NLF-GS: A Mixed Mesh-Gaussian Representation for Generalizable and Drivable Human Avatars

by

Jiayi He

to obtain the degree of Master of Science in Computer Science
at the Delft University of Technology,
to be defended publicly on Thursday July 2, 2026 at 10:00.

Student number:	6180647	
Project duration:	September 1, 2025 – July 2, 2026	
Thesis committee:	Dr. Chirag Raman,	TU Delft, thesis supervisor
	Prof. Pablo Cesar Garcia,	TU Delft / CWI, daily supervisor
	Dr. Petr Kellnhofer	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft Delft University of Technology

NLF-GS: A Mixed Mesh-Gaussian Representation for Generalizable and Drivable Human Avatars

Jiayi He
TU Delft

Figure 1. **Teaser for our NLF-GS avatar.** Our framework is trained on a group of subjects (left) and generalizes to previously unseen identities in a zero-shot manner (middle). The reconstructed avatars can also be driven at real-time speed (right).

Abstract

Immersive telepresence requires avatar representations that preserve a person's identity and appearance while remaining responsive to changes in pose under practical deployment constraints such as low latency, compact transmission, and efficient rendering. Existing high-fidelity avatar methods often rely on person-specific optimization, while generalizable methods may struggle to maintain drivability under novel poses. To address this challenge, we present NLF-GS, a structured multi-view avatar reconstruction framework for learning generalizable and drivable mixed mesh-Gaussian avatars, inspired by Neural Localizer Fields (NLF). The core design of NLF-GS is to anchor Gaussian primitives to the faces of an SMPL-X body model and treat each Gaussian as a localized reconstruction query that predicts its geometry and appearance attributes in a feed-forward manner. This design combines the structural control of parametric human models with the rendering efficiency and local appearance flexibility of Gaussian primitives. Experiments show that NLF-GS achieves competitive reconstruction quality against representative generalizable baselines, while maintaining a compact representation and supporting novel-pose animation. The model generates new avatars at nearly 10 *FPS and drives existing avatars at approximately* 196 *FPS on a RTX 4070 GPU. These results suggest that our mixed mesh-Gaussian representations provide a practical representation-level step toward scalable telepresence avatars.*

1. Introduction

Immersive telepresence is a technique that recreates the sensation of being physically present in a different loca-

tion [33]. It aims to make remote communication feel closer to physical co-presence by allowing users to interact responsively [5]. Thus, an avatar in the telepresence system should not only preserve the user's identity and appearance, but also remain responsive to pose changes under practical runtime constraints [41]. This makes the task more demanding than static 3D human reconstruction, as the representation must balance visual fidelity, generalization to unseen identities, pose drivability, and deployment efficiency. In this context, deployment efficiency includes not only rendering speed, but also inference latency, input-view requirements, representation compactness, and potential transmission costs.

Despite the rapid progress in this area, the requirements mentioned above remain difficult to satisfy simultaneously. Many high-fidelity avatar methods, including implicit neural field based avatars [24, 31, 43] and Gaussian-based avatars [35, 51], often require dense capture data, long training time, or a separate optimization process for each new subject. While such methods can produce more convincing visual quality, they limit scalability in practical telepresence scenarios, where new users need to be reconstructed quickly. Conversely, some generalizable methods improve scalability by learning across multiple identities, but they often struggle to preserve fine appearance details or maintain stable animation under novel poses [8, 18, 28, 49, 52]. This creates a central challenge: an avatar representation must be expressive enough to model subject-specific appearance, but structured enough to support reliable pose drivability and compact enough for practical deployment. This leads to the main research question of this thesis: *How can we design a human avatar representation that supports visual fidelity, generalization to unseen identities, pose drivability, and deployment efficiency for telepresence-oriented applications?*

To explore this direction, we propose NLF-GS, a mixed mesh-Gaussian representation for generalizable and drivable human avatars. The name combines Neural Localizer Fields (NLF) [39], which inspires our localized query design, with 3D Gaussian Splatting (GS) primitives used for explicit appearance modeling. The key idea is to organize Gaussian primitives on an SMPL-X body scaffold, so that the avatar retains an explicit human structure for pose-driven animation while using Gaussian primitives for efficient rendering and local appearance modeling. Instead of optimizing a separate representation for each subject, NLF-GS predicts subject-specific Gaussian attributes in a feed-forward manner from sparse multi-view observations. Following the localized prediction philosophy of NLF, each Gaussian is treated as a local reconstruction query that gathers visual evidence from the input views.

