HFGaussian: Learning Generalizable Gaussian Human with Integrated Human Features

Arnab Dey[1][*]   Cheng-You Lu[2]   Andrew I. Comport[1]   Srinath Sridhar[3]   Chin-Teng Lin[2]

Jean Martinet[1]

[1]I3S-CNRS/Université Côte d'Azur   [2]University of Technology Sydney   [3]Brown University

Abstract

Recent advancements in radiance field rendering show promising results in 3D scene representation, where Gaussian splatting-based techniques emerge as state-of-the-art due to their quality and efficiency. Gaussian splatting is widely used for various applications, including 3D human representation. However, previous 3D Gaussian splatting methods either use parametric body models as additional information or fail to provide any underlying structure, like human biomechanical features, which are essential for different applications. In this paper, we present a novel approach called HFGaussian that can estimate novel views and human features, such as the 3D skeleton, 3D key points, and dense pose, from sparse input images in real time at 25 FPS. The proposed method leverages generalizable Gaussian splatting technique to represent the human subject and its associated features, enabling efficient and generalizable reconstruction. By incorporating a pose regression network and the feature splatting technique with Gaussian splatting, HFGaussian demonstrates improved capabilities over existing 3D human methods, showcasing the potential of 3D human representations with integrated biomechanics. We thoroughly evaluate our HFGaussian method against the latest state-of-the-art techniques in human Gaussian splatting and pose estimation, demonstrating its real-time, state-of-the-art performance.

1. Introduction

Generating virtual photorealistic 3D human avatars is a long-standing challenge in the field of computer vision [1, 23]. These 3D models have diverse applications in fields such as augmented reality, virtual reality [32], entertainment, and the medical domain [6, 53]. The task of reconstructing a complete 3D human model with integrated structural properties [5, 76] in real time from images alone presents significant challenges. Classical approaches rely on complex multiview capture systems and body markers [24, 48] to obtain 3D models of humans, incorporating structural properties such as 3D pose involve fitting parametric body models such as SMPL [34] and STAR [43]. However, these methods require substantial resources and computational effort to generate each 3D model.

In recent years, radiance field rendering becomes significantly popular [55, 63] for the scene representation capabilities. More recently, 3D Gaussian splatting (3DGS) [29] provides a new research direction and demonstrates notable improvements compared to neural rendering-based methods. 3D Gaussian splatting proposes a novel explicit representation that represents the scene using a set of 3D Gaussians for point-based rendering. The efficient representation of Gaussian splatting makes it particularly well-suited for real-time rendering applications. Subsequent researches apply 3DGS to various applications [10, 37, 62] including 3D human reconstructions [22, 73]. The existing methods employing 3D Gaussian Splatting (3DGS) for human avatar reconstruction either rely on parametric body models or fail to incorporate any underlying biomechanical features crucial for downstream applications [20, 65].

In this work, we propose a novel, generalizable approach for estimating a 3D human representation with integrated 3D pose and dense pose in real time, given sparse input images of the human subject. The proposed method, named Human Feature Gaussian (HFGaussian), uses Gaussian splatting to represent the human subject and its associated biomechanical features[1], which include the 3D skeleton, 3D keypoints, and dense pose. These biomechanical properties are essential for recreating natural human movements and interactions in the virtual world [15, 27]. One straightforward approach to representing human features while maintaining real-time rendering speed, is to directly parameterize the 3D Gaussian with additional human features. However, we point out that simply parameterizing

---

[*]Contact email: adey@i3s.unice.fr

[1]In this study, "biomechanical features" refer to components of the human musculoskeletal system, such as bones, muscles, ligaments, and joint locations, which are critical for human movement and function.

the 3D Gaussian with these human features results in sub-optimal performance, as the same parameters like opacity, scaling, and rotation factor are not suitable for different human features. Instead of directly parameterizing the 3D Gaussian with human features, inspired by feature splatting [36], we learn these human features by optimizing additional feature parameters for each 3D Gaussian, which are then decoded into human features after rendering.

Regarding 3D pose estimation, we find that even a subset of 3D Gaussians can serve as an effective point cloud for 3D pose estimation using a novel pose regression network based on DGCNN [60] and PointNet [50].

In conclusion, we propose HFGaussian, a human-centric Gaussian framework that enables real-time representation of human features through 3D Gaussians. Using GPS-Gaussian [73] as the backbone, HFGaussian can generalize to unseen human data without any fine-tuning. Building on this foundation, we additionally introduce feature splatting [36] to overcome the performance constraints of using the same set of Gaussians for various human features which may have different frequencies. Furthermore, HFGaussian employs a novel pose regression network to estimate the 3D pose from a subset of the 3D Gaussians, ensuring efficient estimation. HFGaussian is capable of simultaneously rendering novel poses, corresponding 3D poses, and human features in real time. Although this study focuses on human pose estimation, we believe that the HFGaussian can be extended to include other human features, such as body part segmentation.

To evaluate our proposed method, we train our method in a large amount of human data generated from human scans and evaluate in real-world data. To the best of our knowledge, this is the first method to estimate 3D humans with biomechanics features and 3D pose in real time directly from images. The contributions of this work can be summarized as follows:

- We present a novel generalizable approach named HF-Gaussian that is capable of estimating human features and 3D human pose.
- The proposed method has demonstrated its ability to estimate 3D pose using a novel pose regression network and human features using feature splatting.
- We propose a generalizable approach to predicting human features, 3D pose, photometric, and geometric representations from 2D sparse images in real time.
- Our extensive experimental analysis across 3 datasets validates the applicability and versatility of our method.

2. Related works

The proposed method HFGaussian uses the Gaussian splatting technique to estimate 3D human avatars with integrated biomechanical features in real time from sparse mul-

tiview images. In this section, we review the previous studies relevant to this research.

2.1. Radiance field rendering

Radiance field rendering-based techniques become popular in recent years because of the photorealistic scene representation capability. NeRF [38], introduced in 2020, proposes a coordinate-based neural network to represent a 3D scene. The neural networks based on MLP map 3D coordinates and 2D view directions into density and color. Several follow-up works [3, 4, 44] are proposed and achieved impressive results, further verifying the capability. In addition, further studies are conducted to address the limitations, including long training [12, 16, 33, 41] and inference time [40, 51], scene specificity [66], and static scene constraints [49, 54]. In recent years, Gaussian splatting [29] techniques emerge as an alternative to implicit neural radiance fields by utilizing a set of 3D Gaussians to learn fast explicit scene representations.

2.2. Radiance field rendering for human

Radiance field rendering techniques demonstrate promising results in various applications for 3D human representation. Early studies [54, 64, 71] employ neural radiance fields to produce 3D human avatars from a spare set of images. Subsequent studies [9, 21, 25, 64, 68] improve the generalizability of these models by incorporating the SMPL [34] parameters as input. Likewise, [26, 46, 54, 61] utilize existing skeletal data, state-of-the-art pose estimators, or pose data to generate novel views and poses.

Recently, thanks to the fast rendering speed of Gaussian splatting techniques, radiance field-based methods [22, 30, 73] have been able to represent 3D humans in real-time. More recently, [13] proposes generalized radiance fields for versatile human features that extend beyond RGB rendering by integrating additional human features, although the rendering speed is not real time. Our approach stands out from prior works by combining generalizable Gaussian splatting techniques with feature splitting, which maintains the quality of both low- and high-frequency features, and a dedicated pose regression network to estimate 3D human avatars with integrated biomechanical features in real time.

2.3. Human pose estimation

Human features such as 3D pose and dense pose estimation are a long-standing problem in computer vision research. Many popular 2D and 3D pose estimation from images is based on supervised training. Algorithms for 2D pose estimation [7, 11, 14, 19, 31] employ 2D CNN architecture to estimate 2D poses from images. Works such as [39, 42, 52, 58] employ person detectors to estimate the poses of multiple persons. Bottom-up approaches like [7, 11, 31] identify joints using heatmaps and link body parts, but face

challenges with occluded or partially visible body parts. Techniques for estimating 3D poses can be classified into direct methods and 2D-to-3D lifting methods. [39, 52] concentrate on determining 3D poses directly from images. In this paper, we introduce a new method for directly predicting 3D poses. [17] introduce DensePose estimation from 2D images. Recent techniques [18, 59] employ a multitask learning approach for DensePose estimation. In this study, we present a novel approach for learning DensePose estimation by employing the Gaussian splatting method.

3. Method

We present HFGaussian, a unified framework that leverages Gaussian splatting to estimate human features and 3D pose in real-time.

