IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 October 2023, accepted 6 January 2024, date of publication 17 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3355052

RESEARCH ARTICLE

4D Facial Avatar Reconstruction From Monocular Video via Efficient and Controllable Neural Radiance Fields

JEONG-GI KWAK AND HANSEOK KO, (Senior Member, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Hanseok Ko (hsko@korea.ac.kr)

This work was supported by the ''Development of Cognitive/Response Advancement Technology for AI Avatar Commercialization'' Project funded by the Brand Engagement Network (BEN) under Grant Q2312881.

ABSTRACT We present an efficient approach for monocular 4D facial avatar reconstruction using a dynamic neural radiance field (NeRF). Over the years, NeRFs have been popular methods for 3D scene representation, but lack computational efficiency and controllabilty, thus it is impractical for real world application such as AR/VR, teleconferencing, and immersive experiences. Recent the introduction of grid-based encoding by InstantNGP has enabled the rendering process of NeRF much faster, but it is limited to static 3D scenes. To address the issues, we focus on developing a novel dynamic NeRF that allows explicit control over pose and facial expression, while keeping the computational efficiency. By leveraging a low-dimensional basis from the morphable model (3DMM) with elaborately designed spatial encoding branch and ambient encoding branch, we condition a dynamic radiance field in an ambient space, improving controllability and visual quality. Our model achieves rendering speeds approximately 30x faster at training and 100x faster at inference than the baseline (NeRFace), enabling practical approaches for real world applications. Through qualitative and quantitative experiments, we demonstrate the effectiveness of our approach. The dynamic NeRF exhibits superior controllability, enhanced 3D consistency, and improved visual quality. Our efficient model opens new possibilities for real-time applications, revolutionizing AR/VR and teleconferencing experiences.

INDEX TERMS Neural radiance field (NeRF), monocular facial avatar reconstruction, face reenactment.

I. INTRODUCTION

Recently,4D dynamic facial avatar reconstruction has received remarkable attention as a rapid progress of virtual reality (VR), augmented reality (AR), and teleconferencing. Beyond simply reconstructing the appearance of the target facial identity, it requires the ability to faithfully capture dynamics, where several attributes, e.g., head pose, lips or facial expressions should be controllable according to a given driving video (face reenactment) or conditions given by users (controllable generation).

Several approaches [1], [2], [3] have exploited 2D modules in order to synthesize and edit human portrait videos. They have leveraged image-to-image translation using generative adversarial networks (GANs) and shown photorealistic results. However, they lack 3D understanding, meaning the underlying 3D geometry of an object is ignored in the rendering process.

Due to the complicated structure and diverse rigid and non-rigid motions of the human head, several approaches have exploited 3D morphable models (3DMM) [4], [5], [6] as the prior of the 3D human head. 3DMMs generally represent a target human face with a 3D mean face and linear combination of principal bases obtained from pre-captured 3D scans. Early methods [7], [8], [9] have leveraged explicit surface directly to reconstruct the target head model and they have the advantage of extracting a 3D facial mesh with a lower dimension of input (linear coefficient).However, they fail to generate delicate details that do not exist in template mesh, e.g., hairs and teeth. Several hybrid methods [10],

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

FIGURE 1. Comparison of generated samples and PSNR results according to training time. Red boxes denote the results at the time when training is completed.

[11] have improved visual quality and 3D consistency by combining image synthesis module with 3DMM. While they have shown notable results, they still struggle to represent consistent 3D head motion.

To overcome these limitations, we focus on ''NeRF-based 4D avatar reconstruction'' to obtain both 3D consistency and rendering quality. Recently, neural radiance field (NeRF) [12] has achieved remarkable success in novel-view synthesis tasks and has been applied in various fields since it could render high-quality images with accurate 3D awareness. NeRF has a coordinated-based implicit structure, where it takes a 3D position and view direction as an input and predicts its view-dependent color and density using MLPs. However, the training process of the original NeRF assumes that the target object is static, thus it is not suitable for dynamic objects such as the human face. To handle the problem, several dynamic NeRFs [13], [14], [15] capable of capturing rigid and non-rigid motions of objects have been proposed. They effectively capture the dynamic scenes with reasonable visual quality even in a monocular setting but struggle to explicitly control over pose or appearance of objects, meaning just playback or reconstruction of the given video, not controllability.

NeRFace [16] achieved expression control over the reconstruction of the facial avatar by conditioning the implicit network with 3DMM expression parameters. As a result, NeRFace has the advantages of both rendering quality of NeRF and parametric control of 3DMM. Nevertheless, NeRFace suffers from slow training and inference speed because the naive NeRF has to calculate all querying points sampled from rays including a number of redundant points. Even worse, conditional NeRFs like NeRFace require more training images to faithfully capture the relationship between 3DMM parameters and facial dynamics, resulting in slower convergence. Actually, it takes about 1-2 days to learn scene dynamics from a 512^2 resolution monocular video and about 8-9 seconds to render for each image. Despite the reasonable rendering quality and 3D consistency of NeRF-based 4D avatar reconstruction, computational inefficiency is a major obstacle to real-world application.