Experiments on THuman2.0 dataset [45] show that NLF-GS achieves strong reconstruction quality when compared to representative generalizable baselines, with competitive pixel-level scores and similar visual quality. We further demonstrate that the reconstructed avatars can be driven by novel poses, supporting the proposed representation's potential for pose-drivable telepresence avatars. The code and models for NLF-GS are publicly available.[1]

In summary, this thesis makes the following contributions:

- We propose NLF-GS, a structured mesh-anchored Gaussian representation for telepresence-oriented avatar reconstruction. The representation combines the pose drivability of SMPL-X with the rendering efficiency and local appearance flexibility of 3D Gaussian primitives.
- We introduce a localized pipeline in which each Gaussian primitive acts as a reconstruction query. By projecting mesh-anchored Gaussians into sparse input views, the model predicts subject-specific Gaussian attributes in a feed-forward manner.
- We evaluate the proposed representation on the THuman2.0 dataset in terms of visual fidelity, generalization to unseen identities, pose drivability, and deployment efficiency. Experimental results show that NLF-GS achieves reconstruction accuracy and deployment speed while maintaining competitive perceptual quality and compact representation size.

2. Related Work

Person-specific Human Avatar Representations. Human avatar models can be broadly grouped into parametric, implicit, and explicit representations. Parametric body models such as SMPL [25] and SMPL-X [29] provide compact and articulated human priors, making them widely used as geometric proxies for pose control, animation, and cross-subject alignment. However, a parametric mesh alone cannot faithfully represent subject-specific appearance or visual details. To address this limitation, implicit representations combine geometry and appearance with continuous fields. For example, Neural Body [31] attaches structured latent codes to a deformable SMPL mesh and decodes them into a neural radiance field for sparse-view rendering, showing strong reconstruction fidelity for dynamic humans. Besides human neural rendering with radiance fields [3, 21, 24, 43], other studies also explore signed distance fields [7, 14], and occupancy-based implicit surfaces [37, 38]. Nevertheless, these methods are often computationally expensive, while pose-driven control can remain indirect, since pose changes must be mediated through learned latent or deformation fields.

Recently, 3D Gaussian Splatting [15] has shifted avatar modeling towards explicit primitives. These methods usually represent humans as canonical Gaussian sets with pose-

[1] https://github.com/tapri-lab/NLF-GS

Method Family	Examples	Fidelity	Generalization	Pose Drivability	Deployment Efficiency
Parametric model	SMPL [25], SMPL-X [29]	↓	↑	↑	↑
Person-specific implicit avatar	Neural Body [31]	↑	↓	→	↓
Person-specific Gaussian avatar	3DGS-Avatar [35]	↑	↓	↑	→
Generalizable implicit avatar	SHERF [12], MPS-NeRF [8]	→	↑	→	↓
Generalizable Gaussian avatar	GHG [18], GIGA [52]	↗	↑	↓	↗
NLF-GS (ours)	Mixed mesh-Gaussian	↗	↑	↑	↗

Table 1. Qualitative comparison of representative avatar generation methods under the four requirements of telepresence-oriented avatar reconstruction. (↑ High, ↗ Medium-High, → Medium, ↓ Low/Limited)

dependent deformation. For example, 3DGS-Avatar [35] generates the avatar by optimizing a set of 3D Gaussians in canonical space with deformation signals. Similar Gaussian formulations have been explored for monocular avatars, multi-view human rendering, head avatars, and human-scene compositions by [16, 19, 22, 27, 34, 51]. Moreover, some works have bridged the gap between implicit and explicit modeling to leverage the strengths of different representations. GAvatar [46], for example, uses neural implicit fields to predict Gaussian attributes, while HeadGaS [6] extends 3DGS with learnable latent features blended with parametric head controls. These methods demonstrate the fidelity that avatar representations can reach, but their dependence on per-subject optimization limits scalability. This thesis instead targets feed-forward reconstruction for unseen identities while keeping an explicit body structure for pose drivability.