3.1. Preliminary

In this study, we present a novel method for reconstructing the 3D human model while incorporating biomechanical characteristics through the use of Gaussian splatting. Here, we briefly outline the fundamental concepts behind 3D Gaussian splatting [29] and generalizable Gaussian splatting [73] techniques. 3DGS is introduced as an alternative explicit representation in contrast to continuous NeRF-based approaches. A scene in 3DGS is represented through a collection of 3D Gaussians characterized by certain properties: the 3D position of Gaussian μ, color c represented using spherical harmonics, opacity α, and a 3D covariance matrix Σ. A 3D Gaussian in 3DGS is represented as:

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

The covariance matrix Σ can be broken down into a rotation matrix $\mathbf{R}$ and a scaling matrix $\mathbf{S}$. The 3D Gaussians are projected into 2D space using a view transformation matrix $\mathbf{W}$ and Jacobian of an affine approximation of the projective transform $\mathbf{J}$. The 2D covariance matrix Σ' is represented as $\Sigma' = \mathbf{JW}\Sigma\mathbf{W}^T\mathbf{J}^T$. The alpha blending technique similar to NeRF, is used to rendering final pixel colors from Gaussians:

$$C = \sum_{i \in \mathcal{N}} c_i \alpha'_i \prod_{j=1}^{i-1}(1 - \alpha'_j)$$

where c_i represents the learned color, α'_i denotes the result of the multiplication between the opacity α_i and the 2D Gaussian. The 3DGS technique demonstrates notably faster rendering speeds compared to continuous methods based on NeRF, primarily because it can directly project and blend 3D Gaussian into color.

Although 3DGS is efficient and produces high-quality results, vanilla 3DGS is scene-specific and is not generalizable to new scenes. To address this issue, GPS-Gaussian [73] proposes Gaussian parameter maps on the

source views and directly estimates instant novel views. They focus only on human subjects and train on a large amount of human data. Given sparse source views and a novel target view I_{tar}, GPS-Gaussian selects 2 neighboring views I_r and I_l of the target view. The source views are then passed through an image encoder ε_{img} to extract dense feature maps $f^s \in \mathbb{R}^{H/2^s \times W/2^s \times D_s}$, corresponding to each source image. Using the feature maps from each source view (f_r^s, f_l^s), a 3D correlation volume C is generated. This correlation volume along with the corresponding camera parameters for the source views (K_r, K_l) is used to iteratively estimate depth maps. It can be formulated as:

$$< \mathbf{D}_l, \mathbf{D}_r > = \phi_{depth}(f_l^s, f_r^s, K_l, K_r),$$

where ϕ_{depth} represents the depth estimation module. Later, these depth estimations are used to generate the position of 3D Gaussians.

To predict the Gaussian parameters, they employ a mapping function that formulates the 3D Gaussian from the 2D image plane. When given a foreground pixel coordinate x in the image plane, the Gaussian map can be represented as:

$$\mathbf{G}(x) = \{\mathcal{M}_p(x), \mathcal{M}_c(x), \mathcal{M}_r(x), \mathcal{M}_s(x), \mathcal{M}_\alpha(x)\}$$

The previously estimated depth and camera projection matrix can be utilized to unproject the pixel coordinates x from the image plane to the 3D coordinates, represented as $\mathcal{M}_p(x)$. The color map uses the source image color directly: $\mathcal{M}_c(x) = I(x)$. To estimate the remaining Gaussian parameters, they first generate depth features using a depth encoder. These features are then combined with the image feature and passed through a U-Net like decoder to generate pixel-wise Gaussian features:

$$\Gamma = D_{parm}(\epsilon_{img}(I) \oplus \epsilon_{depth}(I))$$

, where $\oplus$ represents the concatenation operation. Finally, three different prediction heads are used to generate the remaining Gaussian feature maps $\mathcal{M}_r$, $\mathcal{M}_s$, and $\mathcal{M}_\alpha$ represent the rotation, scaling, and opacity head, respectively. The entire method is end-to-end differentiable and optimized using photometric and depth loss.

3.2. Learning human feature with 3DGS

We propose a novel method for real-time 3D human avatar estimation with biomechanic properties. Although recent advancements in 3D Gaussian splatting techniques [29] show remarkable efficiency in 3D scene representation compared to previous approaches, the vanilla method still requires per-subject optimization for scene representation. To address this limitation, GPS-Gaussian [73] introduces a generalizable approach for 3D human representation that takes advantage of 2D Gaussian parameter maps to estimate 3D Gaussians. This approach enables the
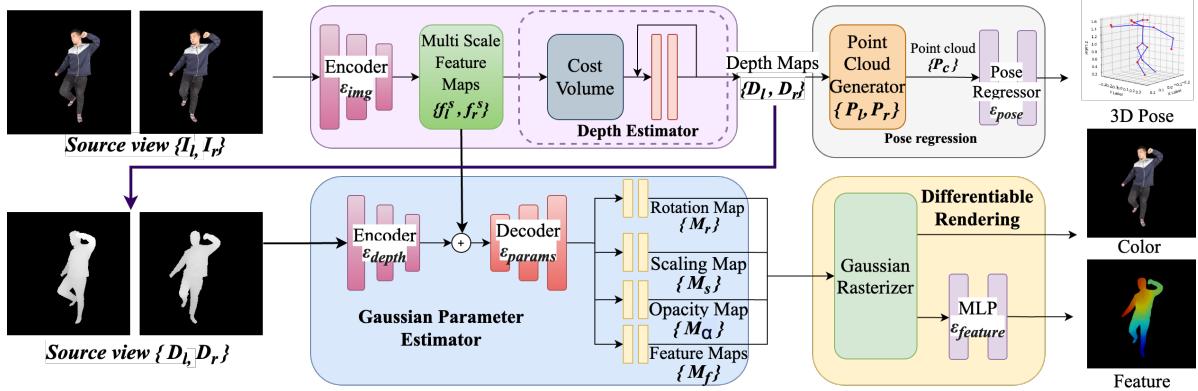
Figure 1. **The HFGaussian pipeline:** Given a target view, the nearest source views I_l and I_r are selected, and passed through an image encoder ϵ_{img} to generate feature maps f_l^s and f_r^s for depth maps D_l and D_r estimation. The depth maps are then encoded using a ϵ_{depth} encoder and combined with the image features before passing through a U-Net based decoder ϵ_{params} to predict Gaussian feature maps $\mathcal{M}_r$, $\mathcal{M}_s$, $\mathcal{M}_\alpha$, and $\mathcal{M}_f$. Finally, the predicted Gaussians are splatted and rasterized to generate the novel view and human features, which are further processed by a smaller MLP $\epsilon_{feature}$ to obtain the final human features.

direct regression of 3D Gaussian properties, facilitating instant novel view synthesis without the need for fine-tuning or optimization.

Building on this, we propose to extend the capabilities of 3D Gaussian splatting by learning a generalizable representation of 3D humans with integrated biomechanic properties. We develop a novel architecture that can predict 3D pose and human features along with photometric and geometric representation for novel views. Similarly to GPS-Gaussian, given a target view, we select the two nearest source views, I_l and I_r, which are RGB images ($H \times W$) corresponding to the left and right views. These two views are then fed into a shared image encoder ϵ_{img} to generate multiscale image features f_l^s and f_r^s where s is the feature scale.

From the feature maps (f_l^s, f_r^s) of each source view, a cost volume C is generated by correlating the two feature maps. Then, an iterative update mechanism is used to estimate depth maps (D_l, D_r) corresponding to each source view. These estimated depth maps are subsequently used as input to the pose regression network and Gaussian parameter estimator, which are discussed in Sec. 3.3 and 3.4. The pose regression network is capable of outputting the 3D pose of human subjects.

The scene is represented using a set of optimized 3D Gaussians, where each Gaussian is characterized by $G = \{X, c, r, s, \alpha, f\}$ where X denotes the 3D position, c represents the color, r signifies the rotation, s corresponds to the scaling, α indicates the opacity, and f encodes the human feature. We estimate the 3D Gaussian parameters on 2D planes using Gaussian parameter maps represented as:

$$G(x) = \{\mathcal{M}_p(x), \mathcal{M}_c(x), \mathcal{M}_r(x), \mathcal{M}_s(x), \mathcal{M}_\alpha(x), \mathcal{M}_f(x)\}$$

where x denotes the coordinates of a foreground pixel within the image plane, $\mathcal{M}_p$, $\mathcal{M}_c$, $\mathcal{M}_r$, $\mathcal{M}_s$, $\mathcal{M}_\alpha$, and $\mathcal{M}_f$ represent the Gaussian parameter maps corresponding to position, color, rotation, scaling, opacity, and feature, respectively. The $\mathcal{M}_p$ function maps the 2D image pixel coordinates x to the 3D space by using the predicted depth information and the known camera parameters. The $\mathcal{M}_c$ function directly uses the RGB color values from the source images.

To estimate the remaining four Gaussian parameters, we use an encoder ϵ_{depth} to encode depth maps and then employ a U-Net-like decoder ϵ_{params} to generate a Gaussian feature Γ from depth encoding and image encoding. Finally, we use four separate prediction heads implemented with convolutional layers to estimate Gaussian parameters. The prediction head for the human feature is defined as $\mathcal{M}_f = Sigmoid(h_f(\Gamma(x)))$, where h_f represents the feature head.