To handle the issue, we exploit a concept of grid-based encoding [17] to reconstruct a controllable 4D facial avatar. Instead of predicting an arbitrary point in a 3D scene implicitly, recently proposed methods [17], [18] divide a 3D scene into a learnable explicit voxel grid or efficient planes [19] that stores scene features. With this setting, the value of a querying point is determined by a combination of features of the surrounding voxels like bilinear interpolation.However, directly applying these explicit methods to our problem is challenging because they capture static 3D scenes or just add time components, without controllability.

To this end, we add a grid-based ambient encoder which encodes the 3DMM expression parameters and provides the condition to NeRF model by slicing the hyper-space of the model inspired by HyperNeRF [15]. Since our model takes the 3DMM expression parameter as a condition, it is capable of face reenactment using a driving video and direct control over facial expressions like NeRFace. Notably, our model is much more efficient than NeRFace, showing about 30x faster speed at training (Fig. 1) and about 100x faster at inference. Therefore, we can obtain a personalized 4D avatar with fast convergence speed, and even when comparing visual quality, our model delivers competitive or superior results. We demonstrate the effectiveness of our method with qualitative and quantitative results.

In summary, our contributions are as follow:

- We propose a novel dynamic NeRF model for 4D avatar reconstruction which is computationally efficient.
- Beyond just reconstructing an input video, our model enables users to control 4D avatar by exploiting 3DMM parameters as additional condition of the NeRF model.
- We introduce explicit grid-based encoding branches that encode spatial and conditioning information, resulting in notable speed up.

II. RELATED WORK

A. FACE RECONSTRUCTION, SYNTHESIS, AND CONTROL

The desire for representing realistic digital human faces has been a long-standing problem in computer vision and graphics communities due to its endless potential applications. However, it is challenging to express the unique characteristics and details of each face and to precisely control them. For this reason, the majority of head reconstruction approaches [7], [8], [9], [10], [11], [20], [21], [22], [23], [24] have utilized several parametric 3D Morphable Models (3DMM) [4], [5], [6] obtained by dimensionality reduction of complex high-dimensional 3D face scans using PCA. Here, the morphable models are used as a 3D prior to head reconstruction. While these methods have shown plausible reconstruction results and the ability to preserve facial details, they have struggled with capturing accurate 3D geometry. Besides, since their 3D prior is heavily based on a template mesh, it is difficult to restore several missing parts in the template, such as hair and teeth.

Another mainstream research is 2D generative model-based approaches. Among them, GAN-based methods have shown remarkable success in photorealistic facial image synthesis. It has been extended to facial attribute editing [25],

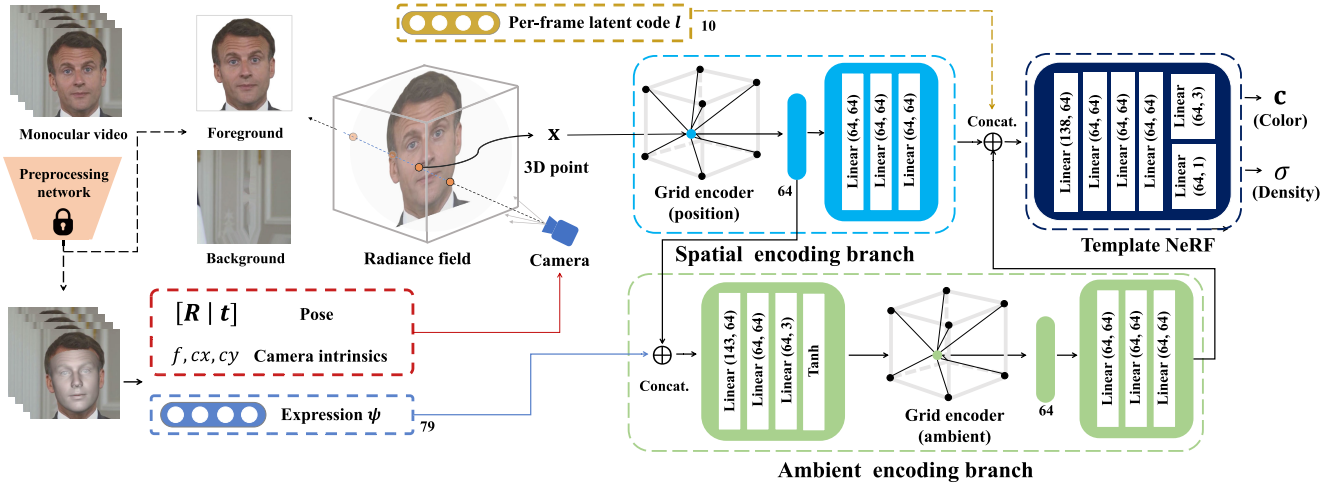

FIGURE 2. Overview of our framework. It represents dynamic neural fields from an input video. Spatial and ambient branches encode spatial and conditioning information, respectively. After that, a template NeRF predicts view-dependent color (c) and its density (σ).

