JEI Journal of Electronic Imaging

RESEARCH PAPER

Temporal residual neural radiance fields for monocular video dynamic human body reconstruction

Tianle Du,[a] Jie Wang,[b] Xiaolong Xie,[b] Wei Li,[a,*] Pengxiang Su,[a] and Jie Liu[a]
[a]Nanchang University, School of Software, Nanchang, China
[b]Nanchang University, Former Lake College, Nanchang, China

ABSTRACT. In the field of computer vision and graphics, high-quality reconstruction of the human body in static scenes has been achieved in recent years by a single multilayer perceptron (MLP) in a number of approaches. However, MLPs have capacity limitations, requiring substantial training time and computational resources for dynamic scene reconstruction, and the quality of reconstruction is significantly constrained. We propose a method for effectively processing complex spatiotemporal signals in dynamic scene human three-dimensional (3D) modeling. The proposed method uses temporal residual neural radiance fields to achieve novel view rendering and new pose synthesis of human bodies. To address the problem of representing temporal signals in video sequences, we construct a temporal residual field that is not related to the MLP architecture. Second, to improve the reconstruction efficiency, we propose an integrated approach that reduces the trainable parameters and accelerates rendering, thereby enhancing the network's feature representation capability. Finally, we design a multi-dimensional loss function to accurately measure the loss between predicted and actual spatial pixel values. The experimental results show that our proposed approach improves the peak signal-to-noise ratio and structural similarity index accuracy metrics compared with the latest representative methods. It maintains a similar accuracy to Anim-NeRF and Neural Body while achieving a nearly 780-fold increase in time efficiency.

 [DOI: 10.1117/1.JEI.33.4.043018]

Keywords: computer vision; human dynamic reconstruction; neural radiance fields; temporal residuals

Paper 240186G received Feb. 26, 2024; revised Jun. 16, 2024; accepted Jun. 18, 2024; published Jul. 16, 2024.

1 Introduction

Neural radiance fields (NeRFs)[1] have gained significant attention in the fields of computer graphics and computer vision due to their ability to learn the density and color attributes of every spatial point in a scene, resulting in astonishing effects in novel view synthesis[2] and 3D reconstruction. The multi-layer perceptron (MLP) neural network structure used by NeRFs can represent continuous spatio-temporal signals, demonstrating powerful capabilities in 3D reconstruction. In recent years, NeRFs have achieved impressive results in synthesizing new views of static scenes. This opens up new possibilities for more realistic virtual reality experiences and scene reconstruction.

Early methods[3,4] relied on multi-view stereo to obtain accurate sequences of 3D meshes. Multi-view stereoscopic methods obtain disparity or depth information by matching and aligning

*Address all correspondence to Wei Li, weili.cs@ncu.edu.cn

1017-9909/2024/$28.00

images from multiple viewpoints.[5] With the acquired disparity or depth information, a point cloud can be created to represent the 3D shape of the human body. However, methods that compute 3D meshes often struggle to accurately depict complex geometric structures and are less robust when faced with complex poses and movements. Additionally, they may struggle with missing texture or occlusion, resulting in limited photorealism. With the introduction of NeRFs, 3D reconstruction based on NeRFs has become a popular research focus. The method is capable of generating new and realistic views by learning a continuous 3D representation of the scene, thus overcoming the limitations of traditional methods.

However, NeRF and its variants[6–8] require a large number of queries to a deep MLP due to the limited capacity of MLP. This time-consuming computation makes them challenging for processing large spatio-temporal signals, such as long videos or dynamic 3D scenes. Furthermore, these models require training times of several hours due to the dual requirements for differentiable deformation modules and volumetric rendering. This also limits their application to scenarios that demand high rendering efficiency. In a previous study, Christian Reiser et al.[6] attempted to expedite rendering by replacing a single large MLP with thousands of smaller MLPs, with each MLP representing only a portion of the scene. However, this approach uses a uniform grid to divide the scene, which may not be optimal for irregularly shaped or detail-varied scenes. To address these challenges, one direct method[9,10] to increase the network complexity is by augmenting the total number of neurons or layers of the MLP. However, this approach can lead to slower inference and rendering speeds, along with increased GPU memory demands. The reason for this is that the inclusion of parameters leads to an increase in both time and memory requirements.[10]

In this paper, we propose an MLP architecture independent building block, TRes-NeRF, to model spatio-temporal fields aimed at solving the fast reconstruction of human performers in monocular videos. Specifically, we incorporate a temporal residual into the neural field and replace the linear layers in the MLP with time-dependent layers. These time-dependent weights are modeled as trainable temporal correlation coefficients and fused with the original layer weights. During the training process, our model is able to learn the time-dependent features in the image sequences, which improves the modeling capability of the spatio-temporal correlation. Compared with traditional MLP architectures,[11] our TRes-NeRF has a distinct advantage in spatio-temporal modeling. By adding additional trainable parameters without increasing the network width, it significantly reduces the training time. In addition, TRes-NeRF is able to capture dynamic changes in temporal sequences and integrate them into the reconstruction process. This ability to assimilate temporal information enables our model to more accurately reproduce the actions and postures of human performers.

To further accelerate rendering, our network integrates several key structures. First, for computing pixel colors, we employ a joint module consisting of rigid transformation and a spatial skipping scheme.[12] This approach extends the reconstruction of input postures into a dynamic canonical space and swiftly bypasses blank spaces by jumping through grid cells, thereby speeding up rendering. Second, to learn typical shapes and appearances, we utilize an efficient variant of NeRFs, Instant-NGP,[13] which replaces the MLP with a more efficient hash table as its data structure. We evaluate our method on both synthetic and real monocular videos of moving human figures. Compared with current state-of-the-art methods, our approach not only demonstrates superior reconstruction quality but also requires significantly shorter training periods. Therefore, our method is more suitable for monocular video human dynamic reconstruction and provides a new idea for human 3D reconstruction.