3.3. Pose regression Network

A key component of the architecture is the pose regression network, which aims to estimate the 3D pose of the human subject from the partial point cloud data generated from depth maps (D_l, D_r), also referred to as the subset of the 3D Gaussians. Estimation of 3D human keypoints from point cloud data is an active research area, with prior work exploring various approaches to address these challenges [8,69,75]. As mentioned in the previous section, the depth estimation models predict depth maps (D_l, D_r) corresponding to each source view (I_l, I_r). These depth maps of the source views and their associated camera parameters can be used to generate point clouds P_l, P_r. The 2D masks of each source view are then used to extract only the point cloud of the human. The point clouds from both views are then combined as they are in the same 3D frame to generate
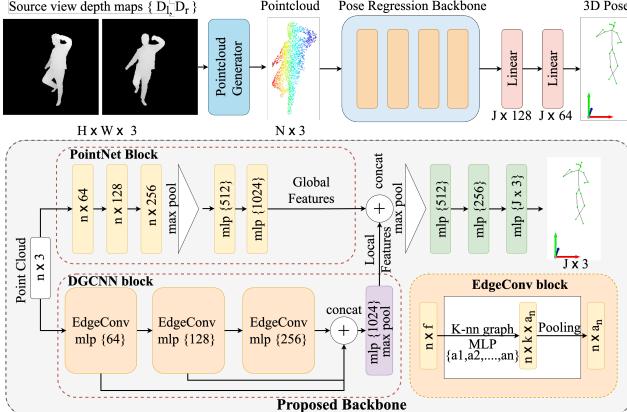
4

Figure 2. **Pose Regression Network Overview:** The network takes point clouds generated from depth maps as input and outputs 3D poses. We compare three point cloud classification backbones and propose a novel architecture combining PointNet and DGCNN architecture for robust feature extraction.

a combined point cloud. From this combined point cloud, 2048 points are randomly sampled, and then this sampled point cloud P_c is used as input to the pose regression network. For the pose regression network, we experiment with three different popular backbone architectures, namely PointNet [50], DGCNN [60], and Point Transformer [72], which are well known for point cloud classification. In this work, we compare the performance of all three models on 3D and 2D pose estimation, using their corresponding classification backbone with added MLP layers at the end for pose estimation, as shown in figure 2. The PointNet architecture used in this work is a customized version based on the implementation described in [8]. In this work, we introduce a novel pose regression backbone that combines global features from PointNet architecture and local features from DGCNN architecture while maintaining computational efficiency similar to PointNet for real-time inference. The proposed architecture for the pose regression network is shown in Figure 2. The proposed architecture achieves comparable performance to the more complex Point Transformer network yet remains as efficient as PointNet, making it suitable for our real-time applications. As the proposed model learns both global and local features jointly, it can provide a more robust pose estimation that is resilient to noisy and incomplete point cloud data.

3.4. Human feature estimation

The proposed method also estimates human features. We demonstrate the capabilities of our approach by estimating dense pose, which involves predicting Continuous Surface Embeddings for the human subject. We include an additional branch in the Gaussian parameter estimator to predict human feature maps $\mathcal{M}_f$ for each Gaussian, similar to

the rotation and opacity maps. Inspired by feature splatting [36], we apply a similar technique that estimates human feature vectors f_p by splatting Gaussian features f_i in the image plane, and then blending the feature vectors using alpha composition:

$$f_p = \sum_{i \in \mathcal{N}} f_i \alpha_i' \prod_{j=1}^{i-1} (1 - \alpha_j')$$

The blended feature vectors f_p are decoded by a MLP consisting of two linear layers with ReLU activation functions, followed by a final layer with a sigmoid activation function, to render the continuous surface embeddings.

3.5. Optimization

The proposed HFGaussian method comprises three key components: generalizable Gaussian splatting, 3D pose estimator, and human feature estimator module. The model is trained in two stages. First, the depth estimator module is trained on both source views. Then, the Gaussian parameter estimator module is trained using the depth maps and image features, along with the feature estimator MLP and the 3D pose estimator. A combined loss function is utilized to train all three parts simultaneously, as described:

$$\mathcal{L} = \mathcal{L}_{image} + \mathcal{L}_{depth} + \mathcal{L}_{pose} + \mathcal{L}_{feature}$$

where $\mathcal{L}_{image}$ is the photometric loss between the ground truth and the rendered image represented as $\mathcal{L}_{image} = \beta \mathcal{L}_{mae} + \gamma \mathcal{L}_{ssim}$ where β and γ are 1.6 and 0.4 respectively. The depth loss $\mathcal{L}_{depth}$ is defined as:

$$\mathcal{L}_{\text{depth}} = \sum_{t=1}^{T} \mu^{T-t} \|\mathbf{d}_{gt} - \mathbf{d}^t\|_1$$

where d represents the depth and μ is set to 0.9 for our experiments. The 3D pose estimation loss $\mathcal{L}_{pose}$ is the L2 loss between ground truth and estimated 3D keypoints. Lastly, the feature estimation loss $\mathcal{L}_{feature}$ is the L1 loss between the ground truth and the predicted human features, which in this case are the continuous surface embeddings.

4. Experimental Results

We conduct a comprehensive evaluation of our model's ability to learn a generalized Gaussian representation of humans, including both 3D human poses and features. Extensive experiments are conducted on a variety of datasets, and we evaluate our results with other state-of-the-art NeRF and Gaussian splatting-based methods.

4.1. Implementation details

The proposed HFGaussian method is implemented using the PyTorch framework and the AdamW optimizer [35]

5

with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5}. First, the depth estimation module is trained for 40,000 iterations, and then all three components of the network are jointly trained for 100,000 iterations with a batch size of 4 for all experiments. The complete training process takes approximately 14 hours on the dataset proposed in this study. For all experiments, we utilize the official versions of GHNeRF [13] and ENeRF [33], training them for 100,000 iterations. Details about the evaluation metrics can be found in the supplementary Section 3.

4.2. Dataset

In this work, we create a custom dataset similar to the one utilized in GPS-Gaussian [73]. The GPS-Gaussian dataset is created using 526 scans from the THuman2.0 [67] dataset and includes images, masks, depth information, and camera parameters. The dataset is insufficient for this study and required additional ground-truth data on various human features, such as keypoints and dense pose, in order to train the proposed method. To this extent, we extend the GPS-Gaussian dataset by generating additional ground-truth information. We start with 526 human scans and SMPLX [34] parameters from THuman2.0. Next, we employ the SMPLX model fitted to the scan to produce accurate 3D poses. We produced 19 key points representing the main joints of the human body (*nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, pelvis, right hip, right knee, right ankle, left hip, left knee, left ankle, right eye, left eye, right ear, left ear*). To produce the 2D keypoints, we map the 3D keypoints into image coordinates utilizing the camera's parameters. We use DensePose [17] to generate ground-truth Continuous Surface Embeddings for each image in the dataset. To render images from human scans, we follow the same convention of GPS-Gaussian, where 8 images are rendered in a circle around the human at 45-degree intervals, and those are used as source views. Additionally, 3 random viewpoints are generated as target views. We generate all images in the dataset at a resolution of 512×512. We also use the real-world dataset captured by [73] to evaluate our methods. This real dataset does not provide any ground-truth 3D or 2D keypoint information. To further evaluate the generalization ability of our model, we also preprocess and utilize the THuman4.0 dataset [74], which contains three clips of real-world subjects, resulting in test sets with 19656, 40464, and 24880 samples, respectively. We will release the preprocessed THuman2.0 and THuman4.0 after the paper is published.

4.3. Baseline

The proposed HFGaussian method is compared with other state-of-the-art generalizable approaches for human subjects. Regarding the benchmarking of human features and pose estimation, many previous generalizable tech-

niques lack the ability to estimate human features and pose simultaneously. We select GHNeRF [13] as a baseline, as it can estimate human features such as 2D keypoints and dense pose, although not in real time.

