HGM: Human Gaussian to Mesh by Adaptive Optimization

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

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

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

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

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

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

Abstract

The rapid development of 3D Gaussian Splatting (3DGS) has positioned efficient mesh extraction from human-centered Gaussian representations as a critical challenge in digital human research. Current methods mainly target at rigid scenes, and still struggle to reconstruct high-quality meshes from vast, disordered human Gaussians because of the visual blind spot of 2D reconstruction and fundamental barrier of Gaussian representation. To address those two limitations, we propose HGM, a self-refinement based unified human mesh extraction framework based on the source Gaussians themselves. Centered on human geometry, the optimization loop in the framework integrates 3D structural information with 2D appearance cues through three key innovations: an epipolar-attention-guided diffusion lifter that enhances multi-view appearance refinement when integrated into standard Gaussian renderers; a geometry-aware depth estimator that generates accurate multi-view depth maps to improve reconstruction precision and detail preservation; and a depth and angle-guided adaptive optimizer that drives Gaussian ellipsoids to closely conform to the human surface for efficient extraction of smooth, detailed human meshes with remarkable speed and scalability. This joint self-optimization framework allows HGM to reconstruct fully editable, realistic human meshes in just minutes, outperforming existing methods in both fidelity and efficiency.

***Keywords:** Human Gaussian, 3D Gaussian Splatting, Diffusion, Extract meshes.*

1. Introduction

Reconstructing high-fidelity and photorealistic digital humans [9, 33, 39] has long been a critical research topic in computer vision and computer graphics, with broad applications in gaming, film production, and virtual/augmented reality. Recently, the emergence of 3DGS [20] has attracted significant attentions from researchers [36,40,52] due to the substantially higher rendering quality and speed of 3DGS than those of other techniques, such as Neural Radiance Fields (NeRF) [31]. However, conventional mesh-based scene representations [28, 39, 47] remain relatively mature and serve as the industrial foundation for downstream tasks such as scene editing, object interaction, animation generation, and relighting. Therefore, efficiently and accurately extracting high-quality human meshes from 3DGS can bridge the rendering advantages of 3DGS with practical applications in existing tools.

Nevertheless, Gaussian distributions model scenes through numerous discrete, unstructured ellipsoids, inherently lacking explicit surface topology or continuous geometric information. Additionally, Gaussians are typically optimized for statistical accuracy rather than geometric precision, resulting in directly extracted geometry that often lacks high fidelity.

Existing methods [5, 10, 35] achieve direct surface modeling by binding Gaussians to mesh triangles. Subsequent approaches, such as Animatable Gaussians [24] and GaussianAvatar [15], learn Gaussian properties from 2D images, leveraging powerful 2D networks for realistic human modeling. However, by relying solely on the SMPL [27] model for guidance, these methods struggle to capture detailed structural motion. They face two primary challenges in practice. First, supervision from only 2D images or videos leads to significant occluded regions, producing blurred and incomplete Gaussian reconstructions. Second, converting Gaussians to meshes often introduces artifacts due to the

(a) GT (b) HGM (c) MeshGS [4]

Figure 1. Example mesh extraction results from different methods. Note the superiority of HGM in reconstructing geometric structures and texture details.

inherent volumetric and unstructured nature of Gaussians, which hinders a precise alignment with the underlying surface geometry.

As show in Figure 1, we proposed a self-refinement based idea to tackle those limitations by optimizing the out meshes based solely on the source Gaussians themselves. In particular, it incorporates three tactics to overcome those two limitations.

First, to address the first challenge on limited supervision with stronger one, an epipolar attention [13, 18] and diffusion [12, 14] guided lifter and a geometry-aware depth estimator are proposed for enhanced high-quality multi-view images and corresponding depth images. The lifter utilizes an epipolar attention mechanism and a diffusion-based image enhancer for augmented high-resolution multi-view synthesis based on the standard Gaussian Splatting renderer [45]. It delivers photorealistic and consistent results across unconstrained viewpoints. The estimator takes the SMPL-X [1] human template as strong geometric prior and fuses with the appearance features of multi-view images, so that accurate corresponding multi-view depth maps can be obtained.

Second, to address the second challenge on fundamental barrier of Gaussian representation. An adaptive Gaussian optimization module guided by the estimated multi-view depth images and corresponding normal maps is proposed for accurate geometry reconstruction. By utilizing the enhanced images from the lifter, it facilitates accurate alignment of the Gaussian distributions to the human surfaces [25], substantially increasing geometric fidelity.

This paper proposed three-stage self-refinement framework, which performs Gaussian optimization and mesh extraction, produces a well-optimized Gaussian representation. This representation is then utilized by a Truncated Signed Distance Function (TSDF) to generate a clothed human mesh. The final mesh exhibits a smooth surface while being rich in high-frequency details.

In summary, our main contributions are as follows:

- An epipolar-attention-guided diffusion lifter, which integrates epipolar attention mechanisms with diffusion models to improve rendered multi-view images to be high realistic and view-consistent.
- A geometry-aware depth estimator, which obtain an accurate SMPL-X human template through multi-view alignment optimization and utilize it as a geometric prior to drive the depth estimation network for estimating multi-view depth maps.
- A self-refinement oriented depth and angle-guided adaptive optimizer, which integrates multi-view depth maps, multi-view RGB images, and the normal vectors of Gaussian ellipsoids, and achieves topologically precise human mesh extraction with rich texture details through adaptive optimization.
- Experimental results demonstrate the superiority of our framework over existing methods. Notably, the proposed modules can serve as independent plugins to enhance other reconstruction pipelines.

2. Related Work

2.1. Mesh-based Human Reconstruction

Traditionally, the geometric representation of human virtual avatars [28,39] has commonly employed explicit polygonal meshes, a format highly compatible with the classical graphics rendering pipeline. Existing studies leverage parametric human models, such as SMPL [1, 27] and optimize the mesh deformation process to reconstruct human mesh models. These methods can generate meshes with topology consistent with the parametric model, ensuring compatibility with existing model libraries and enabling direct defor-

mation control via pose parameters. However, parametric models are constrained by their predefined topology, which limits their ability to represent complex clothing geometry and visual effects. To address this limitation, another line of research [7, 11, 50] focuses on modeling clothing as an independent mesh layer separate from the body. Nevertheless, using a single predefined template mesh to model diverse body shapes and clothing remains challenging, particularly in achieving high-fidelity results with varying materials, folds, and topological changes. In contrast to traditional explicit polygonal mesh-based approaches that emphasize appearance modeling, our method is grounded in implicit representation and employs Gaussian rendering for high-quality modeling of both the human body and clothing.

2.2. NeRF-based Human Reconstruction

Implicit representation methods, particularly NeRF [9], have achieved remarkable progress in 3D human reconstruction due to their ability to model arbitrary topological structures. Leveraging this advantage, NeRF and related methods have been widely adopted for modeling clothed humans in studies [3, 33, 34]. For instance, methods such as Neural Body [34] [34], Animatable NeRF [33], and Neural Actor [25] employ structured latent codes anchored to mesh vertices as conditional inputs. While these approaches can generate photorealistic novel views from sparse multi-view videos, they still exhibit limitations in generalizing to unseen poses. On the other hand, UV Volumes [2] represents dynamic humans using a combination of 3D UV volumes and 2D neural textures, enabling high-fidelity rendering and editing.

