HiAvatar: High-Fidelity Reconstruction of Avatars by Spatial and Temporal Enhancement

Xianyong Fang
Anhui University
Hefei, China
fangxianyong@ahu.edu.cn

Zongxin Shang
Anhui University
Hefei, China
e23201114@stu.ahu.edu.cn

Jiarui Li
Anhui University
Hefei, China
e23201123@stu.ahu.edu.cn

Renlong Dai
Anhui University
Hefei, China
e23301311@stu.ahu.edu.cn

Linbo Wang
Anhui University
Hefei, China
wanglb@ahu.edu.cn

Zhengyi Liu
Anhui University
Hefei, China
liuzywen@ahu.edu.cn

Abstract

This paper presents HiAvatar, a high-fidelity human reconstruction system based on 3D Gaussian Splatting. Existing methods perform well for trained views and seen poses but often suffer from geometric errors and texture blurriness under novel or unseen poses. Those shortcomings are caused by limited motion constraints and limited Gaussian feature abstraction abilities. To address those limitations, this study focuses on exploring motion information in both temporal and spatial domains for rich motion constraints and augmented Gaussian features, respectively. In the temporal domain, HiAvatar adopts dual-dimensional motion modeling: It processes 2D optical flow to improve current-view accuracy, while estimating vertex-level surface motion via a 3D parametric model to ensure the correctness of other views. In the spatial domain, to mitigate the limitation that discrete Gaussian point clouds rely on local neighborhoods, a global feature supplementation mechanism is introduced, along with a pose-adaptive mixture-of-experts strategy. This avoids the regression tendency of a single network toward mean representation and enhances generalization to unseen poses. Experimental results demonstrate that HiAvatar effectively reduces geometric errors and texture distortions in dynamic avatar reconstruction.

***Keywords:** Avatar reconstruction, Mixture of Experts, Spatial Feature, Temporal Feature.*

1. Introduction

Three-dimensional (3D) human avatar reconstruction has wide applications (e.g., film production, simulation) and has advanced rapidly in recent years [33]. Among related tasks, dynamically realistic avatar reconstruction from monocular videos [7, 10] is more practical than static reconstruction from single images [28, 38] and requires simpler setups than multi-view video-based methods [23]. Therefore, it is our primary focus here. However, existing monocular-video-based methods (see Fig. 1(c)-(f)) still fail to recover photorealistic surfaces in untrained frames, where geometric errors, texture misalignment, and blurriness often appear.

Further analysis identifies two key limitations hindering high-fidelity reconstruction:

First, existing methods use either human poses [10, 15] or 2D optical flows [6, 29] as motion constraints for surface deformation. Human surface changes between adjacent frames are inherently 3D, pose parameters alone miss complex surface details, while optical flows operate at the pixel level but lack depth consistency. Consequently, those ideas have limited motion constraints and, therefore, can easily lead to cross-view geometric errors and texture blurriness.

Second, most methods [4, 25] recover geometry via point-wise local features from the current frames. Gaussian point clouds, however, depend on local neighborhoods, while human surface motion is globally driven. Local points are influenced by full-body pose and global dynamics, so relying solely on local features causes inconsistencies between local surfaces and overall structure. Even studies aiming at minimizing local-global gaps [47, 48] often employ a single network or globally shared representation, which tends to learn mean representation over the training pose distribution and thus cannot generalize well to unseen poses. In general, existing methods only abstract Gaussian features with limited descriptive powers.

To address the limited motion constraints, 2D and 3D combined multi-dimensional motion information can be ex-

Figure 1: Reconstruction comparison among SOTA methods and HiAvatar on untrained frames from real-world videos.

plored for improving surface reconstruction accuracy. Especially, 3D motion information is critical for modeling dynamic deformations of untrained views, while fine-grained vertex-level 3D motion is key to accurately guiding motion-aware appearance modeling. Leveraging the effectiveness of parametric models (e.g., SMPLX [21]) in modeling detailed body motions [32, 38, 39], this paper proposes an implicit mapper to estimate 3D surface vertex motion via parametric model-based meshes.

In addition, 2D motion features derived from optical flows [6, 29] remain valuable for precise motions while suppressing noise, thereby improving geometric and texture reconstruction. This article introduces an optical flow mapper which can generate optical flows for arbitrary poses. Hash-encoded feature representations [20, 25] preserve discriminative patterns with $O(1)$ spatial complexity via sparse parametric mapping and fixed-dimensional latent projection. Therefore, a gated filter is further proposed, which adaptively selects robust multi-scale hash-based 2D motion features for high-quality, low-noise representations.

To address the limited Gaussian feature abstraction abilities, novel and effective local and global coupled features of Gaussian point clouds are expected. Especially, local features can be extracted via multi-scale hash encoding to ensure high-fidelity geometric detail representation. For global features, considering that generalization across pose-specific structural differences is important, we introduce a globally adaptive feature extraction structure based on Mixture of Experts (MoE) [2, 9], where multiple experts model distinct pose deformation patterns. It avoids the mean representation issue via pose adaptation and thus ensures generalization for unseen pose reconstruction.

Building on those ideas, we propose HiAvatar, a framework for dynamic human reconstruction from monocular videos. HiAvatar leverages 3D Gaussian Splatting [11] for high-fidelity dynamic surface reconstruction. Experimental results demonstrate state-of-the-art performance, with superior accuracy in novel views and unseen poses.

The main contributions of this work can be summarized as follows.

- A temporal motion embedding strategy for rich motion constraints, which integrates an optical flow mapper, an optical-flow-based gated filter, and a parametric model-driven implicit mapper to capture 2D pixel-level and 3D vertex-level non-rigid motion, ensuring accurate dynamic surface reconstruction in novel views.
- A spatially local-global consistent encoding scheme for augmented Gaussian features, which extracts detailed local features via multi-scale hash encoding and models global dynamic structures using pose-adaptive Mixture of Experts (MoE) for effective generalization to unseen poses.
- A high-fidelity dynamic avatar reconstruction method for monocular videos, HiAvatar, which fuses 2D–3D motion embeddings with local-global consistent feature abstraction for robust spatio-temporal feature encoding. Combined with 3D Gaussian Splatting, it achieves accurate geometry and fine texture reconstruction.

2. Related Work

Avatar reconstruction can be fulfilled by various inputs, including single images [38, 39], monocular videos [29, 36], multi-view videos [14, 22], sparse image sequences [13, 26], RGB-D videos [3, 42], etc. Our method is on monocular videos and is the main focus here. For other related work, you may refer to Wang et al. [31, 33].

2.1. Avatar Reconstruction by Monocular Videos

Dynamic human surface reconstruction from monocular videos has been extensively studied [16,24] and recently advanced rapidly [7,36] due to the capability of deep learning in handling unstructured targets. Neural Radiance Fields (NeRF) [18] have been widely used for surface regression, either via human pose constraints [19,40] or ray tracing in normalized NeRF space [10,36].

More recently, 3D Gaussian Splatting (3DGS) [11] has gained popularity for its high rendering quality and speed [8, 25, 35, 44]. For example, GauHuman [8] refines 3DGS densities via KL divergence, and GoMAvatar [35] integrates mesh representations for precise geometry. These methods rely on transforming normalized space to observation space, which we also adopt.

Most deep learning methods condition on human skeletal poses to model surface motion. Some studies [17,29] relax this constraint by discarding parametric models or extending skeletal structures; e.g., HosNeRF [17] expands object skeletons for interaction learning, and GART [15] learns latent skeletal representations. However, they still rely on pose guidance.