Generalizable Avatar Representation. While the above works demonstrate a convincing visual quality of avatar representation, many of them rely on person-specific optimization, limiting their scalability in telepresence. Generalizable human avatar representations, however, can alleviate this limitation by training across identities and being used to generate unseen subjects. Early works in this direction mainly build on implicit representations by conditioning the radiance field on image features or body priors [8, 12, 17, 28]. For example, MPS-NeRF [8] combines multi-view image features with a canonical human NeRF and body-guided deformation for zero-shot rendering. Although effective, these approaches incur the heavy cost of volume rendering. Later, other research predict 3D Gaussian parameters from sparse views, UV-space features, or template-aligned representations to achieve feed-forward reconstruction and efficient novel-view rendering for unseen humans [18, 30, 49, 52]. In particular, GHG [18] regresses human Gaussians in the UV space of a template over multiple layers, while GIGA [52] uses a texture-space formulation to aggregate sparse-view image evidence, body shape, and Gaussian parameters. Compared with previous methods, these approaches better match the deployment requirements of interactive avatars, especially in terms of feed-forward inference and efficient rendering. While existing generalizable methods improve scalability, they often separate appearance inference from pose-drivable structure. NLF-GS addresses this gap by predicting Gaussian attributes directly on a shared SMPL-X scaffold.

Drivable Avatar Representation. Unlike generalizable avatar reconstruction, which emphasizes zero-shot inference, drivable avatar reconstruction focuses on motion-conditioned control of a specific identity. Early neural rendering methods achieve this by defining avatars in canonical space and mapping between canonical and posed spaces using body-model-guided or skeleton-conditioned deformation fields [3, 21, 24, 31, 43]. For example, Neural Actor [24] uses a coarse body model to canonicalize posed observations and learns pose-dependent residual information for controllable novel-pose rendering. A more structured direction attaches Gaussian primitives to explicit deformation carriers. Namely, D3GA [51] uses tetrahedral cages as a geometrically coherent motion carrier to embed the Gaussian primitives inside. SplattingAvatar [40] instead embeds Gaussians on a triangle mesh using barycentric coordinates and local displacements, letting the mesh describe low-frequency body motion while Gaussians capture high-frequency appearance details. Other mesh-guided Gaussian avatars further extend this idea to expressive body, hand, and face control, such as [11, 26, 32]. These works suggest that drivability benefits from attaching visual primitives to explicit human structure. These mesh-guided Gaussian avatars show the value of explicit motion carriers, but most are still optimized for a specific person or sequence. This motivates bringing the structural idea into a generalizable setting by using SMPL-X as a shared anchor for feed-forward Gaussian prediction.

The discussion above suggests that different avatar representation families satisfy different parts of the telepresence requirement set. Table 1 summarizes these trade-offs.

Dataset	Type	Data Form
THuman [50]	Static	3D scans
THuman2.0 [45]	Static	3D scans + SMPL-X
RenderPeople [36]	Static	Textured scans
PeopleSnapshot [1]	Dynamic	Monocular video
ZJU-MoCap [31]	Dynamic	Multi-view video
DNA-Rendering [4]	Large-scale	Multi-view sequences
HuMMan [2]	Large-scale	Multi-modal sequences
GeneBody [17]	Large-scale	Multi-view sequences
ActorsHQ [13]	Large-scale	High-quality sequences

Table 2. Comparison of representative human avatar datasets. THuman2.0 is selected in this thesis because it provides high-quality clothed human scans with fitted SMPL-X parameters, making it suitable for controlled sparse-view avatar reconstruction.