More recently, NeuS [42] introduced the integration of SDF with volume rendering, pioneering the direct extraction of isosurfaces during optimization. Subsequent research has explored various strategies to improve efficiency and representation quality: Instant-NGP [32] adopted multi-resolution hash encoding to reduce training time from hours to minutes while maintaining visual quality; Plenoxels [8] abandoned neural networks altogether by storing radiance properties in sparse voxel grids, enabling real-time differentiable rendering; and Neus2 [43] utilized a time-varying SDF field to generate dynamic mesh sequences and innovatively combined the optimization of NeRF with parametric meshes to support real-time rendering and editing. Nevertheless, current NeRF-based methods still face several challenges, including limited geometric accuracy in human reconstruction, insufficient capability in modeling diverse clothing, and weak adaptability to topological changes.

2.3. 3DGS-based Human Reconstruction

With the introduction of 3DGS [20], breakthrough progress has been achieved in real-time high-quality neural rendering. This method employs explicit and optimizable 3D Gaussian ellipsoids as fundamental units for scene representation. Through efficient Splatting-based rendering, it maintains high visual fidelity while significantly surpassing the rendering speed of NeRF [9]. This capability has led to its rapid adoption in the challenging task of dynamic human avatar reconstruction.

As a pioneering work, 3DGS-Avatar [36] [36]introduced the canonical space and deformation field paradigm. By binding the deformation of Gaussians in canonical space to the pose parameters of an SMPL model, it enabled the reconstruction of high-quality, drivable human models from sparse multi-view videos, preserving both the training efficiency and rendering quality of 3DGS while addressing dynamic geometry modeling. Subsequently, Gaussian Avatars [15] focused on high-fidelity head reconstruction by introducing the concept of skinnable weights for Gaussian units, allowing them to be directly driven by expression parameters of a FLAME [23] model and generating photorealistic head avatars with pore-level details. HumanGaussian [26] drew inspiration from Instant-NGP's multi-resolution hash encoding and proposed a structured hash Gaussian representation, reducing training time to minutes and enhancing the rationality and stability of the representation by incorporating SMPL geometric priors to optimize Gaussian spatial distribution. Gaustudio [48] further developed a comprehensive framework that systematically compared different reconstruction strategies and explored editing functionalities based on explicit Gaussian representations. MeshGS [4] introduced an adaptive mesh-aligned Gaussian Splatting method that achieves high-quality rendering by aligning Gaussian distributions with mesh surfaces. HeadStudio [51] constructed a text-driven head avatar generation framework that integrates 3D Gaussian Splatting with efficient mesh deformation techniques, enabling controllable avatar animation through a geometric manipulation pipeline.

Despite these methods significantly advancing the field in terms of speed, quality, and controllability, current 3DGS-based dynamic human modeling still faces several challenges. These include insufficient robustness to extreme topological changes and rapid motion, limited physical interpretability in Gaussian deformation, and unresolved difficulties in handling fine clothing materials and complex occlusions.

3. Our Reconstruction Method

For any human 3D Gaussian representation, our method (Figure 2) seeks to achieve the reconstruction of high-fidelity geometry and textures through a path of self-optimization, leveraging the properties of the Gaussians themselves. The method integrates a triple-strategy for enhanced supervision: an epipolar-attention-guided diffusion

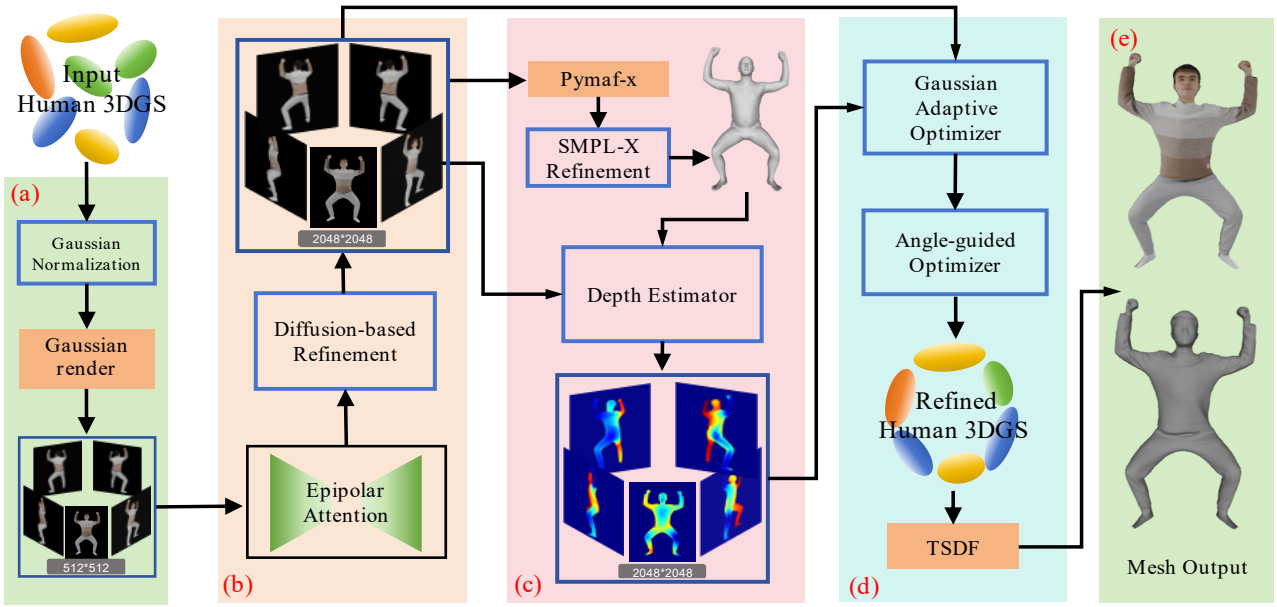

Figure 2. HGM Pipeline: The method begins by obtaining a normalized Gaussian model and multi-view RGB images through the Gaussian Preprocessing Module (a). Subsequently, the epipolar-attention-guided diffusion lifter (b) enhances the multi-view RGB images. These enhanced images, combined with SMPL-X geometric priors, are then processed by the geometry-aware depth estimator (c) to generate multi-view depth maps. Finally, the depth and angle-guided adaptive optimizer (d) iteratively refines the geometric representation of the Gaussian model by integrating depth and normal information, ensuring it closely aligns with the realistic appearance characteristics of the human body.

lifter, a geometry-aware depth estimator, and a depth and angle-guided adaptive optimizer. The first two components are responsible for generating enhanced multi-view images and their corresponding depth maps, which serve as strong supervisory signals to drive the optimizer for the refined reconstruction of the human model's surface and texture.

In particular, given a human Gaussian representation, it is first processed by a Gaussian preprocessing module to obtain normalized Gaussian distributions and multi-view RGB images rendered from specified viewpoints centered on the human body (Section 3.1). Next, a shared image encoder integrated with an epipolar attention mechanism extracts features from the multi-view RGB images. These features are then enhanced by a diffusion model to produce refined images (Section 3.2). Subsequently, the enhanced images are processed by an SMPL-X-based depth estimator to generate corresponding depth maps (Section 3.3). Finally, the previously obtained depth maps and the normals derived from the Gaussians themselves are utilized by a depth-normal adaptive optimizer. This optimizer iteratively integrates human geometry information to progressively refine the geometric representation of the Gaussians, ensuring a better alignment with the realistic appearance characteristics of the human body (Section 3.4).