Other approaches [5, 6] leverage optical flow. Guo et al. [6] used differentiable rendering to enforce flow constraints, while MotionGS [5] separates optical flow into motion and camera components to remove camera effects. Nevertheless, skeletal points and optical flows provide limited 3D information, restricting reconstruction accuracy and leading to potential errors.

Some methods [37, 41] avoid pose or flow constraints, using temporal information instead. For instance, Neus2 [34] uses the first frame as a reference and computes offsets from previous frames, and Wu et al. [37] queried 4D voxel features (XYZ-T) with a network to learn time-dependent deformations. These methods are not tailored for dynamic human reconstruction and only handle slow, small motions, making them less suitable for general dynamic human modeling as achieved in our approach.

2.2. Position Encoding

The positional encoding methodologies have evolved from classical non-parametric designs to advanced frequency-domain and hash encoding techniques. Early non-trainable approaches employ predefined mathematical rules to map spatial positions into fixed vectors, effectively capturing relative positional relationships but lacking adaptability. Subsequently, learnable frequency-domain encoding [30] leverages Fourier series expansion to project spatial coordinates into a multi-frequency sine/cosine function space and thus significantly enhances the capacity of modeling high-frequency geometric and textural details [4,36].

Recently, multi-scale hash encoding [20] can fulfill efficient coordinate-to-feature vector mapping and real-time querying by multi-resolution hash tables for memory-efficient spatial features. It has widely adopted in 3D Gaussian studies [1, 25]. For instance, 3DGS-Avatar [25] employs hash encoding to process Gaussian spheres represented human surfaces. However, existing methods predominantly focus on isolated feature extraction at individual positions, lacking the ability to capture the global motion context of point clouds, such as deformation correlations and dynamic consistency. Instead, HiAvatar integrates multi-scale hash encoding of local details with global motion dynamics, thereby mitigating the disjunction between local and global features.

3. Our Proposed Method

HiAvatar (Fig. 2) reconstructs the geometry and texture of dynamic avatars using 3D Gaussian Splatting. To mitigate geometric errors and texture blur under novel views or unseen poses in existing studies, it models motion features in the temporal domain with rich motion constraints and computes augmented Gaussian features in spatial domain. Integrating those features for final decoding and rendering leads to better reconstruction with rich details than existing methods.

The framework first initializes 3D Gaussian primitives for the current frame I_t, while referencing adjacent frames I_{t-1} and I_{t+1} for motion computation. Two parallel branches perform feature modeling: BEF-Network captures 2D and 3D motion features, and GCG-Encoder extracts locally and globally consistent spatial features. Their outputs, motion embedding $\boldsymbol{m}_t$ and spatial feature $\boldsymbol{f}_t$, serve as conditional inputs for decoding and rendering.

During decoding, the geometric decoder $\mathscr{G}$ receives the spatial feature $\boldsymbol{f}_t$, temporal motion embedding $\boldsymbol{m}_t$, current pose $\boldsymbol{p}_t$ and temporal positional encoding $\gamma(t)$, and outputs geometric features and deformation parameters of Gaussian primitives:

$$\{\boldsymbol{f}_t^g, G_t^\delta\} = \mathscr{G}(\boldsymbol{f}_t, \boldsymbol{m}_t, \boldsymbol{p}_t, \gamma(t)), \tag{1}$$

where $G_t^\delta = \{P_t^\delta, R_t^\delta, S_t^\delta\}$ represent the update amounts of position, rotation, and scale respectively. The Gaussian primitives constrained by pose are then transformed to the target pose via Linear Blend Skinning (LBS).

The geometric feature $\boldsymbol{f}_t^g$ is further input into the texture decoder $\mathscr{T}$ together with the viewing direction $\boldsymbol{r}_t$, initial color $\boldsymbol{c}_t'$, and temporal encoding $\gamma(t)$, outputting the updated color:

$$\boldsymbol{c}_t = \mathscr{T}(\boldsymbol{f}_t^g, \boldsymbol{c}_t', \gamma(t)). \tag{2}$$

Finally, combined with the updated geometry and texture, the 3D Gaussian Splatting renderer generates reconstruction results under different views. During training, in

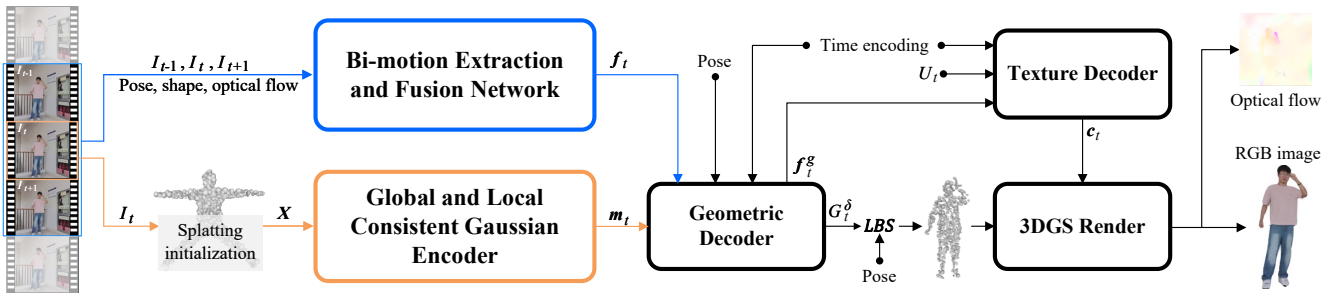


Figure 2: Pipeline of HiAvatar. Two parallel streams incorporating the 2D and 3D combined motion feature $\boldsymbol{m}_t$ and the global and local consistent frame-wise feature $\boldsymbol{f}_t$ are applied first, and then the geometry and texture decoders are applied subsequently for 3D Gaussian Splatting based optimization.

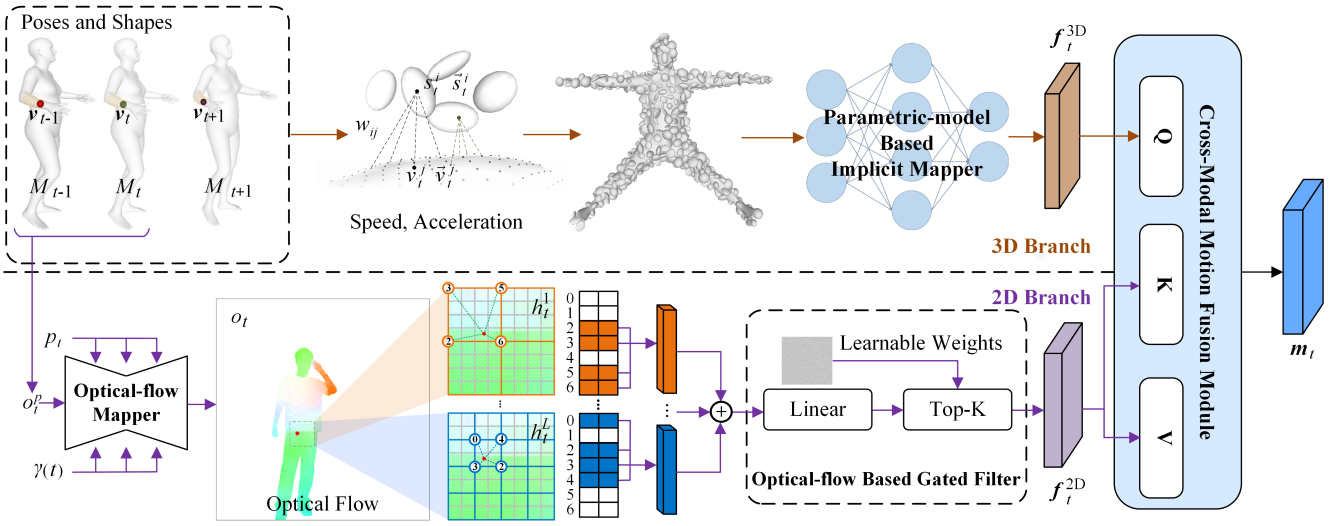


Figure 3: Structure of the bi-motion extraction and fusion network. 3D and 2D branches are independently applied to compute 3D and 2D motion features, f_t^{3D} and f_t^{2D} for the current frame I_t, respectively. These features are finally fused by the cross-modal motion fusion module to form the final motion feature $\boldsymbol{m}_t$.

addition to the RGB reconstruction loss, optical flow supervision is introduced to enhance cross-frame motion consistency. Here, the optical flows are estimated by Neuflow v2 [46], and human parametric meshes are generated by PyMAF-X [45].

