The Visual Computer (2025) 41:6479–6491
https://doi.org/10.1007/s00371-025-03948-z

RESEARCH

Distilling complementary information from temporal context for enhancing human appearance in human-specific NeRF

Renjie Zhang[1] · Xin Wang[1] · George Baciu[1] · Ping Li[1]

Accepted: 21 April 2025 / Published online: 25 May 2025

Abstract

Reconstructing and animating digital avatars with free views from monocular videos have been an interesting research task in the computer vision field for a long time. Recently, some methods have introduced a novel category method of leveraging the neural radiance field to represent the human body in a canonical space with the help of the SMPL model. With the deformation of the points from an observation space into a canonical space, the human appearance can be learned in various poses and viewpoints. However, previous methods highly rely on pose-dependent representation learned from frame-independent optimization and ignore the temporal contexts across the continuous motion video, causing a bad influence on the dynamic appearance texture generation. To overcome these problems, we propose a novel free-viewpoint rendering framework, TMIHuman. It aims at introducing temporal information into NeRF-based rendering and distilling task-relevant information from complex pixel-wise representations. To be specific, we build a temporal fusion encoder that imports timestamps into the learning of non-rigid deformation and fuses the visual features of other frames into human representation. Then, we propose to disentangle the fused features and extract useful visual cues via mutual information objectives. We have extensively evaluated our method and achieved state-of-the-art performance on different public datasets.

Keywords Novel view synthesis · Neural rendering · Mutual information · Avatar reconstruction

1 Introduction

The free-viewpoint synthesis of human motion is a challenging task in the current computer vision field. This task aims to synthesize all views of a human at any frame in a given pose sequence. It can be the fundamental component of other interactive and immersive applications, such as virtual reality, virtual try-on, and video entertainment, which require this technique to provide controllable viewing experiences. Existing methods either require expensive multi-view image-capturing setups, or ignore the effectiveness of information extraction way from temporal contexts. Hence, it is urgent to have a direct method for providing effective time-aware digital avatar reconstruction and corresponding controllable motion animation from a monocular video.

Traditional methods [22, 36, 40, 44] for free-viewpoint rendering of dynamic scenes typically rely on a multi-camera setup to capture and synthesize new perspectives of a human subject. To mitigate the high costs associated with multi-view image capture systems, Peng et al. [21] introduced a technique capable of producing realistic images from a single-camera video. However, this approach falls short in its ability to animate avatars with new movements. In response to this limitation, recent advancements [19, 35] have been made using neural radiance fields (NeRF) to represent the human. These methods deform the rays within the NeRF, allowing for the transformation of visual pixels in accordance with SMPL body model parameters. Consequently, this enables the rendering of human figures in various poses across different frames. Despite this progress, such techniques heavily depend on pose-specific deformations and risk overfitting to the views present in the training dataset. To address this issue, contemporary research has shifted toward

✉ Ping Li
p.li@polyu.edu.hk

Renjie Zhang
renjie.zhang@connect.polyu.hk

Xin Wang
xin1025.wang@connect.polyu.hk

George Baciu
csgeorge@comp.polyu.edu.hk

[1] Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

learning a deformation field that is more generalizable. Some studies have employed root-finding algorithms to determine the backward correspondence for a learned pose-independent forward blend function within a canonical space. However, the computational intensity of root-finding algorithms has led to the development of newer methods [39] that combine forward and backward mapping networks with consistency constraints. These mapping functions are designed to be both frame-specific and generalizable. Nevertheless, existing approaches tend to concentrate solely on pose-dependent deformations, overlooking the temporal variations and context between frames. Given the rapid changes in human motion, the texture of a person's appearance can vary significantly across different time points, even for similar poses. Current NeRF-based human rendering techniques lack effective training strategies to capture this temporal information.

To address the limitations of pose-dependent modeling and better capture temporal dynamics, we introduce timestamps into the deformation field, making time an explicit variable for frame-wise deformation. We further apply information-theoretic supervision to ensure effective extraction of temporal cues across frames.

Specifically, we then introduce TMIHuman, a novel framework that generates free-viewpoint human views from a single-camera video while enabling text- or speech-driven avatar animation as depicted in Fig. 1. Our method starts with a temporal fusion encoder that integrates a bidirectional deformation module. This module decomposes deformations into two parts: a skeleton-based branch that employs shared weights informed by human anatomy and motion patterns, and a non-rigid branch enhanced by a learnable time-dependent embedding to capture frame-specific appearance details. To achieve free-viewpoint synthesis, we construct a template bank that provides high-quality visual details and serves as a reference for rendering occluded body parts. The pixel features extracted from these templates are merged with those of the current frame and fed into the NeRF. Importantly, to exploit the temporal context beyond mere pose information, a representation disentanglement module is used to separate the fused features into task-relevant and irrelevant components. A mutual information-based objective is introduced to guide this separation and ensure that the most pertinent temporal features are retained. Furthermore, to realize controllable, speech-driven animation, we integrate an automatic speech recognition (ASR) system with a human motion generation module. This combination allows us to generate new sequences of human motion that respond directly to speech inputs in a generative manner. Our method achieves significant performance improvements over the state-of-the-art methods on public datasets. We conduct extensive experiments, and the results validate the effectiveness of the proposed components and the information-theoretic objective.

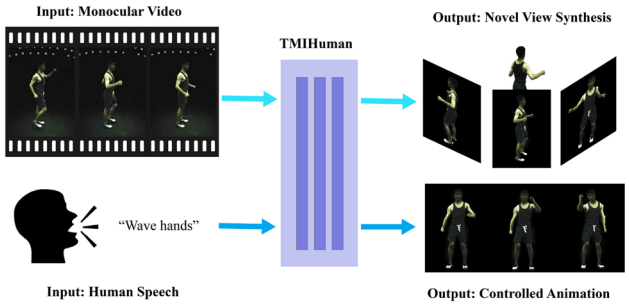


Fig. 1 Overview of TMIHuman. We present TMIHuman, which learns a controllable animatable human neural field from monocular videos. With monocular video input, it can synthesize the novel views of the digital avatar. And with the human speech or text, it can generate corresponding novel human motion animation

Our contributions are summarized as follows:

- We advocate the idea of encoding the temporal information into the learning of deformation for the human-specific NeRF, refining the pose parameters, improving the performance of rendering novel views.
- We propose to disentangle the pixel-wise visual feature in the NeRF into task-relevant and irrelevant components to fully mine the temporal contexts and bring more useful discriminative useful temporal cues for the final human appearance reconstruction and animation.
- Extensive experiments have demonstrated that our method can achieve state-of-the-art results on public benchmark datasets.

2 Related work

2.1 Human performance capture

The free-viewpoint rendering of humans requires exact modeling geometry of human body structure and surface properties like clothing or skin textures. It has been seen as an important task in the computer vision field. Previous works generally leverage multi-view videos or depth cameras and albedo map of surface mesh to do the reconstruction. Traditional methods utilize studio setups to capture these properties physically with multi-view cameras or depth cameras [15, 32]. Carranza et al. [2] proposed the earliest image-based rendering approach with markerless-motion capture, which represents a human body as a parameterized model and uses view-dependent texturing [4] for novel view synthesis of the human. Then, the following works use different techniques to improve the rendering quality. These approaches highly rely on multi-view image capturing studio setups. To tackle this difficulty, some recent works has

introduced human rendering with the single view setup [3, 6]. But they need pre-scanned human body templates, which actually requires expensive equipment. Similarly, monocular RGB-D-based methods [29] adopt the traditional modeling and rendering pipeline for synthesizing different views of humans, and they also suffer from the high prices of RGB-D cameras with depth sensors.

Although the above methods have achieved good rendering performance, they require stable indoor image-capturing environments and high-quality cameras, both of which are uneasily reachable. Our method does not need these settings, and only requires a monocular video with a human performing self-rotating actions.