3.1. Gaussian Preprocessing Module

For a given human Gaussian model, the pipeline begins with a meticulously designed Gaussian pre-processing module, which serves as the foundation for the entire reconstruction workflow by providing normalized and information-rich data. Normalization of the Gaussian distributions is critical in this stage. As Gaussian models from different sources or states may exhibit significant disparities in their parameters (e.g., means, covariance matrices), such variations can severely impact the accuracy and stability of subsequent processing steps. Therefore, we perform mean centralization and covariance scaling. Specifically, the mean of each Gaussian is shifted to the origin, and the covariance matrix is scaled such that its eigenvalues fall within a specified range. This normalization ensures that all Gaussians reside in a standardized statistical space, facilitating unified processing and analysis. Let $G_i = \{\mu_i, \Sigma_i\}$ represent the i-th Gaussian in the model, where μ_i is the mean and Σ_i is the covariance matrix. The normalized parameters $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are computed by:

$$\hat{\mu}_i = \frac{\mu_i - \mu_{\text{global}}}{\sigma_{\text{global}}}, \quad \hat{\Sigma}_i = \frac{\Sigma_i}{\lambda_{\max}} \tag{1}$$

where μ_{global} and σ_{global} are the global mean and standard deviation of all Gaussian means; and $\lambda_{\max}$ is the maximum

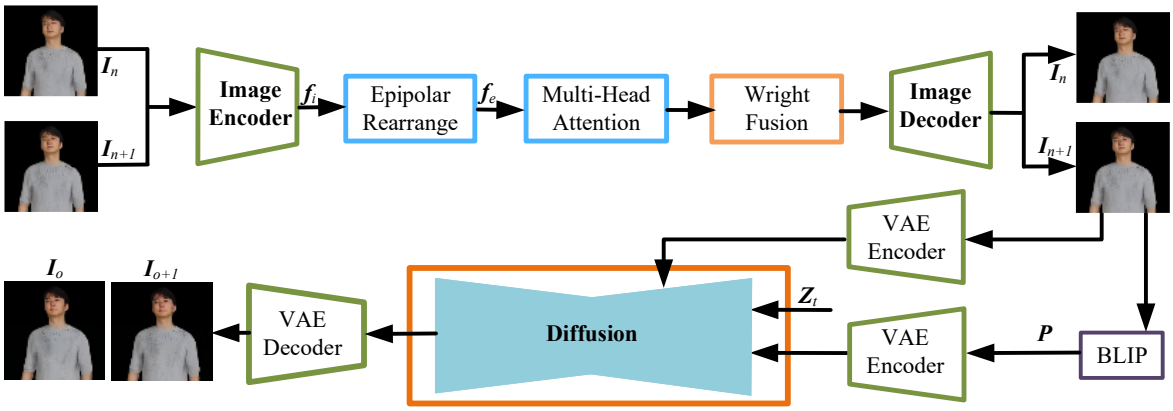

Figure 3. Epipolar-attention-guided Diffusion Lifter: This moudle first enhances multi-view image features through an epipolar attention mechanism, then performs multimodal conditional denoising on the latent representation using a diffusion model, and ultimately generates high-quality human images.

eigenvalue of the covariance matrix across the model.

Subsequently, rendering from multiple viewpoints is essential for comprehensively capturing the subject's appearance. A set of virtual cameras $\{C_n\}_{n=1}^N$ surrounding the subject are defined, so that coverage from back, posterior, and oblique angles is fulfilled to minimize occlusions. Given the normalized Gaussian model $\hat{G}$, multi-view RGB images $\{I_n\}_{n=1}^N$ are rendered via a differential Gaussian Splatting renderer R. The rendering process for a pixel p in view n can be abstracted as:

$$\{I_n\}_{n=1}^N = R(p, \{C_n\}_{n=1}^N, \hat{G}). \tag{2}$$

These rendered images provide a robust and comprehensive data foundation for subsequent tasks such as image feature extraction and depth estimation.

3.2. Epipolar-attention-guided Diffusion Lifter

Given N input images $\{I_n\}_{n=1}^N$ and their corresponding camera poses $\{C_n\}_{n=1}^N$, due to the relative continuity of the preset camera viewpoints, we can sequentially pair the multi-view RGB images to construct stereo image pairs. These image pairs are then fed into a shared image encoder E_{img} to extract dense feature maps f_s. This encoder consists of multiple residual blocks and downsampling layers, which effectively preserve spatial structural information and lay the foundation for subsequent correspondence feature search between the two views.

To further enhance the quality of the feature representation, we introduce an epipolar attention module at the bottleneck features f_S to achieve efficient information interaction between the image pairs. Since the encoder E_{img} processes the left and right views independently, it often struggles to extract features with high symmetry and discriminative power in the absence of explicit supervision. In contrast, the encoder E_{att} is based on the epipolar attention mechanism and can significantly expand the receptive field of the feature extractor. It can output features modeling long-range dependencies and contextual information across views. By explicitly incorporating epipolar geometric constraints, this module effectively enhances the discriminative ability and robustness of the features, thereby mitigating the common matching ambiguities in textureless and occluded regions and improving the overall performance of subsequent depth estimation tasks.

Specifically, we first rearrange the feature maps f_S into $H/2^S$ epipolar line features $f_e \in \mathbb{R}^{W/2^S \times D^S}$, and then apply a multi-head self-attention mechanism along each epipolar line. This process can be expressed as:

$$\hat{f}_k = \text{MultiHeadAttention}\left(Q = f_e^k,\ K = f_e,\ V = f_e\right) \tag{3}$$

where f_e^k denotes the k-th epipolar line feature. The attention mechanism is calculated as follows:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{4}$$

The attention-enhanced line features $\{\hat{f}_k\}_{k=1}^{H/2^S}$ undergo feature fusion and enhancement through a series of convolutional operations. They are then progressively reconstructed by a decoder composed of upsampling layers and convolutional modules. Each upsampling stage incorporates shallow features transferred via skip connections to restore spatial details and texture information. Finally, the decoder outputs high-quality, view-consistent multi-view RGB images $\{\hat{I}_n\}_{n=1}^N$.

Generally, the rendering technique of 3DGS struggles to achieve photorealistic materials. To address this, this paper proposes an enhancement architecture based on a diffusion model, integrating a Variational Autoencoder (VAE)

encoder, BLIP [22], and CLIP [37] models to achieve multimodal information fusion. The network adopts a cascaded encode-diffuse-decode pipeline: The VAE encoder first maps the input image into a latent representation; BLIP generates descriptive text features t, while CLIP extracts semantic embeddings; subsequently, the multimodal features are fused and input into the diffusion model for denoising and refinement.

As shown in Figure 3, the enhancement process utilizes a pre-trained VAE encoder [30, 44] to map the attention-enhanced images $\{\hat{I}_n\}_{n=1}^{N}$ into a low-dimensional latent space, preserving key visual features while reducing computational complexity. To improve image quality, BLIP generates descriptive text prompts P related to the image and camera viewpoint, and the CLIP text encoder extracts their deep semantic information and projects it into the latent space. The image latent representation and text semantic features are then fused and jointly fed into a latent diffusion model for denoising.