Now, let's delve into the details of BEF-Network and GCG-Encoder while discussing the details of the proposed two tactics.

3.1. Bi-motion Extraction and Fusion Network

As shown in Fig. 3, BEF-Network consists of a 3D branch, a 2D branch, and a cross-modal fusion module. The 3D branch derives vertex-level velocity and acceleration from parametric model meshes, and learns the nonlinear mapping from these motion quantities to Gaussian primitives through an implicit mapper. The 2D branch generates pose-related optical flow maps via an optical flow mapper, then extracts robust optical flow features in the multi-scale hash space combined with an optical flow-based gated filter. Finally, the two branches are fused via cross-modal fusion to form a spatio-temporally consistent motion feature representation.

3.1.1 3D Motion Features: Parametric-model Based Implicit Mapper

The 3D branch computes vertex-level velocity V_t and acceleration $\vec{V}_t$ using parametric meshes M_{t-1}, M_t, M_{t+1} of adjacent frames to model non-rigid surface dynamics:

$$\begin{cases} V_t = \frac{M_t - M_{t-1}}{\Delta t}, \\ \vec{V}_t = \frac{V_t - V_{t-1}}{\Delta t}. \end{cases} \tag{3}$$

Subsequently, a weighted mapping is established between each Gaussian primitive and its neighboring Q mesh ver-

tices (the weight w_{ij} is inversely proportional to the distance between the vertex and the primitive) to transmit vertex motion information:

$$\begin{cases} \boldsymbol{s}_t^i = \sum_{j=1}^{Q} w_{ij} \cdot \boldsymbol{v}_t^j, \\ \vec{\boldsymbol{s}}_t^i = \sum_{j=1}^{Q} w_{ij} \cdot \vec{\boldsymbol{v}}_t^j. \end{cases} \tag{4}$$

To further model the nonlinear correlation between motion and Gaussian primitives, an implicit mapper $\mathscr{I}$ is designed to embed the above motion field into a high-dimensional space, resulting in the final 3D motion features:

$$\boldsymbol{f}_t^{3D} = \mathscr{I}(S_t, \vec{S}_t). \tag{5}$$

3.1.2 2D Motion Features: Optical-flow Mapper and Gated Filter

The 2D branch mainly consists of two components: an optical flow mapper and a gated filter. It requires optical flows of the current pose.

First, the optical flow mapper $\mathscr{O}$ generates the optical flow corresponding to the pose. Its inputs include the pose parameters $\boldsymbol{p}_t$, temporal encoding $\gamma(t)$, and optical flow o_t^p between adjacent frames of parametric model renderings. It outputs the optical flow map o_t of the moving human surface, with the formula given as:

$$o_t = \mathscr{O}(\boldsymbol{p}_t, \gamma(t), o_t^p). \tag{6}$$

Next, o_t is mapped to the multi-scale hash space to obtain pixel-level initial features $\boldsymbol{h}_t^i \in \mathbb{R}^d$:

$$\boldsymbol{h}_t^i = \bigoplus_{l=0}^{L-1} \bigoplus_{j=1}^{4} \mathscr{H}(o_t^l, j), \tag{7}$$

where $\mathscr{H}(o_t^l, j)$ denotes the hash feature of the j-th neighboring pixel in the l-th layer of the multi-scale optical flow map o_t^l.

To suppress optical flow noise interference, a gated mechanism is introduced to weight pixels. For each pixel p_i, the normalized weight is calculated as:

$$a_t^i = \sigma(\mathscr{W}(o_t)). \tag{8}$$

The top K pixel features (where K is set to the number of Gaussian primitives) are selected as the final 2D motion features:

$$\boldsymbol{f}_t^{2D} = \bigoplus \{\boldsymbol{h}_t^k \mid a_k \in \mathscr{T}_K(A)\}. \tag{9}$$

3.1.3 Cross-modal Motion Fusion

The 2D feature $\boldsymbol{f}_t^{2D}$ and 3D feature $\boldsymbol{f}_t^{3D}$ are fused via a multi-head cross-attention mechanism to achieve modal information complementarity:

$$\boldsymbol{m}_t = \mathscr{A}(\boldsymbol{f}_t^{3D} W^Q, \boldsymbol{f}_t^{2D} W^K, \boldsymbol{f}_t^{2D} W^V), \tag{10}$$

where W^Q, W^K, W^V are learnable parameters. This mechanism strengthens the synergy between 2D pixel-level details and 3D depth consistency through attention weight assignment, ultimately forming a spatio-temporally unified motion embedding $\boldsymbol{m}_t$.

3.2. Global-and-local Consistent Gaussian Encoder

This module (Fig. 4) models local details and globally consistent features simultaneously in the spatial domain to address the lack of global constraints for Gaussian primitives and poor generalization across different poses. It consists of three steps: first, local features $\boldsymbol{f}_t^l$ are extracted via multi-scale hashing; second, global features $\boldsymbol{f}_t^g$ are adaptively aggregated based on Mixture-of-Experts (MoE); finally, these features are fused with temporal encoding to obtain the final spatially consistent feature $\boldsymbol{f}_t$.

In the local feature extraction stage, initial Gaussian primitives G_t are generated by downsampling a standard T-pose parametric model, and mapped to H multi-scale 3D grid spaces G_t^h $(1 \le h \le H)$. At each scale, features are extracted via hash encoding $\mathscr{H}^h$, and concatenated to form the local feature representation:

$$\boldsymbol{f}_t^l = \bigoplus_{i=1}^{M} \left(\bigoplus_{h=1}^{H} \mathscr{H}^h(G_t^h, i) \right), \tag{11}$$

where M is the number of Gaussian primitives.

In the global feature extraction stage, the local features $\boldsymbol{f}_t^l$ first undergo a pooling operation $\mathscr{P}$ to obtain a global feature representation. To avoid the mean representation issue caused by single global abstraction and improve adaptability to surface structures under different poses, an MoE module $\mathscr{M}$ is introduced. This module consists of multiple expert networks, so that adaptive expert combinations are generated based on input poses via a gating mechanism to dynamically generate global features:

$$\boldsymbol{f}_t^g = \mathscr{M}\left(\mathscr{P}(\boldsymbol{f}_t^l)\right). \tag{12}$$

In experiments, the MoE contains 3 expert networks, and the gating module selects 2 of them for feature combination each time.