In summary, our contributions in this study are twofold:

- We proposed a residual building block to model the spatio-temporal field that is not related to the MLP architecture, allowing the use of smaller MLPs without sacrificing reconstruction quality and facilitating accelerated training with reduced GPU memory requirements.
- We integrate a method to accelerate the volumetric rendering process that enables fast real-time reconstruction of human performers in monocular videos. We validate the effectiveness of our method on a number of challenging tasks.

2 Related Work

2.1 Neural Radiance Fields

A NeRF[1] is a fully connected neural network oriented toward 3D implicit space modeling, designed to reconstruct high-quality 3D scenes from multi-view 2D images; it has become a mainstream method in novel view synthesis and 3D reconstruction. Its core idea is to model the color and density of each 3D point in the scene as a function, without the intermediate process of 3D reconstruction; based only on the positional parameters with the image, it can be used for new-view image synthesis. However, a NeRF requires the evaluation and rendering of a large number of 3D sampling points using a deep MLP and, for each sample point, millions of queries to an MLP to obtain the density and brightness. Therefore, rendering temporally and spatially complex scenes efficiently through a NeRF presents a challenge. In further research, a key approach has been the use of positional encoding[14,15] to increase the effective capacity of the MLP, transforming the modeling of radiance and density into functions of position and viewing direction. However, this approach is not sufficient for accelerated training on large-scale data and complex scenarios due to the increased amount of encoding computation.

2.2 Monocular Human Reconstruction

In recent research,[16,17] the use of neural representation has become a key technology for reconstructing high-quality 3D human models, especially for generating free-view videos of human performers. Peng et al.[18] implemented human model animation through neural blending weight fields and skeletal-driven deformation. The method improves the constraints of the optimization algorithm and is able to better normalize the learning of the deformation field to accomplish the reconstruction of a dynamic human body in monocular videos. Noguchi et al.[19] proposed a deformable 3D representation based on articulated objects with an algorithm that formulates an implicit representation of a 3D articulated object by taking into account the rigidity transformations of the most relevant parts of the object, which in turn renders the changes related to the pose. Although these methods can learn human models from monocular videos and achieve realistic reconstructions, the slow speed of canonical representation and deformation algorithms results in relatively slow training and rendering speeds.[20,21] By contrast, our method addresses this issue and achieves a realistic reconstruction quality.

2.3 Temporal Field

In recent years, researchers have attempted various ways to accelerate the inference of traditional NeRFs, some of which have achieved real-time rendering performance. Yu et al.[22] used sparse 3D grids and spherical harmonics to represent scenes, enhancing the optimization speed in reconstruction by adjusting through gradients and regularization methods from image calibration. However, these methods[23,24] are only applicable to static scenes and still pose challenges in modeling dynamic scenes. To address this problem, researchers introduced the concept of temporal fields,[11] which are functions that change over time and describe the motion and deformation of objects in the scene. By introducing the temporal dimension, NeRF can model objects in the scene over time, thus capturing the details and subtle changes in dynamic scenes. The key to using temporal fields in our TRes-NeRF architecture for reconstruction is to treat time as an additional input dimension and input it into the neural network along with spatial coordinates for rendering. In this way, the neural network can learn the correlations between time and space and thus infer temporal scene changes.

2.4 Residual Connection

Residual connection[25] is a technique to address the vanishing and exploding gradient problems in neural network training;[26] it was proposed and applied to residual neural networks by Kaiming He et al.[27] It passes the residuals from the intermediate layers to the subsequent layers by expressing the output of each layer as a linear superposition of the input and the input after a single nonlinear transformation. This type of connection allows the network to learn the difference between the input and the target and thus to train and optimize more efficiently. It can significantly enhance the training performance in the modeling process of NeRFs. The recent ReRF[28] is a method to model the residual information between adjacent timestamps in the modeling space

using a compact motion mesh and a residual feature mesh and exploiting the similarity of inter-frame features, thus effectively modeling dynamic scenes. Our TRes-NeRF layers model the residuals of MLP weights, which is different from simple addition of residuals to MLP layer outputs[29,30] and methods that optimize the residuals of model parameters.[31,32] This modeling approach enhances the capability of TRes-NeRF's MLP linear layers in neural field representations, making it more suitable for modeling complex real-world spatiotemporal signals.

2.5 Accelerating Neural Radiance Fields

With the development of neural representations, many methods based on differentiable rendering[1,33] have optimized individual neural representations for each scene. However, this optimization process typically takes several hours on a GPU, resulting in high computational costs. Currently, to optimize training and rendering speeds, some methods, such as TensoRF[34] and NSVF,[35] focus on replacing the architecture in neural representations with more efficient representations for faster training. Recently, Instant-NGP[13] used a smaller MLP structure and achieved faster training using a more efficient hash encoding to replace trigonometric function frequency encoding as the data structure for storing feature grids at different coarse scales. We integrated this multi-resolution hash encoding structure with our TRes-NeRF network to achieve a higher quality and faster speed 3D reconstruction of the human body.

3 Method

Given a monocular video of a dynamic human performer, the goal of our model is to achieve fast human model reconstruction with a low-cost computational resource. In this section, we detail the architecture of our network framework, as shown in Fig. 1. We begin by introducing the foundation of our framework (Sec. 3.1), a linear transformation of the basic temporal field. Subsequently, we describe two main components within our framework: a residual neural field for spatiotemporal signals (Sec. 3.2), which enhances network complexity through the addition of trainable parameters, and an integrated method for accelerated training (Sec. 3.3), which completed the reconstruction in a few minutes.