2.2 Human neural rendering from monocular video

To tackle the difficulty of expensive multi-view setups, many current methods attempt to explore the human reconstruction based on a individual image or a monocular video. Some works [23, 24] proposed to use pixel-aligned implicit function to learn the accurate 3d human surface reconstruction from a single image. When it comes to the dynamic human, with the development of differentiable neural rendering techniques, a lot of methods adapted the neural rendering manner to obtain high-quality rendered results based on different data representations [27]. Among them, NeRF [16] and its extensions [1, 8] has achieved the most attractive performance on novel view synthesis. Liu et al. [13] proposed to utilize a pre-scanned body model and modeled time-dependent dynamic textures via NeRF. Wu et al. [37] and Peng et al. [21] learn a embedding for the human model based on point clouds or reposed mesh vertices. Zhang et al. [41] and Jiang et al. [10] decomposed a scene into the static background and foreground moving human object, and represented them with separated NeRFs thus enabling human pose or scene editing. Xu et al. [38], Su et al. [28] and Noguchi et al. [18] introduce learning implicit geometry via a deformable NeRF. Recently, Weng et al. [35] propose a free-viewpoint rendering pipeline of human in monocular videos. In the pipeline, the rendering of human is done in a canonical space, and the human representation will be transformed form the observation space for each frame by warping. Based on this method, Yu et al. [39] present a framework that achieve generalizable consistent forward and backward deformation to query correspondence appearance features to guide rendering.

These methods focus on the learning of a single frame, just targeting at the geometry transformation and rendering functions. In fact, they ignore the feature-level knowledge hidden in the intermediate representation of NeRF. Our method leverages the information-theoretic objective to extract useful temporal information from a continuous sequence to refine the rendering.

2.3 Animatable digital human

The pursuit of animatable human avatars has been seen a significant task in computer vision field, and it has shown great potential based on recent intricate neural approaches guided by 3D human models. Habermann et al. [7] introduced a weak multi-view supervision to provide a deformable pre-scanned humans. Liu et al. [12] proposes to use a coarse body model as a proxy and learn a deformable radiance field with SMPL as a guidance with incorporating 2D texture maps to render human with novel poses. Peng et al. [20] introduce neural blend weight fields to produce the deformation fields and then generate the animatable human model. Recently, kinds of deformable neural fields [25, 35] show progression from coarse rendering to high-quality results. Then by incorporating human prior-like SMPL models, the 3D text-to-human generation techniques have made a great success. For example, Hong et al. [9] utilize SMPL and Neus [33] to generate 3D human based on the guidance of CLIP. Kolotouros et al. [11] utilizes a pose-conditioned NeRF to obtain density fields. However, these methods neglect the control of the motion of the animatable human model. Recently, Yu et al. [39] proposes to use text prompts to guide the pose-dependent NeRF, obtaining a novel motion sequence of a human avatar which is learned from a monocular video.

However, these methods all ignore the task-relevance of the learned information for the avatar reconstruction. In contrast, we focus on the development of the extraction way for useful information. Our proposed method can effectively obtain the task-relevant information for improving rendering performance of the digital avatar reconstruction. Besides, as far as we know, there is no any work for speech-driven 3D avatar animation generation. To address this problem, in this paper, we propose to use the speech recognition technique to generate a motion sequence from a speech, guiding the deformation of points in the NeRF and the synthesis of a human motion video.

3 Method

We present our TMIHuman for reconstructing an animatable avatar from a monocular video. In this section, first, we review some techniques of the previous human rendering approaches. Second, we introduce the novel NeRF-based human rendering framework which has a novel representation disentanglement module which introduces the mutual information-based objectives for extracting complementary task-relevant information from other frames in the same video. Finally, we give details of the volume rendering and the overall training objective of the framework. The illustration of the pipeline is shown in Fig. 2.

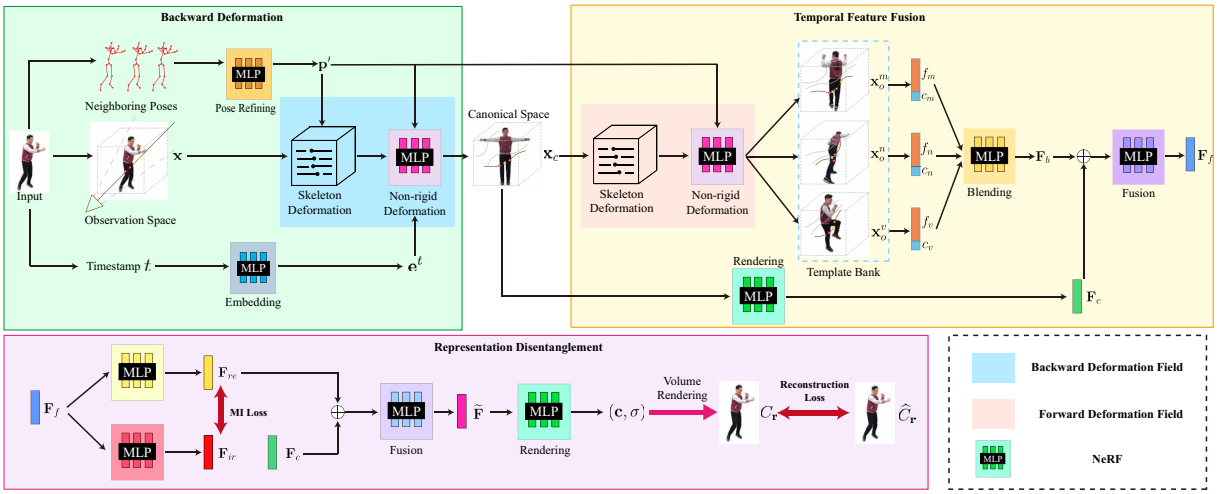


Fig. 2 Illustration of the TMIHuman. Our framework takes a single frame as input for each optimization, and learns a backward deformation field mapping points $\mathbf{x}$ from observation space into canonical space $\mathbf{x}_c$ with the guidance of the temporal information. We utilize the poses of neighboring frames to refine the pose detections, and apply a MLP to learn a time-dependent embedding $\mathbf{e}^t$ for importing the temporal information. With the symmetrical forward deformation, the points can be deformed back to the canonical space. In the temporal feature fusion, inputting the corresponding time-dependent embeddings, $\mathbf{x}_c$ can be deformed to $\mathbf{x}_o^m$, $\mathbf{x}_o^n$ and $\mathbf{x}_o^v$ in different observation spaces that are selected previously from the monocular video to obtain the features f_m, f_n, f_v and the colors c_m, c_n, c_v. Through a blending network, these features are used to obtain the feature $\mathbf{F}_b$. Aggregating $\mathbf{F}_b$ and the learned feature in the canonical space $\mathbf{F}_c$, we can get a fused visual feature $\mathbf{F}_f$ for rendering. Further, in the representation disentanglement, $\mathbf{F}_f$ is disentangled into task-relevant representation $\mathbf{F}_{re}$ and task-irrelevant representation $\mathbf{F}_{ir}$ via two separate MLPs . $\mathbf{F}_{re}$ is used for the final rendering. Besides the direct loss between the rendered ray C_r and the ground truth $\widehat{C}_r$ obtained by the input image, we introduce an additional mutual information objectives to guarantee the useful information distillation

3.1 Preliminaries

Given a series of monocular video frames of a person, denoted as $\mathbb{I} = \{\mathbf{I}^t \in \mathbb{R}^{H\times W\times 3} | t \in \{1, ..., T\}\}$, along with corresponding human poses $\mathbb{P} = \{\mathbf{p}^i = (J^t, \Omega^t) | t \in \{1, ..., T\}\}$and segmentation masks $\mathbb{M} = \{\mathbf{M}^t \in \mathbb{R}^{H\times W} | t \in \{1, ..., T\}\}$, we follow the approach of HumanNeRF [35] to synthesize free-viewpoint rendering results of a human by representing the human via a neural radiance field (NeRF). Here, H and W represents the resolution of the input image. T is the total number of frames in the monocular video. And J^t and Ω^t indicate K standard 3D joint locations and local joint rotations of the human body at frame $\mathbf{I}^t$.