Finally, the local features, global features, and temporal encoding $\boldsymbol{f}_t^P$ are fused first and then processed via linear projection $\mathscr{L}$ to obtain the final spatially consistent feature:

$$\boldsymbol{f}_t = \mathscr{L}(\boldsymbol{f}_t^l \bigoplus \boldsymbol{f}_t^P \bigoplus \boldsymbol{f}_t^g). \tag{13}$$

This local-global consistent encoding strategy not only retains the detail expression capability of Gaussian point clouds but also introduces global dynamic constraints. It maintains the overall consistency of features during pose changes, providing robust spatial feature support for subsequent dynamic surface reconstruction.

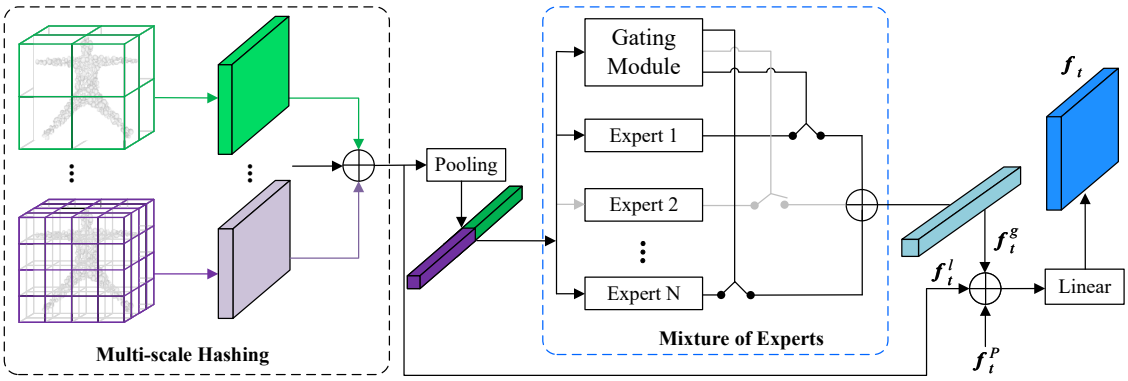


Figure 4: Global-and-local consistent Gaussian encoder. The multi-scale hashing is first applied to compute the local features $\boldsymbol{f}_t^l$ for the Gaussian primitives initialized by the standard T-posed parametric model, which are also adaptively processed by MOE for global features $\boldsymbol{f}_t^g$. Combining local and global features and the encoded time $\boldsymbol{f}_t^P$ leads to the comprehensive features f_t.

4. Loss

The overall loss of our method is defined as follows:

$$\begin{aligned}\mathscr{L} =& \lambda_1 \mathscr{L}_{\text{RGB}} + \lambda_2 \mathscr{L}_{\text{mask}}^{\text{image}} + \lambda_3 \mathscr{L}_{\text{mask}}^{\text{parametric}} \\ &+ \lambda_4 \mathscr{L}_{\text{pose}} + \lambda_5 \mathscr{L}_{\text{skin}} + \lambda_6 \mathscr{L}_{\text{flow}}^{\text{mapper}} \\ &+ \lambda_7 \mathscr{L}_{\text{flow}}^{\text{Gaussian}},\end{aligned} \tag{14}$$

where $\lambda_i\,(i = 1,\ldots,7)$ represent the weight coefficients ($\lambda_i\,(1 \le i \le 3)$, λ_5 and λ_7 are set to $1e-3$ and λ_4 and λ_6 are set to $1e-4$ in our experiment). $\mathscr{L}_{\text{RGB}}$ denotes the pixel-level loss between the rendered image and the ground truth image. $\mathscr{L}_{\text{mask}}^{\text{image}}$ represents the loss associated with the mask image with $\mathscr{L}_{\text{mask}}^{\text{parametric}}$ for the loss by the mask from the projected parametric model. $\mathscr{L}_{\text{mask}}^{\text{parametric}}$ and $\mathscr{L}_{\text{pose}}$ are for eliminating inaccurate pose parameters, while $\mathscr{L}_{\text{skin}}$ is for the forward skinning network. $\mathscr{L}_{\text{flow}}^{\text{mapper}}$ and $\mathscr{L}_{\text{flow}}^{\text{Gaussian}}$ are the motion supervision losses applied to the optical flow mapper and the Gaussian primitives, respectively.

Especially, $\mathscr{L}_{\text{mask}}^{\text{parametric}}$ measure the pixel-level discrepancy of the masks ($\mathscr{S}$) between the reconstructed parametric model M_t, and the ground truth images $\hat{M}_t$:

$$\mathscr{L}_{\text{mask}}^{\text{parametric}} = \|\mathscr{S}(M_t) - \mathscr{S}(\hat{M}_t)\|_1. \tag{15}$$

$\mathscr{L}_{pose}$ optimizes the rotation matrices of the human joints by the differences between rotation matrices computed from the parametric model R_t and their ground truths $\hat{R}_t$ and thus constrains the accuracy of the body posture:

$$\mathscr{L}_{\text{pose}} = \|R_t - \hat{R}_t\|_2^2. \tag{16}$$

$\mathscr{L}_{\text{flow}}^{\text{mapper}}$ is used to optimize the optical flow mapper:

$$\mathscr{L}_{\text{flow}}^{\text{mapper}} = \|o_t - \hat{o}_t\|_1, \tag{17}$$

where $o_t = \mathscr{O}(\boldsymbol{p}_t, \gamma(t), o_t^p)$ denotes the predicted optical flow from the mapper for the current pose, and $\hat{o}_t$ is the corresponding ground truth optical flow.

$\mathscr{L}_{\text{flow}}^{\text{Gaussian}}$ is measured by 2D optical flow based mapped motion of Gaussian primitives by opacity weighting to supervise the motion of the Gaussian primitives [6]:

$$\mathscr{L}_{\text{flow}}^{\text{Gaussian}} = \left\| \hat{o}_t - \sum_{i=1}^{N} \omega_i (\mu_{ni} - \mu_{ci}) \right\|_1, \tag{18}$$

where ω_i is the opacity of the i-th Gaussian primitive, and μ_{ni} and μ_{ci} represent the center positions of the Gaussian sphere in the next and current frames, respectively.

5. Experiments

5.1. Evaluation Datasets

ZJU-MoCap dataset [23] and real-world in-the-wild videos are used for training and testing. ZJU-MoCap provides a multi-camera, multi-subject benchmark for human rendering evaluation, comprising nine dynamic human performance videos captured by 23 synchronized cameras. In-the-wild videos are additionally collected to qualitatively validate the generalization ability of our method.

For fair comparison, six sequences (377, 386, 387, 392, 393, 394) are selected from ZJU-MoCap and split into training and testing subsets following the protocol in HumanNeRF [36]. Six real-world videos are also included. SAM [12] is used to generate subject masks, while PyMAF-X [45] is employed to estimate SMPLX parameter models. The first 70% of frames are used for training with the remaining for testing.

5.2. Baseline Methods and Evaluation Metrics

Eight monocular video-based avatar reconstruction methods are compared, including NeuralBody [23], Hu-

(a) Novel View GT (b) HiAvatar (Ours) (c) GauHuman[8] (d) 3DGS-Avatar[25] (e) GoMAvatar[35] (f) R3-Avatar[44]

Figure 5: Qualitative comparison of novel views on six subjects from ZJU-MoCap.

manNeRF [36], MonoHuman [43], GoMAvatar [35], 3DGS-Avatar [25], GauHuman [8], TE-NeRF [19], and R3-Avatar [44]. Reconstruction quality is evaluated using PSNR, SSIM, and LPIPS, where LPIPS is scaled by 1000 ($LPIPS = LPIPS' \times 1000$).