[26], [27], [28], [29], [30], [31], [32], stylization [33], [34], [35], [36], [37], and animation [38]. Recent works have focused on leveraging powerful pretrained generators, e.g. StyleGAN [39], [40]. They have demonstrated that semantic editing and animation can be achieved by several manipulations in the latent space of StyleGAN. Combined with GAN inversion methods, they have shown powerful performance for real image editing [41], [42] and animation [37], [43]. Despite the expressiveness of 2D GANs, they struggle with preserving multi-view consistency due to a lack of 3D understanding in synthesis. Recently proposed 3D-aware generative models and several applications [44], [45], [46], [47], [48] have alleviated the problem, but they have limitations in capturing diverse facial dynamics.

B. DYNAMIC NEURAL RADIANCE FIELD

Neural radiance field (NeRF) incorporates implicit neural representation (INR) where it takes 3D coordinates as an input and predicts the corresponding information with a neural network. The output image is rendered into a 2D grid with a classic volumetric rendering [49].

Since NeRF is a method initially proposed for the novel-view synthesis of a static object, it fails to represent dynamic scenes. To address the issue, several dynamic NeRF models have been proposed to capture the changes in the scene over time. Nerfies [14] introduced a deformation field to model deformable dynamic objects, and HyperNeRF [15] proposed a hyperspace to handle topological variations inspired by level set methods. NeRFace [16] combines NeRF network with a low dimensional morphable model to provide explicit control over pose or expressions in the synthesis process. Our work is most closely related to NeRFace [16], on the commonality of introducing 3DMM expression parameters in dynamic NeRF to obtain controllability. Compared to NeRFace, our method greatly improves the computational efficiency and rendering quality with the introduction of a novel architecture and loss function.

C. EXPLICIT GRID STRUCTURE FOR EFFICIENT NERF

After the advent of NeRF, many efforts have been made to improve the computational efficiency of NeRF-based rendering. Recently, several hybrid approaches that combine NeRF with sparse explicit representations such as voxels or planes have been proposed and shown dramatic speed up in both training and inference. DVGO [50] replaces MLPs in the original NeRF with a voxel grid and shows fast convergence by optimizing the voxel grid directly. Plenoxel [18] prunes redundant space and represents a scene as a sparse 3D grid with spherical harmonics. InstantNeRF [17] leverages a multi-resolution encoding of feature grids using hash tables. Tensorf [51] proposed low-rank tensor decomposition to reduce the computational complexity of NeRF-based scene representation. These methods open up the possibility that NeRF-based methods can be applied to real-world applications requiring real-time or equivalent speed e.g., teleconference, VR/AR, and the metaverse. Our method is also inspired by InstantNeRF and leverages a multi-resolution feature grid to improve computational efficiency. Unlike InstantNeRF, which aims to represent a static scene, our approach is applied to a dynamic NeRF which should be able to capture various facial features.

III. PROPOSED METHOD

In this section, we first explain several backgrounds of neural radiance field (NeRF) briefly. After that, we propose our 4D avatar reconstruction framework, by first introducing control over facial expressions by slicing a hyperspace with 3DMM parameters, and then by explaining an efficient structure using explicit grid-based NeRF.

A. PRELIMINARIES: NEURAL RADIANCE FIELDS

Before introducing our method, we briefly explain a volumetric rendering process of NeRF [12] and its variants. Original NeRF consists of several MLPs, where it takes a 3D position $\mathbf{x} \in \mathbb{R}^3$ and an additional 2D view direction

$\mathbf{d} \in \mathbb{R}^2$ as input and predicts volume density $\sigma(\mathbf{x}) \in \mathbb{R}$ and the view-dependent RGB color $\mathbf{c}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$. To render the image, it utilizes a classic volumetric technique (ray-tracing) [49], accumulating the information of querying points on each ray from start t_n to end t_f. It can be formulated as,

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \cdot \sigma(\mathbf{o} + t\mathbf{d}) \cdot \mathbf{c}(\mathbf{o} + t\mathbf{d}, \mathbf{d}) dt,$$
$$\text{where} \quad T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s)) ds\right). \tag{1}$$

Here, $T(t)$ represents the accumulated transmittance along the ray from the starting point to t. Therefore, the process of predicting color and density using NeRF network can be decribed as,

$$f_{\text{NeRF}} : (\mathbf{x}, \mathbf{d}) \mapsto (\mathbf{c}, \sigma), \tag{2}$$

One of the limitations of the original NeRF is that it can not handle dynamic scenes or objects that move over time or other conditions. To address the issue, several methods, called dynamic NeRF [13], [14], [15], [16], [52], have been proposed in order to capture the motion and deformation of objects over time or several conditions. Therefore, We can express this by using an additional variable in Eq. 2, i.e.,

$$f_{\text{D.NeRF}} : (\mathbf{x}, \mathbf{d}, \alpha) \mapsto (\mathbf{c}, \sigma), \tag{3}$$

where α denotes some additional conditions which enable the network to model the dynamics of the object.

B. 4D AVATAR CONTROL WITH 3DMM PARAMETERS

Beyond just the reconstruction from a monocular portrait video (like "replaying video"), we aim to enable users to control facial expressions and poses. To this end, we need to modulate NeRF architecture according to conditions. We selected 3DMM expression parameters [5] as input condition (α in Eq. 3) since we can extract 3DMM parameters from a monocular facial video using pretrained face tracking methods [53], [54], thus it is possible to train the NeRF model with regression-based optimization.