4.4. Novel view synthesis

Our proposed methods are compared with other state-of-the-art generalizable radiance field rendering techniques for human representation. The quantitative results for novel
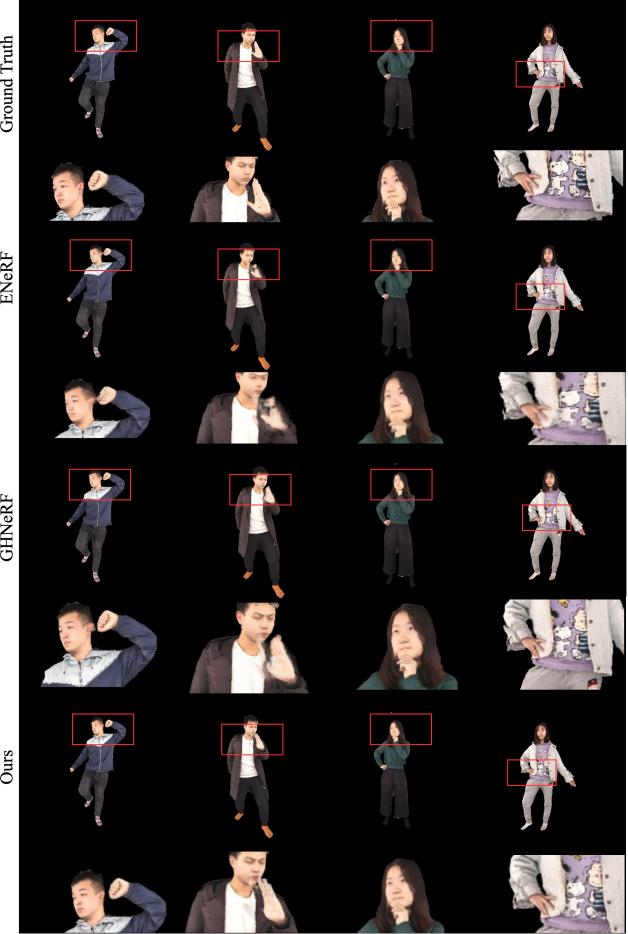

Figure 3. Qualitative comparison of novel view synthesis results on THuman2.0 test set.

view synthesis presented in Table 1 indicate that our methods are competitive with other state-of-the-art approaches while also estimating human 3D joints and poses, going beyond mere RGB image synthesis. Among all the methods evaluated, GPS-Gaussian achieves the highest PSNR of 32.55. Our method closely matches this performance, achieving a PSNR of 32.43, while also estimating other human features simultaneously. The qualitative results for novel view synthesis are shown in Figure 3.

To demonstrate the generalization capability of the proposed method, we evaluate our pre-trained model directly

	Image			Dense Pose	3D Pose	2D Pose	Inference Time
Method	PSNR ↑	SSIM ↑	LPIPS ↓	MSE ↓	MPJPE ↓	PCK ↑	FPS ↑
ENeRF [33]	32.51	0.9823	0.0245	-	-	-	28.97
GPS-Gaussian [73]	32.55	0.9737	0.0300	-	-	-	32.02
GHNeRF [13]	32.98	0.9839	0.0210	0.0020	-	0.6767	11.22
Ours	32.43	0.9734	0.0303	0.0017	0.0704	0.8707	24.37

Table 1. Quantitative results for novel view synthesis on the THuman2.0 dataset. Photometric quality is assessed via PSNR, SSIM, and LPIPS metrics, while dense pose estimation accuracy is measured using mean square error (MSE).
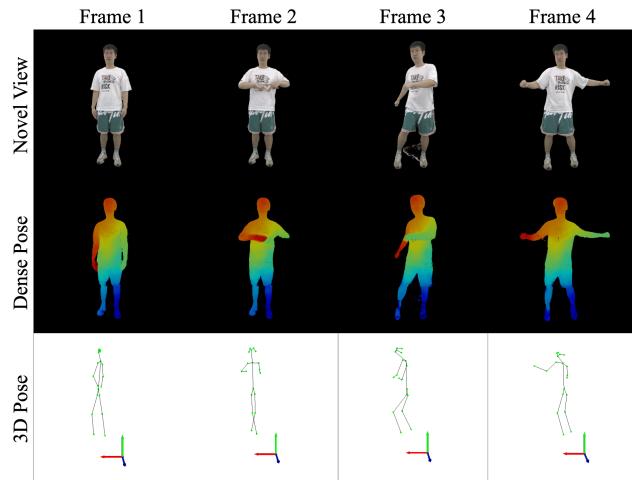

Figure 4. Qualitative result of the proposed method of real human dataset. The figure illustrates random frames from real human data and their corresponding novel view, dense pose, and 3D pose predicted by the HFGaussian. The 3D pose images are generated by projecting the 3D pose in a 2D plane from a fixed viewing angle.

on the real-world dataset by [73] and THuman4.0 real-world dataset without any fine-tuning. The qualitative results present in Figure 4 and Figure S4 demonstrate the robustness of our approach, generalizing unseen data in real time and accurately estimating biomechanical features. The quantitative results in Table S2 demonstrate that our approach is much more robust across datasets and outperforms all baselines. Furthermore, the results indicate that our method can reliably estimate biomechanical features even in challenging and self-occlusion scenarios in real time.

4.5. 3D/2D Pose Estimation

The proposed HFGaussian method estimates 3D keypoints using a pose regression network, which takes the point cloud generated by the position of 3D Gaussians as input. In contrast, many state-of-the-art approaches do not provide 3D keypoints or any human biomechanical features. Our method is the first to directly estimate 3D pose without relying on prior supervision or parametric body models. The quantitative results of our approach are presented in Table 2, and the qualitative results are shown in Figure S3. As

mentioned earlier, we experiment with different backbone architectures for the pose regression network. We then propose a novel architecture that combines elements of Point-Net [50] and DGCNN [60]. The results demonstrate that the proposed backbone architecture achieves the best MPJPE score among all backbones tested, as it effectively combines global and local features to provide robust 3D pose estimation. This is also evident in the visual comparison presented in Figure S3.

In addition to 3D pose estimation, we also evaluate 2D pose estimation using the same pose regression network, with a modified final linear layer for 2D keypoint prediction. We compare our 2D pose estimation performance with that of GHNeRF [13], which is also capable of estimating 2D keypoints. The quantitative results are presented in Table 1, and the qualitative comparisons are shown in Figure S1. The experiments demonstrate that our method significantly outperforms GHNeRF in the estimation of 2D key points.
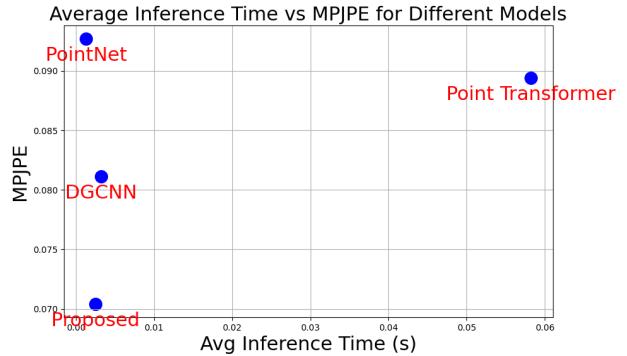

Figure 5. The figure illustrates the performance and average inference time of different 3D pose estimation backbone architectures. The 3D pose estimation performance is measured in terms of MPJPE, which is plotted on the y-axis. The x-axis represents the average inference time of the respective backbone models.

4.6. Dense Pose Estimation

The proposed HFGaussian method is capable of estimating various human features using the feature splatting technique discussed previously. To showcase the feature

7

	Image			Dense Pose	3D Pose	2D Pose
Method	PSNR ↑	SSIM ↑	LPIPS ↓	MSE ↓	MPJPE ↓	PCK ↑
Ours + PointNet + 3D pose	32.4087	0.9730	0.0303	0.00176	0.0927	-
Ours + DGCNN + 3D pose	32.4131	0.9732	0.0303	0.00174	0.0811	-
Ours + PointTransfomer + 3D pose	32.4141	0.9730	0.0303	0.00174	0.0894	-
Ours + Proposed pose regressor + 3D pose	32.4300	0.9734	0.0303	0.00173	0.0704	-
Ours + PointNet + 2D pose	32.4300	0.9728	0.0311	0.00175	-	0.8309
Ours + DGCNN + 2D pose	32.4020	0.9731	0.0304	0.00173	-	0.8707
Ours + PointTrans + 2D pose	32.4000	0.9729	0.0304	0.00177	-	0.7721
Ours + Proposed pose regressor + 2D pose	32.4300	0.9731	0.0304	0.00176	-	0.8680

Table 2. The table shows quantitative results of our approach on the Thuman2.0 dataset. The results are categorized into four groups: novel view, dense pose estimation, 3D pose estimation, and 2D pose estimation.

prediction capabilities of the proposed approach, we conduct a dense pose estimation task as an illustrative example. Specifically, we estimate the Continuous Surface Embedding to represent the dense pose. The quantitative results, presented in Table 1, compare the dense pose estimation performance of our method against the GHNeRF baseline. Furthermore, the qualitative results, depicted in Figure S2, further demonstrate that our method outperforms GHNeRF in learning the dense pose representation. Additionally, we showcase our model's ability to estimate dense pose on real-world data, as illustrated in Figure 4.

4.7. Real Time Performance

One of the primary objectives of this work is to achieve real-time performance during inference and estimate novel views with biomechanical features on unseen data. We have compared the real-time performance of our method with GPS-Gaussian [73] and GHNeRF [13], among which only GHNeRF is capable of estimating additional human features such as 2D keypoints. The results presented in Table 3 demonstrate the performance of different methods in terms of frames per second. In our method, we experiment with various backbones for 3D pose estimation. The relationship between the different backbones, the inference speed, and the precision is shown in Figure 5. The figure indicates that the proposed architecture for the 3D pose estimation backbone achieves the best balance between speed and accuracy.