Within the latent space, a Denoising Diffusion Probabilistic Model (DDPM) [14] is employed for progressive refinement. During the forward diffusion process, Gaussian noise is gradually added to the latent representation according to a predefined noise schedule, perturbing its structural information, expressed as:

$$q(z_t \mid z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t, \{\hat{I}_n\}_{n=1}^{N}\right) \quad (5)$$

where t is the diffusion timestep, and β_t controls the amount of noise at each step. During the reverse denoising process, the denoising network predicts the noise based on the current noisy latent variable z_t, the time step t, and conditional signals such as text features c_t and image features c_i. The reverse transition is modeled as:

$$p_\theta(z_{t-1} \mid z_t, c_\mathrm{t}, c_\mathrm{i}) = \mathcal{N}\left(z_{t-1}; \mu_\theta(z_t, t, c_\mathrm{t}, c_\mathrm{i}), \Sigma_\theta(z_t, t)\right) \quad (6)$$

where μ_θ and Σ_θ are the mean and covariance predicted by the denoising network, respectively. Through iterative conditional denoising, a refined and detail-enhanced latent representation is gradually recovered.

Furthermore, We use a Stable Diffusion-style U-Net DDPM with 4 residual blocks and 16×16 attention layers. The model, initialized from a public latent-diffusion checkpoint, is fine-tuned on 30k multi-view RGB patches from THuman2.0 using AdamW (lr=2×10^{-4}, wd=0.01) and a combined L1-perceptual loss over 50 epochs. This DDPM module remains fixed during HGM optimization and acts only as a post-processing texture enhancer, without joint training with 3D Gaussians.

Finally, the VAE decoder decodes the denoised latent representation into a set of high-quality images $\{I_o\}_{o=1}^{N}$

3.3. Geometry-aware Depth Estimator

For a given Gaussian human model, this paper takes the image $I_{o=1}$ as the foundational input. For an accurate and ergonomic 3D human representation, the regression-based PyMAF-X [49] method is is employed to estimate a set of SMPL-X model parameters from this single monocular RGB image—including shape parameters $\beta \in \mathbb{R}^{200}$, pose parameters $\theta \in \mathbb{R}^{55}$, and expression parameters $\psi \in \mathbb{R}^{100}$, thereby obtaining an initial, parametric SMPL-X standardized human template S_i. However, 3D geometry recovered from a single viewpoint inherently suffers from ambiguity, which can lead to implausible deformations in unseen views.

To overcome this limitation, we fully leverage the 3D consistency constraints embedded within the multi-view image sequence $\{I_o\}_{o=1}^{N}$. Accordingly, the paper designs an iterative optimization process: the initial template S_i is placed in 3D space, and using the PyTorch3D differentiable renderer. S_i is rendered into 2D semantic segmentation maps or silhouette maps from each known camera pose corresponding to view. Subsequently, these rendered results are aligned and compared with their corresponding input images, and the discrepancies between them are calculated. This discrepancy is backpropagated through the differentiable renderer, directly guiding gradient updates to the SMPL-X model parameters (primarily the global pose and shape parameters β). The optimization objective at each iteration can be summarized as minimizing the total multi-view reprojection error:

$$\Theta^* = \arg\min_{\Theta} \sum_{o=1}^{N} L_{\text{align}}(R_o(S(\Theta)), I_o) \quad (7)$$

where Θ^* represents the SMPL-X parameters, and L_{align} is the alignment loss function (IoU and Chamfer Distance). Through multiple iterations, the model S is continuously adjusted until its 2D projections across all available viewpoints achieve optimal alignment with the corresponding input images, ultimately yielding a precise SMPL-X human template S that is highly consistent with the multi-view imagery.

After obtaining the optimized and accurate human template S, we utilize it as a strong geometric prior to drive subsequent depth estimation tasks. The template S is rendered into the corresponding camera view with the given accurate camera parameters and then the associated SMPL-X depth map $\{D_s^o\}_{o=1}^{N}$ is generated.

Finally, a depth estimation network G_d is constructed. This network takes the concatenated RGB image I_o and its corresponding SMPL-X depth map $\{D_s^o\}_{o=1}^{N}$ as input. Its design purpose is to enable the network, guided by the strong prior of the known basic human geometry (provided by $\{D_s^o\}_{o=1}^{N}$), to learn to infer the true surface depth of the

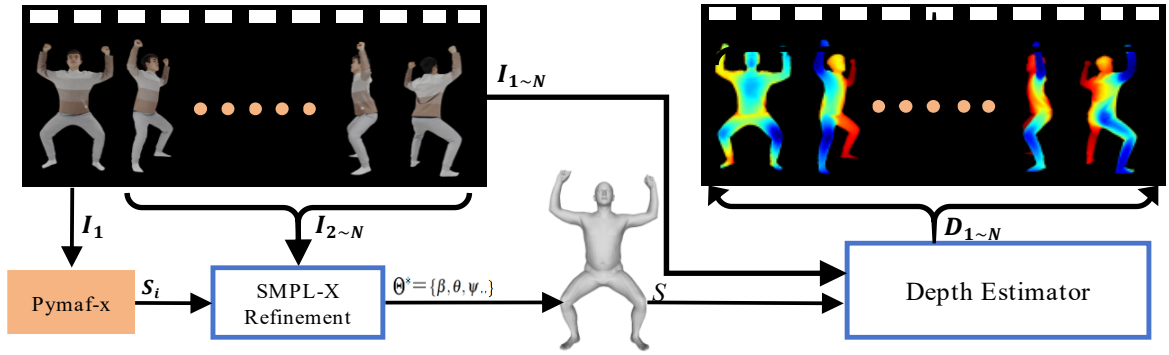

Figure 4. Geometry-aware Depth Estimator: This module achieves high-fidelity human reconstruction by optimizing the SMPL-X human template through multi-view constraints. The template is then used to drive the depth estimation network and finally adaptive optimization is performed on the 3D Gaussian model with a multi-source loss function.

clothed human from RGB textures, lighting, and shadows. The network G_d need parse and fuse information from these two modalities, ultimately outputting a precise final depth map $\{D_o\}_{o=1}^N$. The complete depth estimation process is defined as:

$$D_o = G_d(I_o \oplus D_s^o; \Omega) \tag{8}$$

where $\oplus$ denotes the channel-wise concatenation operation, and Ω represents the learnable parameters of the network G_d. This process effectively combines the geometric consistency of the parametric model with the detail-recovery capability of the data-driven model, thereby achieving robust estimation of high-quality 3D depth information from monocular images.

The loss function for our depth estimation need account for the characteristics of geometric information (such as scale sensitivity and geometric consistency). Therefore, the following combination is used:

$$L_D = L_{\text{pixel}} + \lambda_{\text{grad}} L_{\text{grad}} + \lambda_{\text{VGG}} L_{\text{VGG}} \tag{9}$$

where:

$$L_{\text{pixel}} = \|D_c - \hat{D}_c\|_1. \tag{10}$$

This is the absolute difference loss between the ground truth depth map $\hat{D}_c$ and the predicted depth map D_c.

$$L_{\text{grad}} = \|\nabla_x(D_c - \hat{D}_c)\|_1 + \|\nabla_y(D_c - \hat{D}_c)\|_1. \tag{11}$$

This loss enforces the predicted depth map to maintain gradient similarity with the ground truth at edges, helping to preserve clear geometric structures.

$$L_{\text{VGG}} = \|\phi(\hat{D}_c) - \phi(D_c)\|_2 \tag{12}$$

Here, ϕ is the feature extractor of a pre-trained VGG network. The perceptual loss aids in recovering details and improving visual quality.