One of the basic methods to condition a NeRF model involves concatenating 3DMM parameters with the input 3D position. NeRFace [16] employs this approach, yet it occasionally falls short in capturing the nuances of facial dynamics due to the limited effectiveness of this concatenation method in representing intricate and dynamic scenes. To handle this problem, we adopt a concept of hyper-space [15] where each training image can be obtained by slicing a higher dimensional space using 3DMM parameters. It can be represented as,

$$\mathbf{w} = H(\mathbf{x}, \boldsymbol{\psi}), \tag{4}$$

where $H, \boldsymbol{\psi}, \mathbf{w}$ denote the slicing surface field, 3DMM expression parameter, and the point of ambient coordinate space which determines how to slice the higher dimensional hyper-space, respectively. Therefore, the model is capable of constructing a higher dimension space including various expression information. Finally, our dynamic radiance field conditioned by the 3DMM parameter can be formulated by,

$$F_\theta : (\mathbf{x}, \mathbf{d}, \mathbf{w}, l) \mapsto (\mathbf{c}, \sigma), \tag{5}$$

where θ denotes the learnable parameters of our model. In addition, we add a learnable per-frame latent code l which is widely used for dynamic NeRFs to compensate for missing information.

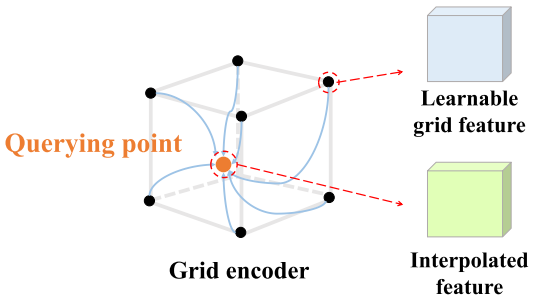

FIGURE 3. Structure of the grid encoder. The feature of an arbitrary 3D querying point is determined by the interpolation of the surrounding grid features.

The overall framework is described in Fig. 2. As explained above, a conditioning vector (3DMM expression parameter) extracted from a source image is encoded with a branch of ambient slicing surface instead of simply applying concatenation. After that, a template NeRF which consists of shallow MLPs takes position and view direction as input and predicts view-dependent color and its density Finally, the output image is obtained with a classic volumetric rendering technique [49]. Note that we do not utilize a positional encoding but exploit grid-based parametric encoding. Because our network has a grid-based explicit structure and it is described in the following section (Sec. III-C).

C. GRID-BASED STRUCTURE FOR EFFICIENT RENDERING

Another problem of the original NeRF and its variants lies in their slow speed at both the training and inference phases. Besides, conditional or dynamic NeRF methods like our model require more training images to faithfully capture the scene dynamics and the change of input images according to conditions. The reason for the inefficiency is due to its implicit structure, where the MLP-based NeRF network should independently perform calculations for all queried arbitrary points in 3D space for volumetric rendering. Inspired by recent remarkable advances in computational efficiency led by explicit grid-based methods [17], [18], [50], [51], we replace the MLP-based implicit NeRF network with an explicit grid [17]. Here, discrete voxels on the grid store their own learnable parametric feature. Encoding of each point is determined by combinations of grid features, e.g., bilinear interpolation, instead of calculating every querying point on rays (Fig. 3). Unlike other grid-based methods that handle a static 3D scene, our model needs to capture dynamic

TABLE 1. Quantitative comparisons with the competitive methods.

	FOMM	NeRFace	HyperNeRF + Cond.	Ours
PSNR (↑)	24.64	26.71	27.06	**29.81**
LPIPS(↓)	0.21	0.13	0.14	**0.09**
SSIM(↑)	0.89	0.93	0.94	**0.96**
Frame/sec. (↑)	**14**	0.12	0.16	11
Training time (hr.) (↓)	-	36	32	**1**

scenes. RAD-NeRF [55], which is audio-based NeRF, solves this problem by exploiting an additional 2D audio grid. Inspired by RAD-NeRF, we introduce an additional grid encoder, which encodes a 3D point in ambient coordinates conditioned by the extracted 3DMM parameter. Our approach dramatically improves the computational efficiency in both training and inference phases (about 30x and 100x faster than NeRFace, respectively) while maintaining competitive visual quality or even better performance.

D. MODEL OBJECTIVES

Our network takes a facial image from a monocular video with its extracted 3DMM expression parameter and camera parameter inputs and is trained to reconstruct the input image precisely with pixel-level reconstruction loss (MSE), i.e.,

$$\mathcal{L}_{\text{rec}} = \sum_{\boldsymbol{r}\in\mathcal{F}} \left\| \boldsymbol{C}(\boldsymbol{r}) - \boldsymbol{C}_{\text{gt}}(\boldsymbol{r}) \right\|_2^2, \tag{6}$$

where $\boldsymbol{C}$ and $\boldsymbol{C}_{\text{gt}}$ denote predicted color by model and ground-truth color, respectively. $\mathcal{F}$ means face regions including part of the upper torso. Although this regression-based optimization is a core objective of NeRF-based methods it may hurt the expressiveness of the dynamic NeRF. In other words, a network tends to produce similar results, rather than producing various results depending on input conditional parameters. This is especially noticeable in the region of the mouth because it is a small area but requires sophisticated movements. To alleviate the problem, we utilize facial landmarks to guide the network to more accurately capture facial structure. This is formulated as,

$$\mathcal{L}_{\text{lmk}} = \sum_{i=1}^{L} \|\mathbf{m}_i - \mathbf{n}_i\|_1, \tag{7}$$

where m_i and n_i denote i-th facial landmark of ground-truth and reconstructed image, and L is a number of target landmarks (we used 68). Finally, the full objective to train the proposed model can be formulated as,

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{lmk}}\mathcal{L}_{\text{lmk}} \tag{8}$$

where λ_{rec} and λ_{lmk} are hyper-parameters that balance each loss term.