Method	FPS
GHNeRF + ResNet	11.22
GHNeRF + DINO	4.08
GPS-Gaussian	32.02
Ours + PointNet	25.07
Ours + DGCNN	24.18
Ours + Point Transformer	10.29
Ours + Proposed	24.37

Table 3. Average rendering speed in FPS(Frame per second). GPS-Gaussian represent the baseline method.

4.8. Ablation Studies

To assess the importance of feature splatting for estimating human features, we conduct an ablation study. We compare the performance of our method in estimating dense pose, a key human feature, using feature splatting versus directly predicting feature values similar to RGB color. In both cases, we keep the other Gaussian parameters constant. The qualitative results presented in Figure 6 demonstrate that the method without feature splatting is unable to accurately estimate dense pose, as the same Gaussians are unable to effectively represent both RGB values and feature attributes simultaneously.
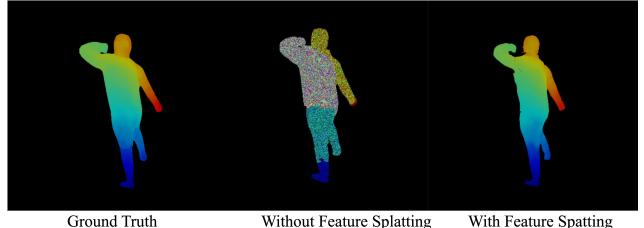

Ground Truth   Without Feature Splatting   With Feature Spatting

Figure 6. The significance of feature splatting methods in estimating human features.

5. Conclusion

This paper presents a novel framework, HFGaussian, for the real-time rendering of human avatars with biomechanical properties. The proposed method utilizes Gaussian splatting to generate novel views from sparse source views of human subjects in real time. Extensive experimentation demonstrates the effectiveness of this approach, which represents a significant improvement over previous Gaussian splatting-based methods for human representation. While the primary focus of this work is on 3D human pose estimation and feature splatting-based dense pose estimation, we believe that feature splatting can be leveraged to learn other human-centric features, such as body part segmentation. In summary, this work presents promising results and opens up new avenues for research in 3D human representation.

References

[1] Kairat Aitpayev and Jaafar Gaber. Creation of 3d human avatar using kinect. *Asian Transactions on Fundamentals of Electronics, Communication & Multimedia*, 1(5):12–24, 2012. 1

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *CVPR*, pages 5855–5864, 2021. 2

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVRP*, 2022. 2

[5] JM Buades, FJ Perales, M Gonzalez, Antoni Aguiló, and P Martinez. Human body segmentation and matching using biomechanics 3d models. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pages 79–84. IEEE, 2004. 1

[6] Yusuf Ozgur Cakmak, Ben Kei Daniel, Niels Hammer, Onur Yilmaz, Erdem Can Irmak, and Prashanna Khwaounjoo. The human muscular arm avatar as an interactive visualization tool in learning anatomy: medical students' perspectives. *IEEE Transactions on Learning Technologies*, 13(3):593–603, 2020. 1

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2019. 2

[8] Jiaan Chen, Hao Shi, Yaozu Ye, Kailun Yang, Lei Sun, and Kaiwei Wang. Efficient human pose estimation via 3d event point cloud. In *2022 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2022. 4, 5

[9] Jianchuan Chen, Wentao Yi, Liqian Ma, Xu Jia, and Huchuan Lu. Gm-nerf: Learning generalizable model-based neural radiance fields from multi-view images. In *CVPR*, 2023. 2

[10] Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. Text-to-3d using gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21401–21412, 2024. 1

[11] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 2

[12] A. Dey, Y. Ahmine, and A.I. Comport. Mip-NeRF RGB-D: Depth Assisted Fast Neural Radiance Fields. *Journal of WSCG*, 30:34–43, 2022. 2

[13] Arnab Dey, Di Yang, Rohith Agaram, Antitza Dantcheva, Andrew I. Comport, Srinath Sridhar, and Jean Martinet. Gh-nerf: Learning generalizable human features with efficient neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2812–2821, June 2024. 2, 6, 7, 8, 4

[14] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 2

[15] Luca Fortini, Mattia Leonori, Juan M Gandarias, Elena De Momi, and Arash Ajoudani. Markerless 3d human pose tracking through multiple cameras and ai: Enabling high accuracy, robustness, and real-time performance. *arXiv preprint arXiv:2303.18119*, 2023. 1

[16] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 2

[17] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3, 6

[18] Yuyu Guo, Lianli Gao, Jingkuan Song, Peng Wang, Wuyuan Xie, and Heng Tao Shen. Adaptive multi-path aggregation for human densepose estimation in the wild. In *Proceedings of the 27th ACM International conference on multimedia*, pages 356–364, 2019. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[20] Karl H Höhne, Henry Fuchs, and Stephen M Pizer. *3D imaging in medicine: algorithms, systems, applications*, volume 60. Springer Science & Business Media, 2012. 1

[21] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. *arXiv preprint arXiv:2303.12791*, 2023. 2

[22] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:2312.02973*, 2023. 1, 2

[23] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):1–14, 2015. 1

[24] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1, 2

[25] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16922–16932, 2023. 2

[26] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*. Springer, 2022. 2

[27] Yifeng Jiang, Tom Van Wouwe, Friedl De Groote, and C Karen Liu. Synthesis of biologically realistic human motion using joint torque actuation. *ACM Transactions On Graphics (TOG)*, 38(4):1–12, 2019. 1

[28] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, page 5. Aberystwyth, UK, 2010. 2

[29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 1, 2, 3

[30] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 505–515, 2024. 2

[31] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, 2019. 2

[32] Zhong Li, Lele Chen, Celong Liu, Yu Gao, Yuanzhou Ha, Chenliang Xu, Shuxue Quan, and Yi Xu. 3d human avatar digitization from a single image. In *Proceedings of the 17th International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 1–8, 2019. 1

[33] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 2, 6, 7

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34:248:1–248:16, 2015. 1, 2, 6

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[36] T Berriel Martins and Javier Civera. Feature splatting for better novel view synthesis with low overlap. *arXiv preprint arXiv:2405.15518*, 2024. 2, 5

[37] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 1

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Comm. of the ACM*, 65(1):99–106, 2021. 2

[39] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2019. 2, 3

[40] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[41] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H Mueller, Chakravarty R Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In *Computer Graphics Forum*, volume 40, pages 45–59. Wiley Online Library, 2021. 2

[42] G. Ning, Z. Zhang, and Z. He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE TMM*, 2017. 2

[43] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 598–613. Springer, 2020. 1

[44] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *CVPR*, pages 5865–5874, 2021. 2

[45] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[46] Sida Peng, Chen Geng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Implicit neural representations with structured latent codes for human body modeling. *IEEE TPAMI*, 2023. 2

[47] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2

[48] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4):120:1–120:14, Aug. 2015. 1

[49] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[50] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 5, 7

[51] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14335–14345, 2021. 2

[52] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE TPAMI*, 2019. 2, 3

[53] Vivek Singh, Kai Ma, Birgi Tamersoy, Yao-Jen Chang, Andreas Wimmer, Thomas O'Donnell, and Terrence Chen. Darwin: Deformable patient avatar representation with deep image network. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pages 497–504. Springer, 2017. 1

[54] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in neural information processing systems*, 34:12278–12291, 2021. 2

[55] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022. 1

[56] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[57] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 2

[58] Haoqian Wang, W. P. An, Xingzheng Wang, Lu Fang, and Jiahui Yuan. Magnify-net for multi-person 2d pose estimation. *ICME*, 2018. 2

[59] Xuanhan Wang, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Ktn: Knowledge transfer network for multi-person densepose estimation. In *Proceedings of the 28th ACM International conference on multimedia*, pages 3780–3788, 2020. 3

[60] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 2, 5, 7

[61] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16210–16220, 2022. 2

[62] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 1

[63] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, pages 1–30, 2024. 1

[64] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 2

[65] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13284–13293, 2021. 1

[66] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[67] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 6, 2

[68] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[69] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. Hum3dil: Semi-supervised multi-modal 3d humanpose estimation for autonomous driving. In *Conference on Robot Learning*, pages 1114–1124. PMLR, 2023. 4

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[71] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 2

[72] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 5

[73] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. *arXiv preprint arXiv:2312.02155*, 2023. 1, 2, 3, 6, 7, 8

[74] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 6, 2

[75] Yufan Zhou, Haiwei Dong, and Abdulmotaleb El Saddik. Learning to estimate 3d human pose from point cloud. *IEEE Sensors Journal*, 20(20):12334–12342, 2020. 4

[76] Xinxin Zuo, Sen Wang, Jiangbin Zheng, Weiwei Yu, Minglun Gong, Ruigang Yang, and Li Cheng. Sparsefusion: Dynamic human avatar modeling from sparse rgbd images. *IEEE Transactions on Multimedia*, 23:1617–1629, 2020. 1

Supplementary Material: HFGaussian: Learning Generalizable Gaussian Human with Integrated Human Features

1. Author Statement

In this supplementary material, we present information regarding the preprocessed dataset, evaluation metrics, additional visualization results, and analysis. We are committed to the ongoing maintenance and support of the preprocessed dataset. The dataset is distributed under an MIT license, allowing use, redistribution, and citation in accordance with the license terms.