3.1 Typical Temporal Neural Radiance Field

In the temporal NeRF, a neural network parameterized by Φ_θ[1] is used to encode the scene signal: $\mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^r$. Specifically, the neural network takes a spatiotemporal coordinate pair $(x \in \mathbb{R}^n, t \in \mathbb{R})$ as input and maps it to a scalar field $f \in \mathbb{R}^r$. The architectural design of the neural network enables it to model viewpoint-related effects. In other words, the definition of the temporal NeRF[36] is described as

$$\Phi_\theta(x,t) = \upsilon_n(W_n(h_1 \circ h_2 \circ h_3 \cdots \circ h_{n-1})(x,t) + b_n), \tag{1}$$

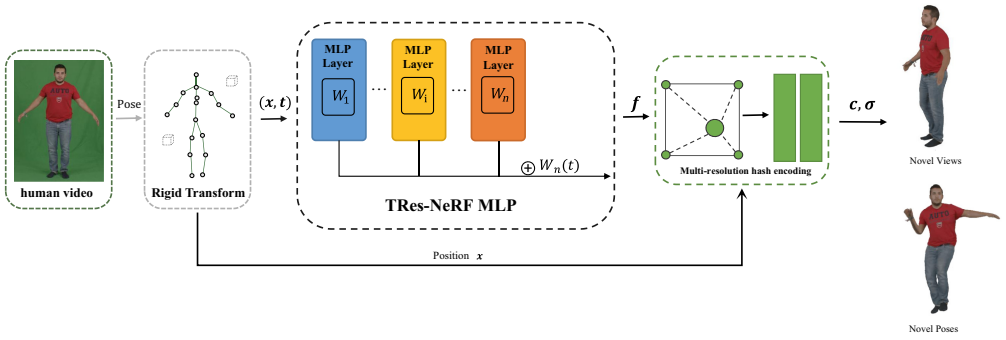

Fig. 1 Overview of TRes-NeRF network structure. For a given sequence of human videos, after applying rigid transformations to the human joints, we utilize an occupancy grid to filter out points in the empty space to reduce the computational load. The remaining points are then extracted by the TRes-NeRF network to extract time-dependent features among the image sequences and are input into a multiresolution hash coding structure along with spatial location features to evaluate their color and density.

$$h_i(x_i, t) = \upsilon_i(W_i x_i + b_i). \tag{2}$$

In this definition, $h_i : \mathbb{R}_i^a \rightarrow \mathbb{R}_i^b$ represents the i'th layer in the neural network's MLP. The encoding of each layer at spatial position $x_i \in \mathbb{R}^{b_i}$ is transformed by a weight matrix $W_i \in \mathbb{R}^{a_i \times b_i}$ that has undergone linear transformation, a bias value $b_i \in \mathbb{R}^{a_i}$, and a nonlinear activation function υ_i. Through the learning and optimization of the neural network Φ_θ, the temporal neural field is capable of learning the spatiotemporal patterns and dynamic regularities within the input data. This enables the modeling and prediction of continuous signals.

3.2 Residual Neural Radiance Field

Traditional MLP structures primarily focus on the fully connected mapping between input features,[37] with the number of parameters in each layer growing exponentially with the increase in input dimensions. This leads to a large scale of the network model and constrains its capability in modeling complex spatiotemporal signals. To overcome this challenge, this study introduced a residual NeRF for spatiotemporal signals (Fig. 1), which enhanced the network's ability to model spatiotemporal signals by inserting a special residual connection into the MLP structure. Specifically, we replaced the linear layer in the MLP with a temporal residual layer, which is defined as follows:

$$h_i(x_i, t) = \upsilon_i((W_i + W_i(t))x_i + b_i). \tag{3}$$

In this structure, $W_i(t) : \mathbb{R} \rightarrow \mathbb{R}^{a_i \times b_i}$ are residual parameters related to the time series. This simple substitution enhances the capacity of the MLP by increasing the trainable temporal parameters instead of directly increasing the number of neurons while maintaining the implicit regularization characteristics of the neural network.

3.2.1 *TRes-NeRF factor decomposition*

After replacing the linear layers of the MLP, if the temporal residual $W_i(t)$ optimization is simply added to the MLP as a trainable weights dictionary, the model will introduce many additional parameters to capture temporal dependencies, which in turn increases the training and inference computational costs. To address this spatiotemporal signal segmentation issue,[38] we decompose this temporal residual and optimize the $\mathbb{R}_i$-dimensional generative set shared by the entire spatiotemporal signal's residual network weights. The optimized definition of the temporal residual is as follows:

$$W_i(t) = \sum_{k=1}^{R_i} g_i(t)[k] \cdot V_i[k]. \tag{4}$$

In this approach, $g(t) \in \mathbb{R}^{\mathbb{R}_i}$ and the generative set $V \in \mathbb{R}^{R_i \times a_i \times b_i}$ are identified as trainable parameters related to temporal features. [] is used to denote the selection of elements. By setting k as a matrix, temporal sequence data can be divided into multiple sub-intervals, and linear interpolation is applied to fill the parameter values between these sub-intervals. This method[39] not only allows the model to utilize the continuity and dynamic characteristics of data more effectively in the temporal dimension but also significantly reduces the total number of trainable parameters.

3.3 Additional Accelerated Rendering Methods

3.3.1 *Input stage*

Due to the presence of numerous open areas or obstructions around the human body in the input monocular video scenes, computing these areas is a waste of resources. Therefore, to save GPU computational resources, we adopt a rigid transformation and empty voxel skipping Scheme 12 at the input stage. This method divides the space around the human body into discrete grid cells by maintaining an occupancy grid. For point samples within unoccupied cells, we directly set their density to zero without querying the hypothetical radiance fields, reducing unnecessary

Table 1 Empty voxel skipping.

