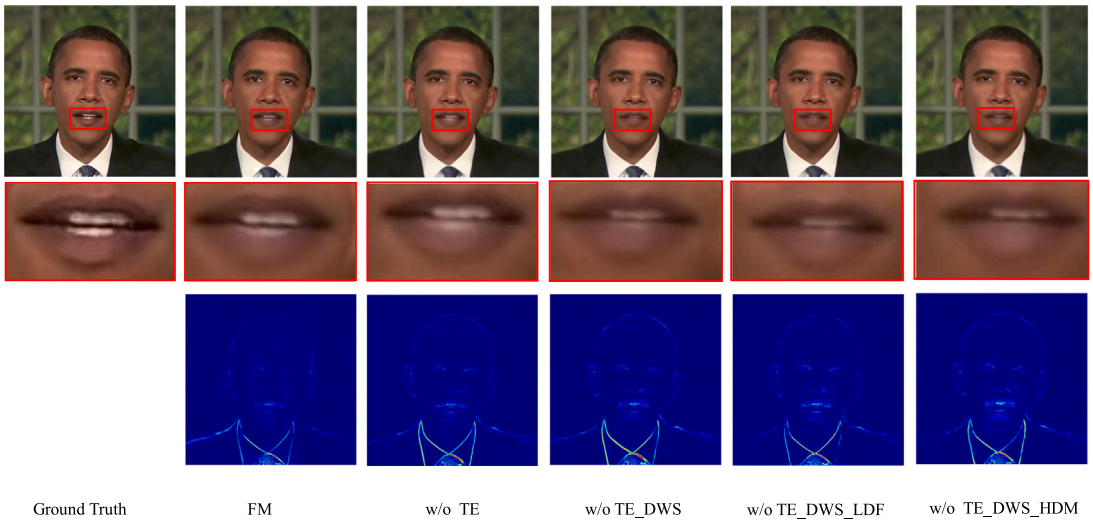

Fig. 5. The qualitative ablation results on audio-driven face reenactment.

squared error image displayed in Fig. 4, we observe that the torso is well-reconstructed through the hyper-space deformation module, mitigating the issue of unnatural torso movements. In summary, these results demonstrate the effectiveness of each component.

In addition to the previous ablation studies, we further investigate the influence of sampling weight sizes specifically for the audio-driven face reenactment task. We adjust the sampling weight sizes for lips and mouth to 1, 1.1, 1.5, 5, and 10. Each variation undergoes testing under identical experimental conditions to ensure consistency and comparability. The results, as presented in Table 5, illustrate how different sampling weight sizes affect the quality and realism of face reenactment, emphasizing the critical role they play in enhancing facial reenactment outcomes.

4. Conclusion and future work

We propose a semantic-aware hyper-space deformable NeRF-based framework for reconstructing high-fidelity facial avatars from monocular videos, which can be driven by either 3DMM coefficients or audio input. The proposed framework effectively captures the complex motion dynamics across the head, torso, and facial features by introducing dynamic weighted ray sampling and a hyper-space deformation module. Extensive experiments demonstrate that the proposed framework surpasses the existing state-of-the-art methods, achieving remarkable results in reconstructing realistic and expressive facial avatars.

Table 3
Quantitative comparison of the ablation study on 3DMM-driven face reenactment.

Methods	Metrics			
	$L_1 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	**0.017**	**25.75**	**0.953**	**0.059**
w/o TE	0.019	25.32	0.947	0.093
w/o TE_DWS	0.019	25.46	0.947	0.094
w/o TE_DWS_LDF	0.019	25.21	0.946	0.094
w/o TE_DWS_HDM	0.021	24.36	0.946	0.101

Table 4
Quantitative comparison of the ablation study on audio-driven face reenactment.

Methods	Metrics				
	$L_1 \downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓	Sync ↑
Ours	**0.013**	28.62	**0.964**	**0.056**	**2.789**
w/o TE	0.013	28.57	0.960	0.062	2.737
w/o TE_DWS	0.013	28.41	0.959	0.058	1.742
w/o TE_DWS_LDF	0.013	28.38	0.959	0.062	1.620
w/o TE_DWS_HDM	0.013	**28.85**	0.961	0.068	0.807

Table 5
Quantitative assessment of varying sampling weight sizes for lips and mouth in audio-driven face reenactment.

Methods	Metrics				
	L_1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Sync ↑
weight=1	0.016	29.29	0.960	0.057	4.553
weight=1.1	**0.015**	29.41	**0.961**	0.055	**4.985**
weight=1.5	0.016	29.23	0.960	**0.054**	4.719
weight=5	0.015	29.40	0.960	0.054	4.656
weight=10	0.015	**29.44**	0.960	0.057	4.335

While our framework shows superiority over existing methods, certain limitations remain. Firstly, reproducing finer facial details poses a challenge, which we aim to address through further exploration of image enhancement methodologies. Secondly, the training and inference time of our NeRF renderer is prolonged due to its vanilla implementation. Future work will focus on reducing this time through the integration of recent advancements in model optimization techniques.

CRediT authorship contribution statement

Kaixin Jin: Writing – original draft, Methodology. **Xiaoling Gu:** Writing – review & editing, Methodology. **Zimeng Wang:** Visualization, Validation, Software. **Zhenzhong Kuang:** Conceptualization. **Zizhao Wu:** Writing – review & editing. **Min Tan:** Writing – review & editing. **Jun Yu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21F020019, LZ23F020007, LY22F020028 and the National Science Foundation of China under Grants 62125201, 62372147, U21B2040, 61802100 and 61972119. This work was also supported by the Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2314 and No. A2306), Zhejiang University. This work was also partially supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LDT23F02025F02.