2. Dataset

In this work, we extend the GPS-Gaussian dataset, which includes 526 scans from the THuman2.0 dataset, by generating additional ground-truth data, such as 3D poses, 2D keypoints, and DensePose embeddings. Starting with SMPLX parameters fitted to the scans, we extract 19 keypoints representing major human joints and project the 3D keypoints into 2D image coordinates using the camera parameters. Images are rendered from 8 source views at 45-degree intervals around each scan, along with 3 random target viewpoints, all at a resolution of 512x512. Additionally, we preprocess the THuman4.0 dataset, which includes three clips of real-world subjects captured by 24 cameras arranged in a circle. This results in test sets containing 19,656, 40,464, and 24,880 samples, where the two source views of each sample are also separated by 45 degrees. The key difference between our THuman2.0 and THuman4.0 datasets is that THuman4.0 is based on real-world RGB images with real-world camera poses, which may not form a perfect circle. This difference can verify the generalization ability of the proposed method (see supplementary Section 5).

Table S1 demonstrates that our THuman2.0 and THuman4.0 datasets cover a wide range of human features that are lacking in existing datasets. We plan to release our datasets after publication to encourage further development in this field.

3. Metrics

We use six distinct metrics to evaluate the quality of the predicted RGB images, 3D pose, and dense pose estimation. For RGB image reconstruction, we use the peak signal-to-noise ratio (PSNR) and Structural Similarity Index (SSIM) to compare the quality, with higher values indicating better quality. Additionally, the Learned Perceptual Image Patch Similarity (LPIPS) [70] metric is utilized to measure the similarity between image patches, with lower values indicating greater similarity between two patches. For dense pose estimation, the Mean Squared Error (MSE) is used to measure the distance between the ground truth and the pre-

dicted dense pose, with lower values indicating better quality. For 2D keypoints, we adopt the Percentage of Correct Keypoints (PCK) which measures the percentage of predicted 2D keypoints that are located within a certain distance threshold from the ground truth. Specifically, we set the distance to $0.2\times$ torso diameter (PCK@0.2). For 3D human pose estimation, we use the Mean Per Joint Position Error (MPJPE), which calculates the mean Euclidean distance between the estimated and actual 3D joint positions. Specifically, it is defined as:

$$\text{MPJPE} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{J}_i^{\text{pred}} - \mathbf{J}_i^{\text{gt}}\|_2, \qquad \text{(S1)}$$

where N is the total number of joints, $\mathbf{J}_i^{\text{pred}}$ is the predicted 3D coordinate of the i-th joint, $\mathbf{J}_i^{\text{gt}}$ is the ground truth 3D coordinate of the i-th joint, and $\|\cdot\|_2$ denotes the Euclidean distance (L2 norm). The MPJPE value is obtained by computing the Euclidean distance between the predicted and actual joint positions for each joint and then determining the average distance across all joints. Lower MPJPE indicates better performance because it indicates that the predicted joint positions are closer to the ground truth.

4. Limitations

Although the current method provides substantial advancements compared to the state-of-the-art, it is limited to single-human representations and is primarily applicable to human subjects. Future research can explore extending this method to handle multi-subject representations and developing more generalized models applicable to a broader range of articulated objects or animals.

5. Additional Results

Figure S1 shows the 2D keypoint results comparing GH-NeRF, the only existing method capable of rendering both RGB and human features simultaneously, with HFGaussian. HFGaussian is competitive and even outperforms GH-NeRF in certain cases (e.g., Subject 4), while still maintaining real-time rendering speed. For dense pose results, HFGaussian clearly outperforms GHNeRF in Figure S2. The left hand of Subject 1, both hands of Subject 2, the right hand of Subject 3, and the hands and waist of Subject 4 are much closer to the ground truth. This level of detail in the hands highlights the superior capability of HFGaussian in accurately capturing fine human features.

Figure S3 demonstrates that the proposed pose regression network can maintain good quality and rendering speed

Dataset	Multi-View	RGB	2D Keypoints	3D Keypoints	Dense Pose	Real
LSP [28]	✗	✓	✓	✗	✗	✓
MPII [2]	✗	✓	✓	✗	✗	✓
Smplify-X [45]	✗	✓	✓	✓	✓	✓
THuman2.0 [67]	✓	✓	✗	✗	✗	✗
THuman4.0 [74]	✓	✓	✗	✗	✗	✓
Human3.6M [24]	✓	✓	✓	✓	✗	✗
SURREAL [56]	✓	✓	✓	✓	✓	✗
3DPW [57]	✓	✓	✓	✓	✗	✓
ZJU-MoCap [47]	✓	✓	✓	✓	✗	✓
THuman2.0 (Ours)	✓	✓	✓	✓	✓	✗
THuman4.0 (Ours)	✓	✓	✓	✓	✓	✓

Table S1. Comparison between our preprocessed dataset and other human-centric datasets. In this table, Real denotes high-quality, realistic images of actual human scenes captured by physical cameras instead of virtual cameras.

Subject	Method	Image			Dense Pose	3D Pose	2D Pose
		PSNR ↑	SSIM ↑	LPIPS ↓	MSE ↓	MPJPE ↓	PCK ↑
S00	ENeRF [33]	31.121	0.978	0.029	-	-	-
	GPS-Gaussian [73]	32.505	0.977	0.0196	-	-	-
	GHNeRF [13]	29.117	0.969	0.049	-	-	0.3714
	Ours	33.610	0.981	0.017	0.0051	0.104	0.0018
S01	ENeRF [33]	29.204	0.966	0.041	-	-	-
	GPS-Gaussian [73]	30.324	0.966	0.0272	-	-	-
	GHNeRF [13]	27.805	0.962	0.050	-	-	0.4241
	Ours	30.968	0.968	0.026	0.0065	0.119	0.0029
S02	ENeRF [33]	27.993	0.968	0.044	-	-	-
	GPS-Gaussian [73]	28.560	0.967	0.0324	-	-	-
	GHNeRF [13]	26.165	0.964	0.050	-	-	0.3981
	Ours	29.051	0.967	0.033	0.0109	0.143	0.0041

Table S2. Quantitative results for novel view synthesis on the THuman4.0 dataset.

(see Figure 5). This further verifies the statement that running the pose regression network on a subset of the 3D Gaussian is sufficient.

