MaintaAvatar: A Maintainable Avatar Based on Neural Radiance Fields by Continual Learning

Shengbo Gu[1,3], **Yu-Kun Qiu**[1,3], **Yu-Ming Tang**[1,3], **Ancong Wu**[1,3*], **Wei-Shi Zheng**[1,2,3*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{gushb3,qiuyk,tangym9}@mail2.sysu.edu.cn, wuanc@mail.sysu.edu.cn, wszheng@ieee.org

Abstract

The generation of a virtual digital avatar is a crucial research topic in the field of computer vision. Many existing works utilize Neural Radiance Fields (NeRF) to address this issue and have achieved impressive results. However, previous works assume the images of the training person are available and fixed while the appearances and poses of a subject could constantly change and increase in real-world scenarios. How to update the human avatar but also maintain the ability to render the old appearance of the person is a practical challenge. One trivial solution is to combine the existing virtual avatar models based on NeRF with continual learning methods. However, there are some critical issues in this approach: learning new appearances and poses can cause the model to forget past information, which in turn leads to a degradation in the rendering quality of past appearances, especially color bleeding issues, and incorrect human body poses. In this work, we propose a maintainable avatar (MaintaAvatar) based on neural radiance fields by continual learning, which resolves the issues by utilizing a Global-Local Joint Storage Module and a Pose Distillation Module. Overall, our model requires only limited data collection to quickly fine-tune the model while avoiding catastrophic forgetting, thus achieving a maintainable virtual avatar. The experimental results validate the effectiveness of our MaintaAvatar model.

1 Introduction

Free-viewpoint rendering of 3D scenes has attracted significant academic and industrial attention, with effective solutions such as neural radiance fields (NeRF) and 3D Gaussian Splatting (3DGS) demonstrating impressive performance in this field. (Mildenhall et al. 2021; Barron et al. 2021; Martin-Brualla et al. 2018, 2021; Wang et al. 2021; Zhang et al. 2020; Kerbl et al. 2023; Qian et al. 2024; Li et al. 2024; Zhou et al. 2024; Sun, Wu, and Gao 2024). Among them, human-specific methods (Peng et al. 2021; Liu et al. 2020; Weng et al. 2022; Sun et al. 2022; Hu, Hu, and Liu 2024; Pan et al. 2023; Zhao et al. 2022; Li et al. 2024) model dynamic human bodies by establishing a deformable radiance field, achieving the driving and free-viewpoint rendering of dynamic human bodies.

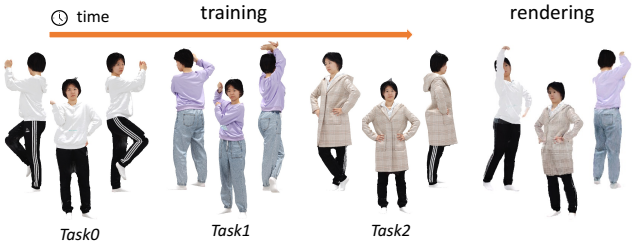

Figure 1: In reality, a person's pose and appearance constantly update. Our maintainable avatar is designed to continuously learn from sequential data, enabling it to render any previously encountered viewpoint, pose, or appearance.

In this paper, we focus on creating a maintainable avatar that models a certain character with ever-changing poses and appearances (see Figure 1). In the real world, the poses and appearances of a person could be updated frequently, and consequently, the corresponding human avatar should also be updatable. Employing conventional static human-nerf methods to track real-world trends will lead to a high demand for computation, training time, and extra storage room, let alone the need for abundant training samples from new styles and poses. We raise such a question: *Is it possible to maintain a dynamic human avatar using only a few shots from the updated character, with the ability to trace back to any point of time?*

The main challenge of such a task lies in the old avatar re-rendering with old poses and appearances, which is related to the well-known catastrophic forgetting (French 1999; Yang et al. 2023) of deep neural networks. Specifically, if the model directly learns from new data, it tends to forget the information about past poses and appearances, resulting in poor rendering outcomes. Further, we identify two major issues that employ the popular 'replay' strategy that stores or generates (fake) old inputs for future task training (Chung et al. 2022; Wu et al. 2024; Cai and Müller 2023; Shin et al. 2017) into a human avatar: (1) The appearances between different tasks can interfere with each other, resulting in color bleeding and affecting the geometric rendering quality of each task. (2) During the process of continual learning of novel appearances and poses, the model may render incor-

*Corresponding Authors.

rect poses in past tasks, consequently affecting the driving of the virtual avatar.

To address these problems, we propose two key solutions. First, the Global-Local Joint Storage Module stores the distinctions between global and local information of different appearances separately in global embeddings and Triplanes. This allows the model to better differentiate the variations among different appearances. Second, the Pose Distillation Module extracts the pose information from past tasks as supervision signals for novel tasks, enabling the model to better retain the pose information of past tasks. Additionally, we create a pretrained model trained on multiple human bodies as the initialization for training, which can adjust the surface of the avatar model with only a few images, and quickly adapt to new human bodies. We summarize our main contributions as follows:

- We study the problem of modeling a maintainable virtual avatar, which only needs a few shots to quickly update the character's appearance and also maintain the ability to render the old appearance. This maintainable virtual avatar can adapt in real-time to updates in the appearance of real-world individuals.
- We analyze the two critical issues in the continual learning of maintainable virtual avatars. First, in order to prevent color bleeding between different appearances, we introduce the Global-Local Joint Storage Module that can precisely model the differences between various appearances.
- Besides, to alleviate the information loss in the human pose, we introduce the Pose Distillation Module, which is capable of preserving the correct pose information of past appearances.

Experimental results in two datasets demonstrate the effectiveness of our model, achieving state-of-the-art performance.