Algorithm 1 Epipolar-attention Diffusion Lifter

Require:
Multi-view images: $\{I_n\}_{n=1}^N$.
Camera poses: $\{C_n\}_{n=1}^N$.
Ensure: Enhanced multi-view images: $\{I_o\}_{o=1}^N$.
for $n = 1$ to N **do**
  $f_s^n \leftarrow E_{\text{img}}(I_n)$ (Section 3.2).
end for
// Epipolar Attention
for $k = 1$ to $H/2^S$ **do**
  $f_e^k \leftarrow \text{Rearrange}(f_s, k)$.
  $\hat{f}_k \leftarrow \text{MultiHeadAttention}(f_e^k, f_e, f_e)$(Eq. 3).
end for
// Diffusion Enhancement
$\{\hat{I}_n\}_{n=1}^N \leftarrow \text{Decoder}(\{\hat{f}_k\})$.
$z_0 \leftarrow \text{VAE\_Encoder}(\{\hat{I}_n\}), c_t \leftarrow \text{CLIP}(\text{BLIP}(\{\hat{I}_n\}))$.
for $t = T$ down to 1 **do**
  $\varepsilon_\theta \leftarrow \text{Denoiser}(z_t, t, c_t, c_i)$(Eq. 5 and 6).
  $z_{t-1} \leftarrow \text{ReverseStep}(z_t, \varepsilon_\theta)$.
end for
$\{I_o\}_{o=1}^N \leftarrow \text{VAE\_Decoder}(z_0)$.
return $\{I_o\}_{o=1}^N$.

3.4. Depth and Angle-guided Adaptive Optimizer

As is well known, 3DGS suffers from inherent shortcomings in the coupling of geometry and appearance: The independent optimization of Gaussian covariance, color, and opacity parameters often leads to local minima, resulting in missing thin human body structures, over-smoothed surfaces, and blurred clothing boundaries. To fundamentally mitigate these issues, this paper reformulates the representation-observation alignment task as an iterative optimization problem constrained by multi-source information. The core idea is that, in each iteration, the Gaussian parameters are simultaneously supervised by visual, geo-

metric, and semantic cues, enabling adaptive adjustments to their spatial distribution, covariance shape, and spherical harmonic coefficients. This ensures convergence to a state that is both faithful to the multi-view inputs and exhibits extractable geometric quality.

To obtain reliable geometric supervision, we employ a Vision Transformer–based monocular depth estimator with a multi-scale decoder, pre-trained on diverse indoor/outdoor scenes and fine-tuned on metric depth rendered from THuman2.0 scans. Crucially, it predicts absolute depth in meters(not relative values) and aligns end-to-end with the 3D Gaussian coordinate system via an explicit camera-intrinsics embedding. The resulting depth maps $D_{o}{}_{o=1}^{N}$ serve as strong geometric constraints during optimization.

Specifically, this module takes the following inputs: multi-view RGB images $\{I_o\}_{o=1}^{N}$, multi-view depth images $\{D_o\}_{o=1}^{N}$, the 3D human Gaussian model $\hat{G}$, and selected camera parameters $\{C_n\}_{n=1}^{N}$. Every optimization iteration follows a render-align-update cyclic process.

First, an improved tile-based rasterizer is employed to simultaneously output RGB images and depth images:

$$\{\hat{I}_o\}_{o=1}^{N},\ \{\hat{D}_o\}_{o=1}^{N} = \mathrm{Render}\big(\hat{G}, \{C_n\}_{n=1}^{N}\big) \tag{13}$$

where π_i denotes the i-th camera parameters. The entire pipeline is fully differentiable, allowing gradients to flow back to every 3D Gaussian attribute.

Next, multi-source consistency is enforced through a compound loss composed of four terms: Visual reconstruction loss:

$$L_{\mathrm{rgb}} = \sum_{o} \lambda_{\mathrm{rgb}} \big\|\hat{\mathbf{I}}_o - \mathbf{I}_o\big\|_1 + \lambda_{\mathrm{lpips}}\,\mathrm{LPIPS}\big(\hat{\mathbf{I}}_o, \mathbf{I}_o\big) \tag{14}$$

Depth consistency loss:

$$L_{\mathrm{depth}} = \sum_{o} \lambda_{\mathrm{d}} \big\|\hat{\mathbf{D}}_o - \mathbf{D}_o\big\|_1 \tag{15}$$

Structured sparsity regularizer:

$$L_{\mathrm{regularization}} = \lambda_{\mathrm{scale}} \sum_{k} \|\mathbf{S}_k\|_F^2 + \lambda_{\mathrm{alpha}} \sum_{i \in \mathcal{N}(k)} (\alpha_k - \alpha_i)^2 \tag{16}$$

The total loss is the weighted sum:

$$L = L_{\mathrm{rgb}} + L_{\mathrm{depth}} + L_{\mathrm{reg}} \tag{17}$$

where the weights λ maintain a dynamically adaptive variation during training: λ_{d} are increased to emphasize geometric constraints when the visual error plateaus; otherwise, λ_{rgb} is raised to recover appearance details. This mechanism enables autonomous switching between the appearance-first, geometry-follow and geometry-first, appearance-compensate operational modes. Following each parameter update, the system performs real-time density control, where Gaussian primitives are split, cloned, or pruned based on metrics including gradient magnitude, covariance trace, and opacity. This ensures the representation maintains both compactness and expressiveness throughout the entire optimization process.

Subsequently, upon completion of model training (i.e., when the loss function converges or the preset maximum number of iterations is reached), a dense 3D point cloud can be extracted from the constructed Gaussian model. The acquired point cloud data is then converted into a continuous 3D voxel field representation using a TSDF-based voxel fusion technique. By applying the Marching Cubes algorithm to this voxel field, an isosurface can be extracted, generating a surface mesh model characterized by geometric continuity and topological consistency. This ultimately enables the reconstruction of a structurally accurate and detail-rich 3D human model.

In this process, to address the blocky artifacts commonly encountered in traditional voxel fusion, this paper proposes a weighted fusion strategy based on normal vector consistency constraints. This method fully considers the directional consistency of the normal vectors of Gaussian points by introducing an angle threshold to filter the contributions of normal vectors within a voxel unit: They are incorporated into the TSDF calculation of the voxel unit only when the angle between the normal vectors of adjacent Gaussian points is smaller than a preset threshold. Specifically, for a voxel unit V_i, its fusion weight w_i is determined by the following formula:

$$w_i = \sum_{j=1}^{N} \delta(\theta_{ij} < \theta_t) \cdot \exp\Big(-\frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{2\sigma^2}\Big) \tag{18}$$

where N is the number of Gaussian points influencing the voxel; θ_{ij} denotes the angle between the normal vector of the voxel unit and that of the j-th Gaussian point; θ_t is the angle threshold; $\mathbf{p}_i$ and $\mathbf{p}_j$ represent the spatial positions of the voxel center and the Gaussian point, respectively; and σ is the spatial decay coefficient. This weighting approach maintains spatial continuity while effectively suppressing interference from points with inconsistent surface normals on the fusion result through normal vector angle constraints. Consequently, it significantly improves the sharpness of feature edges and geometric fidelity, enabling the reconstructed model to retain sharp geometric features while maintaining smooth surfaces.