Table S2 demonstrates that HFGaussian is more robust to unseen data. It outperforms all baselines by at least 0.5 PSNR across all subjects on RGB images, and shows a clear advantage over GHNeRF in capturing human features. Figure S4 and Figure S5 further highlight HFGaussian's robust generalization ability across all three subjects. Note that the ground truth 3D pose does not perfectly align with the predicted 3D pose due to the real-world camera poses not forming a perfect circle, which may cause slight shifts between them. However, the 3D poses are still relatively well-matched.
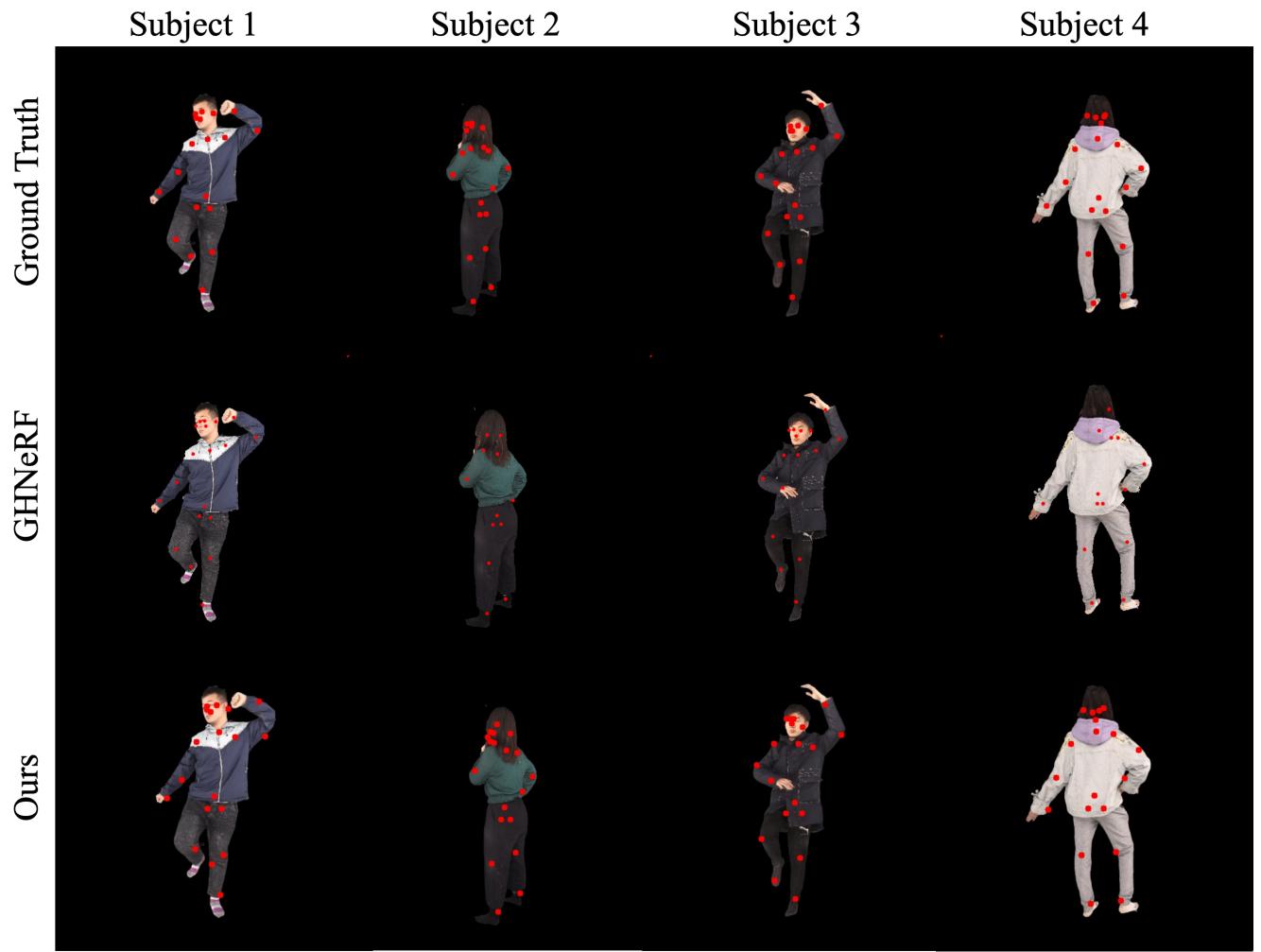
Figure S1. Comparison of qualitative results for the 2D keypoint estimation task on the THuman2.0 dataset.
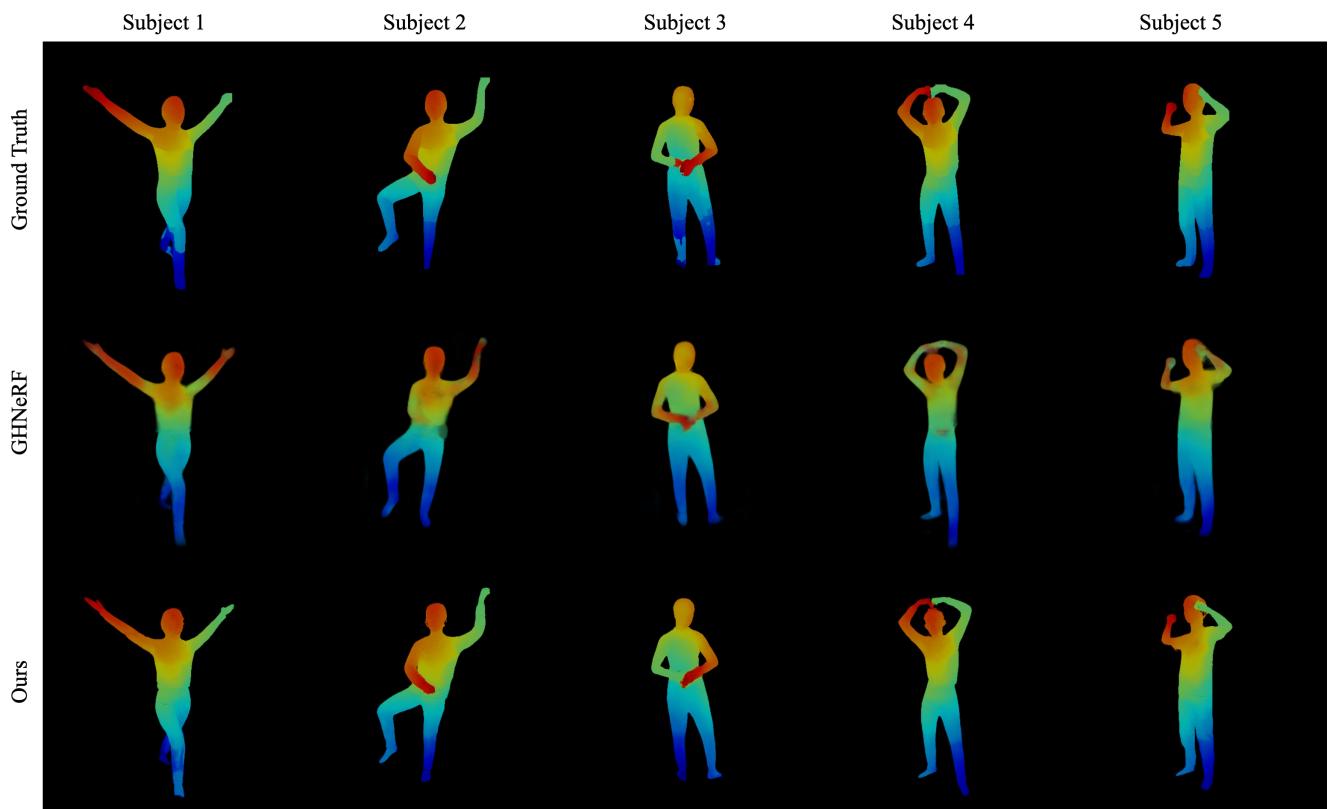