	Mean training ↓	Mean rendering ↑
w/o empty voxel skipping	3m16s	7 FPS
w/ empty voxel skipping	**1m38s**	**13 FPS**

We compare training and rendering speeds with and without the use of empty voxel skipping. We report the average training time for 50 epochs in PeopleSnapshot. Training and rendering are significantly faster using the occupied grid jumping scheme. Note: bold values indicates which method has better performance on the data set.

computations. The effect of whether or not to use this empty voxel skipping scheme on rendering speed is shown in Table 1.

3.3.2 *Rendering stage*

During the training stage, we enhanced the network's ability to model spatiotemporal signals using TRes-NeRF building blocks. To further achieve rapid rendering and inference, we adopted a recent Instant-NGP[13] to parameterize scene signals. Using hash tables to store feature grids at different coarse scales, we eliminate hash collisions and reduce computations. Our method utilizes TRes-NeRF to capture temporal features from image sequences during volumetric rendering. These features are then combined with the original positional features to compute the color and density properties of pixels in space. Such integration enables our model to improve the reconstruction quality while achieving faster rendering speeds, thereby optimizing the process of neural rendering.

3.4 Training Loss

We adopted a comprehensive loss function framework to train our model, aimed at accurately capturing the 3D structure of the human body in dynamic scenes. This framework takes into account the accuracy of color prediction, transparency handling, and regularization of the model.

To evaluate the accuracy of color prediction, we calculate the mean squared error (MSE) between the predicted RGB values and the target RGB values, assessing the discrepancy between them as

$$L_{\text{color}} = \text{MSE}(R, R_{oi}), \tag{5}$$

where R symbolizes the predicted RGB values and R_{oi} represents the corresponding true RGB values of the image.

To reduce floating artifacts in space, we apply an MSE[40] loss to the predicted transparency or visibility, which is given as

$$L_{\text{trs}} = \text{MSE}(\alpha, \alpha_{oi}). \tag{6}$$

In our model, we constrain and optimize the processing of neural network outputs for time-involved fine details (tif), thereby enhancing the model's stability and accuracy in handling subtle spatial variations. We apply L2 regularization to the outputs of the tif neural network as

$$L_{L_2} = \|\text{sigmoid}(tif_{op}) \times \beta\|_2^2, \tag{7}$$

where tif_{op} represents the output of the neural network and β is a scaling factor used to regulate the impact of the tif neural network output on subsequent operations.

To further enhance the accuracy and efficiency of the model, we employ a combination of multi-dimensional loss functions, calculated as

$$L = \omega_1 L_{\text{color}} + \omega_2 L_{trs} + \omega_3 L_{L_2}. \tag{8}$$

With this multi-dimensional loss function combination, our model is capable of efficiently handling dynamic human body modeling tasks in complex 3D scenes, improving the rendering quality while maintaining high computational performance.

4 Experiment

To validate the speed and effectiveness of our method, we conducted experiments on dynamic human body reconstruction from monocular videos and compared them with other state-of-the-art methods, demonstrating the efficacy of our method.

4.1 Set the Dataset

4.1.1 *PeopleSnapshot*

This dataset is a widely used benchmark for human monocular video modeling,[33] which includes dynamic videos of humans rotating in front of a camera and provides parameters such as human body masks and keypoints. We evaluated our method on this dataset and conducted a fair comparison with baseline methods.

4.1.2 *NeuMan*

Due to the limitations of the PeopleSnapshot dataset, which provides imperfect posture parameters and only includes simple human rotational movements, there are some constraints in assessing the competitiveness of our method. Therefore, to more accurately evaluate the performance in complex postures, we introduced the NeuMan dataset.[41] It offers a richer variety of human postures and movements across different scenes. For each video sequence, we utilized the Openpose algorithm[42] for 2D keypoint estimation and employed the Segment-Anything method[43] for scene segmentation. In addition, we used the ROMP algorithm[44] to estimate camera parameters and SMPL model parameters, which were then incorporated into the training of our model.

4.2 Baseline

To validate our method, we compared it with the following reconstruction techniques.

4.2.1 *Anim-NeRF*

This baseline uses an MLP-based NeRF for human body reconstruction from videos of individual people.[45] It takes a frame video sequence containing posture changes as input. The SMPL parameters for each frame are first estimated; then sample points along the camera rays in the observation space are rendered. Finally, using posture-guided deformation, these sampling points are transformed back into the canonical space.

4.2.2 *Neural body*

This baseline utilizes a statistical human body model to integrate temporal information.[46] By feeding structured latent codes into the network, which are attached to the human body model, the latent code for each point can be obtained by trilinear interpolation of its neighboring points and mapping the density and color values using the MLP network.

4.2.3 *InstantAvatar*

This baseline takes spatial positions as input and samples along the rays in the located space for each frame.[12] The points are then transformed into normalized space after using the occupancy grid to filter the points in empty voxel space. Finally, the remaining points are deformed into the canonical space using a joint module to evaluate the color and density.

4.3 Compared with SoTA Methods

4.3.1 *Reconstruction quality*

For comparison, we animated and rendered the human model for each frame of pose in the PeopleSnapshot dataset. Table 2 lists the quantitative analysis of the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) across four datasets of the PeopleSnapshot dataset

Table 2 Qualitative comparison with SoTA methods on the PeopleSnapshot[33] dataset.