IV. EXPERIMENTS

This section firstly describes several experimental settings and then presents qualitative and quantitative comparisons with state-of-the-art methods [2], [15], [16] Finally, we analyze the effectiveness of each component in our method by ablation study.

A. EXPERIMENTAL SETUP

1) DATASET

We used 1-2 min. of person-specific monocular RGB videos containing rigid and non-rigid motion of the leader's face. This public dataset is widely used for several tasks, e.g., dynamic NeRF and talking head generation. We used 9/10 of the frames for training and the others for testing. Each image is cropped to 512 × 512 to include the face area. Through the experiments, we aim to validate two key aspects of each method: visual quality and the ability to represent dynamics. To this end, we selected two datasets. First, we used a talking head dataset [54], which contains several talking head videos including natural head motions. With this dataset, we compare reconstruction (self-driving) results. In addition, we also use the NeRFace dataset [16], which consists of several monocular person-specific videos. Compared to the talking head dataset, the NeRFace dataset contains more dynamic head poses or expressions. Therefore, we utilize it for the face reenactment experiment.

2) IMPLEMENTATION

We utilize a pretrained face parsing model [56] to extract facial region in each image, face-alignment network (FAN) [53] to estimate 68 facial landmarks, and face tracker to extract 3DMM expression parameter and head pose. In training, we set coefficients of the loss functions in Eq. 8 as $\lambda_{\text{rec}} = 1$ and $\lambda_{\text{lmk}} = 0.1$. For NeRF setting, we sample 256^2 rays for each step and at most 16 points are sampled per ray. We adopt ADAM [57] solver with an initial learning rate of 0.0005. We implement our network with PyTorch [58] and all experiments are conducted on a single 24GB RTX 3090 GPU.

3) BASELINES

We compare our method against several state-of-the-art methods. Two NeRF-based methods are used in both qualitative and quantitative comparisons, i.e., NeRFace [16] and HyperNeRF [15] As described above, NeRFace is the closest and strongest competitor to our method in that it leverages 3DMM parameters for facial avatar reconstruction. HyperNeRF is capable of modeling non-rigid facial motions from a monocular video. However, it is impossible to control facial expressions or mouth with a driving video. Hence, we utilize a modified framework that takes the 3DMM expression parameter as an additional condition (i.e., HyperNeRF + Cond.). In addition, we also use First Order Motion Model (FOMM) [2] to compare with the 2D-based reenactment model.

B. RECONSTRUCTION (SELF-DRIVING) RESULTS

First, we conduct a self-reenactment analysis by synthesizing the output video using the same identity as a driving video. We present the qualitative output in Fig. 4. The top row denotes the driving sequence images that are unseen during the training phase. FOMM delivers reasonable results

FIGURE 4. Qualitative comparison on self-reenactment using talking head videos. The row from top to down is ground-truth images, FOMM [2], NeRFace [16], HyperNeRF [15] and ours, respectively. Please zoom in to see more details.

according to the driving images, but it fails to capture facial details, especially mouth, and lips. NeRFace shows more accurate face reenactment results compared to FOMM, but it loses high-frequency details. This limitation comes from the strategy of conditioning expression parameter, where conditioning vector is simply concatenated to input position vector. HyperNeRF (HyperNeRF + Cond.) preserves the facial details with good visual quality, but it struggles to capture the relation between fine regions of the face and input condition, especially the mouth area. As shown in right (Obama), it fails to control the mouth area and is expressed ambiguously. However, our method shows precise self-reenactment results without losing global and local details of the face. Instead of direct concatenation, the ambient branch elaborately encodes 3DMM expression parameter to ambient vector that slices the hyper-space to control over facial dynamics. We also report the quantitative comparisons of the methods in Table. 1. Here, we use Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [59], and Structure Similarity Index (SSIM) [60] to measure the quality of reconstructed images. Our method achieves superior results compared to the competitive models. As explained in previous sections, one of the main contributions of our method is computational efficiency. Hence, we also measure frame per second (Frame/Sec.). Compared to other NeRF-based methods (NeRFace and HyperNeRF), our model achieves overwhelming results in efficiency by adopting explicit grid-encoding scheme, where explicit representation has learnable fixed grid features and updates the features at each iteration, thus it does not consider an arbitrary position. As a result, this strategy significantly decreases the computational cost and our method shows competitive results with 2D-based FOMM. Lastly, we evaluate the total time to train each model with a single video. Here, FOMM is excluded because it does not adopt identity-specific training. To demonstrate the effectiveness of our method in the training phase, we additionally present PSNR and generated images at training time steps in Fig. 6. Our model shows an overall higher PSNR in the same iteration. We also mark the time to reach 50k iteration, our method shows fast training and convergence.