5.3. Training Settings

The method is implemented in Python and trained on an NVIDIA® 4090 GPU. Training on ZJU-MoCap takes $15k$ iterations with $30k$ iterations for in-the-wild videos. The Adam optimizer is adopted with learning rates of 1×10^{-4} for the forward skinning network and 1×10^{-3} for others. An exponential scheduler reduces learning rates by a factor of 0.1, and a decay weight of 0.05 is applied to the time encoding.

During the first $1k$ iterations, only the LBS process is performed. Optimization of 3D Gaussians starts after $1k$ iterations, with BEF-Network optimization after $3k$, and pose correction after $5k$ iterations.

For the Mixture-of-Experts (MoE) module in the GCG-Encoder, we adopt the *Noisy Top-K Gating* mechanism proposed by Shazeer et al. [27] to enable sparse expert activation. Specifically, learnable Gaussian noise is added to the outputs of the gating network, and only the top-K experts are selected for each input. The gating network and expert networks are jointly optimized via standard backpropagation. Following [27], we incorporate the importance loss and load loss defined in that work as auxiliary regularization terms to encourage balanced expert utilization by reducing the variation of expert importance and computational load across the batch. In addition, the gating weights are initialized to zero, as suggested in [27], to promote a uniform expert assignment at the early stage of training.

Table 1: Quantitative results on ZJU-MoCap. The best results are shown in **bold**.

Methods	377			386			387		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeuralBody [23]	29.11	0.9674	40.95	30.54	0.9678	46.43	27.00	0.9518	59.47
HumanNeRF [36]	30.41	0.9743	24.06	33.20	0.9752	28.99	28.18	0.9632	35.58
MonoHuman [43]	29.12	0.9727	26.58	32.94	0.9695	36.04	27.93	0.9601	41.76
GoMAvatar [35]	30.56	0.9765	24.06	32.97	0.9749	30.64	28.31	0.9636	35.88
3DGS-Avatar [25]	30.64	0.9774	20.88	33.63	0.9773	25.77	28.33	0.9642	34.24
GauHuman [8]	31.03	0.9716	24.98	33.64	0.9687	33.34	28.50	0.9564	41.82
TE-NeRF [19]	29.79	0.9656	29.41	32.73	0.9605	36.30	27.86	0.9596	39.87
R3-Avatar [44]	30.48	0.9682	28.83	33.56	0.9661	37.35	28.20	0.9521	46.12
HiAvatar (Ours)	**31.76**	**0.9897**	**19.79**	**34.24**	**0.9836**	**24.61**	**30.01**	**0.9794**	**33.01**
Methods	392			393			394		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeuralBody [23]	30.10	0.9642	53.27	28.61	0.9590	59.05	29.10	0.9593	54.55
HumanNeRF [36]	31.04	0.9705	32.12	28.31	0.9603	36.72	30.31	0.9642	32.89
MonoHuman [43]	29.50	0.9635	39.45	27.64	0.9566	43.17	29.15	0.9595	38.08
GoMAvatar [35]	31.09	0.9707	35.36	28.80	0.9622	37.77	30.24	0.9641	34.17
3DGS-Avatar [25]	31.66	0.9730	30.14	28.88	0.9635	35.26	30.54	0.9661	31.21
GauHuman [8]	31.54	0.9630	35.07	29.48	0.9544	40.07	30.59	0.9571	35.84
TE-NeRF [19]	30.23	0.9540	41.54	26.66	0.9284	65.16	29.68	0.9476	41.22
R3-Avatar [44]	31.35	0.9610	38.99	29.12	0.9516	45.59	30.41	0.9547	41.31
HiAvatar (Ours)	**32.27**	**0.9861**	**28.56**	**30.03**	**0.9875**	**32.45**	**31.92**	**0.9886**	**29.39**

(a) Novel pose (b) HiAvatar (Ours) (c) 3DGS-Avatar [25] (d) GauHuman [8]

Figure 6: Qualitative comparison of novel poses on three subjects from ZJU-MoCap.

5.4. Qualitative Results

Fig. 5 gives visual comparisons of multiple methods on ZJU-MoCap under novel views. For novel poses (Fig. 6), we compare our method specifically with 3DGS-Avatar [25] and GauHuman [8]. Our proposed HiAvatar achieves the best results, exhibiting the least blurriness and reconstruction artifacts among all methods.

The generalization ability of our method on in-the-wild videos is also evaluated (Fig. 7), which further demonstrate the superiority of our approach in modeling photo-realistic human surfaces.

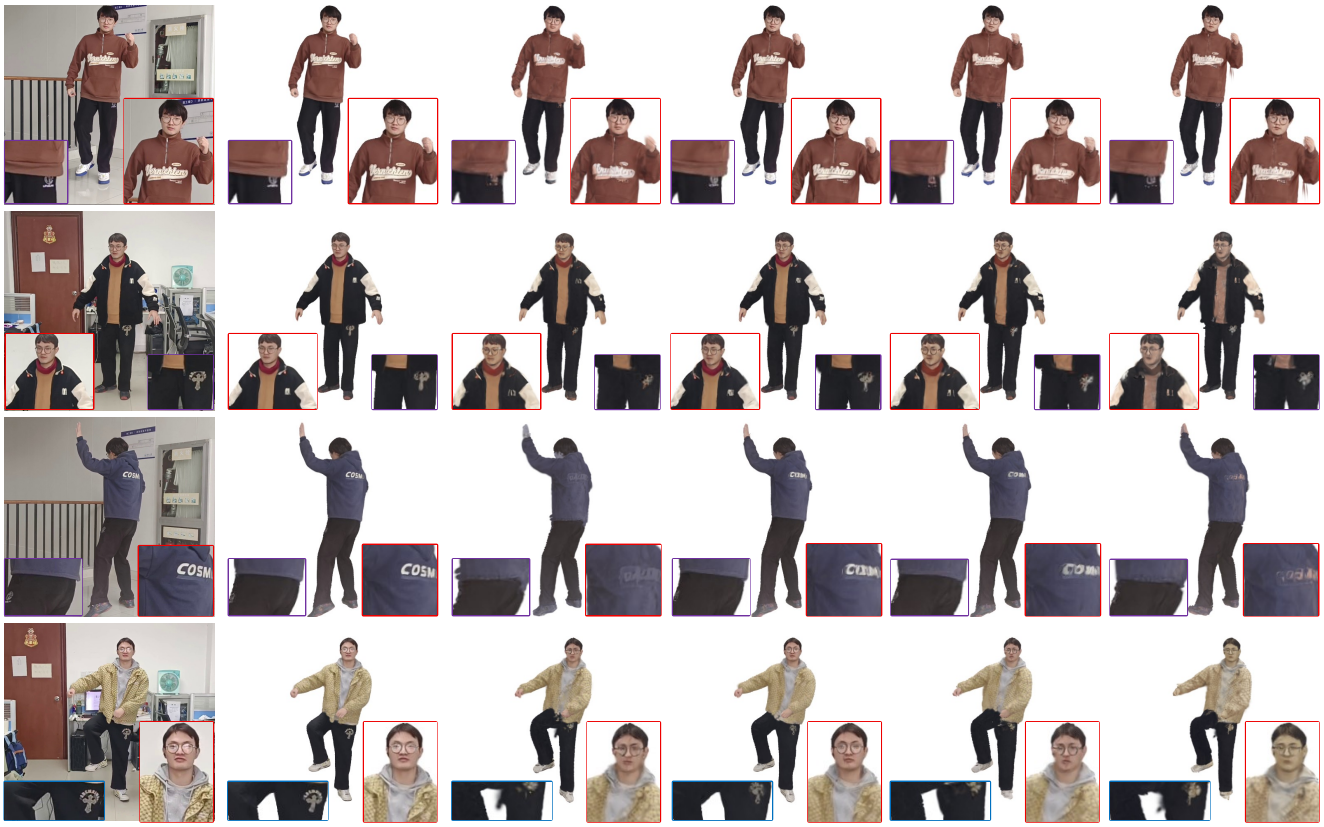
5.5. Quantitative Results

Statistical comparison among different methods are presented (Table 1). Our method outperforms all competing baselines in terms of PSNR, SSIM and LPIPS.