Human Avatar Datasets. Existing human avatar datasets can be divided into three categories based on whether they emphasize static geometry, movement, or large-scale generalization, as seen in Table 2. Static scan datasets such as THuman [50], THuman2.0 [45], and RenderPeople [36] provide high-quality clothed human geometry and texture and are widely used for image-based human reconstruction. Dynamic datasets PeopleSnapShot [1] and ZJU-MoCap [31] offer monocular or multi-view videos of people under different motions, making them suitable for person-specific optimization or novel pose synthesis. Large-scale datasets including DNA-Rendering [4], HuMMan [2], GeneBody [17] and ActorsHQ [13] further increase identity diversity, motion variation, and annotation richness. Among them, THuman2.0 best matches our sparse-view reconstruction setting. It contains hundreds of different identities with clothing textures and fitted SMPL-X parameters, which allows us to render multi-view observations while keeping a consistent body alignment.

3. Method

3.1. Problem Setting

This thesis focuses on the representation layer of telepresence-oriented avatar reconstruction. Given sparse multi-view RGB observations, camera parameters, and corresponding SMPL-X body parameters, the goal is to reconstruct a subject-specific avatar representation that can be rendered from free viewpoints and remains drivable with novel poses.

3.2. Design Principles

To address the requirements discussed in section 1, we adopt a mixed mesh-Gaussian representation. We choose this instead of relying on a purely implicit field (like NeRF-based methods) or an unconstrained Gaussian set, which are both discussed in section 2. In our design, the SMPL-X mesh template provides a shared human scaffold with a consistent body structure, while Gaussian primitives provide efficient rendering and local appearance flexibility. Therefore, the template is not treated as the final reconstructed geometry, but as an organizing structure for the Gaussian attributes.

The second principle is to predict avatar attributes from localized multi-view evidence instead of relying on subject-specific optimization. Inspired by the query-based philosophy of NLF [39], each Gaussian acts as a local reconstruction query: it projects into the sparse input views, samples image features at its corresponding locations, and predicts its own attributes through a shared decoder. This design connects the mixed representation directly to the four requirements: structure supports pose drivability, local feature prediction supports visual fidelity, feed-forward inference supports generalization, and compact Gaussian rendering supports deployment efficiency.

3.3. Overview

The overall framework of our proposed method is illustrated in Fig. 2. Our goal is to learn a generalizable encoder-decoder-based model that reconstructs a subject-specific Gaussian avatar from sparse multi-view images. During training, the model is optimized over many subjects so that it can infer a structured avatar for unseen identities in a feed-forward manner at test time.

In this framework, we build the representation on a canonical mesh-anchored Gaussian template, where each Gaussian serves as a query point for appearance reconstruction. Given the input views of a subject, we first extract dense per-view image features using a Feature Pyramid Network[23] composed of ResNet50[9] as the backbone. The 3D Gaussian primitives are then projected into each view using the corresponding camera parameters, and local features are sampled at the projected image locations. Since the sampled features are not equally reliable across views, we compute per-view weights based on geometric cues to obtain an aggregated feature for each Gaussian, which is then passed through a shared MLP decoder to predict attributes. Finally, the predicted Gaussians are rendered with a differentiable Gaussian renderer, and the entire framework is trained end-to-end using composite losses.

3.4. Mesh-anchored Gaussian Representation

SMPL-X parametric model. SMPL-X [29] is a parametric human body model with fully articulated hands and facial expressions, whereas SMPL only covers the body. SMPL-X uses standard vertex-based linear blend skinning [20] with $N = 10475$ vertices and $K = 54$ joints. It is defined by a function $M(\theta, \beta, \psi)$, where θ denotes pose parameters, β denotes shape parameters and ψ denotes facial expression parameters.