C. CONTROL OVER POSE AND EXPRESSION

We present face reenactment results driven by unseen videos. By extracting pose and expression information from a driving video and injecting them into models, we can control pose and expression. In this scenario, it is important to accurately represent the pose and expression of the driving video. The outputs are presented in Fig. 5. As shown in the results, the shortcomings of the 2D-based approach (FOMM) become evident due to its absence of 3D perception during the generation process. It can be seen that FOMM struggles to

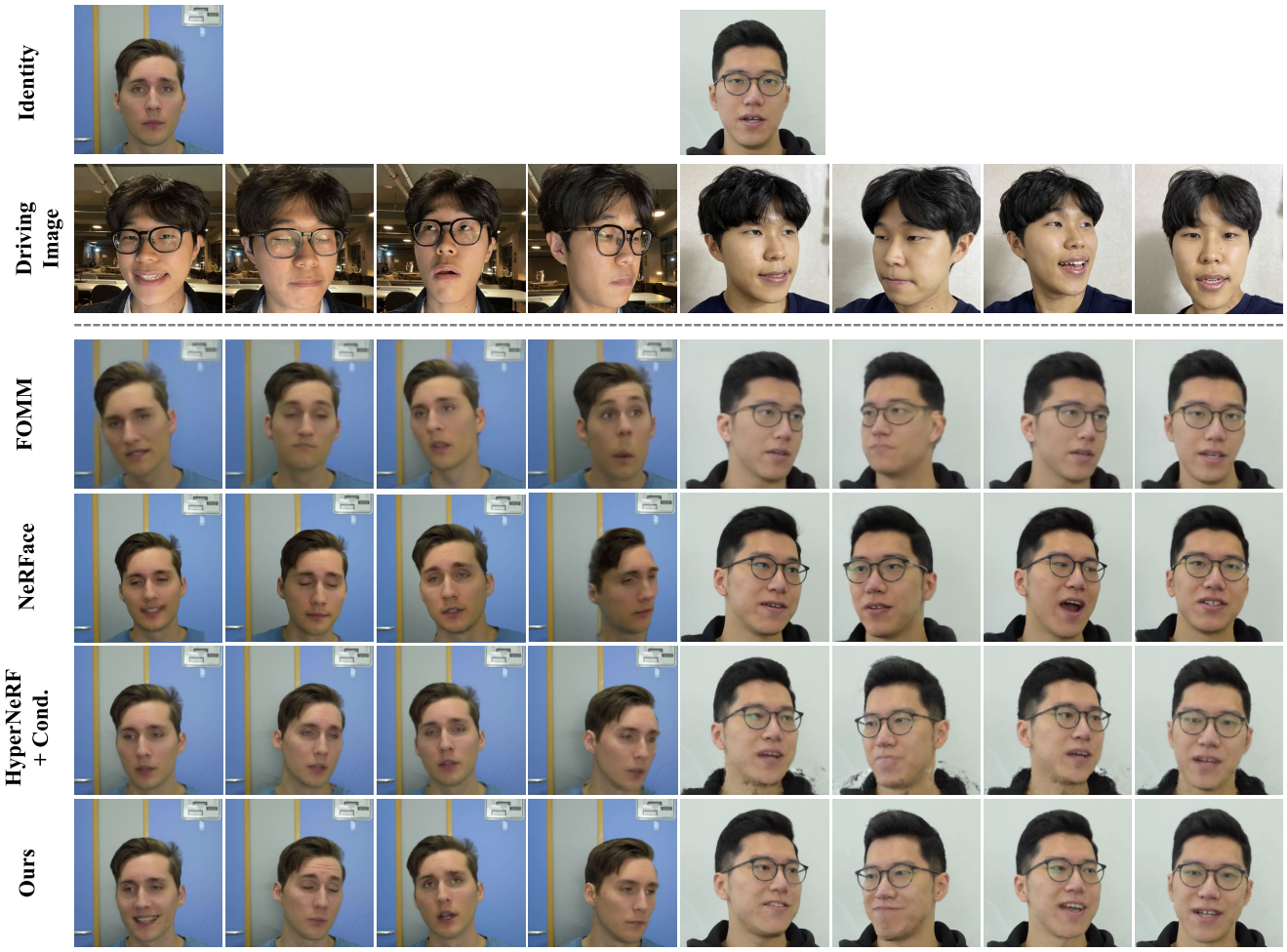

FIGURE 5. Qualitative comparison by face reenactment with novel facial expression and head-pose. Please zoom in to see more details.