2 Related Work

Neural Radiance Fields. NeRF (Mildenhall et al. 2021) is a widely acclaimed 3D technology that primarily focuses on synthesizing novel views. It represents a 3D environment as an implicit neural radiance field. Given a camera pose, it can generate images from any viewpoint using ray tracing. Specifically, NeRF (Mildenhall et al. 2021) utilizes a Multilayer Perceptron (MLP) (Taud and Mas 2018) to map spatial coordinates and view directions to their corresponding colors and opacity. NeRF has sparked significant interest and extensive research across a range of domains, including applications in autonomous driving (Hu et al. 2024b; Chen et al. 2024; Cheng et al. 2023; Feldmann et al. 2024), controllable human avatars (Weng et al. 2022, 2023; Peng et al. 2021; Sun et al. 2022; Jiang et al. 2022), large urban scenes (Wang et al. 2023; Xu et al. 2023; Turki et al. 2023), and text-driven 3D scene editing (Bao et al. 2023) and so on.

Human-specific Neural Representation. Previous work (Weng et al. 2022, 2023; Zhao et al. 2022; Hu, Hu, and Liu 2024; Hu et al. 2024a; Peng et al. 2021) expressed the human body as an implicit neural field, learning dynamic human figures by transforming bodies of different poses into a T-pose in a standard space. (Weng et al. 2023) learns human bodies with multiple appearances from unstructured data, addressing sparse observations in multi-outfit datasets by fusing appearance features via a shared network. However, it cannot create a virtual avatar that continually updates appearances while retaining the ability to render old ones. (Mu et al. 2023; Gao et al. 2022; Hu et al. 2023; Pan et al. 2023; Kwon et al. 2023; Zhao et al. 2022) propose a human body NeRF model that can quickly generalize to unseen human bodies in a few-shot setting without the need for continual learning. However, their performance is poor. (Liu et al. 2024; Kolotouros et al. 2024) generate virtual avatars through text guidance, and (Zhang et al. 2023a; Mendiratta et al. 2023) further provide the capability of controllable editing, but they cannot create virtual avatars corresponding to real-world characters. Recent works have leveraged Gaussian Splatting (Wu et al. 2023; Chen, Wang, and Liu 2023; Kerbl et al. 2023; Tang et al. 2023) for human representation, achieving impressive results. (Hu, Hu, and Liu 2024; Qian et al. 2024; Hu et al. 2024a) further use it for fast training and real-time rendering.

NeRF for Continual Learning. NeRF for continual learning is a new research hotspot. (Cai and Müller 2023; Chung et al. 2022; Po et al. 2023) propose a replay-based algorithm to establish continual learning NeRF, which can continuously learn new perspectives or dynamic scenes and achieve impressive results. (Wu et al. 2024) propose a lightweight expert adaptor to adapt to scene changes and a knowledge distillation learning objective to retain the invariant model. (Zhang et al. 2023b) focusing on real-world scenarios with unknown camera poses. However, they cannot handle dynamic human bodies as they lack body modeling.

In contrast, MaintaAvatar is the first to propose maintaining dynamic virtual characters, enabling sequential learning of diverse poses, viewpoints, and appearances.

3 MaintaAvatar

Our approach, MaintaAvatar, maintains a virtual avatar that can continuously learn novel poses and novel appearances without the need for retraining, saving both training and data collection costs. The pipeline is shown in Figure 2. We introduce a deformation field to model the dynamic avatar (Section 3.1) and employ a strategy based on generative replay to enable continual learning of the model, avoiding catastrophic forgetting (Section 3.1). Building on this, we propose the Global-Local Joint Storage Module to model global and local appearance variations (Section 3.2). Furthermore, we propose a Pose Distillation Module to mitigate the rendering of incorrect poses during the continual learning process (Section 3.3).

3.1 Preliminaries

Deformable NeRF Based on SMPL Model. Following PersonNeRF (Weng et al. 2023), we model NeRF (Mildenhall et al. 2021) as a deformable neural radiance field. Based on the observed SMPL (Loper et al. 2023) model parameters, we can warp the canonical volume F_c to the observed volume F_o (Weng et al. 2022, 2023).