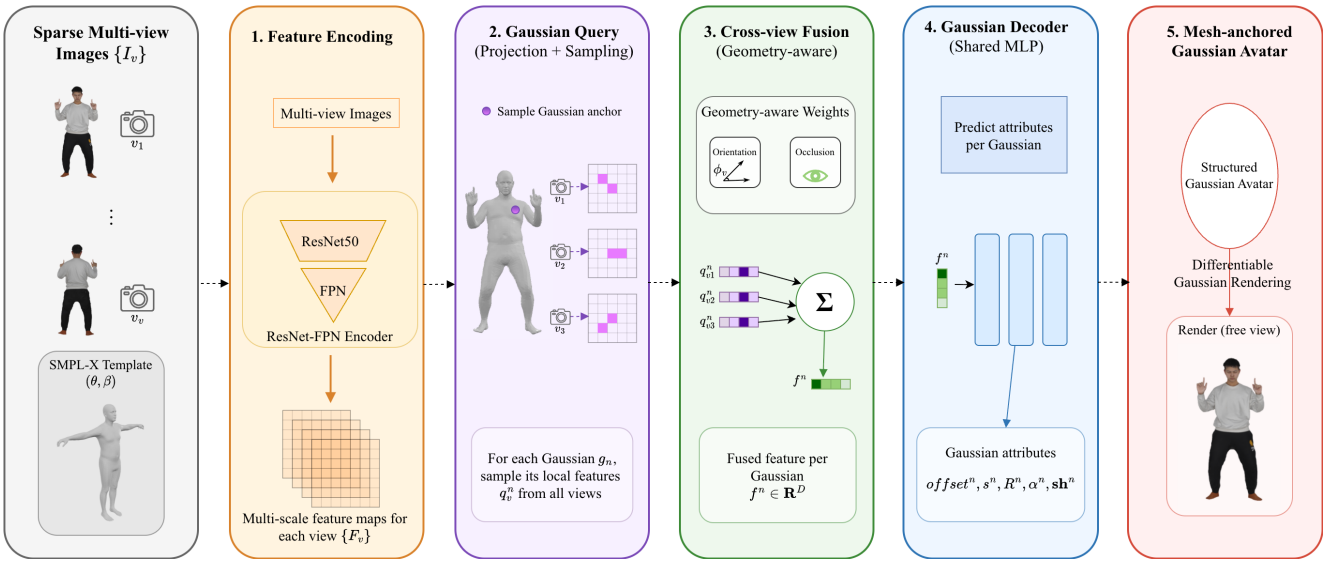


Figure 2. Overview of the proposed pipeline. Using a SMPL-X template for spatial anchoring, our framework processes sparse multi-view images through a ResNet-FPN encoder to generate dense feature maps. We then project each Gaussian into all views to sample local features, fusing them via a geometry-aware weighting to generate a unified descriptor. The unified descriptor is decoded into Gaussian attributes by a shared MLP, and the predicted Gaussian avatar is rendered with differentiable Gaussian splatting.

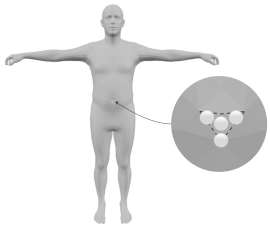


Figure 3. Illustration of the Mesh-anchored canonical Gaussian template. The scaled part demonstrates how Gaussian primitives (the white spheres) are attached to the SMPL-X surface.

Canonical template. We use a canonical SMPL-X mesh in T-pose as a shared structural scaffold for the avatar representation. This template is not the final reconstructed avatar itself; instead, it provides a consistent reference frame that organizes Gaussian primitives over the human body surface.