Following HumanNeRF [35], our goal is to render the human appearance in a canonical space. To achieve this, we first deform the points in the observation space into the canonical space. For each input image $\mathbf{I}_t$, we represent the person by a volume V_c in a canonical space, which is warped to an observation space V_o, i.e., $V_o = V_c(D(\mathbf{x}, \mathbf{p}))$, where $V_c = \Phi(\gamma(\mathbf{x}))$ is actually represented as a neural radiance field Φ. Φ takes the position of point $\mathbf{x} \in \mathbb{R}^3$ as input and outputs color $\mathbf{c}$ and density σ. γ is a standard sinusoidal positional encoding function. The warping function is defined as a learnable deformation field D that maps points from the observed space to the canonical, and D consists of two main components:

$$D(\mathbf{x}, \mathbf{p}) = (D_S(\mathbf{x}, P(\mathbf{p})), D_N(D_S(\mathbf{x}, P(\mathbf{p})), \mathbf{p})), \quad (1)$$

where P represents a pose correction function for refining previously estimated 3D human pose, D_S is skeleton-based motion weight deformation, and D_N represents an non-rigid deformation field. Like traditional HumanNeRF methods [35, 39], we use MLPs to represent P and D_N and D_S is set as an inverse linear blend skinning function. Specifically, in D_S, the skinning weight ω can be represented with a single learnable volume $W_c(\mathbf{x}_o)$, where $\mathbf{x}_o$ is the points in the observation space.

But individual images cannot provide enough visual information of different aspects of human appearance for the training of NeRF. To obtain multi-view visual information for helping the rendering, we follow MonoHuman [39] to find correspondences in different time stamps for the current observed frame. We use the technique proposed by MonoHuman [39] to build a template bank $\mathbb{B} = \mathbf{I}_b \mid b \in \{1, ..., T\}$ by picking up several key frame images $\mathbf{I}_b$. The selection of the templates is based on the pose similarity and texture map complementarity. Considering the traditional deformation field D as a backward deformation D_B, we additionally build another forward deformation D_F. D_B can map the points in different observation space $\mathbf{x}_o$ into $\mathbf{x}_c$ in the same

canonical space: $D_B : (\mathbf{x}_o, \mathbf{p}) \rightarrow \mathbf{x}_c$,, and D_F can deform back $\mathbf{x}_c$ to $\mathbf{x}_o$: $D_F : (\mathbf{x}_c, \mathbf{p}) \rightarrow \mathbf{x}_o$. By learning all the above network parameters, the final rendering step can be done as volume rendering in the canonical space.

3.2 Temporal fusion encoder

It is obvious that the visual contexts in different timestamps can not be simply regarded as the other views of the human, because of the temporal influence, such as the different poses, texture changes due to the wind or light conditions. To mine the temporal correspondences fully, we propose a module called temporal fusion encoder (TFE), which combines a learned time-dependent embedding and the bidirectional deformation modules proposed by MonoHuman [39] to get visual features with temporal information.

Temporal pose refining and embedding. For each frame $\mathbf{I}_t$, first, we propose to refine the detected pose $\mathbf{p}$ with temporal contexts. We combine the poses of adjacent frames with the current frame together as the input of a multilayer perceptron (MLP) to compute a new $\mathbf{p}'$:

$$\mathbf{p}' = MLP_p(\mathbf{p}_t, \mathbf{p}_{t-1}, \mathbf{p}_{t+1}). \tag{2}$$

To leverage the temporal information, we need encode the timestamp to obtain a latent code which can be utilized for the following deformation. We use a MLP to learn $\mathbf{e}_t$ as:

$$\mathbf{e}_t = MLP_e(\mathbf{p}', t). \tag{3}$$

Time-dependent deformation. We implement the deformation of the points. As described previously, the deformation D_B and D_F can be both decomposed into skeleton-driven deformation and non-rigid deformation. And the motion weight for the skeleton-driven deformation obtained in D_B and D_F with different ways.

For D_B, the computation of the skeleton-driven deformation D_S^B is linear blend skinning:

$$D_S^B(\mathbf{x}_o, \mathbf{p}') = \sum_{i=1}^{K} \omega_o^i(\mathbf{x}_o)(R_i\mathbf{x}_o + l_i), \tag{4}$$

where ω_o^k is the blend weight of k^{th} bone, and R_i, l_i are the rotation and translation of the bone's coordinates in $\mathbf{p}'$. $\omega_o^k(\mathbf{x}_o)$ can be computed by a set of motion weight volumes ω_c^i which are in canonical space:

$$\omega_o^k(\mathbf{x}_o) = \frac{\omega_c^i(\mathbf{x}_o)(R_i\mathbf{x}_o + l_i)}{\sum_{k=1}^{K} \omega_c^k(\mathbf{x}_o)(R_i\mathbf{x}_o + l_i)}, \tag{5}$$

By solving the parameters of a CNN, we can obtain a volume representation for the skeleton-driven deformation , i.e., $W_c(\mathbf{x}_o) = CNN_{\theta_S}(\mathbf{x}; \mathbf{z})$, for getting ω_o^k, where $\mathbf{z}$ is a constant random code and θ_S represents the parameters of the CNN. For D_F, the motion weight is queried by $\mathbf{x}_o$ as:

$$D_S^F(\mathbf{x}_c, \mathbf{p}') = \sum_{i=1}^{K} \omega_c^i(\mathbf{x}_c)\mathbf{x}_c. \tag{6}$$

As for non-rigid deformation, MLP is always the choice. We build two individual MLPs to compute the backward and forward deformations respectively. Different from conventional methods, we add a latent embedding $\mathbf{e}_t$ encoding temporal information as the additional input of the MLP:

$$D_N(\mathbf{x}, \mathbf{p}) = MLP_N(D_s^F(\mathbf{x}, \mathbf{p}'), \mathbf{p}', \mathbf{e}_t) + D_s^F(\mathbf{x}, \mathbf{p}'). \tag{7}$$

Note that the non-rigid deformation in the forward and backward deformation can be both computed as Eq. (7).

Visual feature fusion. We take three templates $\mathbf{I}_m, \mathbf{I}_n, \mathbf{I}_v$ from $\mathbb{B}$ randomly. We regard the visual features of a same point in observation spaces corresponding to these frames as the correspondence features in canonical space. With deforming $\mathbf{x}_c$ to $\mathbf{x}_o^m, \mathbf{x}_o^n, \mathbf{x}_o^v$ by D_F. For example, we compute the projected points $\mathbf{x}m$ of the image coordinate as:

$$\mathbf{x}_o^m = K_m E_m D_F(\mathbf{x}_c, \mathbf{p}_m'), \tag{8}$$

where K_m and E_m are the intrinsic and extrinsic camera parameters of $\mathbf{I}_m$. Then, we sample to obtain the feature f_m and color c_m indicated by the pixel location $\mathbf{x}_m$. We use the U-Net to extract f_m. After obtaining f_m, f_n, f_v and c_m, c_n, c_v, respectively, we use a MLP to map features to the blend weights. And to import the temporal information into the learning for blending, we explicitly add the learned embedding $\mathbf{e}_t$ corresponding to each template as an extra input of the MLP:

$$\mathbf{w} = MLP_b((f_m; c_m; \mathbf{e}_t^m), (f_n; c_n; \mathbf{e}_t^n), (f_v; c_v; \mathbf{e}_t^v)), \tag{9}$$

where (;) indicates the concatenation. And $\mathbf{w}$ is a vector containing the blend weights for the m^{th}, n^{th}, and v^{th} templates. Then, the blended feature of these templates can be computed as:

$$\mathbf{F}_b = \mathbf{w}((f_m; c_m); (f_n; c_n); (f_v, c_v)), \tag{10}$$

where $\mathbf{F}_b$ is the blended feature to provide guidance for the training of network.