References

[1] Y. Chen, R. Xia, K. Yang, K. Zou, DARGS: Image inpainting algorithm via deep attention residuals group and semantics, J. King Saud Univ. Comput. Inf. Sci. 35 (6) (2023) 101567.
[2] Y. Chen, R. Xia, K. Yang, K. Zou, MFMAM: Image inpainting via multi-scale feature module with attention module, Comput. Vis. Image Underst. 238 (2024) 103883.
[3] J. Zhang, X. Li, Z. Wan, C. Wang, J. Liao, FDNeRF: Few-shot dynamic neural radiance fields for face reconstruction and expression editing, in: SIGGRAPH, 2022, pp. 12:1–12:9.
[4] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in: SIGGRAPH, 1999, pp. 187–194.
[5] S. Athar, Z. Shu, D. Samaras, FLAME-in-NeRF: Neural control of radiance fields for free view face animation, in: 17th IEEE International Conference on Automatic Face and Gesture Recognition, 2023, pp. 1–8.
[6] Y. Chen, R. Xia, K. Yang, K. Zou, GCAM: lightweight image inpainting via group convolution and attention mechanism, Int. J. Mach. Learn. Cybern. 15 (5) (2024) 1815–1825.
[7] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, NeRF: Representing scenes as neural radiance fields for view synthesis, in: ECCV, Vol. 12346, 2020, pp. 405–421.
[8] G. Gafni, J. Thies, M. Zollhöfer, M. Nießner, Dynamic neural radiance fields for monocular 4D facial avatar reconstruction, in: CVPR, 2021.
[9] Y. Bai, Y. Fan, X. Wang, Y. Zhang, J. Sun, C. Yuan, Y. Shan, High-fidelity facial avatar reconstruction from monocular video with generative priors, 2022, CoRR abs/2211.15064 arXiv:2211.15064.
[10] Y. Guo, K. Chen, S. Liang, Y. Liu, H. Bao, J. Zhang, AD-NeRF: Audio driven neural radiance fields for talking head synthesis, in: ICCV, 2021, pp. 5764–5774.
[11] X. Liu, Y. Xu, Q. Wu, H. Zhou, W. Wu, B. Zhou, Semantic-aware implicit neural audio-driven video portrait generation, in: ECCV, in: Lecture Notes in Computer Science, vol. 13697, 2022, pp. 106–125.
[12] S. Shen, W. Li, Z. Zhu, Y. Duan, J. Zhou, J. Lu, Learning dynamic facial radiance fields for few-shot talking head synthesis, in: ECCV, Vol. 13672, 2022, pp. 666–682.
[13] R. Wu, Y. Yu, F. Zhan, J. Zhang, X. Zhang, S. Lu, Audio-driven talking face generation with diverse yet realistic facial animations, Pattern Recognit. 144 (2023) 109865.
[14] X. Liu, Y. Xu, Q. Wu, H. Zhou, W. Wu, B. Zhou, Semantic-aware implicit neural audio-driven video portrait generation, 2022, arXiv preprint arXiv:2201.07786.
[15] K. Park, U. Sinha, P. Hedman, J.T. Barron, S. Bouaziz, D.B. Goldman, R. Martin-Brualla, S.M. Seitz, HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields, ACM Trans. Graph. 40 (6) (2021) 238:1–238:12.
[16] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, J. Wang, FENeRF: Face editing in neural radiance fields, in: CVPR, 2022, pp. 7672–7682.
[17] Y. Chen, R. Xia, K. Yang, K. Zou, DNNAM: Image inpainting algorithm via deep neural networks and attention mechanism, Appl. Soft Comput. 154 (2024) 111392.
[18] Y. Chen, R. Xia, K. Yang, K. Zou, MICU: Image super-resolution via multi-level information compensation and U-net, Expert Syst. Appl. 245 (2024) 123111.
[19] W. Li, L. Zhang, D. Wang, B. Zhao, Z. Wang, M. Chen, B. Zhang, Z. Wang, L. Bo, X. Li, One-shot high-fidelity talking-head synthesis with deformable neural radiance field, in: CVPR, IEEE, 2023.
[20] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos, in: CVPR, 2016, pp. 2387–2395.
[21] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., Deep speech 2: End-to-end speech recognition in english and mandarin, in: ICML, 2016, pp. 173–182.
[22] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: ECCV, 2018, pp. 334–349.
[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E.Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: NeurIPS, 2019, pp. 8024–8035.
[24] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), ICLR, 2015.
[25] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe, First order motion model for image animation, in: NeurIPS, 2019.
[26] W. Zielonka, T. Bolkart, J. Thies, Instant volumetric head avatars, in: CVPR 2023, IEEE, 2023, pp. 4574–4584.
[27] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, F. Wang, SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation, in: CVPR 2023.
[28] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[29] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR, 2018, pp. 586–595.
[30] J.S. Chung, A. Zisserman, Out of time: automated lip sync in the wild, in: Workshop on Multi-View Lip-Reading, ACCV, 2016.
[31] Y. Zheng, W. Yifan, G. Wetzstein, M.J. Black, O. Hilliges, PointAvatar: Deformable point-based head avatars from videos, in: CVPR, 2023.