The quantitative results on in-the-wild videos also show the same findings (Table 2), demonstrating that our method maintains consistent advantages over existing methods.

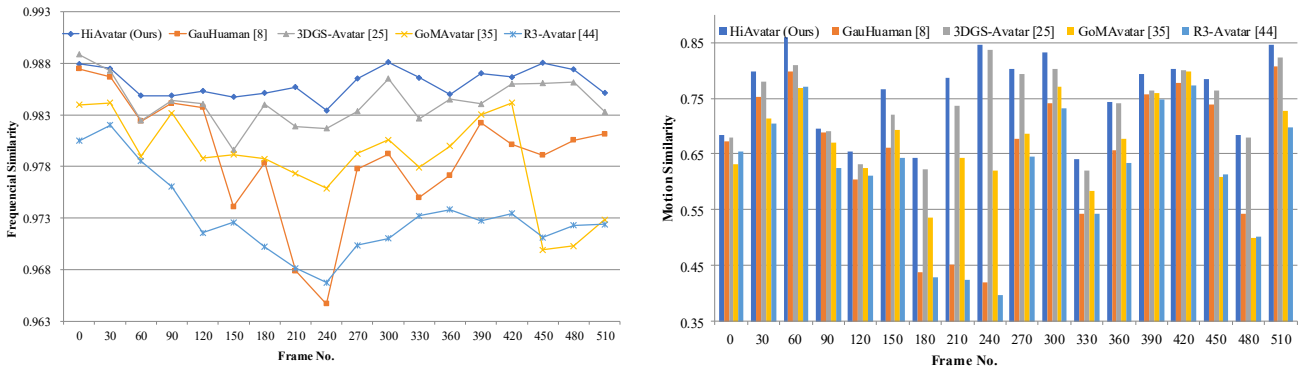
In addition to reconstruction quality, we further compare the training efficiency and rendering speed of different methods, as summarized in Table 3. While HiAvatar does not achieve the fastest rendering speed, it maintains competitive inference performance with a moderate training time compared to other state-of-the-art methods. This indicates a reasonable trade-off between reconstruction quality and computational efficiency.

The reconstruction capability for local details can be captured by the frequency distribution and the accuracy of the recovered spatial motions. Therefore, statistical comparisons of the reconstruction accuracy among SOTA methods and ours are conducted by directly comparing the frequencial and motion similarities between the recovered images with the input frames (GT) (Fig. 8). Here, every 30



(a) Test GT (b) HiAvatar (Ours) (c) GauHuman [8] (d) 3DGS-Avatar [25] (e) GoMAvatar [35] (f) R3-Avatar [44]

Figure 7: Qualitative comparison on one in-the-wild video.



(a) Frequencial similarity (b) Motion similarity

Figure 8: Statistical comparison on the reconstruction capability for local details.

frames from the with two subjects (386, 387) from ZJU-MoCap are taken. The frequencial similarity is computed by the cosine similarity of the magnitudes of the Fourier transformed images. The motion similarity is computed with the optical flow based ratio of dynamic foreground areas in the estimated result to its GT frames, i.e., those pixels having both directions and magnitudes of optical flows bigger than a pre-defined threshold (it is set to 0.85 in our experiment). Both results show that our proposed method can achieve the best performances among all methods.

5.6. Ablation Study

This subsection evaluates the two main components of HiAvatar: the Bi-motion Extraction and Fusion Network

Table 2: Quantitative results on in-the-wild videos.The best results are shown in **bold**.

Methods	In-the-wild video 1			In-the-wild video 2			In-the-wild video 3		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GoMAvatar [35]	27.70	0.9498	49.72	30.51	0.9571	45.02	27.89	0.9513	44.83
3DGS-Avatar [25]	28.43	0.9501	48.02	30.97	0.9603	44.65	28.04	0.9580	44.32
GauHuman [8]	30.03	0.9432	52.94	31.04	0.9647	45.26	30.17	0.9202	53.21
R3-Avatar [44]	28.04	0.9477	54.21	30.94	0.9589	44.98	27.97	0.9190	54.84
HiAvatar (Ours)	**31.74**	**0.9741**	**44.32**	**32.34**	**0.9892**	**40.21**	**32.79**	**0.9854**	**41.01**
Methods	In-the-wild video 4			In-the-wild video 5			In-the-wild video 6		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GoMAvatar [35]	28.32	0.9532	47.34	32.01	0.9490	41.43	27.02	0.9379	48.90
3DGS-Avatar [25]	29.98	0.9634	45.43	32.43	0.9588	42.45	28.71	0.9412	47.32
GauHuman [8]	30.43	0.9341	49.32	32.98	0.9532	48.32	28.21	0.9201	52.73
R3-Avatar [44]	29.86	0.9485	51.74	32.34	0.9417	48.96	28.28	0.9245	53.56
HiAvatar (Ours)	**31.87**	**0.9767**	**42.21**	**34.01**	**0.9878**	**40.18**	**30.17**	**0.9778**	**40.12**

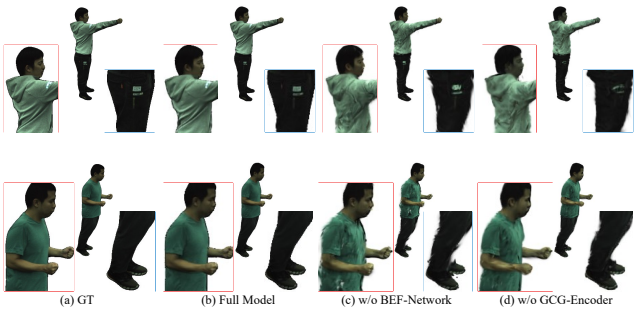


Figure 9: Effects of different ablation configurations for our proposed two tactics.

Table 3: Comparison of training time and rendering speed among different methods.

Methods	Train Time	FPS
GoMAvatar [35]	1020 min	43
3DGS-Avatar [25]	30min	27
GauHuman [8]	**20min**	**189**
R3-Avatarr [44]	90min	60
HiAvatar (Ours)	45min	25

(BEF-Network) and the Global-and-local Consistent Gaussian Encoder (GCG-Encoder).

First, the ablation with the same two subjects (386, 387) from ZJU-MoCap by averaging their measurements is conducted. Consequently, two additional configurations are obtained except the full model. (1) w/o GCG-Encoder: GCG-Encoder is replaced with traditional frequency-domain encoding; and (2) w/o BEF-Network: The entire motion-conditioned input of BEF-Network is removed. Fig. 9 and Table 4 show that the full model outperforms the other configurations in all ablation experiments, which demonstrate the importance of the proposed strategies.

Table 4: Ablation study on the proposed two tactics. Best results are shown in **bold**.