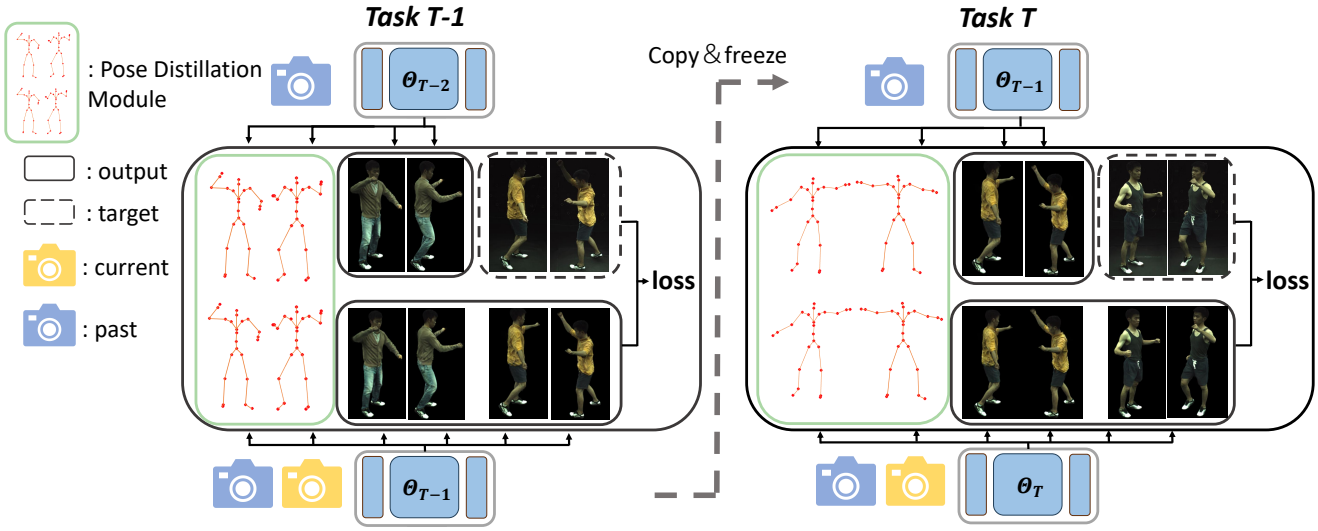

Figure 2: **MaintaAvatar Pipeline.** In this paper, we propose a continual learning strategy pipeline primarily based on the replay method. During the training of *Task T*, we replicate and freeze the network Θ_{T-1} from the past *Task T-1*. Given the camera parameters from *Task T-1*, the network Θ_{T-1} can generate corresponding patches and the residual human body pose of one randomly selected past appearance, which are utilized to supervise the training of *Task T*. Simultaneously, Θ_T is trained using images from the new appearance. In addition to image supervision, we incorporate a Pose Distillation Module to enhance the memory of past pose information, thereby improving rendering quality. Ultimately, our model is capable of continuously learning the novel appearance without forgetting past appearances.

Specifically, similar to PersonNeRF (Weng et al. 2023), with SMPL (Loper et al. 2023) model, we establish a rigid deformation field. By employing inverse linear skinning, we transform the vertices from the canonical volume F_c into their corresponding positions in the observed volume F_o:

$$\mathrm{F}_o(\mathbf{x}, \mathbf{p}) = \mathrm{F}_c(\mathrm{T}_{skel}(\mathbf{x}, \mathbf{p})). \tag{1}$$

where $\mathbf{p} = (J, \Omega)$ represents body pose, including joint positions J and local joint rotations Ω. $\mathbf{x}$ represents the coordinates of the sampling points. T_{skel} is defined as rigid body motion, which maps an observed position to the canonical space:

$$\mathrm{T}_{skel}(\mathbf{x}, \mathbf{p}) = \sum_{i=1}^{K} w_o^i(\mathbf{x})(R_i\mathbf{x} + \mathbf{t}_i). \tag{2}$$

where (R_i,$\mathbf{t}_i$) is the mapping of the i-th bone from the observation volume to the canonical volume, which can be computed through body pose $\mathbf{p}$. $w_o^i(\mathbf{x})$ and $w_c^i(\mathbf{x})$ are the corresponding weight in the observed volume and the canonical volume (Weng et al. 2022, 2023), respectively. We derive the formula for computing $w_o^i(\mathbf{x})$ as:

$$w_o^i(\mathbf{x}) = \frac{w_c^i(\mathbf{x})(R_i\mathbf{x} + \mathbf{t}_i)}{\sum_{k=1}^{K} w_c^k(\mathbf{x})(R_k\mathbf{x} + \mathbf{t}_k)}. \tag{3}$$

To address the issue of inaccuracies in the body pose parameters from the dataset, some work utilizes a plug-and-play module MLP_p (Weng et al. 2022, 2023; Hu, Hu, and Liu 2024). This module predicts the residual between the current body pose parameters and the true body pose parameters, correcting the body pose. As follow:

$$\Delta_\Omega(\mathbf{p}) = \mathrm{MLP}_{\theta_{\mathrm{pose}}}(\Omega). \tag{4}$$

Then the pose $\mathbf{p}$ could be updated to:

$$P_{\mathrm{pose}}(\mathbf{p}) = (J, \Delta_\Omega(\mathbf{p}) \otimes \Omega). \tag{5}$$

Continual Learning for NeRF. The primary approach for continual learning for NeRF is to integrate a continual learning strategy with the NeRF network, parameterized by Θ. And Θ_T represents the network parameter of the current task, and Θ_{T-1} represents the network structure of the past task. At every time step t of the continual NeRF:

1. Freeze and copy the model Θ_{T-1} from *Task T-1*.

2. Given camera parameters, and pose parameters from the current task, current model Θ_T firstly transforms the human body from the canonical volume F_c to the observed volume F_o and then utilizes an MLP network to render the color corresponding to each ray $\widehat{C}(\mathbf{r})$.

3. Utilizing the saved human body pose and camera parameters from past tasks, the network Θ_{T-1} executes the process for step 2, and obtains the color supervision signal $\widetilde{C}(\bar{\mathbf{r}})$ of the past tasks.

4. The supervision signals $\widetilde{C}(\bar{\mathbf{r}})$ generated by model Θ_{T-1} and the images obtained from the current task are used to supervise the training of model Θ_T.

3.2 Global-Local Joint Storage Module

Most current researchs model dynamic scene variations using global geometry and color embeddings. However, our experiments show that this approach causes mutual interference between the novel and past appearances, leading