3.3 Representation disentanglement

Traditionally, the blended feature $\mathbf{F}_b$ is concatenated directly to the pixel feature of canonical space as the input of a MLP

in the NeRF Θ to obtain the color $\mathbf{c}(\mathbf{x})$ and density $\sigma(\mathbf{x})$:

$$\mathbf{c}(\mathbf{x}), \sigma(\mathbf{x}) = \Theta(\gamma(\mathbf{x}), \mathbf{F}_b), \tag{11}$$

where γ is a sinusoidal positional encoding. Specifically, regarding $\gamma(\mathbf{x})$ as feature $\mathbf{F}_c$, in the Θ, $\mathbf{F}_c$ and $\mathbf{F}_b$ are fused to $\mathbf{F}_f$ as:

$$\mathbf{F}_f = MLP_f((\mathbf{F}_c; \mathbf{F}_b)). \tag{12}$$

But this direct process inevitably brings a lot of task-irrelevant pixel features in $\mathbf{F}_f$ from other frames. To tackle this difficulty, we propose to utilize the mutual information-based representation learning to mine the useful information for the rendering from the $\mathbf{F}$.

Factorization and rendering. We follow TDMI [5] to build a representation disentanglement module (RD) for disentangling the representation $\mathbf{F}_f$ into two compositions which are relevant part $\mathbf{F}_{re}$ and irrelevant part $\mathbf{F}_{ir}$. $\mathbf{F}_{re}$ contains useful information for the rendering of current frame, while $\mathbf{F}_{ir}$ is a contrastive landmark to help the useful knowledge distillation of $\mathbf{F}_{re}$ from $\mathbf{F}_f$. To factorize $\mathbf{F}_f$ into $\mathbf{F}_{re}$ and $\mathbf{F}_{ir}$, we employ two separate MLPs to process, respectively:

$$\mathbf{F}_{re} = MLP_{re}(\mathbf{F}_f), \mathbf{F}_{ir} = MLP_{ir}(\mathbf{F}_f). \tag{13}$$

The relevant feature composition $\mathbf{F}_f^{re}$ and the original pixel feature $\mathbf{F}_c$ can be combined as the input of a feature aggregation MLP to output a final feature $\widetilde{\mathbf{F}}$. $\widetilde{\mathbf{F}}$ is fed into the rest MLP in the Θ to yield the color and density. Then, the formula Eq. 11 can rewritten as:

$$\mathbf{c}(\mathbf{x}), \sigma(\mathbf{x}) = \Theta(\mathbf{F}_c, \widetilde{\mathbf{F}}). \tag{14}$$

For the final rendering, we use the volume rendering [16]. We compute the color $C_\mathbf{r}$ of a ray $\mathbf{r}$ as:

$$C_\mathbf{r} = \sum_{j=1}^{N} (e^{-\sum_{i=1}^{j-1} \sigma_i \delta_i})(1 - e^{-\sigma_j \delta_j})\mathbf{c}_j, \tag{15}$$

where δ_j and δ_i are the distance between two adjacent samples in the ray. And we apply the stratified sampling approach proposed to sample points from the rays.

Information-theoretic objectives. To obtain rendering-relevant information, we treat the visual features as representations and design a information-theoretic objective to extract useful knowledge from the fusion feature $\mathbf{F}_f$. With the disentangled representations $\mathbf{F}_{re}$ and $\mathbf{F}_{ir}$, it is naive for us to consider minimizing the similarity or the amount of the shared information between $\mathbf{F}_{re}$ and $\mathbf{F}_{ir}$:

$$\mathcal{L}_{diff} = cos[\mathbf{F}_{re}, \mathbf{F}_{ir}] + \mathcal{M}(\mathbf{F}_{re}; \mathbf{F}_{ir}), \tag{16}$$

where $cos[\cdot]$ represents the cosine similarity between two variables and $\mathcal{M}(;)$ is a mutual information term. Through this loss function, with the training, the features encoded by these two representations become different. Considering that the $\mathbf{F}_{re}$ will be used in the subsequent rendering process, the unique features encoded into the representation $\mathbf{F}_{ir}$ can be regarded as task-irrelevant. And to retain the useful information in $\mathbf{F}_{re}$ from $\mathbf{F}_f$, we build another objective as:

$$\mathcal{L}_{reta} = -\mathcal{M}(\mathbf{F}_f; \mathbf{F}_{re}). \tag{17}$$

With minimizing this loss function, the effective visual cues from other timestamps can be encoded into the feature of the final task-relevant representation.

With achieving $\mathcal{L}_{diff}$ and $\mathcal{L}_{reta}$ together, we can maximize the additional task-relevant information. Besides, considering the possible information vanishing in the learning, wo propose to add $\mathcal{L}_{alle}$ to alleviate this bad influence:

$$\mathcal{L}_{alle} = \mathcal{M}(\mathbf{F}_{re}; \widehat{C}_\mathbf{r}|\widetilde{\mathbf{F}}) + \mathcal{M}(\mathbf{F}_c; \widehat{C}_\mathbf{r}|\widetilde{\mathbf{F}}), \tag{18}$$

where $\widehat{C}_\mathbf{r}$ is the ground truth color of the rendered ray $\mathbf{r}$ described in Eq. (15). $\mathcal{M}(\mathbf{F}_{re}; \widehat{C}_\mathbf{r}|\widetilde{\mathbf{F}})$ and $\mathcal{M}(\mathbf{F}_c; \widehat{C}_r|\widetilde{\mathbf{F}})$ represent the dropped task-relevant information loss in $\mathbf{F}_{re}$ and $\mathbf{F}_c$. With minimizing $\mathcal{L}_{alle}$, the useful information can be furthest retained. Following the previous work [14], the mutual information terms can be simplified as:

$$\begin{matrix} \mathcal{M}(\mathbf{F}_{re}; \widehat{C}_\mathbf{r}|\widetilde{\mathbf{F}}) \rightarrow \mathcal{M}(\mathbf{F}_{re}; \widehat{C}_\mathbf{r}) - \mathcal{M}(\mathbf{F}_{re}; \widetilde{\mathbf{F}}) \\ \mathcal{M}(\mathbf{F}_c; \widehat{C}_\mathbf{r}|\widetilde{\mathbf{F}}) \rightarrow \mathcal{M}(\mathbf{F}_c; \widehat{C}_\mathbf{r}) - \mathcal{M}(\mathbf{F}_c; \widetilde{\mathbf{F}}) \end{matrix}. \tag{19}$$

Finally, the overall information-theoretic objective $\mathcal{L}_{info}$ for useful temporal information extraction can be formulated as:

$$\mathcal{L}_{info} = \mathcal{L}_{diff} + \mathcal{L}_{reta} + \mathcal{L}_{alle}. \tag{20}$$

Variational self-distillation (VSD) [31] can be used to achieve the computation of these mutual information terms by minimizing the KL divergence between two variables in each mutual information terms. In addition, in the computation, we use the NeRF Θ to generate rays for each representation $\mathbf{F}_{re}$, $\widetilde{\mathbf{F}}$ and $\mathbf{F}_c$ separately for the computation of losses.

Training objective. For the network optimization, we first compute the rendering loss $\mathcal{L}_r$ in the NeRF to measure the difference between the rendered color $C_\mathbf{r}$ and the ground truth color $\widehat{C}_\mathbf{r}$ of the ray $\mathbf{r}$:

$$\mathcal{L}_r = \sum_{\mathbf{r} \in \mathbb{Y}} (||C_\mathbf{r} - \widehat{C}_\mathbf{r}||_2^2), \tag{21}$$

where $\mathbb{Y}$ represents all query rays for the rendering. And following HumanNeRF [35], we employ the LPIPS [42] loss

function for improving the reconstruction performance and robustness of the image synthesis. Considering the consistency between the forward and backward deformation, we add the constraint for regularization of the forward and backward deformation. Inspired by MonoHuman [39], we build a consistency loss $\mathcal{L}_c$ as:

$$\mathcal{L}_c = ||(\mathbf{x}_o, D_F(D_B(\mathbf{x}_o, \mathbf{p}'), \mathbf{p}'))||_2^2. \tag{22}$$

Then combining $\mathcal{L}_c$ with the information-theoretic objective described in Sect. 3.3, the overall objective $\mathcal{L}$ of our framework can be formulated as:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_c + \mathcal{L}_{lpips} + \mathcal{L}_{info}. \tag{23}$$

We minimize $\mathcal{L}$ during the training.