Metric	PSNR↑	SSIM↑	LPIPS↓
Full model	**32.13**	**0.9815**	**28.81**
w/o GCG-Encoder	30.36	0.9503	30.87
w/o BEF-Network	29.43	0.9421	33.49

Table 5: Ablation experiments on the key components in the two proposed tactics. Best results are shown in **bold**.

Metric	PSNR↑	SSIM↑	LPIPS↓
Full model	**32.13**	**0.9815**	**28.81**
w/o MOE	30.51	0.9621	30.61
w/o Implicit Mapper	29.84	0.9541	32.06
w/o Gated Filter	31.34	0.9706	29.62

Then, the key components of GCG-Encoder and BEF-Network is evaluated, which adopts the averaging results over subjects 386 and 387 from ZJU-MoCap. Specifically, we test: (1) w/o MOE: pooled local features used as global features; (2) w/o Parametric-Model Implicit Mapper (referred to as Implicit Mapper): motion from neighboring points fused with 2D motion features via a linear layer; (3) w/o Optical-Flow Gated Filter (referred to as Gated Filter): feature selection removed and features mapped directly to 3D motion dimensions. Fig. 10 shows the full model outperforms all variants, confirming the effectiveness of these

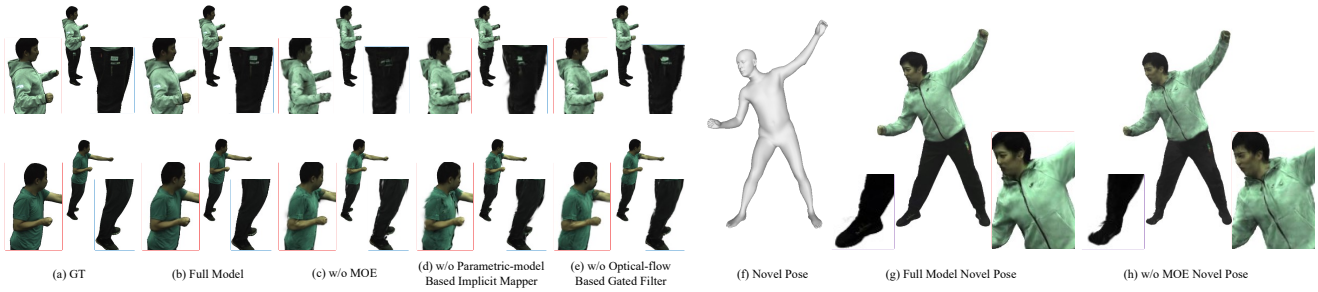
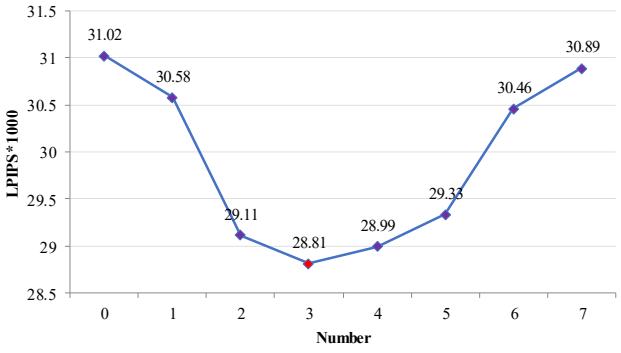


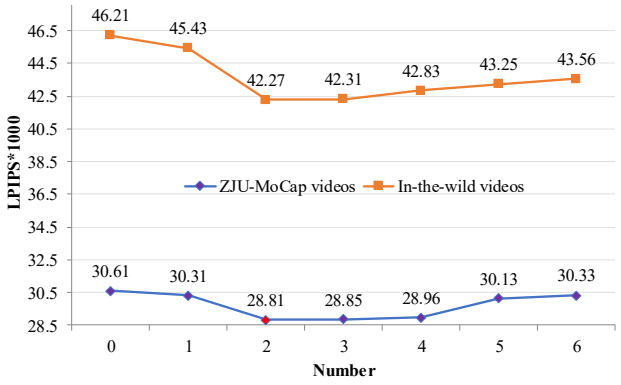
Figure 10: Performance of key components in our two proposed tactics.

Table 6: Ablation study on multi-scale hash encoding configurations. Best results are shown in **bold**.

Setting	Base Resolution	Max Resolution	PSNR ↑	SSIM ↑	LPIPS ↓
16 levels	16	2048	**32.13**	**0.9815**	**28.81**
8 levels	16	2048	31.83	0.9742	29.90
1 level	2048	2048	30.40	0.9547	30.62



(a) Number of Neighoring Vertices in BEF-Network



(b) Number of Experts in GCG-Encoder

Figure 11: Ablation studies on the number of neighboring mesh vertices and the number of experts in MOE.

components.

In addition, we analyze the effect of the multi-scale hash encoding used in the GCG-Encoder, as summarized in Table 6. Multi-level configurations share the same base and maximum resolutions, whereas the single-level setting employs a single hash grid at the maximum resolution. The results indicate that using multiple hash levels substantially improves reconstruction quality compared to the single-level design, with performance gradually increasing as more levels are introduced. This suggests that multi-scale hash encoding enables more effective spatial feature modeling, while an appropriate number of levels achieves a favorable trade-off between reconstruction accuracy and computational cost.

Finally, we study the influence of two key hyper-parameters: the number of experts in the MoE module and the number of neighboring mesh vertices Q used for surface motion estimation (Eq. 4), as shown in Fig. 11. The ablation on Q follows the same protocol as the previous experiments, where results are averaged over the two ZJU-MoCap subjects (386 and 387). For the MoE expert number, we report results separately by averaging over the two ZJU-MoCap subjects and over two in-the-wild videos (In-the-wild video 1 and In-the-wild video 2), respectively. This separation enables a clearer evaluation of the expert configuration under both controlled benchmark data and real-world scenarios. We further observe a performance drop when using an excessive number of experts, which suggests that overly fine-grained expert partitioning may limit the amount of effective training data available to each expert.

5.7. Limitations

Although HiAvatar has demonstrated significant progress in high-fidelity human reconstruction, several limitations still remain.

First, efficiently processing high-resolution videos and long temporal sequences remains challenging due to high computational and memory costs. While the proposed motion-aware representations improve reconstruction quality, they introduce additional temporal dependencies that increase training and inference overhead. Future work may explore more compact representations through feature sparsification or temporal redundancy reduction, as well as adaptive multi-scale or keyframe-based temporal modeling strategies to balance reconstruction accuracy and efficiency.

Second, scenes with rapid motion, severe motion blur, or heavy occlusion continue to pose difficulties, as incomplete or ambiguous observations can degrade motion estimation and appearance reconstruction. Potential directions include incorporating explicit motion deblurring modules, leveraging stronger temporal priors to propagate reliable information across frames, or adopting more robust tracking and occlusion reasoning mechanisms to better handle temporarily invisible regions.

6. Conclusion

This paper addresses geometric errors and texture blurriness in existing monocular video-based avatar reconstruction caused by limited motion constraints and Gaussian feature abstraction abilities, respectively, and proposes a novel system HiAvatar with two core strategies to overcome those two limitations. Temporally, a multi-dimensional motion embedding strategy is designed to enhanced feature encoding with rich motion constraints: It uses an optical flow mapper and an optical-flow-based gated filter to capture 2D pixel-level motion, and takes a parametric model-driven implicit mapper to estimate 3D vertex-level non-rigid motion. Spatially, a local-global consistent encoding scheme is adopted for augmented Gaussian features: It extracts high-fidelity local features through multi-scale hash encoding, and models adaptive global dynamic structures with a pose-adaptive Mixture of Experts (MoE) to avoid mean representation, providing reliable support for 3D Gaussian-based surface regression. Experimental results demonstrate the superiority of HiAvatar in monocular video based dynamic avatar reconstruction, especially for novel views and unseen poses.