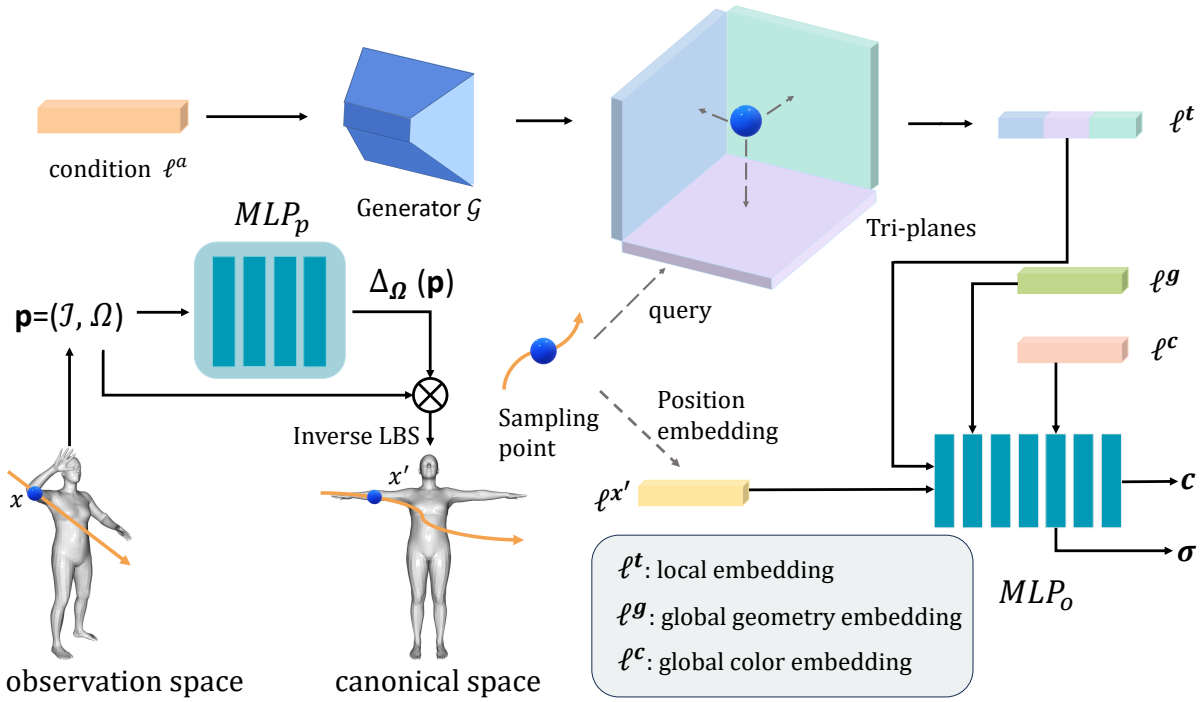

Figure 3: **Pipeline for MaintaAvatar Network Θ Structure.** For any given human body pose, we utilize skeletal motion based on the SMPL model to transform the body from the observation space to the canonical space. Meanwhile, we employ a network MLP_p to predict the residual $\Delta_\Omega(\mathbf{p})$ between the current pose parameters and the true pose parameters. Subsequently, our Global-Local Joint Storage Module generates Tri-plane-based local embedding and global embedding for each appearance. These embedding, along with coordinate embedding, are fed into MLP_o to predict color and opacity.

to color blending (Figure 4). Unlike the continual simple changes in dynamic scenes (Cai and Müller 2023), changes in human appearance usually involve drastic variations in color and geometry (Figure 1). It is challenging to represent these significant spatial variations through global embeddings alone in continual learning, making global embeddings inadequate for accurately representing human appearance variations.

To address this challenge, we propose the Global-Local Joint Storage Module to model both global and local appearance variations. Global embeddings can represent the overall changes in human appearance, while local embeddings is used to represent the fine-grained variations based on these global changes. As shown in the Figure 3, we randomly generate the global geometry embedding ℓ^g and color embedding ℓ^c to model global variations in appearance. On this basis, we introduce local embedding ℓ^t to model local variations superimposed on the global changes. To capture the local information of each appearance, we randomly generate a trainable condition embedding ℓ^a for each appearance, which is fed into the generator $\mathcal{G}$ to generate a Tri-plane. Then, for each sampling point on the ray, the local embeddings can be obtained by querying its position coordinates x on the corresponding appearance Tri-plane.

$$\ell^t = q(\mathcal{G}(\ell^a), x) \tag{6}$$

Finally, we use the following formula to obtain the opacity and color of each sampling point:

$$MLP_o : (\gamma(x), \ell^g, \ell^c, \ell^t) \longrightarrow (c, \sigma). \tag{7}$$

where MLP_o is shown in Figure 3. $\gamma(x)$, ℓ^g, ℓ^c, ℓ^t respectively represent the coordinate embedding, global geometry embedding, global color embedding and local embedding.

3.3 Pose Distillation Module

Human pose is an important topic in the field of human body rendering. Inaccuracies in poses can lead to a significant degradation in rendering quality, resulting in unrealistic effects. MLP_p is a plug-and-play method to correct pose errors in datasets, and is suitable for most human-specific neural representation. However, we found that current continual learning models for the human body may introduce incorrect poses from past tasks because MLP_p overfits to learning new poses. Therefore, maintaining the accuracy of poses from past tasks during continual learning is a challenge to address. In this section, we propose a Pose Distillation Module. As shown in the Figure 2. Similar to multi-layer distillation loss (Sun et al. 2019), we distill the outputs of MLP_p that perform pose correction.

$$\mathcal{L}_{\text{POSE}} = \| \widehat{\Delta}_\Omega(\mathbf{p}) - \widetilde{\Delta}_\Omega(\mathbf{p}) \|_2^2 . \tag{8}$$

where $\widehat{\Delta}_\Omega(\mathbf{p})$ represents the output of the MLP_p for the novel model and $\widetilde{\Delta}_\Omega(\mathbf{p})$ represents the output of the MLP_p

Method	novel view			novel pose		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Joint	29.507	0.9652	34.579	29.626	0.9702	27.329
CLNeRF	12.3	0.927	152	-	-	-
MEIL-NeRF	25.810	0.9429	71.282	25.902	0.9465	69.822
PersonNeRF	28.199	0.9626	44.354	29.008	0.9665	36.724
PersonNeRF$_{CL}$	28.605	0.9632	38.826	28.678	0.9665	32.889
Ours	**29.495**	**0.9663**	**36.497**	**29.630**	**0.9702**	**30.166**

Table 1: The overall free-viewpoint and novel pose rendering experimental results on ZJU-MoCap (Peng et al. 2021), which represent the evaluation metrics across all previously trained tasks using the weights trained on the final task. The best results are marked in bold.

for the previously frozen model, as shown in the Figure 3. we calculate the loss using the L2 norm. In practice, our training process is divided into two phases. The first phase focuses on optimizing rendering quality, during which we do not activate the pose distillation loss function, as doing so could result in blurred rendering outcomes. However, in the second phase of training, which begins at t_0 as indicated in Equation 12, we shift our focus to pose correction. In this phase, we activate the pose distillation loss function while freezing all components of the network except for the network MLP_p and the color embedding ℓ^c. This strategic shift allows the model to better capture and memorize the pose information from previous appearances. Additionally, the continual learning in color embedding will prompt the model to find a balance between the colors of the novel and past appearances.