4. Experiments

Section 4.1 describes the experimental datasets, including their sources and scene types. Section 4.2 details the experimental setup, encompassing the network architecture, optimization strategies, and core hyperparameters. The evaluation in Section 4.3 and Section 4.4 validates the effectiveness of our proposed method through multi-dimensional

comparisons against various baselines. These include methods based on neural implicit representations (NeuS [42] and Neural Body [34]), the foundational 3DGS [20], the hybrid 2D-3D method 2D-GS [17], along with 3DGS-Avatar [36], Gaussian Avatars [15], HumanGaussian [26], and SuGaR [10].

For a fair evaluation of geometric reconstruction quality, the mesh surfaces for baseline methods using implicit representations were uniformly extracted from their density or occupancy fields using the Marching Cubes algorithm.

Additionally, a systematic ablation study is conducted in Section 4.5. Through controlled experiments, the impact of key modules on reconstruction accuracy, geometric fidelity, and rendering efficiency is quantitatively assessed, with all findings supported by quantitative metrics and visual comparisons.

4.1. Datasets

To comprehensively evaluate the generalization capability and practical utility of the proposed method, we validated our approach on three datasets. The People Snapshot [19] dataset provides monocular video sequences designed to test a method's generalization capability for 3D reconstruction from sparse, casually captured "in-the-wild" data. The ZJU-Mocap [6] dataset offers high-quality multi-view videos with 3D ground truth, primarily for evaluating a method's robustness under complex dynamic poses and severe occlusions. The Thuman [41] dataset comprises high-fidelity 3D human scans for assessing a method's accuracy in recovering clothing details and fine-grained geometry.

4.2. Implementation Details

For human rendering quality evaluation, this study employs Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [38] for quantitative assessment. This evaluation framework covers multidimensional measurements ranging from pixel-level accuracy to perceptual similarity. For human geometry reconstruction, we utilize Chamfer Distance (CD), Normal Consistency (NC), and F-score (F-S) to quantify the spatial consistency between the reconstructed human mesh and the ground truth. The experimental setup follows standard data partitioning protocols [4]: 1/8 of the images were randomly held out as the test set with the remaining 7/8 for training. All experiments were executed on an RTX 4090 GPU to ensure consistent hardware conditions, thereby enabling reproducible and fair comparisons across all evaluated methods.

4.3. Qualitative Results

Qualitative comparisons (Figure 5 and 6) demonstrate that HGM outperforms representative methods including 3DGS, 3DGS-Avatar, and 2D-GS in both rendering quality and geometric reconstruction accuracy. Particularly in scenarios involving complex clothing and challenging poses, HGM generates surfaces with physical realism, producing geometric features characterized by sharp boundaries and strong continuity. The method exhibits exceptional capability in reconstructing high-frequency details and microscopic structures of the human body. Furthermore, localized comparisons more distinctly highlight HGM's superior performance in preserving detail integrity and structural consistency, yielding visually natural reconstruction results that better align with real-world geometric priors.

4.4. Quantitative Results

For a comprehensive benchmark evaluation of HGM, two categories of representative methods are selected as baseline comparisons. Specifically, the first category comprises traditional techniques that convert NeRF or 3DGS into meshes. These methods are typically designed for indoor or outdoor scenes, with no existing techniques dedicated to human mesh extraction. The second category consists of explicit optimization methods for human representation based on Gaussian Splatting.

Quantitative evaluation (Table 1 and 2) demonstrates that HGM surpasses existing methods across multiple metrics. While maintaining high-fidelity texture reconstruction, it improves geometric alignment accuracy by 10%-15% and reduces reconstruction errors by over 15% in complex scenarios. Traditional methods relying solely on RGB photometric loss lack explicit geometric constraints, often resulting in surface noise, artifacts, and structural discontinuities. In contrast, HGM establishes a self-refining mechanism integrating 2D and 3D representations, creating bidirectional feedback between Gaussian rendering and geometric features: using rendering to supervise geometric detail generation while employing geometry to constrain physical plausibility, thereby achieving synergistic optimization of appearance and geometry. This approach provides an innovative solution for joint human mesh optimization, significantly advancing the integration of 3D human reconstruction and rendering.

4.5. Ablation study

In the ablation study, we evaluated eight configuration schemes (Table 3) against a baseline method without the EAD, GDE, and DAAO modules. Experiments demonstrate that our proposed modular approach significantly outperforms the baseline across all evaluation metrics. Specifically, the EAD module remarkably enhances texture quality and visual details, while the normal regularization submodule in DAAO shows outstanding contribution to geometric surface optimization as measured by the CD metric. It should be noted that although DAAO improves overall geometric reconstruction quality, quantitative analysis based

(a) GT (b) SuGaR [10] (c) MaGS [28] (d) HGM (Ours)

Figure 5. Visual comparison among HGM, SuGaR, and MaGS on THuman.

Table 1. Statistical performance comparison among different methods. The best scores are shown in **bold**, with the second-best underlined.

Methods	People Snapshot [19]			ZJU-Mocap [6]			Thuman [41]		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
Neus [42]	2.35	0.92	28.95	2.33	0.91	27.34	2.31	0.93	28.01
Neural Body [34]	2.21	0.95	29.48	2.20	0.97	29.34	2.23	0.93	29.34
Neus2 [43]	2.29	0.93	27.01	2.31	0.94	27.17	2.23	0.94	29.33
3DGS [20]	2.24	0.95	29.32	2.22	0.95	29.01	2.26	0.92	28.64
2DGS [17]	2.27	0.94	28.35	2.23	0.94	28.21	2.24	0.94	28.68
3DGS-Avatar [36]	2.21	0.96	29.85	2.22	0.97	29.68	2.19	0.97	29.83
HuGS [21]	2.21	0.97	30.13	2.16	0.97	31.29	2.14	0.96	29.75
GauHuman [16]	2.17	0.97	30.11	2.16	0.97	30.64	2.13	0.96	30.03
HumanGaussian [26]	2.16	0.98	30.06	2.21	0.97	30.01	2.17	0.96	29.96
D3GA [52]	2.18	0.98	30.44	2.17	**0.98**	30.74	2.15	0.96	30.19
HGM (Ours)	**2.14**	**0.99**	**31.69**	**2.13**	0.97	**31.65**	**2.12**	**0.99**	**32.51**

on the LPIPS metric reveals a slight degradation in texture fidelity.

(a) GT (b) SuGaR [10] (c) MaGS [28] (d) HGM (Ours)

Figure 6. Visual comparison among HGM, SuGaR, and MaGS on the wild image.

Table 2. Statistical performance comparison among different methods on the THuman test dataset. The best scores are shown in **bold**, with the second-best underlined.