	Male-3-casual		Male-4-casual		Female-3-casual		Female-4-casual	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Neural Body[46] (14 h)	24.94	0.9428	24.71	0.9496	23.87	0.9504	24.37	0.9451
Anim-NeRF[45] (13 h)	29.37	0.9703	**28.37**	0.9605	**28.91**	**0.9743**	28.90	0.9678
InstantAvatar[12] (1 min)	29.65	0.9730	27.97	0.9649	27.90	0.9722	28.92	0.9692
Ours **(1 min)**	**29.89**	**0.9735**	28.12	**0.9652**	28.42	0.9728	**29.56**	**0.9714**

The PSNR and SSIM metrics of images generated by TRes-NeRF were compared with those produced by three state-of-the-art methods: Neural Body,[46] Anim-NeRF,[45] and InstantAvatar.[12] This comparison evaluated the speed and accuracy of these four methods. TRes-NeRF achieved a training speed comparable to that of InstantAvatar, while obtaining superior reconstruction quality.
Note: bold values indicates which method has better performance on the data set.

for four of the latest representative methods. The experimental results indicate that, compared with the current state-of-the-art InstantAvatar, our proposed method achieves optimal results in terms of PSNR and SSIM. In addition, in comparison with Anim-NeRF and neural body, our method maintains a similar accuracy while enhancing the operational efficiency by nearly 780 times. Overall, our proposed method achieves a state-of-the-art performance.

With the aid of TRes-NeRF, we successfully completed the reconstruction of monocular videos at an exceptionally high rendering speed while maintaining the quality of the reconstruction. The results of the image reconstruction quality are shown in Fig. 2, which indicates that our reconstruction is closer to real images. Our method excels at reconstructing details (as shown in Fig. 2) and is more effective in eliminating artifacts post-reconstruction (as shown in Fig. 3). By contrast, InstantAvatar experiences reconstruction failures in detail-rich areas, such as the face and legs, and exhibits more artifacts. The TRes-NeRF network can more effectively utilize the continuity and dynamic characteristics of data in the time dimension (Sec. 3.1). And this

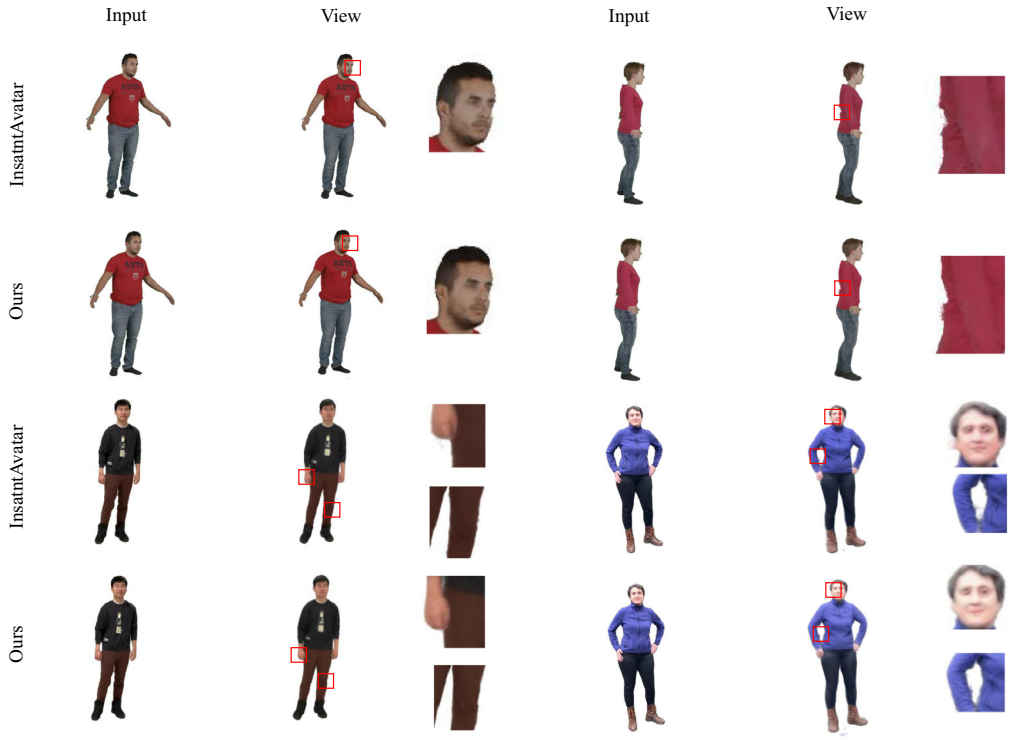

Fig. 2 Results of human body reconstruction on the PeopleSnapshot dataset[33] (above) and the Neuman dataset[41] (below) demonstrate that TRes-NeRF excels in reconstructing details.