3.4 Optimization

Loss Function. The overall loss function is composed as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{CR}} + \lambda_p \mathcal{L}_{\mathrm{CL}} + \lambda_\beta \mathcal{L}_{\mathrm{POSE}}. \tag{9}$$

$\mathcal{L}_{\mathrm{POSE}}$ is defined by Equation 8. $\mathcal{L}_{\mathrm{CR}}$ represents the loss function for learning novel tasks, while $\mathcal{L}_{\mathrm{CL}}$ represents the loss function for memorizing past tasks. $\mathcal{L}_{\mathrm{CR}}$ and $\mathcal{L}_{\mathrm{CL}}$ are respectively defined by the following formulas:

$$\mathcal{L}_{\mathrm{CR}} = \sum_{\mathbf{r}\in\mathcal{R}} [\| \widehat{C}(\mathbf{r}) - C(\mathbf{r}) \|_2^2] + \lambda_2 LPIPS(\widehat{C}(\mathbf{r}), C(\mathbf{r})). \tag{10}$$

$$\mathcal{L}_{\mathrm{CL}} = \sum_{\bar{\mathbf{r}}\in\bar{\mathcal{R}}} [\| \widehat{C}(\bar{\mathbf{r}}) - \widetilde{C}(\bar{\mathbf{r}}) \|_2^2] + \lambda_2 LPIPS(\widehat{C}(\bar{\mathbf{r}}), \widetilde{C}(\bar{\mathbf{r}})). \tag{11}$$

where $\mathbf{r}$ refer to the current rays of the set of rays $\mathcal{R}$ in each batch, while $\bar{\mathrm{r}}$ refer to the past rays of the set of rays $\bar{\mathcal{R}}$ in each batch. C, $\widehat{C}$ and $\widetilde{C}$ are the ground truth RGB colors, current network predicted RGB colors, and past network predicted RGB colors as supervisory signals.

We set λ_2 to 0.2. λ_1, λ_p refer to weights for controlling the trade-off of current rays and past rays (Chung et al. 2022). In which, λ_1 = 0.2 and λ_p is defined by the following formula:

$$\lambda_p = \begin{cases} \sin\left(-\frac{\pi}{2} + \frac{\pi(t-t_{\text{init}})}{t_{\max}-t_0-t_{\text{init}}}\right) + 1 & \text{if } t < t_{\max} - t_0 \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

where t_{init}, $t_{\max}$ respectively represent the initial and final iteration of the current task. t_0 represents the iteration to activate the Pose Distillation Module. For λ_β, if $t < t_{\max} - t_0$, we set it to 0; otherwise, we set it to 800. It is worth noting that if $t < t_{max} - t_0$, we will freeze all networks except for network MLP_p and the color embedding ℓ^c.

4 Experiments

4.1 Datasets

Our model is evaluated on ZJU-MoCap (Peng et al. 2021) (Peng et al. 2021) and THuman2.0 dataset (Yu et al. 2021).

ZJU-MoCap (Peng et al. 2021). For ZJU-MoCap (Peng et al. 2021), we select subjects (377, 392, 393, 394) for our dataset, as they all feature the same individual in different sets of clothing. This dataset includes one camera assigned for training and the other 22 cameras for evaluation. For each task, we choose only five images with different viewpoints (ensuring a wide distribution of viewpoints as much as possible.) and different poses for training.

THuman2.0 (Yu et al. 2021). For Thuman2.0 (Yu et al. 2021), we select subjects (262, 220, 207, 125) as the dataset. Thuman2.0 (Yu et al. 2021) provides SMPL (Loper et al. 2023) but does not offer rendered images or corresponding camera parameters. We use PyTorch3D to obtain rendered images from different viewpoints. To fully utilize the limited dataset, we render images from four viewpoints (0, 90, 180, 270 degrees) for each pose as the training set and render images from nine viewpoints (0, 40, 80, ..., 280, 320 degrees) for evaluation.

4.2 Implementation Details

The random seed is set to 42. The network MLP_o and the MLP_p have 8 and 4 layers respectively. The global color embedding ℓ^c (length 48), global geometry embedding ℓ^g (length 16), and condition embedding ℓ^a (length 16) are all optimized during training. The Tri-plane has dimensions of 3*512*512*8. We set the learning rates for the MLP_o and

Method	subject 262			subject 220			subject 207			subject 125		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
CLNeRF	12.9	0.94	236	12.9	0.938	235	14.2	0.913	123	10.8	0.907	224
MEIL-NeRF	18.907	0.9344	78.161	17.142	0.9210	97.789	18.411	0.9409	72.456	15.484	0.9215	96.626
PersonNeRF	20.094	0.9298	82.099	20.485	0.9338	78.517	21.560	0.9493	59.248	18.930	0.9390	71.145
PersonNeRF$_{CL}$	20.905	0.9415	63.568	20.130	0.9365	72.227	22.951	0.9624	45.814	19.448	0.9449	65.084
Our method	**22.316**	**0.9491**	**56.023**	**21.642**	**0.9449**	**64.173**	**23.998**	**0.9664**	**41.271**	**20.209**	**0.9488**	**60.884**

Table 2: The overall free-viewpoint experimental results on Thuman2.0 (Yu et al. 2021). The best results are displayed in boldfaced font.

both the ℓ^c and ℓ^g to 5×10^{-4}, and the rest to 5×10^{-5}. Adam is adopted as the optimizer. For the current task, we sample 6 patches of 32×32 size, whereas for past tasks, we sampled one patch of 64×64 size. 128 points are sampled from each ray. In the ZJU-MoCap dataset (Peng et al. 2021), each task is trained for 12,000 iterations, with the pose distillation loss (Equation 8) inactive for the first 10,000 iterations and activated for the final 2,000. In contrast, in the Thuman2.0 dataset (Yu et al. 2021), tasks undergo 80,000 iterations, divided into two phases: the pose distillation loss remains inactive for the initial 70,000 iterations and becomes active for the last 10,000. To enable the model to quickly fine-tune to new human bodies, as well as to cope with a small amount of data, we constructed a pretrained model of other similarly sized human bodies as initialization to train the new human.