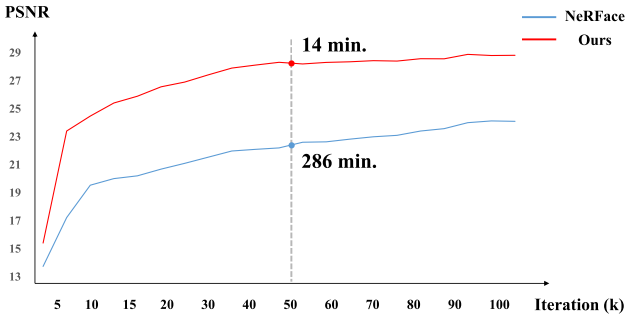

FIGURE 6. Comparison of PSNR according to training step up to about 100k iterations (vs. NeRFace). The time for each model to reach 50k is indicated.

elaborately follow the poses of driving video compared to NeRF-based methods. Besides, this incorrect recognition of facial geometry leads to reconstruction failure of expression in several cases. Compared to FOMM, NeRF-based methods faithfully capture pose (camera) information. NeRFace effectively tracks the driving video, but it faces challenges in faithfully reconstructing intricate facial expressions, due to its conditioning strategy. HyperNeRF encounters difficulties in capturing a wide range of facial expressions and head positions, often exhibiting artifacts like blurring and halo effects in the generated images. Furthermore, it struggles to accurately model changes around the mouth region. Our model demonstrates a remarkable ability to faithfully replicate both the poses and expressions of the driving video with exceptional visual fidelity. In addition, it is important to highlight that our approach significantly outpaces the NeRF-based baseline models in terms of training and inference speed.

D. ABLATION STUDY

To demonstrate the effectiveness of each part in the proposed method, we evaluate the performance of our model by excluding or replacing key components. We compare our model with three different versions, i.e., (i) adopting a direct conditioning method (like NeRFace) instead of using hyperspace concept (w.o. ambient), (ii) training without landmark loss (w.o. $\mathcal{L}_{\text{lmk}}$), and (iii) using only 1/10 dataset for training (1/10 training). Fig. 7 represents the output of each method. When we discard the ambient encoding

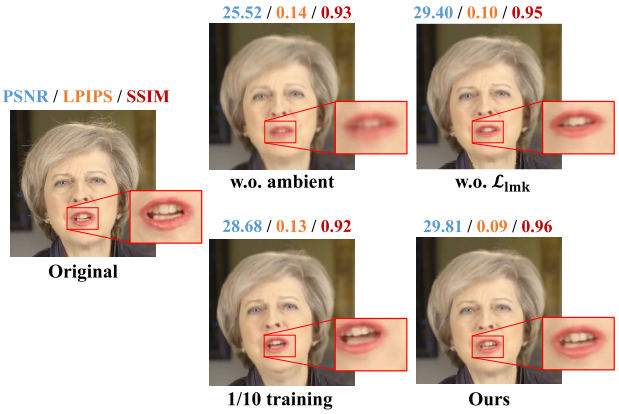

FIGURE 7. Results of the ablation study. All the same areas in each image are cropped and zoomed in (red box).

module and apply simple direct concatenation for expression conditioning, the model fails to capture facial expression and the visual quality of the outputs drops significantly. Although $\mathcal{L}_{\text{lmk}}$ does not prominently affect image quality, fine details of the face are lost when our model is trained without $\mathcal{L}_{\text{lmk}}$. When we use only 1/10 sequences for training, the model does not sufficiently learn the conditions, thus it shows some errors in control head pose or expression.

V. CONCLUSION

In this research paper, we have introduced a novel and efficient dynamic neural field for reconstructing 4D avatars with controllable features. Our approach leverages an ambient encoder to effectively encode the low-dimensional basis of the 3D Morphable Model (3DMM) for precise control over facial expressions. In a comparison with the conventional direct conditioning method (concatenation), our technique consistently outperformed in both qualitative and quantitative evaluations. Additionally, through the incorporation of explicit grid encoding, our model achieved significant improvements in computational efficiency, being 100 times faster during inference and 30 times faster during training when contrasted with existing methods. We believe that our approach holds great potential for practical applications in the realm of 4D facial avatar reconstruction. However, it's worth noting that a limitation of our method, which is also a shared challenge in NeRF-based approaches, is the necessity for independent training for each distinct identity. Recently, there has been a growing trend among researchers to combine NeRF with generative models. Although 3D-aware generative models currently have limited applicability, the prospect of extending the versatility of NeRF is an intriguing avenue for future research, and we defer this exploration to our forthcoming work.

REFERENCES

[1] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9458–9467.

[2] A. Siarohin, S. Lathuiliére, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[3] Z. Chen, C. Wang, B. Yuan, and D. Tao, "PuppeteerGAN: Arbitrary portrait animation with semantic-aware appearance transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13515–13524.

[4] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 1999.

[5] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Sep. 2009.

[6] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1–17, Dec. 2017.

[7] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 641–650, 2003.

[8] S. Bouaziz, Y. Wang, and M. Pauly, "Online modeling for realtime facial animation," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–10, Jul. 2013.

[9] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3D shape regression for real-time facial animation," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–10, Jul. 2013.

[10] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, Aug. 2018.

[11] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019.

[12] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Dec. 2020, pp. 405–421.

[13] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-NeRF: Neural radiance fields for dynamic scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10313–10322.

[14] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5845–5854.

[15] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz, "HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–12, Dec. 2021.

[16] G. Gafni, J. Thies, M. Zollhöfer, and M. Nießner, "Dynamic neural radiance fields for monocular 4D facial avatar reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8645–8654.