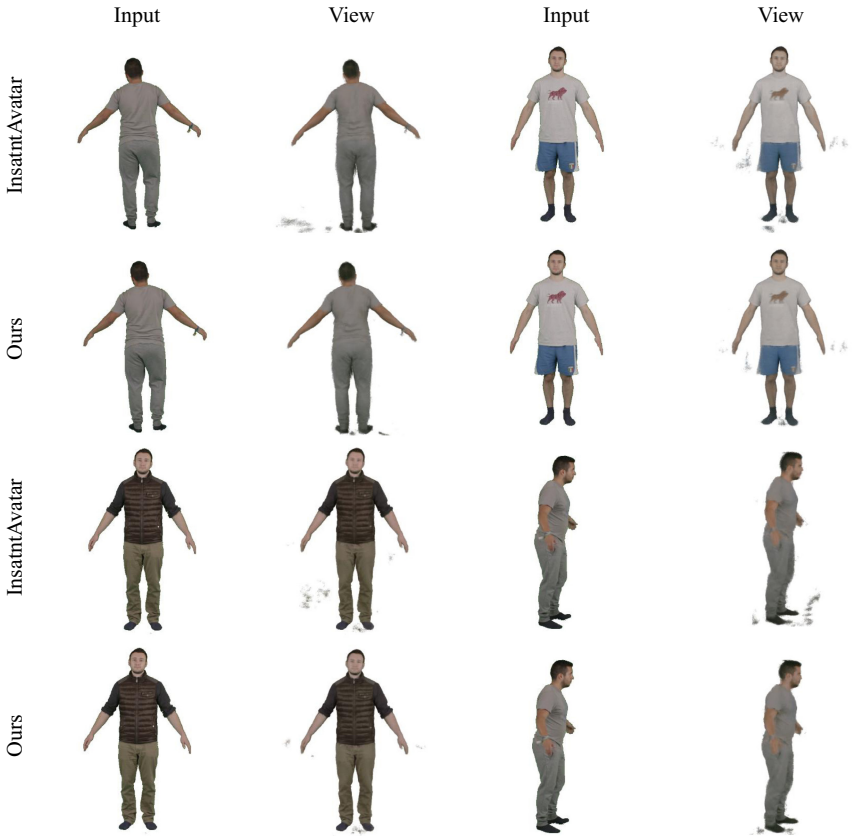

Fig. 3 More challenging human body reconstruction results on the PeopleSnapshot[33] dataset. The reconstruction results of TRes-NeRF and InstantAvatar on challenging human models in the PeopleSnapshot dataset. The images demonstrate that TRes-NeRF is capable of effectively eliminating artifacts in the reconstruction.

multi-dimensional loss function framework can improve the stability and accuracy of the model when dealing with subtle spatial changes, making it better at reconstructing details (Sec. 3.4). Therefore, our method outperforms the baseline.

4.3.2 *Computational resources and speed*

Compared with Anim-NeRF[45] and Neural Body,[46] our method achieves significant optimization in the training time. For instance, while training on an RTX 3090, Anim-NeRF requires 13 h on 2×RTX 3090 to complete the reconstruction task, and Neural Body needs up to 14 h on 4×RTX 2080. By contrast, our method requires only 1 min on a single RTX 3090, as opposed to several hours, and the reconstruction quality is noticeably superior to these two baseline methods (as indicated in Tables 2–4). Similar to InstantAvatar,[12] TRes-NeRF adopts a white space hopping

Table 3 Qualitative results on the Neuman dataset.[41]

	Bike		Citron		Jogging		Lab		Parkinglot		Seattle	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
InstantAvatar[12]	24.18	**0.9498**	24.95	0.9491	23.40	0.9338	27.34	**0.9731**	24.14	0.9520	26.57	**0.9674**
Ours	**24.32**	0.9487	**24.99**	**0.9496**	**23.49**	**0.9339**	**27.54**	0.9697	**24.45**	**0.9528**	**26.66**	0.9659

On the Neuman dataset,[41] a comparison between images generated by TRes-NeRF and InstantAvatar in terms of PSNR and SSIM was conducted. The data indicate that TRes-NeRF achieved the best PSNR metrics, particularly under more complex motion poses, thereby optimizing the reconstruction quality.
Note: bold values indicates which method has better performance on the data set.

Table 4 Qualitative comparison of more challenging data in the PeopleSnapshot dataset.[33]

	Male-3-sport		Male-2-sport		Male-2-casual		Female-4-sport	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
InstantAvatar[12] (1 min)	23.67	0.9216	**24.99**	0.9277	**25.25**	**0.9306**	22.04	0.8930
Ours (1 min)	**23.98**	**0.9281**	24.92	**0.9304**	25.15	0.9270	**22.30**	**0.9021**

A comparison was made between the PSNR and SSIM of images generated by TRes-NeRF and InstantAvatar. The data show that TRes-NeRF exhibits competitive results in challenging datasets, demonstrating its effectiveness and reliability in processing highly complex and dynamic scenes.
Note: bold values indicates which method has better performance on the data set.

scheme as well as multi-resolution hash encoding in the input stage and rendering stage, so the reconstruction index may be stronger. However, TRes-NeRF introduces temporal detail-related information, allowing the network to more accurately capture detailed information in the 3D structure of the human body. This allows our method to require less GPU memory and a smaller MLP structure under the same training time budget, as well as achieve a noticeably better detailed image quality. These advantages make our method more efficient and practical for human body reconstruction tasks and enable training to be completed at a lower cost.

4.3.3 *Performance in new pose synthesis*

To verify the effectiveness of the method proposed in this study in the field of new pose synthesis, we conducted challenging experiments on new pose synthesis using our method on the PeopleSnapshot dataset, as shown in Fig. 4. The results demonstrate that our method performs excellently in handling complex new poses as it is capable of generating high-quality human animation reconstruction results.