4.3 Comparison with Other Methods

Up to now, there has been no ongoing research regarding the continual learning of multi-appearance human bodies. Therefore, we compare our model with CLNeRF, MEIL-NeRF, PersonNeRF and PersonNeRF$_{CL}$, as shown in Table 1. Specifically, we respectively augment CLNeRF (Cai and Müller 2023), MEIL-NeRF (Chung et al. 2022) and PersonNeRF (Weng et al. 2023) with a pretrained model as well as global geometry and color embeddings. Additionally, we apply the continual learning strategy of MEIL-NeRF to PersonNeRF as the PersonNeRF$_{CL}$. As a reference of performance upper bound, we report the results of training on the dataset of all tasks together denoted as "Joint". The evaluation metrics we use are PSNR (Hore and Ziou 2010), SSIM (Hore and Ziou 2010), and LPIPS.

Results on ZJU-MoCap (Peng et al. 2021). We first compare our method with others and "Joint" on ZJU-MoCap (Peng et al. 2021). The free-viewpoint and novel pose rendering results are shown in Table 1. In all metrics, our method approaches the performance of "Joint" but does not fully reach it, while still outperforming all other methods.

The visualization results of free-viewpoint rendering on ZJU-MoCap (Peng et al. 2021) are illustrated in Figure 4. We use the model of final task ($Task3$). The PersonNeRF$_{CL}$'s rendering of past tasks demonstrates issues with poor rendering quality and incorrect pose. Specifically, there are problems with missing details in color rendering and overall rendering inaccuracies in poses of the head, arms, and so on. Our proposed Global-Local Joint Storage Module and Pose Distillation Module effectively address these issues.

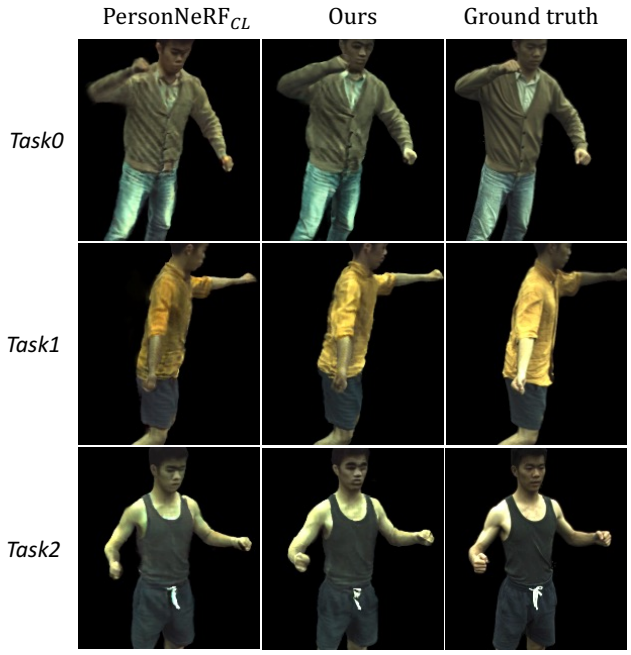

Figure 4: The visualization comparison results for the free-viewpoint rendering of our method and the PersonNeRF$_{CL}$ on ZJU-MoCap (Peng et al. 2021) for the past tasks. Our method demonstrates superior rendering quality, especially in terms of color and human pose.

Results on Thuman2.0 (Yu et al. 2021). We further evaluated our model on Thuman2.0 (Yu et al. 2021). The results in Table 2 show that our model significantly outperforms the other models. At the same time, Figure 5 demonstrates that our model with Global-Local Joint Storage Module can learn new appearances more quickly by better understanding the differences between different appearances.

4.4 Ablation Studies

Global-Local Joint Storage Module. "w/o G-L" denotes the results without using the Global-Local Joint Storage Module. "full model" denotes our full model. Table 3 and Figure 6 demonstrate the rendering improvements brought

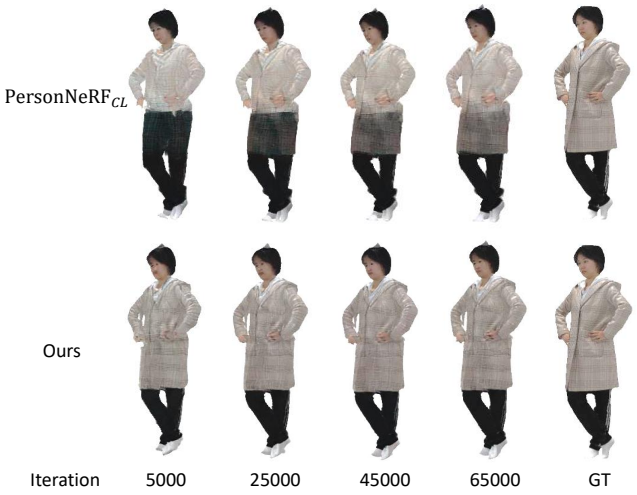

Figure 5: The Global-Local Joint Storage Module can help the model learn new appearances more quickly.