4 Experimental results

4.1 Dataset and evaluation metrics

Our method is evaluated on the ZJU-MoCap dataset [21] and the HumanNeRF-data dataset [43]. Additionally, we collect in-the-wild videos from the Internet to validate the generalizability of our method. For the ZJU-MoCap dataset, we select six subjects (377, 386, 387, 392, 393, and 394) as our data sets. Frames from camera 1 are used as the training set, while images from the other 22 cameras form the test set. For the HumanNeRF-data dataset, we select three subjects (cjm, batman, law) as the rendering objects. This dataset includes six cameras capturing different views. We use the view from camera 1 for training and the others for evaluation. Both datasets are carefully collected in indoor environments, and all subjects provide annotations. We adopt the peak signal-to-noise ratio (PSNR) [34], structural similarity index (SSIM) [34], and learned perceptual image patch similarity (LPIPS) [42] as evaluation metrics. While PSNR favors smooth results but may neglect the visual quality of the rendered image, LPIPS measures the perceived distance between the synthesized image and the ground truth image. Additionally, we provide visual results of the 3D human reconstruction.

4.2 Implementation details

We implement our framework on the PyTorch platform. We use Nvidia GeForce RTX^{TM} 3090 GPU with 24GB memory to train and test the network. We train our models using Adam optimizer with a learning rate $5e^{-4}$ for the NeRF Θ, and $5e^{-5}$ for all the other network components. We use patch-based ray sampling for the calculation of LPIPS. And we sample 2 patches of rays with size 20×20. One hundred and twenty-eight samples are rendered per ray. For the selection of the template bank, we follow the keyframe selection of MonoHuman [39] according to the pose similarity and texture map complementarity. For speech-driven motion generation, we use the speech-to-text API of google [26] to translate the speech into text prompts, and then utilize the MDM [30] model to generate the SMPL parameters of a novel motion sequence.

4.3 Comparison

The main target of our method is the rendering of the human in a monocular video. Therefore, we only provide the comparisons between our method with neural body [21], HumanNeRF [35] and MonoHuman [39] on ZJU-MoCap datasets in terms of qualitative results and quantitative results. And we provide additional visual results in HumanNeRF-data dataset. Table 1 shows the results of novel view synthesis on the ZJU-MoCap dataset. Our method clearly outperforms the others for almost all subjects, achieving improvements across all evaluation metrics. This demonstrates that our method is capable of generating views that are closest to the real images in various aspects. By introducing temporal information into the non-rigid deformation and maximizing the useful information from template images, the rendering quality is significantly enhanced. We also provide qualitative result comparisons on the ZJU-MoCap dataset in Fig. 3. For an individual frame, we apply MonoHuman [39] and our method to synthesize views from different directions. It is evident that our method exhibits high-fidelity details such as textures and geometries, while MonoHuman produces blurred or faulty results. Additionally, we present novel view synthesis result comparisons on the HumanNeRF-data dataset in Fig. 4. For the challenging cases in HumanNeRF-data, our method more accurately renders visual details like textures. Moreover, we provide the controllable animation synthesis ability evaluation of our method by giving the qualitative comparison between our method and MonoHuman. Here, we give the order of "jump" via speech. The results are shown in Fig. 5. It can be seen that MonoHuman deforms the human representation wrongly and generates multiple artifacts, while our method can render more realistic results.

4.4 Ablation study

We conduct ablation study on ZJU-MoCap dataset to evaluate the proposed components and objectives.

Time-dependent embedding. To validate the effectiveness of our time-dependent embedding, we conduct different experiments where the framework is with or without the embedding. As shown in Table 2, we conduct several exper-

Table 1 Novel view synthesis quantitative comparison on ZJU-MoCap dataset

	Subject **377**			Subject **386**			Subject **387**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [21]	29.29	0.9693	39.40	30.71	0.9661	45.89	26.36	0.9520	62.21
HumanNeRF [35]	30.41	0.9743	24.06	33.20	0.9752	28.99	28.18	0.9632	35.58
MonoHuman [39]	30.77	**0.9787**	21.67	32.97	0.9733	32.7	27.93	0.9633	33.45
Ours	**31.01**	0.9801	**20.22**	**33.74**	**0.9764**	**28.43**	**28.86**	**0.9709**	**32.26**
	Subject **392**			Subject **393**			Subject **394**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [21]	28.97	0.9615	57.03	27.82	0.9577	59.24	28.09	0.9557	59.66
HumanNeRF [35]	31.04	0.9705	32.12	28.31	0.9603	36.72	30.31	0.9642	32.89
MonoHuman [39]	31.24	0.9715	31.04	28.46	0.9622	34.24	28.94	0.9612	35.90
Ours	**31.77**	**0.9757**	**29.84**	**29.63**	**0.9698**	**32.97**	**29.92**	**0.9665**	**32.53**

We bold values cells that have the best metric value. LPIPS* = LPIPS $\times 10^3$