Methods	487			500			503		
	CD↓	NC↑	F-S↑	CD↓	NC↑	F-S↑	CD↓	NC↑	F-S↑
3DGS [20]	3.04	0.696	30.69	2.97	0.721	31.94	2.93	0.707	31.63
SuGaR [10]	2.52	0.785	34.95	2.47	0.774	35.06	2.43	0.757	35.64
2DGS [17]	2.35	0.763	35.16	2.39	0.767	35.68	2.33	0.759	36.13
MeshGS [4]	2.28	0.774	36.95	2.20	0.765	35.59	2.34	0.767	37.03
GaussianAvatars [35]	2.68	0.749	34.13	2.54	0.735	33.98	2.65	0.719	35.31
D3GA [52]	2.49	0.754	34.05	2.43	0.746	35.68	2.39	0.763	36.38
MaGS [29]	2.26	0.801	36.88	2.24	0.791	37.28	2.23	0.801	37.34
HGM(Ours)	**2.10**	**0.811**	**38.98**	**2.04**	**0.814**	**39.21**	**2.07**	**0.817**	**38.64**
	513			521			526		
3DGS [20]	3.16	0.689	31.57	3.03	0.714	32.04	2.968	0.715	31.97
SuGaR [10]	2.49	0.765	36.05	2.42	0.763	35.16	2.37	0.759	35.97
2DGS [17]	2.30	0.761	36.03	2.29	0.755	36.07	2.31	0.760	36.06
MeshGS [4]	2.26	0.759	36.13	2.21	0.768	36.12	2.25	0.771	36.17
GaussianAvatars [35]	2.66	0.712	35.13	2.56	0.734	34.87	2.48	0.50	34.93
D3GA [52]	2.40	0.733	36.36	2.29	0.787	36.10	2.47	0.738	35.96
MaGS [29]	2.21	0.795	36.97	2.19	0.803	37.61	2.29	0.781	36.95
HGM(Ours)	**2.12**	**0.827**	**39.37**	**2.11**	**0.818**	**39.01**	**2.10**	**0.813**	**38.44**

4.6. Extending to other methods

The proposed strategy demonstrates excellent portability and can be conveniently applied to enhance the performance of existing 3D Gaussian reconstruction and geometry generation methods. In this study, we selected 2DGS, a represen-

Table 3. Statistical performance comparison with different ablation configurations. The best scores are shown in **bold**.

Method									
Base	EAD	GDE	DAAO	CD↓	NC↑	F-S↑	LPIPS↓	SSIM↑	PSNR↑
✓				3.06	0.698	31.68	2.21	0.95	29.09
✓	✓			2.76	0.694	33.54	2.14	0.97	30.17
✓		✓		2.49	0.786	37.17	2.19	0.95	29.34
✓			✓	2.54	0.763	35.61	2.18	0.95	29.36
✓	✓	✓		2.41	0.791	37.34	2.16	0.96	30.39
✓	✓		✓	2.49	0.793	36.65	2.14	0.97	31.43
✓		✓	✓	2.27	0.804	38.05	2.18	0.95	30.37
✓	✓	✓	✓	**2.09**	**0.812**	**39.56**	**2.13**	**0.98**	**32.07**

Table 4. Addition results on integrating our strategies into 2DGS [17]

Method	CD↓	NC↑	F-S↑	PSNR↑
2DGS	2.28	0.765	35.33	28.30
2DGS+EAD	2.25	0.764	35.35	30.95
2DGS+GDE+DAAO	2.19	0.793	37.64	29.04
2DGS+EAD+GDE+DAAO	2.16	0.806	38.14	31.21

tative Gaussian Splatting-based reconstruction method, for experimental validation. As shown in Table 4, the proposed Gaussian self-optimization module significantly improves the performance of both methods, fully demonstrating its exceptional generalization capability.

Importantly, our DAAO module is architecture-agnostic and can be seamlessly integrated as a plug-and-play refinement module into any 3D Gaussian Splatting pipeline, without relying on SMPL priors or human-specific assumptions (e.g., by using Depth Anything v2 [46] as the depth estimator), as shown in Figure 8

4.7. Discussion

HGM has shown considerable advantages in extracting meshes from 3D human Gaussian representations. However, several limitations remain to be addressed.

First, due to the inherent characteristics of Gaussian functions, subtle deviations inevitably exist between the depth maps generated during Gaussian rendering and the actual scene geometry. Influenced by the intrinsic smoothness of Gaussian functions, fine details in image edges and textured regions are often inadequately preserved, leading to suboptimal reconstruction quality.

Second, the method faces challenges in processing highly ambiguous 3D Gaussian Splatting data, which commonly occurs in dynamic video sequences or noisy scene captures. In such cases, Gaussian smoothing effects restrict

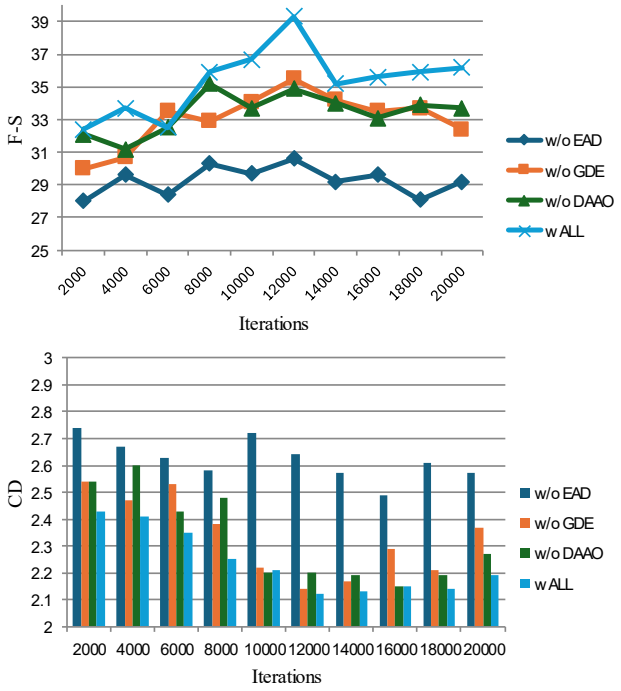

Figure 7. Validate the performance of each component of HGM on No.510 of THuman with four different experimental settings. The horizontal axis represents the number of iterations during Gaussian optimization, and the vertical axis represents the geometric performance metric F-S and CD.

the recovery of high-frequency details in clothing and edge information and, therefore, over-smoothed surfaces or misaligned boundaries in the reconstructed human outputs can be obtained.

5. Conclusion

The rapid advancement of 3DGS has opened new possibilities for digital human modeling, yet efficiently extracting human meshes from unstructured Gaussian representa-

(a) (b) (c) (d)

Figure 8. Visualization of the ablation study without SMPL-X: (a) Ground truth; (b) w/EAD only; (c) w/EAD + DAAO; (d) w/EAD + DAAO + Depth Anything v2. The results illustrate the individual contributions and combined effectiveness of the proposed modules when integrated with the SMPL-X-free Depth Anything v2 model.

tions remains challenging. This article introduces HGM, a unified self-refinement framework that achieves breakthrough performance via optimizing the source Gaussians themselves through three key tactics: an epipolar-attention-guided diffusion lifter for enhanced multi-view detail reconstruction, a geometry-aware depth estimator for improved geometric constraints, and a geometry-aware depth estimator for precise surface fitting. These components form an integrated optimization loop that effectively combines 2D appearance cues with 3D geometry. Experimental results show HGM can generate fully editable, high-fidelity human meshes. This work not only expands 3DGS applications in digital human modeling but also provides a foundation for future dynamic scene research.