	Novel View			Novel Pose		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
w/o pose	29.21	0.9652	37.52	29.59	0.9698	30.83
w/o G-L	28.89	0.9646	36.60	29.15	0.9686	30.99
Full	**29.50**	**0.9663**	**36.50**	**29.63**	**0.9702**	**30.17**

Table 3: Ablations in ZJU-MoCap: "w/o G-L" omits the Global-Local Joint Storage Module, "w/o pose" omits the Pose Distillation Module.

by the Global-Local Joint Storage Module. We use the old model with green clothing to learn yellow clothing. Without using this module, the generated yellow appearance is influenced by the dark green appearance. Figure 5 demonstrates the training speed improvements.

Pose Distillation Module. "w/o pose" denotes the results of removing the Pose Distillation Module from our model. The ablation study regarding pose distillation is shown in Figure 7 and Table 3. We present the visualization experimental results on Thuman2.0 (Yu et al. 2021). The Pose Distillation Module is capable of rendering a human figure with the correct pose more effectively.

λ_p **Hyperparameter.** The hyperparameter λ_p controls the trade-off between current and past tasks, as demonstrated in the ablation experiment shown in Figure 8. λ_p is defined as the Equation 12. Without λ_p, the model struggles to balance the current task and past tasks, especially in cases where there is a significant color difference between current and past appearances.

5 Limitations and Conclusion

Limitations. Our method shows performance drops with significant clothing shape changes and struggles with pose generalization in complex poses due to limited exposure in the few-shot dataset.

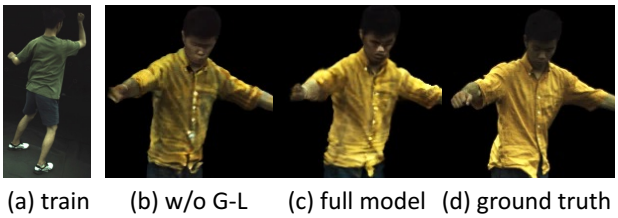

Figure 6: The rendering result for the past task (yellow apperance) after training on the current task (green apperance). The Global-Local Joint Storage Module effectively prevents color bleeding between different appearances.

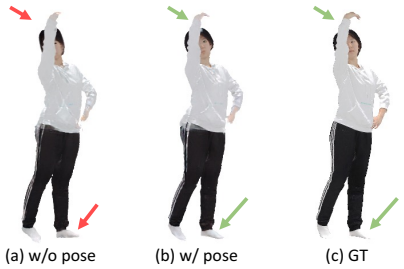

Figure 7: The pose distillation loss function improves the quality of novel view synthesis. The Pose Distillation Module is referred to as "pose".

Conclusion. MaintaAvatar is the first work to propose a maintainable virtual avatar to address the issue of continual changes in human appearance. We introduce two main components: a Global-Local Joint Storage Module to prevent color bleeding between different appearances and a Pose Distillation Module to solve the pose forgetting problem in continual learning. MaintaAvatar requires only a minimal set of training images to fine-tune to a novel appearance within a fixed, shorter training time, without forgetting the past appearances.

Acknowledgments

This work was supported partially by NSFC (92470202, U21A20471), Guangdong NSF Project (No. 2023B1515040025).

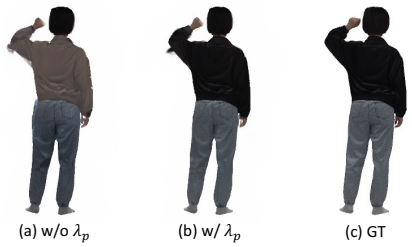

Figure 8: The hyperparameter λ_p enables the model to find balance between current appearance and past appearance.

References

Bao, C.; Zhang, Y.; Yang, B.; Fan, T.; Yang, Z.; Bao, H.; Zhang, G.; and Cui, Z. 2023. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20919–20929.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Cai, Z.; and Müller, M. 2023. CLNeRF: Continual Learning Meets NeRF. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23185–23194.

Chen, Y.; Zhang, J.; Xie, Z.; Li, W.; Zhang, F.; Lu, J.; and Zhang, L. 2024. S-NeRF++: Autonomous Driving Simulation via Neural Reconstruction and Generation. *arXiv preprint arXiv:2402.02112*.

Chen, Z.; Wang, F.; and Liu, H. 2023. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585*.

Cheng, K.; Long, X.; Yin, W.; Wang, J.; Wu, Z.; Ma, Y.; Wang, K.; Chen, X.; and Chen, X. 2023. UC-NeRF: Neural Radiance Field for Under-Calibrated multi-view cameras in autonomous driving. *arXiv preprint arXiv:2311.16945*.

Chung, J.; Lee, K.; Baik, S.; and Lee, K. M. 2022. MEIL-NeRF: Memory-Efficient Incremental Learning of Neural Radiance Fields. *arXiv preprint arXiv:2212.08328*.

Feldmann, C.; Siegenheim, N.; Hars, N.; Rabuzin, L.; Ertugrul, M.; Wolfart, L.; Pollefeys, M.; Bauer, Z.; and Oswald, M. R. 2024. NeRFmentation: NeRF-based Augmentation for Monocular Depth Estimation. *arXiv preprint arXiv:2401.03771*.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135.

Gao, X.; Yang, J.; Kim, J.; Peng, S.; Liu, Z.; and Tong, X. 2022. Mps-nerf: Generalizable 3d human rendering from multiview images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*, 2366–2369. IEEE.

Hu, L.; Zhang, H.; Zhang, Y.; Zhou, B.; Liu, B.; Zhang, S.; and Nie, L. 2024a. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 634–644.

Hu, S.; Hong, F.; Pan, L.; Mei, H.; Yang, L.; and Liu, Z. 2023. SHERF: Generalizable Human NeRF from a Single Image. *arXiv preprint arXiv:2303.12791*.

Hu, S.; Hu, T.; and Liu, Z. 2024. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20418–20431.