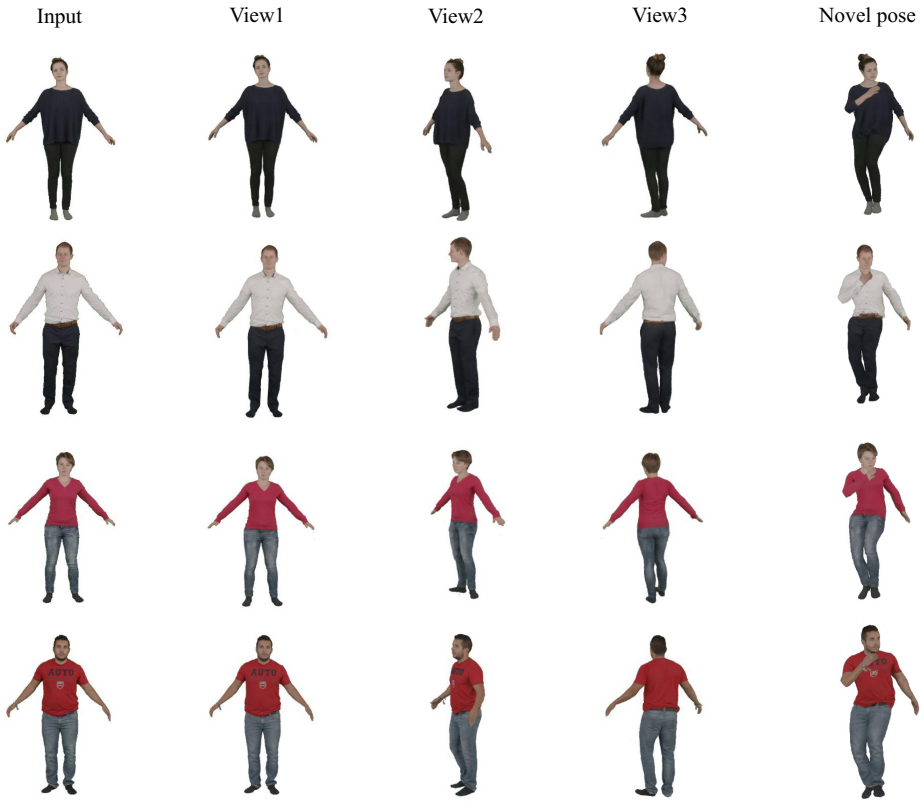

Fig. 4 More human body reconstruction results of TRes-NeRF on the PeopleSnapshot dataset,[33] including the synthesis effects of new poses.

Table 5 Time residual network.

	Mean PSNR↑	Mean SSIM↑	GPU↓
w/o time residual	25.24	0.9540	3.8G
w/ time residual	**25.62**	**0.9586**	**2.4G**

A comparison was made on the Neuman dataset between results with and without the temporal residual network. The inclusion of the temporal residual significantly improved the reconstruction quality while noticeably reducing the GPU memory consumption.
Note: bold values indicates which method has better performance on the data set.

Table 6 TRes-NeRF factor decomposition.

	Mean training↓	Mean rendering↑
w/o TRes factorization	4m24s	1 FPS
w/ TRes factorization	**1m36s**	**13 FPS**

In the PeopleSnapshot dataset, a comparison was made of the training and rendering speeds with and without factor decomposition. After decomposing time-related features, both training and rendering speeds were significantly accelerated.
Note: bold values indicates which method has better performance on the data set.

4.4 Ablation Study

4.4.1 *Temporal residual network*

We investigated the impact of the proposed temporal residual layer on the reconstruction effect, as shown in Table 5. After integrating time-related features into the network, it was observed that the network could learn details better and improve the quality of reconstruction. Moreover, in terms of computational resource consumption, the use of temporal residuals effectively saved GPU memory.

4.4.2 *Decomposition of TRes-NeRF*

We compared the reconstruction speed when only time-related features were added without employing decomposition techniques, as shown in Table 6. The introduction of time-related features brings in more dependencies, and performing decomposition can significantly accelerate both training and rendering speeds.

5 Limitations

Although our method can quickly reconstruct high-quality human models from monocular video at a low hardware cost, it still has some limitations. Our approach can be helpful when the challenge is cost and network capacity. However, when it comes to addressing the lack of unsupervised constraints, our method may not show advantages in challenging scenarios.

6 Conclusion

In this paper, we proposed a method for modeling complex spatio-temporal signals and accelerating rendering, with application to the fast reconstruction of human monocular videos. Our key idea was to introduce temporal residuals into the neural field, which enhanced the ability of the neural network to model spatiotemporal signals by inserting a special residual connection within the MLP structure. This allowed us to increase the network complexity without adding more neurons, while improving the reconstruction quality with smaller MLPs and reducing the GPU memory usage. In addition, to achieve faster rendering speeds, we applied multi-resolution hash coding[13] during the rendering phase, which simultaneously mapped temporal and spatial

features to improve our rendering speed. Compared with baseline methods, our approach achieved faster inference and training times with lower GPU memory requirements, resulting in higher quality reconstructions.

Code and Data Availability

The code used to generate the results and figures is available in a Github repository. It can be accessed via the following link: https://github.com/tianledu/TRes-NeRF.

The data utilized in this study were obtained from [People Snapshot Dataset] and [anuragranj]. Data are available from the authors upon request and with permission from [People Snapshot Dataset] and [anuragranj]. It can be accessed via the following link: https://graphics.tu-bs.de/people-snapshot and https://github.com/apple/ml-neuman.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62361041; 62262040).

References

1. B. Mildenhall et al., "NERF: representing scenes as neural radiance fields for view synthesis," *Commun. ACM* **65**(1), 99–106 (2021).
2. N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.* **35**(6), 1–10 (2016).
3. A. Collet et al., "High-quality streamable free-viewpoint video," *ACM Trans. Graph.* **34**(4), 1–13 (2015).
4. R. L. Cook, "Shade trees," in *Proc. 11th Annu. Conf. Comput. Graph. and Interactive Tech.*, pp. 223–231 (1984).
5. C. R. Qi et al., "Volumetric and multi-view CNNS for object classification on 3D data," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit.*, pp. 5648–5656 (2016).
6. C. Reiser et al., "KiloNeRF: speeding up neural radiance fields with thousands of tiny MLPs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 14335–14345 (2021).
7. S. J. Garbin et al., "FastNeRF: high-fidelity neural rendering at 200fps," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 14346–14355 (2021).
8. A. Yu et al., "PlenOctrees for real-time rendering of neural radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 5752–5761 (2021).
9. C.-Y. Weng et al., "HumanNeRF: free-viewpoint rendering of moving people from monocular video," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 16210–16220 (2022).
10. S.-Y. Su et al., "A-Nerf: articulated neural radiance fields for learning human shape, appearance, and pose," in *Adv. in Neural Inf. Process. Syst.*, Vol. 34, pp. 12278–12291 (2021).
11. V. Sitzmann et al., "Implicit neural representations with periodic activation functions," in *Adv. in Neural Inf. Process. Syst.*, Vol. 33, pp. 7462–7473 (2020).
12. T. Jiang et al., "InstantAvatar: learning avatars from monocular video in 60 seconds," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 16922–16932 (2023).
13. T. Müller et al., "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.* **41**(4), 1–15 (2022).
14. M. Tancik et al., "Fourier features let networks learn high frequency functions in low dimensional domains," in *Adv. Neural Inf. Process. Syst.*, Vol. 33, pp. 7537–7547 (2020).
15. A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, Vol. 30 (2017).
16. B. Jiang et al., "SelfRecon: self reconstruction your digital avatar from monocular video," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 5605–5615 (2022).
17. Z. Zheng et al., "PaMIR: parametric model-conditioned implicit representation for image-based human reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(6), 3170–3184 (2021).
18. S. Peng et al., "Animatable neural radiance fields for human body modeling," arXiv:2105.02872 (2021).
19. A. Noguchi et al., "Neural articulated radiance field," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 5762–5772 (2021).
20. X. Gao et al., "MPS-Nerf: generalizable 3D human rendering from multiview images," *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–12 (2022).
21. H. Choi et al., "MonoNHR: monocular neural human renderer," in *Int. Conf. 3D Vis. (3DV)*, IEEE, pp. 242–251 (2022).