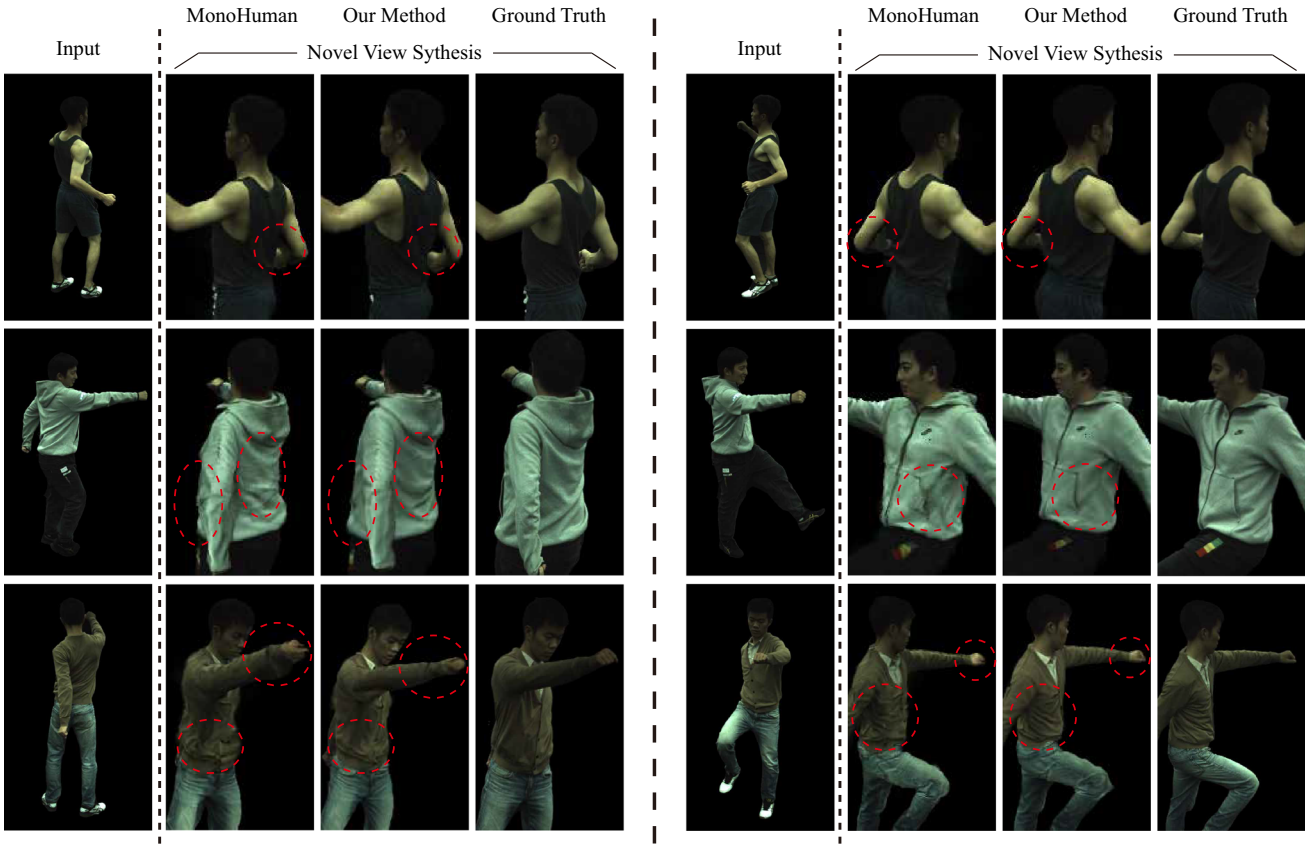


Fig. 3 Qualitative comparison on ZJU-MoCap dataset. The area circled by red dash lines indicates the part where we render better photo-realistic results compared with MonoHuman [39]. The comparison between our results and MonoHuman demonstrates the effectiveness of our proposed method for helping human-specific rendering

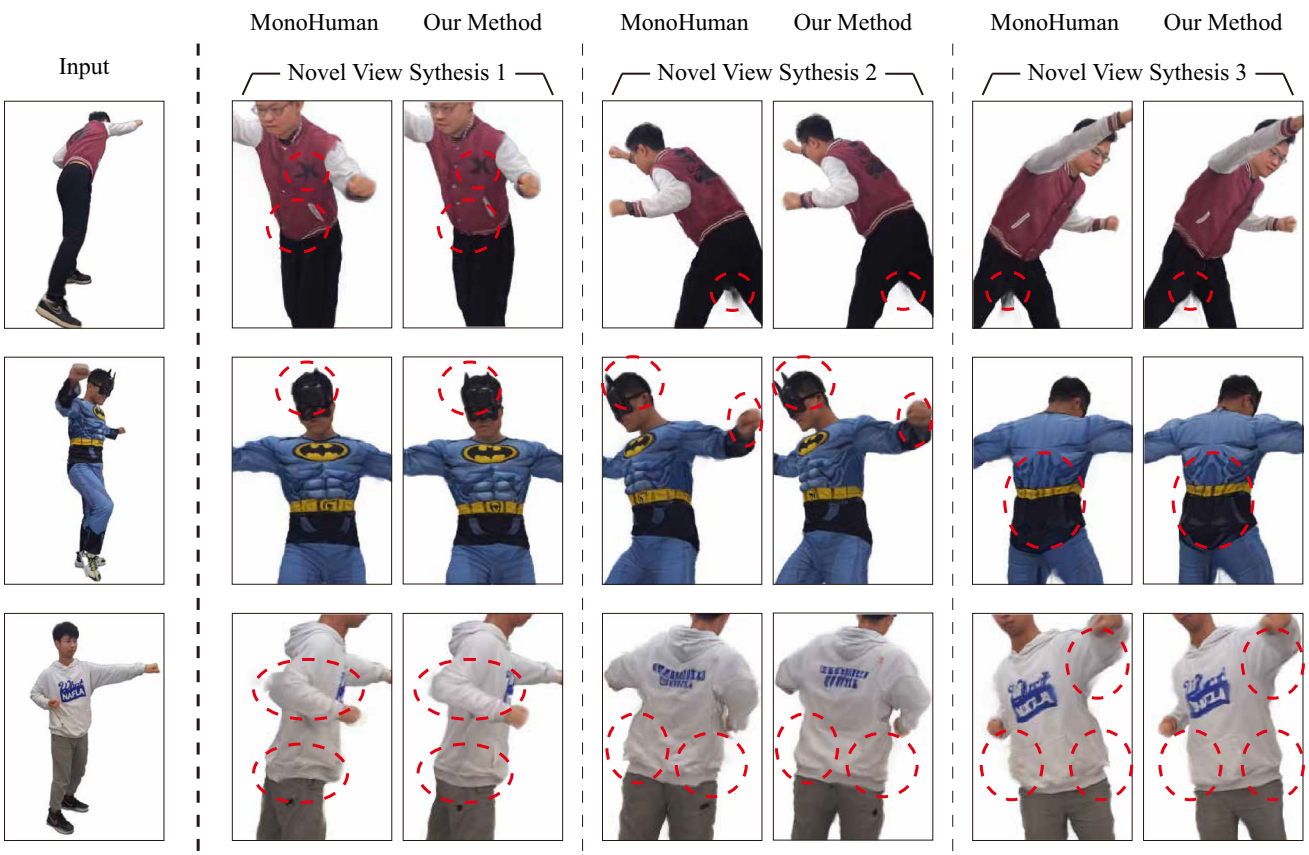


Fig. 4 Qualitative comparison on HumanNeRF-data dataset. The area circled by red dash lines indicates the part where we render better photo-realistic results compared with MonoHuman [39]. The clearly better performance of our method demonstrates the effectiveness of our framework

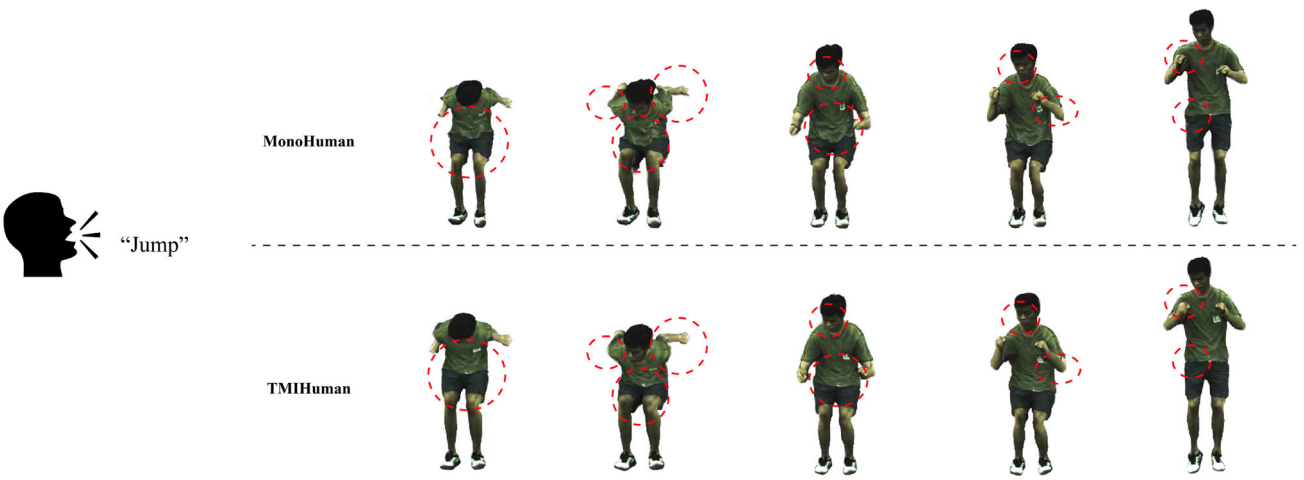


Fig. 5 Qualitative comparison on challenge poses controlled by human speech. We evaluate our method driven by challenge pose sequence generated by MDM model

Table 2 Ablation study for the usage of the time-dependent embedding on ZJU-MoCap

Method	PSNR ↑	SSIM ↑	LPIPS* ↓
HumanNeRF [35]	30.24	0.9679	31.73
Ours-a (w/o $\mathbf{e}^t$ in non-rigid field)	30.45	0.9691	31.52
Ours-b (w/o $\mathbf{e}^t$ in NeRF)	30.52	0.9698	30.31
Ours-f (full model)	**30.82**	**0.9732**	**29.38**

We compute averages over 6 sequences. We bold cells with best metric values. LPIPS* = LPIPS $\times 10^3$