References

[1] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. In *NeurIPS*, volume 36, pages 11454–11468, 2023. 2

[2] Y. Chen, X. Wang, X. Chen, Q. Zhang, X. Li, Y. Guo, J. Wang, and F. Wang. UV volumes for real-time rendering of editable free-view human performance. In *CVPR*, pages 16621–16631, 2023. 3

[3] Y.-Z. Chen, J. Zheng, H. Liu, H. Bao, and Y. Guo. Monocular neural human renderer with generalizable illumination and shadow. In *ECCV*, 2022. 3

[4] J. Choi, Y. Lee, H. Lee, H. Kwon, and D. Manocha. MeshGS: Adaptive mesh-aligned Gaussian Splatting for high-quality rendering. In *ACCV*, pages 3310–3326, 2024. 3, 9, 11

[5] P. Dai, J. Xu, W. Xie, X. Liu, H. Wang, and W. Xu. High-quality surface reconstruction using Gaussian surfels. In *ACM SIGGRAPH 2024*, pages 1–11, 2024. 1

[6] Q. Fang, Q. Shuai, J. Dong, H. Bao, and X. Zhou. Reconstructing 3D human pose by watching humans in the mirror. In *CVPR*, 2021. 9, 10

[7] Y. Feng, J. Yang, M. Pollefeys, M. J. Black, and T. Bolkart. Learning dynamic surface deformations by 3d processing of point cloud sequences. In *3DV*, 2021. 3

[8] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, pages 5501–5510, 2022. 3

[9] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li. NeRF: Neural radiance field in 3D vision, a comprehensive review. In *arXiv preprint arXiv:2210.00379*, 2022. 1, 3

[10] A. Guédon and V. Lepetit. SuGaR: Surface-aligned Gaussian Splatting for efficient 3D mesh reconstruction and high-quality mesh rendering. In *CVPR*, pages 5354–5363, 2024. 1, 9, 11

[11] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. DeepCap: Monocular human performance capture using weak supervision. In *IEEE T-PAMI*, 2021. 3

[12] C. He, Y. Shen, C. Fang, F. Xiao, L. Tang, Y. Zhang, W. Zuo, Z. Guo, and X. Li. Diffusion models in low-level vision: A survey. In *IEEE T-PAMI*. IEEE, 2025. 2

[13] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu. Epipolar transformers. In *CVPR*, pages 7779–7788, 2020. 2

[14] J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion probabilistic models. In *arXiv*, pages 89–110, 2020. 2, 6

[15] 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, 9

[16] S. Hu and Z. Liu. GauHuman: Articulated Gaussian Splatting from monocular human videos. *arXiv preprint arXiv:*, 2023. 10

[17] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao. 2D Gaussian Splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024*, pages 1–11, 2024. 9, 10, 11, 12

[18] Z. Huang, H. Wen, J. Dong, Y. Wang, Y. Li, X. Chen, Y.-P. Cao, D. Liang, Y. Qiao, B. Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *CVPR*, pages 9784–9794, 2024. 2

[19] T. Jiang, X. Chen, J. Song, and O. Hilliges. Peoplesnapshot, 2025. 9, 10

[20] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. In *SIGGRAPTH*, volume 42, pages 139–1, 2023. 1, 3, 9, 10, 11

[21] M. Kocabas, J.-H. R. Chang, J. Gabriel, O. Tuzel, and A. Ranjan. Hugs: Human Gaussian splats. In *CVPR*, pages 505–515, 2024. 10

[22] J. Li, D. Li, C. Xiong, and S. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 6

[23] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. In *ACM TOG*, volume 38. ACM, 2019. 3

[24] Z. Li, Z. Zheng, L. Wang, and Y. Liu. Animatable Gaussians: Learning pose-dependent Gaussian maps for high-fidelity human avatar modeling. In *CVPR*, pages 19711–19722, 2024. 1

[25] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. In *ACM TOG*, volume 40, pages 1–16. ACM New York, NY, USA, 2021. 2, 3

[26] X. Liu, X. Zhan, J. Tang, Y. Shan, G. Zeng, D. Lin, X. Liu, and Z. Liu. HumanGaussian: Text-driven 3d human generation with gaussian splatting. In *arXiv preprint arXiv:2311.17061*, 2023. 3, 9, 10

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. In *SIGGRAPH*, pages 851–866. 2023. 1, 2

[28] Q. Ma, S. Saito, J. Yang, S. Tang, and M. J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *CVPR*, pages 16082–16093, 2021. 1, 2

[29] S. Ma, Y. Luo, W. Yang, and Y. Yang. MaGS: Reconstructing and simulating dynamic 3D objects with mesh-adsorbed gaussian splatting. 2025. 11

[30] H. W. L. Mak, R. Han, and H. H. Yin. Application of variational autoencoder (VAE) model and image processing approaches in game design. In *Sensors*, volume 23, page 3457. MDPI, 2023. 6

[31] 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. In *ECCV*, volume 65, pages 99–106. ACM New York, NY, USA, 2021. 1

[32] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *ACM ACM TOG*, volume 41, pages 1–15. ACM New York, NY, USA, 2022. 3

[33] 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. 1, 3

[34] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, pages 9054–9063, 2021. 3, 9, 10

[35] S. Qian, T. Kirschstein, L. Schoneveld, D. Davoli, S. Giebenhain, and M. Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians. In *CVPR*, pages 20299–20309, 2024. 1, 11

[36] 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, 3, 9, 10

[37] A. Radford, J. Wook, H. Aditya, R. Gabriel, G. Sandhini, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, et al. CLIP: Learning transferable visual models from natural language supervision. In *arXiv preprint*, 2019. 6

[38] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li. Deep image spatial transformation for person image generation. In *CVPR*, pages 7690–7699, 2020. 9

[39] S. Saito, T. Simon, J. Saragih, and H. Joo. PifuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, pages 84–93, 2020. 1, 2

[40] L. A. Stuart and M. P. Pound. 3DGS-to-PC: Convert a 3D Gaussian Splatting scene into a dense point cloud or mesh. In *arXiv preprint arXiv:2501.07478*, 2025. 1

[41] Z. Su, T. Yu, Y. Wang, and Y. Liu. DeepCloth: Neural garment representation for shape and style editing. In *IEEE T-PAMI*, volume 45, pages 1581–1593, 2023. 9, 10

[42] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *arXiv preprint arXiv:2106.10689*, 2021. 3, 9, 10

[43] 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 *CVPR*, pages 3295–3306, 2023. 3, 10

[44] Z. Wang, C. Pei, M. Ma, X. Wang, Z. Li, D. Pei, S. Rajmohan, D. Zhang, Q. Lin, H. Zhang, et al. Revisiting VAE for unsupervised time series anomaly detection: A frequency perspective. In *ACM2024*, pages 3096–3105, 2024. 6

[45] T. Wu, Y.-J. Yuan, L.-X. Zhang, J. Yang, Y.-P. Cao, L.-Q. Yan, and L. Gao. Recent advances in 3D Gaussian Splatting. In *Computational Visual Media*, volume 10, pages 613–642. TUP, 2024. 2

[46] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2, 2024. 12

[47] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, volume 33, pages 2492–2502, 2020. 1

[48] C. Ye, Y. Nie, J. Chang, Y. Chen, Y. Zhi, and X. Han. GauStudio: A modular framework for 3D Gaussian Splatting and beyond. In *arXiv preprint arXiv:2403.19632*, 2024. 3

[49] 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. In *IEEE T-PAMI*, volume 45, pages 12287–12303. IEEE, 2023. 6

[50] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 3

[51] Z. Zhou, F. Ma, H. Fan, Z. Yang, and Y. Yang. HeadStudio: Text to animatable head avatars with 3D Gaussian Splatting. In *ECCV*, pages 145–163. Springer, 2024. 3

[52] W. Zielonka, T. Bagautdinov, S. Saito, M. Zollhöfer, J. Thies, and J. Romero. Drivable 3D Gaussian Avatars. In *3DV*, pages 979–990. IEEE, 2025. 1, 10, 11