22. S. Fridovich-Keil et al., "Plenoxels: radiance fields without neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 5501–5510 (2022).
23. P. Hedman et al., "Baking neural radiance fields for real-time view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 5875–5884 (2021).
24. M. Niemeyer and A. Geiger, "Giraffe: representing scenes as compositional generative neural feature fields," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 11453–11464 (2021).
25. J. Orbach, "Principles of neurodynamics. Perceptrons and the theory of brain mechanisms," *Arch. Gen. Psychiatry* **7**(3), 218–219 (1962).
26. S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertainty Fuzz. Knowl. Based Syst.* **6**(2), 107–116 (1998).
27. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 770–778 (2016).
28. L. Wang et al., "Neural residual radiance fields for streamably free-viewpoint videos," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 76–87 (2023).
29. R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," arXiv:1505.00387 (2015).
30. K. He et al., "Identity mappings in deep residual networks," *Lect. Notes Comput. Sci.* **9908**, 630–645 (2016).
31. R. Karimi Mahabadi, J. Henderson, and S. Ruder, "Compacter: efficient low-rank hypercomplex adapter layers," in *Adv. Neural Inf. Process. Syst.*, Vol. 34, pp. 1022–1035 (2021).
32. T. Dettmers et al., "QLORA: efficient finetuning of quantized LLMS," arXiv:2305.14314 (2023).
33. T. Alldieck et al., "Detailed human avatars from monocular video," in *Int. Conf. 3D Vis. (3DV)*, IEEE, pp. 98–109 (2018).
34. A. Chen et al., "TensorF: tensorial radiance fields," *Lect. Notes Comput. Sci.* **13692**, 333–350 (2022).
35. L. Liu et al., "Neural sparse voxel fields," in *Adv. Neural Inf. Process. Syst.*, Vol. 33, pp. 15651–15663 (2020).
36. A. Raj, P. Verma, and S. Nagarajan, "Structure-function models of temporal, spatial, and spectral characteristics of non-invasive whole brain functional imaging," *Front. Neurosci.* **16**, 959557 (2022).
37. I. O. Tolstikhin et al., "MLP-mixer: an all-MLP architecture for vision," in *Adv. Neural Inf. Process. Syst.*, Vol. 34, pp. 24261–24272 (2021).
38. R. Shao et al., "Tensor4D: efficient neural 4D decomposition for high-fidelity dynamic reconstruction and rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 16632–16642 (2023).
39. Z. Pan, Y. Hu, and B. Cao, "Construction of smooth daily remote sensing time series data: a higher spatiotemporal resolution perspective," *Open Geosp. Data Softw. Std.* **2**(1), 25 (2017).
40. R. F. Gunst and R. L. Mason, "Biased estimation in regression: an evaluation using mean squared error," *J. Amer. Stat. Assoc.* **72**(359), 616–628 (1977).
41. W. Jiang et al., "NeuMan: neural human radiance field from a single video," *Lect. Notes Comput. Sci.* **13692**, 402–418 (2022).
42. Z. Cao et al., "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 7291–7299 (2017).
43. A. Kirillov et al., "Segment anything," arXiv:2304.02643 (2023).
44. Y. Sun et al., "Trace: 5D temporal regression of avatars with dynamic cameras in 3D environments," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 8856–8866 (2023).
45. J. Chen et al., "Animatable neural radiance fields from monocular RGB videos," arXiv:2106.13629 (2021).
46. S. Peng et al., "Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 9054–9063 (2021).

Tianle Du is currently a junior undergraduate student at Nanchang University and a member of Dr. Wei Li's lab. He has won three competition awards in artificial intelligence. His current research interests include image reconstruction and point cloud alignment.

Jie Wang is currently a third-year undergraduate student at Nanchang University, and her research direction is three-dimensional reconstruction.

Xiaolong Xie is currently a third-year master's student at Nanchang University. He is about to graduate and go to Tongji University to study for a doctoral degree. His research direction is three-dimensional reconstruction of the human body.

Wei Li is a teacher at the Software School of Nanchang University. His research areas include three-dimensional vision, spatial perception, point cloud intelligent analysis, deep learning, virtual augmented reality, etc.

Pengxiang Su is currently a teacher at Nanchang University. His research interests are mainly in the fields of intelligent understanding of human movement and intelligent image analysis.

Jie Liu is currently an associate professor at Nanchang University and serves as a master's tutor, mainly engaged in research on virtual reality.