Hu, X.; Xiong, G.; Zang, Z.; Jia, P.; Han, Y.; and Ma, J. 2024b. PC-NeRF: Parent-Child Neural Radiance Fields Using Sparse LiDAR Frames in Autonomous Driving Environments. *arXiv preprint arXiv:2402.09325*.

Jiang, W.; Yi, K. M.; Samei, G.; Tuzel, O.; and Ranjan, A. 2022. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, 402–418. Springer.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

Kolotouros, N.; Alldieck, T.; Zanfir, A.; Bazavan, E.; Fieraru, M.; and Sminchisescu, C. 2024. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems*, 36.

Kwon, Y.; Kim, D.; Ceylan, D.; and Fuchs, H. 2023. Neural Image-based Avatars: Generalizable Radiance Fields for Human Avatar Modeling. *arXiv preprint arXiv:2304.04897*.

Li, Z.; Zheng, Z.; Wang, L.; and Liu, Y. 2024. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19711–19722.

Liu, L.; Xu, W.; Habermann, M.; Zollhöfer, M.; Bernard, F.; Kim, H.; Wang, W.; and Theobalt, C. 2020. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *arXiv preprint arXiv:2001.04947*.

Liu, X.; Zhan, X.; Tang, J.; Shan, Y.; Zeng, G.; Lin, D.; Liu, X.; and Liu, Z. 2024. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6646–6657.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Martin-Brualla, R.; Pandey, R.; Yang, S.; Pidlypenskyi, P.; Taylor, J.; Valentin, J.; Khamis, S.; Davidson, P.; Tkach, A.; Lincoln, P.; et al. 2018. Lookingood: Enhancing performance capture with real-time neural re-rendering. *arXiv preprint arXiv:1811.05029*.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7210–7219.

Mendiratta, M.; Pan, X.; Elgharib, M.; Teotia, K.; Tewari, A.; Golyanik, V.; Kortylewski, A.; and Theobalt, C. 2023. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Transactions on Graphics (ToG)*, 42(6): 1–18.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mu, J.; Sang, S.; Vasconcelos, N.; and Wang, X. 2023. ActorsNeRF: Animatable Few-shot Human Rendering with Generalizable NeRFs. *arXiv preprint arXiv:2304.14401*.

Pan, X.; Yang, Z.; Ma, J.; Zhou, C.; and Yang, Y. 2023. Transhuman: A transformer-based human representation for generalizable neural human rendering. In *Proceedings of*

the IEEE/CVF International conference on computer vision, 3544–3555.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. 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*, 9054–9063.

Po, R.; Dong, Z.; Bergman, A. W.; and Wetzstein, G. 2023. Instant continual learning of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3334–3344.

Qian, Z.; Wang, S.; Mihajlovic, M.; Geiger, A.; and Tang, S. 2024. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5020–5030.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30.

Sun, J.-M.; Wu, T.; and Gao, L. 2024. Recent advances in implicit representation-based 3d shape generation. *Visual Intelligence*, 2(1): 9.

Sun, M.; Yang, D.; Kou, D.; Jiang, Y.; Shan, W.; Yan, Z.; and Zhang, L. 2022. Human 3d avatar modeling with implicit neural representation: A brief survey. In *2022 14th International Conference on Signal Processing Systems (ICSPS)*, 818–827. IEEE.

Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*.

Taud, H.; and Mas, J. 2018. Multilayer perceptron (MLP). *Geomatic approaches for modeling land change scenarios*, 451–455.

Turki, H.; Zhang, J. Y.; Ferroni, F.; and Ramanan, D. 2023. SUDS: Scalable Urban Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12375–12385.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Wang, Z.; Shen, T.; Gao, J.; Huang, S.; Munkberg, J.; Hasselgren, J.; Gojcic, Z.; Chen, W.; and Fidler, S. 2023. Neural Fields meet Explicit Geometric Representations for Inverse Rendering of Urban Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8370–8380.

Weng, C.-Y.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 16210–16220.

Weng, C.-Y.; Srinivasan, P. P.; Curless, B.; and Kemelmacher-Shlizerman, I. 2023. PersonNeRF: Personalized Reconstruction from Photo Collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 524–533.

Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2023. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*.

Wu, X.; Dai, P.; Deng, W.; Chen, H.; Wu, Y.; Cao, Y.-P.; Shan, Y.; and Qi, X. 2024. CL-NeRF: Continual Learning of Neural Radiance Fields for Evolving Scene Representation. *Advances in Neural Information Processing Systems*, 36.

Xu, L.; Xiangli, Y.; Peng, S.; Pan, X.; Zhao, N.; Theobalt, C.; Dai, B.; and Lin, D. 2023. Grid-guided Neural Radiance Fields for Large Urban Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8296–8306.

Yang, Y.; Cui, Z.; Xu, J.; Zhong, C.; Zheng, W.-S.; and Wang, R. 2023. Continual learning with Bayesian model based on a fixed pre-trained feature extractor. *Visual Intelligence*, 1(1): 5.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 2021. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5746–5756.

Zhang, H.; Feng, Y.; Kulits, P.; Wen, Y.; Thies, J.; and Black, M. J. 2023a. Text-guided generation and editing of compositional 3d avatars. *arXiv preprint arXiv:2309.07125*.

Zhang, K.; Riegler, G.; Snavely, N.; and Koltun, V. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*.

Zhang, L.; Li, M.; Chen, C.; and Xu, J. 2023b. IL-NeRF: Incremental Learning for Neural Radiance Fields with Camera Pose Alignment. *arXiv preprint arXiv:2312.05748*.

Zhao, F.; Yang, W.; Zhang, J.; Lin, P.; Zhang, Y.; Yu, J.; and Xu, L. 2022. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7743–7753.

Zhou, S.; Chang, H.; Jiang, S.; Fan, Z.; Zhu, Z.; Xu, D.; Chari, P.; You, S.; Wang, Z.; and Kadambi, A. 2024. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21676–21685.