Table 3 Ablation study for network components on ZJU-MoCap. We compute averages over six sequences

Method	TFE	RD	PSNR ↑	SSIM ↑	LPIPS* ↓
HumanNeRF [35]			30.24	0.9679	31.73
(a)	✓		30.49	0.9698	31.00
(b)		✓	30.47	0.9692	30.62
(c)	✓	✓	**30.82**	**0.9732**	**29.38**

We bold cells with best metric values. LPIPS* = LPIPS $\times 10^3$

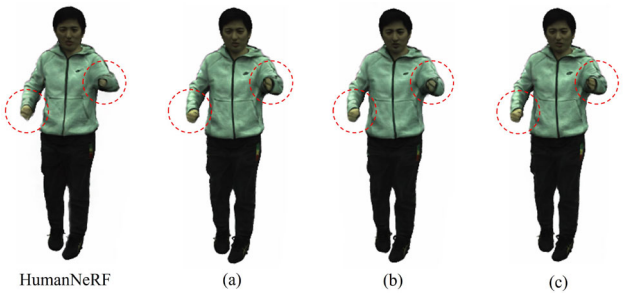


Fig. 6 Qualitative comparison for network components on ZJU-MoCap. The area circled by red dash lines indicates the part where the models with the proposed components achieve better results

iments with inputting the embedding into the non-rigid deformation field, the NeRF separately and both, respectively. The results of ours-a and ours-b are a little higher than the baseline (HumanNeRF), which demonstrates that the temporal information can provide meaningful help for the deformation of rays and the final NeRF-based rendering. And the full body achieves the best rendering performance which validates the effectiveness of the designed framework.

TFE and RD. To investigate the effectiveness of the proposed temporal fusion encoder and the representation disentanglement, we modify the framework with the settings of TFE and RD: (a) only with TFE, (b) only with RD (d) and the full body with both two components. The results are shown in Table 3. From Table 3(a)–(b), we can see that utilizing the TFE and RD can help the model outperforms slightly the baseline HumanNeRF, respectively, which validates the significance of importing temporal information and the novel supervision at feature-level correspondingly. And the result of (c) demonstrates that these two components can bring positive influence to each other, improving the rendering performance eventually. We provide visual results of all experimental settings about HumanNeRF and (a), (b) and (c) in Fig. 6. It is easily to tell that the models with TFE or RD in (a) and (b) can synthesize better details for the human hands compared with the HumanNeRF [35], but still remains distortion and blurs. The model with TFE and RD in (c) can render more photo-realistic textures. These results validate that each proposed component works for improving rendering performance and their collaboration can lead to better outcomes.

Information-theoretic objective. Table 4 illustrates the influence of the proposed objectives to the rendering performance. We conduct experiments with different settings of objective $\mathcal{L}_{info}$.: (a) setting without $\mathcal{L}_{info}$, (b) only applying $\mathcal{L}_{diff}$, (c) applying $\mathcal{L}_{diff}$ and $\mathcal{L}_{reta}$ and (d) the full model with all objectives. From the result of Table 4(a), without the overall information-theoretic objective, the result is slightly better than the HumanNeRF. But compared with the result of Table 3(a), which set the framework without the whole RD component, the result is even worse. It means the simply application of the RD structure into a network cannot bring meaningful help for the performance improvement. The results of Table 4(b) is almost same with the baseline model Table 4(a), which means that only making the

Table 4 Ablation study for loss functions on ZJU-MoCap

Method	$\mathcal{L}_{diff}$	$\mathcal{L}_{reta}$	$\mathcal{L}_{alle}$	PSNR ↑	SSIM ↑	LPIPS* ↓
HumanNeRF [35]				30.24	0.9679	31.73
(a)				30.28	0.9685	31.11
(b)	✓			30.39	0.9694	30.78
(c)	✓	✓		30.44	0.9701	30.21
(d)	✓	✓	✓	**30.82**	**0.9732**	**29.38**

We compute averages over 6 sequences. We bold cells with best metric values. LPIPS* = LPIPS $\times 10^3$

Table 5 Few-shot generalization comparison for novel view synthesis on the ZJU-MoCap dataset

Settings	Method	PSNR ↑	SSIM ↑	LPIPS ↓
5-shot	ActorsNeRF	27.52	0.9566	44.94
	TMIHuman	**27.59**	**0.9572**	**44.33**
10-shot	ActorsNeRF	27.76	0.9590	39.79
	TMIHuman	**27.87**	**0.9592**	**39.42**
30-shot	ActorsNeRF	28.08	0.9600	36.85
	TMIHuman	**28.53**	**0.9613**	**36.01**
100-shot	ActorsNeRF	28.1	0.9602	36.58
	TMIHuman	**28.67**	**0.9615**	**35.87**
300-shot	ActorsNeRF	28.06	0.9599	36.57
	TMIHuman	**28.64**	**0.9614**	**35.46**

We compute averages over 3 sequences. We bold cells with best metric values

two representation different is not meaningful for the useful information extraction. The improvements of Table 4(c)–(d) demonstrate that the $\mathcal{L}_{reta}$ amd $\mathcal{L}_{alle}$ are useful to facilitate the learning of discriminative task-relevant visual details. In particular, $\mathcal{L}_{alle}$ can significantly improve the rendering quality, which evidence the importance of alleviating the information vanishing in the learning.

Low FPS Frames. To evaluate the few-shot generalization ability of our model, we follow the experimental settings of ActorsNeRF [17] to conduct few-shot novel view synthesis experiments. We compute average over 3 sequences (Subjects 387, 393 and 394) of ZJU-Mocap dataset, and the results are shown in the Table 5. It is easy to see that our method can obtain better results because our method leverages the information-theoretic objectives to extract rendering-relevant information from the temporal contexts, obtaining useful knowledge and helping the final rendering.

5 Conclusion

We propose a novel framework TMIHuman which aims at synthesizing novel views of dynamic humans in a monocular video, and outputting video with novel controllable human pose sequence. We propose a novel temporal fusion encoder for encoding temporal information into the deformation field, making the deformation of the points acquiring more guidance from other frames. In this encoder, we first refine the human pose with temporal contexts from detected poses information of adjacent frames. Then, we use a learnable embedding which encodes temporal information from time stamps as the input of the deformation from observation space, providing specific temporal and pose information for each frame. Furthermore, we build a representation disentanglement module and leverage mutual information objectives to extract useful information from the correspondence features of image templates. We generate SMPL pose sequences based on human speech and animate the learned avatar. The high-quality results have shown that our framework is able to generate photo-realistic novel views of human. In addition, the recent Gaussian splatting techniques can be regarded as the better alternatives for the NeRF, and our proposed methods can be utilized in it. The accurate and fine-grained rendering based on designed Gaussian splatting framework with our proposed mechanism might be the most significant future research direction.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00371-025-03948-z.

Acknowledgements The authors would like to thank the editors and anonymous reviewers for their insightful comments and suggestions. This work was supported by The Hong Kong Polytechnic University under Grants P0044520, P0048387, P0050657, and P0049586.