References

[1] Y. Chen, Q. Wu, W. Lin, M. Harandi, and J. Cai. HAC: Hash-grid Assisted Context for 3D Gaussian Splatting Compression. In *ECCV*, pages 422–438, 2024. 3

[2] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, et al. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *arXiv preprint arXiv:2401.06066*, 2024. 2

[3] Z. Dong, C. Guo, J. Song, X. Chen, A. Geiger, and O. Hilliges. PINA: Learning a Personalized Implicit Neural Avatar from a Single RGB-D Video Sequence. In *CVPR*, pages 20470–20480, 2022. 2

[4] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges. Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition. In *CVPR*, pages 12858–12868, 2023. 1, 3

[5] X. Guo, W. Zhang, R. Liu, P. Han, and H. Chen. MotionGS : Compact Gaussian Splatting SLAM by Motion Filter. In *RCAE*, pages 685–692, 2024. 3

[6] Z. Guo, W. Zhou, L. Li, M. Wang, and H. Li. Motion-aware 3D Gaussian Splatting for Efficient Dynamic Scene Reconstruction. *TCSVT*, 2024. 1, 2, 3, 6

[7] L. Hu, H. Zhang, Y. Zhang, B. Zhou, B. Liu, S. Zhang, and L. Nie. GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians. In *CVPR*, pages 634–644, 2024. 1, 3

[8] S. Hu, T. Hu, and Z. Liu. GauHuman: Articulated Gaussian Splatting from Monocular Human Videos. In *CVPR*, pages 20418–20431, 2024. 3, 7, 8, 10

[9] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive Mixtures of Local Experts. *Neural computation*, 3(1):79–87, 1991. 2

[10] T. Jiang, X. Chen, J. Song, and O. Hilliges. InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds. In *CVPR*, pages 16922–16932, 2023. 1, 3

[11] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM ToG*, 42(4):139–1, 2023. 2, 3

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment Anything. In *ICCV*, pages 4015–4026, 2023. 6

[13] M. Kocabas, J.-H. R. Chang, J. Gabriel, O. Tuzel, and A. Ranjan. HUGS: Human Gaussian Splats. In *CVPR*, pages 505–515, 2024. 2

[14] Y. Kwon, D. Kim, D. Ceylan, and H. Fuchs. Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering. *NeurIPS*, 34:24741–24752, 2021. 2

[15] J. Lei, Y. Wang, G. Pavlakos, L. Liu, and K. Daniilidis. GART: Gaussian Articulated Template Models. In *CVPR*, pages 19876–19887, 2024. 1, 3

[16] Z. Li, Z. Zheng, H. Zhang, C. Ji, and Y. Liu. AvatarCap: Animatable Avatar Conditioned Monocular Human Volumetric Capture. In *ECCV*, pages 322–341, 2022. 3

[17] J.-W. Liu, Y.-P. Cao, T. Yang, Z. Xu, J. Keppo, Y. Shan, X. Qie, and M. Z. Shou. HOSNeRF: Dynamic Human-Object-Scene Neural Radiance Fields from a Single Video. In *ICCV*, pages 18483–18494, 2023. 3

[18] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[19] S. Mubashshira and K. Desai. TE-NeRF: Triplane-Enhanced Neural Radiance Field for Artifact-Free Human Rendering. In *WACV*, pages 238–247, 2025. 3, 7, 8

[20] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM ToG*, 41(4):1–15, 2022. 2, 3

[21] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*, 2019. 2

[22] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *ICCV*, pages 14314–14323, 2021. 2

[23] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*, pages 9054–9063, 2021. 1, 6, 8

[24] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. ClothCap: seamless 4D clothing capture and retargeting. *ACM ToG*, 36(4):1–15, 2017. 3

[25] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang. 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting. In *CVPR*, pages 5020–5030, 2024. 1, 2, 3, 7, 8, 10

[26] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *ICCV*, pages 2304–2314, 2019. 2

[27] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 7

[28] D.-Y. Song, H. Lee, J. Seo, and D. Cho. DIFu: Depth-Guided Implicit Function for Clothed Human Reconstruction. In *CVPR*, pages 8738–8747, 2023. 1

[29] J. Tan, D. Xiang, S. Tulsiani, D. Ramanan, and G. Yang. DressRecon: Freeform 4D Human Reconstruction from Monocular Video. In *3DV*, 2025. 1, 2, 3

[30] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *NeruIPS*, 33:7537–7547, 2020. 3

[31] Y. Tian, H. Zhang, Y. Liu, and L. Wang. Recovering 3D Human Mesh From Monocular Images: A Survey. *IEEE TPAMI*, 45(12):15406–15425, 2023. 2

[32] L. Wang, X. Zhao, T. Yu, S. Wang, and Y. Liu. NormalGAN: Learning Detailed 3D Human from a Single RGB-D Image. In *ECCV*, pages 430–446, 2020. 2

[33] R. Wang, Y. Cao, K. Han, and K.-Y. K. Wong. A Survey on 3D Human Avatar Modeling–From Reconstruction to Generation. *arXiv preprint arXiv:2406.04253*, 2024. 1, 2

[34] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu. NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction. In *ICCV*, pages 3295–3306, 2023. 3

[35] J. Wen, X. Zhao, Z. Ren, A. G. Schwing, and S. Wang. GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh. In *CVPR*, pages 2059–2069, 2024. 3, 7, 8, 10

[36] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video. In *CVPR*, pages 16210–16220, 2022. 2, 3, 6, 7, 8

[37] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *CVPR*, pages 20310–20320, 2024. 3

[38] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, pages 512–523, 2023. 1, 2

[39] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, pages 13286–13296, 2022. 2

[40] H. Xu, T. Alldieck, and C. Sminchisescu. H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion. *NeurIPS*, 34:14955–14966, 2021. 3

[41] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction. In *ICCV*, pages 20331–20341, 2024. 3

[42] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *CVPR*, pages 5746–5756, 2021. 2

[43] Z. Yu, W. Cheng, X. Liu, W. Wu, and K.-Y. Lin. MonoHuman: Animatable Human Neural Field from Monocular Video. In *CVPR*, pages 16943–16953, 2023. 7, 8

[44] Y. Zhan, W. Xu, Q. Zhu, M. Niu, M. Ma, Y. Liu, Z. Zhong, X. Sun, and Y. Zheng. R3-Avatar: Record and Retrieve Temporal Codebook for Reconstructing Photorealistic Human Avatars. *arXiv preprint arXiv:2503.12751*, 2025. 3, 7, 8, 10

[45] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu. PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images. *IEEE TPAMI*, 2023. 4, 6

[46] Z. Zhang, A. Gupta, H. Jiang, and H. Singh. NeuFlow v2: High-Efficiency Optical Flow Estimation on Edge Devices. *arXiv preprint arXiv:2408.10161*, 2024. 4

[47] Z. Zheng, H. Huang, T. Yu, H. Zhang, Y. Guo, and Y. Liu. Structured Local Radiance Fields for Human Avatar Modeling. In *CVPR*, pages 15893–15903, 2022. 1

[48] Z. Zheng, X. Zhao, H. Zhang, B. Liu, and Y. Liu. AvatarReX: Real-time Expressive Full-body Avatars. *ACM ToG*, 42(4):1–19, 2023. 1