[17] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–15, Jul. 2022.

[18] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022.

[19] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023.

[20] Y.-L. Chen, H.-T. Wu, F. Shi, X. Tong, and J. Chai, "Accurate and robust 3D facial capture using a single RGBD camera," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3615–3622.

[21] P. Garrido, L. Valgaert, C. Wu, and C. Theobalt, "Reconstructing detailed dynamic face geometry from monocular video," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–10, Nov. 2013.

[22] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormaehlen, P. Perez, and C. Theobalt, "Automatic face reenactment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 4217–4224.

[23] A. E. Ichim, S. Bouaziz, and M. Pauly, "Dynamic 3D avatar creation from hand-held video input," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–14, Jul. 2015.

[24] M. C. Bühler, A. Meka, G. Li, T. Beeler, and O. Hilliges, "VariTex: Variational neural face textures," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13870–13879.

[25] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[26] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "STGAN: A unified selective transfer network for arbitrary image attribute editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3668–3677.

[27] J.-g. Kwak, D. K. Han, and H. Ko, "CAFE-GAN: Arbitrary face attribute editing with complementary attention feature," in *Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 524–540.

[28] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.

[29] Y. Gao, F. Wei, J. Bao, S. Gu, D. Chen, F. Wen, and Z. Lian, "High-fidelity and arbitrary face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16110–16119.

[30] D. Yoon, J. Kwak, Y. Li, D. Han, Y. Jin, and H. Ko, "Reference guided image inpainting using facial attributes," 2023, *arXiv:2301.08044*.

[31] D. Yoon, J.-G. Kwak, Y. Li, D. Han, and H. Ko, "DIFAI: Diverse facial inpainting using StyleGAN inversion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 1141–1145.

[32] J. G. Kwak and H. Ko, "Unsupervised generation and synthesis of facial images via an auto-encoder-based deep generative adversarial network," *Appl. Sci.*, vol. 10, no. 6, p. 1995, Mar. 2020.

[33] J. N. M. Pinkney and D. Adler, "Resolution dependent GAN interpolation for controllable image synthesis between domains," 2020, *arXiv:2010.05334*.

[34] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2019, *arXiv:1907.10830*.

[35] G. Song, L. Luo, J. Liu, W.-C. Ma, C. Lai, C. Zheng, and T.-J. Cham, "AgileGAN: Stylizing portraits by inversion-consistent transfer learning," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–13, Aug. 2021.

[36] J.-g. Kwak, Y. Li, D. Yoon, D. Han, and H. Ko, "Generate and edit your own character in a canonical view," 2022, *arXiv:2205.02974*.

[37] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "VToonify: Controllable high-resolution portrait video style transfer," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–15, Dec. 2022.

[38] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10034–10044.

[39] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[40] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[41] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[42] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1532–1540.

[43] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny, "StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3616–3626.

[44] J. Gu, L. Liu, P. Wang, and C. Theobalt, "StyleNeRF: A style-based 3D-aware generator for high-resolution image synthesis," 2021, *arXiv:2110.08985*.

[45] Y. Deng, J. Yang, J. Xiang, and X. Tong, "GRAM: Generative radiance manifolds for 3D-aware image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10663–10673.

[46] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. de Mello, O. Gallo, L. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3D generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022.

[47] J.-g. Kwak, Y. Li, D. Yoon, D. Kim, D. Han, and H. Ko, "Injecting 3D perception of controllable nerf-gan into stylegan for editable portrait image synthesis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 236–253.

[48] J. Xie, H. Ouyang, J. Piao, C. Lei, and Q. Chen, "High-fidelity 3D GAN inversion by Pseudo-multi-view optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023.

[49] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," in *Proc. 11th Annu. Conf. Comput. Graph. Interact. Techn.*, Jan. 1984.

[50] C. Sun, M. Sun, and H.-T. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5449–5459.

[51] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "TensoRF: Tensorial radiance fields," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 333–350.

[52] Z. Li, Q. Wang, F. Cole, R. Tucker, and N. Snavely, "DynIBaR: Neural dynamic image-based rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023.

[53] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial Landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.

[54] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "AD-NeRF: Audio driven neural radiance fields for talking head synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5764–5774.

[55] J. Tang, K. Wang, H. Zhou, X. Chen, D. He, T. Hu, J. Liu, G. Zeng, and J. Wang, "Real-time neural radiance talking portrait synthesis via audio-spatial decomposition," 2022, *arXiv:2211.12368*.

[56] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5548–5557.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

JEONG-GI KWAK received the B.S. and M.S. degrees from the School of Electrical Engineering, Korea University, Seoul, South Korea, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include computer vision and graphics, specifically, 3D-aware image/video synthesis, digital human, and neural rendering.

HANSEOK KO (Senior Member, IEEE) received the B.S. degree in electrical engineering from Carnegie Mellon University, in 1982, the M.S. degree in electrical engineering from Johns Hopkins University, in 1988, and the Ph.D. degree in electrical engineering from CUA, in 1992. At the onset of his career, he was with WOL, MD, USA, where his work involved signal and image processing. In March 1995, he joined the Department of Electronics and Computer Engineering, Korea University, where he is currently a Professor.