Funding Open access funding provided by The Hong Kong Polytechnic University

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

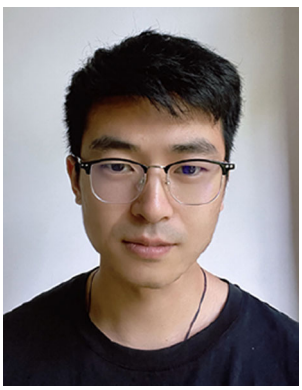
References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields. In: IEEE International Conference on Computer Vision, pp. 5835–5844 (2021)
2. Carranza, J., Theobalt, C., Magnor, M.A., Seidel, H.P.: Free-viewpoint video of human actors. ACM Trans. Graph. **22**(3), 569–577 (2003)
3. Chen, X., Pang, A., Yang, W., Ma, Y., Xu, L., Yu, J.: SportsCap: monocular 3d human motion capture and fine-grained understanding in challenging sports videos. Int. J. Comput. Vis. **129**, 2846–2864 (2021)
4. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: a hybrid geometry-and image-based approach. In: SIGGRAPH, pp. 11–20 (1996)

5. Feng, R., Gao, Y., Ma, X., Tse, T.H.E., Chang, H.J.: Mutual information-based temporal difference learning for human pose estimation in video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 17,131–17,141 (2023)
6. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: LiveCap: Real-time human performance capture from monocular video. ACM Trans. Graph. 38(2), pp. 1–17, 14 (2019)
7. Habermann, M., Xu, W., Zollhofer, M., Pons-Moll, G., Theobalt, C.: DeepCap: monocular human performance capture using weak supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
8. Hedman, P., Srinivasan, P.P., Mildenhall, B., Barron, J.T., Debevec, P.: Baking neural radiance fields for real-time view synthesis. In: IEEE International Conference on Computer Vision, pp. 5855–5864 (2021)
9. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: AvatarCLIP: zero-shot text-driven generation and animation of 3d avatars. ACM Trans. Graph. **41**(4), pp. 1–19, 161 (2022)
10. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: NeuMan: neural human radiance field from a single video. In: European Conference on Computer Vision, pp. 402–418. Springer (2022)
11. Kolotouros, N., Alldieck, T., Zanfir, A., Bazavan, E.G., Fieraru, M., Sminchisescu, C.: DreamHuman: animatable 3d avatars from text. In: Advances on Neural Information Processing Systems (2023)
12. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: neural free-view synthesis of human actors with pose control. ACM Trans. Graph. **40**(6), 1–16, 219 (2021)
13. Liu, L., Xu, W., Habermann, M., Zollhöfer, M., Bernard, F., Kim, H., Wang, W., Theobalt, C.: Learning dynamic textures for neural rendering of human actors. IEEE Tran. Vis. Comput. Graph. **27**, 4009–4022 (2021)
14. Liu, Z., Feng, R., Chen, H., Wu, S., Gao, Y., Gao, Y., Wang, X.: Temporal feature alignment and mutual information maximization for video-based human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10,996–11,006 (2022)
15. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. In: SIGGRAPH, pp. 369–374 (2000)
16. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**, 99–106 (2022)
17. Mu, J., Sang, S., Vasconcelos, N., Wang, X.: ActorsNeRf: animatable few-shot human rendering with generalizable nerfs. In: IEEE International Conference on Computer Vision, pp. 18,345–18,355 (2023)
18. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: IEEE International Conference on Computer Vision, pp. 5742–5752 (2021)
19. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: IEEE International Conference on Computer Vision, pp. 14,314–14,323 (2021)
20. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: IEEE International Conference on Computer Vision, pp. 14,314–14,323 (2021)
21. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural Body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9050–9059 (2021)
22. Rabich, S., Stotko, P., Klein, R.: Fpo++: efficient encoding and rendering of dynamic neural radiance fields by analyzing and enhancing Fourier plenoctrees. Vis. Comput. **40**(7), 4777–4788 (2024)
23. Saito, S., Huang, Z., Natsume, R., Morishima, S., Li, H., Kanazawa, A.: PIFu: pixel-aligned implicit function for high-resolution clothed human digitization. In: IEEE International Conference on Computer Vision, pp. 2304–2314 (2019)
24. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 81–90 (2020)
25. Saito, S., Yang, J., Ma, Q., Black, M.J.: SCANimate: weakly supervised learning of skinned clothed avatar networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2886–2897 (2021)
26. Schuster, M.: Speech recognition for mobile devices at google. In: Pacific Rim International Conference on Artificial Intelligence, pp. 8–10 (2010)
27. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhöfer, M.: DeepVoxels: learning persistent 3d feature embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2432–2441 (2019)
28. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-NeRF: articulated neural radiance fields for learning human shape, appearance, and pose. In: Advances in Neural Information Processing Systems pp. 12,278–12,291 (2021)
29. Su, Z., Xu, L., Zheng, Z., Yu, T., Liu, Y., Fang, L.: RobustFusion: human volumetric capture with data-driven visual cues using a RGBD camera. In: European Conference on Computer Vision, pp. 246–264 (2020)
30. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: International Conference on Learning Representations (2022)
31. Tian, X., Zhang, Z., Lin, S., Qu, Y., Xie, Y., Ma, L.: Farewell to mutual information: Variational distillation for cross-modal person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1522–1531 (2021)
32. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: ACM Trans. Graph. 27(3), pp. 1–9 (2008)
33. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: NeuS: learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: Advances in Neural Information Processing Systems pp. 27,171–27,183 (2021)
34. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing pp. 600–612 (2004)
35. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: free-viewpoint rendering of moving people from monocular video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 16,189–16,199 (2022)
36. Wirth, T., Rak, A., von Buelow, M., Knauthe, V., Kuijper, A., Fellner, D.W.: NeRF-FF: a plug-in method to mitigate defocus blur for runtime optimized neural radiance fields. Vis. Comput. **40**(7), 5043–5055 (2024)
37. Wu, M., Wang, Y., Hu, Q., Yu, J.: Multi-view neural human rendering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1679–1688 (2020)
38. Xu, H., Alldieck, T., Sminchisescu, C.: H-NeRF: neural radiance fields for rendering and temporal reconstruction of humans in motion. In: Advances in Neural Information Processing Systems pp. 14,955–14,966 (2021)
39. Yu, Z., Cheng, W., Liu, X., Wu, W., Lin, K.Y.: MonoHuman: animatable human neural field from monocular video. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 16,943–16,953 (2023)

40. Yuan, J., Fan, M., Liu, Z., Han, T., Kuang, Z., Pan, C., Ding, J.: Collaborative neural radiance fields for novel view synthesis. Vis. Comput. 41(2), pp. 991–1006 (2025)
41. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Trans. Graph. 40(4), pp. 1–18, 149 (2021)
42. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
43. Zhao, F., Yang, W., Zhang, J., Lin, P., Zhang, Y., Yu, J., Xu, L.: HumanNeRF: efficiently generated human radiance field from sparse inputs. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7743–7753 (2022)
44. Zhu, L., Zhou, H., Wu, S., Cheng, T., Sun, H.: Polynomial for real-time rendering of neural radiance fields. Vis. Comput. pp. 1–14 (2024)

Renjie Zhang received the BEng degree in software engineering from the Sun Yat-sen University, Guangzhou, China, in 2019. He is currently pursuing the PhD degree in computing with The Hong Kong Polytechnic University, Hong Kong. His current research interests include human skeleton establishment, neural architecture search, pose estimation, deep learning, and 3D human reconstruction.



Xin Wang received the BEng degree in computer science and technology from the Dalian University of Technology, Dalian, China, in 2017. He is currently pursuing the Ph.D. degree in computing with The Hong Kong Polytechnic University, Hong Kong. His current research interests include image synthesis, deep learning, diffusion models, and computer graphics.

George Baciu received the PhD degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1992. He was a Professor with the Department of Computing (COMP), The Hong Kong Polytechnic University (PolyU), Hong Kong. He was also a Member of the Waterloo Computer Graphics Laboratory and the Pattern Analysis and Machine Intelligence Laboratory. He is the founding Director of the GAME Laboratory, The Hong Kong University of Science and Technology, Hong Kong, in 1993, and the Graphics and Multimedia Applications Laboratory, COMP, PolyU, in 2000. He has authored or coauthored extensively in computer graphics, image processing, and VR journals and conferences, and was as Chair of many international conferences. His current research interests include information visualization, virtual reality and computer graphics, with applications to cognitive digital agents, digital twins, motion synthesis, animation, collision detection, geometric modeling, and image analysis.

Ping Li received the PhD degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Assistant Professor with the Department of Computing and an Assistant Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. He has published over 260 top-tier scholarly research articles, pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the *ACM TechNews*, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts to research fields, addressing societal needs and contributed tremendously to the people concerned. His current research interests include image/video stylization, colorization, artistic rendering and synthesis, computational art, and creative media.