Figure S2. Visual comparison of dense pose estimation results. The findings indicate that our proposed method considerably outperforms GHNeRF [13] in dense pose estimation.
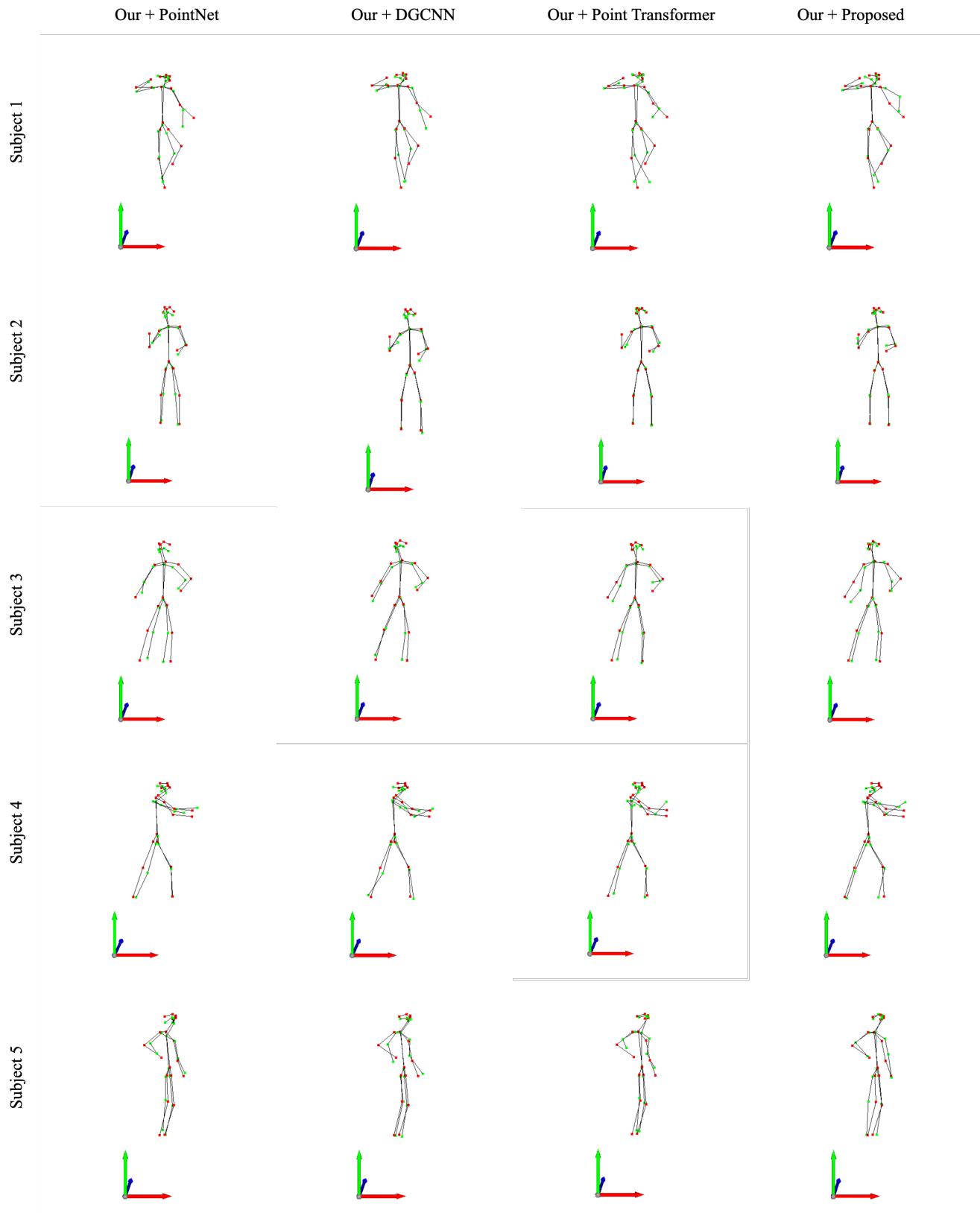
Figure S3. Qualitative results of 3D pose estimation using various pose regression network. The red markers show the actual 3D poses, while the green markers show the predicted 3D poses.
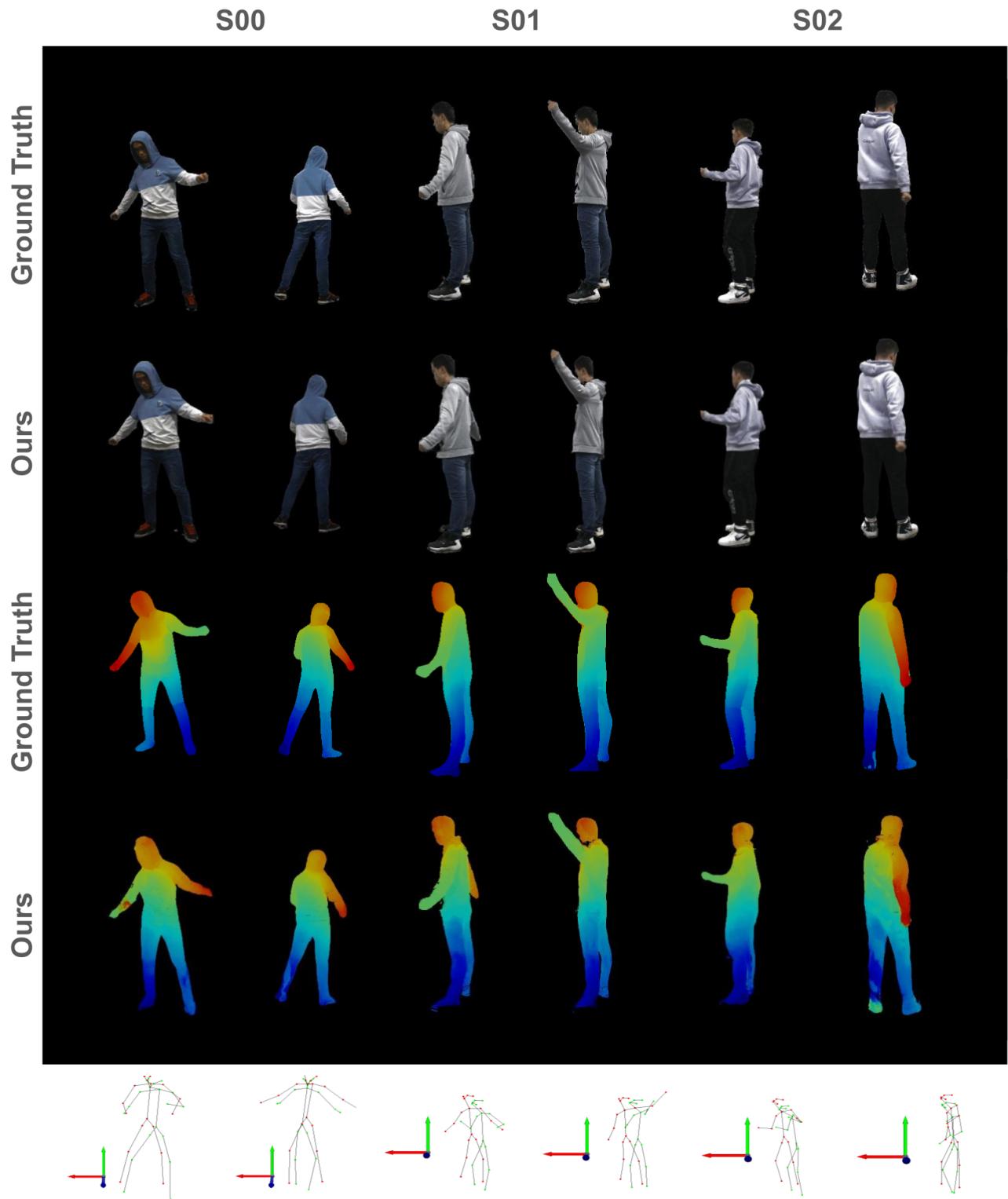
Figure S4. Qualitative results of HFGaussian on THuman4.0 dataset. For 3D pose estimation, the red markers show the actual 3D poses, while the green markers show the predicted 3D poses.
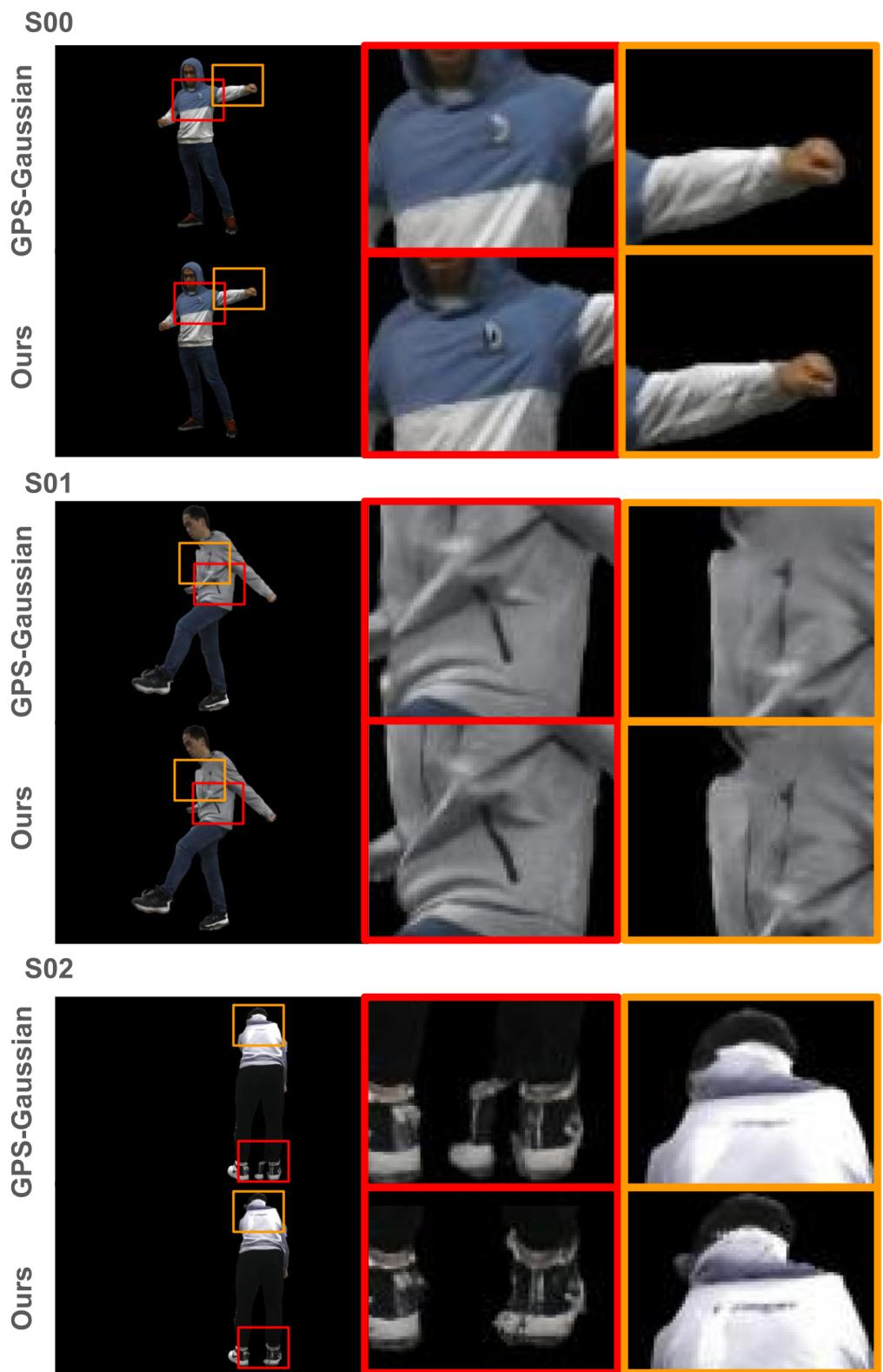
Figure S5. Qualitative results of HFGaussian and GPS-Gaussian on THuman4.0 dataset. GPS-Gaussian produces over-smoothed rendering results on the THuman4.0 dataset.