Mesh-anchored Gaussian parameterization. Let $\mathbf{f}_f = (\mathbf{v}_{f,1}, \mathbf{v}_{f,2}, \mathbf{v}_{f,3})$ denote the f-th face of the canonical mesh. We associate each face with a fixed set of K Gaussian primitives, as seen in Figure 3. The k-th Gaussian on face f is defined as

$$g_{f,k} = \left\{ \mathbf{x}_{f,k}^{\text{can}}, \mathbf{r}_{f,k}, \mathbf{s}_{f,k}, \alpha_{f,k}, \mathbf{c}_{f,k}^{\text{SH}}, \mathbf{p}_{f,k} \right\},$$

where $\mathbf{x}_{f,k}^{\text{can}} \in \mathbb{R}^3$ is its positional center, $\mathbf{r}_{f,k} \in \mathbb{R}^4$ is the rotation quaternion, $\mathbf{s}_{f,k} \in \mathbb{R}^3$ is the anisotropic scale, $\alpha_{f,k} \in \mathbb{R}$ is the opacity, $\mathbf{c}_{f,k}^{\text{SH}}$ denotes the spherical harmonics coefficients and $\mathbf{p}_{f,k}$ refers to the parent face index to preserve the connection between the Gaussians and the mesh. In our representation, we fix $K = 4$ for a simple, but robust structure.

The positional center is parameterized relative to the associated mesh face:

$$\mathbf{x}_{f,k}^{\text{can}} = \sum_{j=1}^{3} \beta_{f,k,j} \mathbf{v}_{f,j} + \Delta \mathbf{x}_{f,k}, \tag{1}$$

where $\beta_{f,k,j}$ are fixed barycentric weights and $\Delta \mathbf{x}_{f,k}$ is a predicted local offset. This parameterization preserves mesh-based spatial organization while allowing local geometric adaptation.

The Gaussian attributes for the target subject relative to this canonical scaffold are predicted by our model. In this way, the template defines the structure of the representation, while the decoded Gaussian parameters instantiate a subject-specific avatar.

3.5. Feature Encoding and Query

Per-view feature encoding. Given sparse multi-view input images, we extract dense feature maps independently for each view using a ResNet50-FPN encoder. The encoder provides multi-scale image features while preserving spatial correspondence, which is necessary for the subsequent Gaussian-level feature query. Since the encoder architecture follows a standard Feature Pyramid Network (FPN) design, we only include its role here. For implementation details, please refer to Appendix A.

Template-guided Gaussian feature query. To associate each Gaussian primitive with visual evidence, we query the per-view feature maps at the projected image location of the Gaussian center. This design is inspired by NLF [39], whose pose reconstruction is driven by the features sampled at spatially grounded query locations. In our case, the query locations are provided by the structured Gaussian anchors defined on the canonical SMPL-X template.

Let $\mathbf{x}_{f,k} \in \mathbb{R}^3$ denote the 3D center of the k-th Gaussian attached to face f. For view v, with camera intrinsics $\mathbf{K}_v$ and extrinsics $(\mathbf{R}_v, \mathbf{t}_v)$, the Gaussian center is projected into the image plane as

$$\tilde{\mathbf{u}}^v_{f,k} = \mathbf{K}_v \left(\mathbf{R}_v \mathbf{x}_{f,k} + \mathbf{t}_v\right), \tag{2}$$

where $\tilde{\mathbf{u}}^v_{f,k}$ is the homogeneous image coordinate.

We then convert this image-space coordinate to the feature-space normalized coordinate $\mathbf{u}^v_{f,k}$ and use bilinear grid sampling to obtain the per-view Gaussian feature for each feature map $\mathbf{F}_v$:

$$\mathbf{q}^v_{f,k} = \mathrm{GridSample}\big(\mathbf{F}_v, \mathbf{u}^v_{f,k}\big) \in \mathbb{R}^C. \tag{3}$$

Here, $\mathbf{q}^v_{f,k}$ denotes the feature of Gaussian (f,k) observed from view v. Since the sampled descriptor is tied to the projected position of an individual Gaussian, the model can recover spatially varying appearance cues in a localized manner. The per-view queried features are then passed to the subsequent cross-view fusion module to obtain a unified Gaussian-level descriptor.

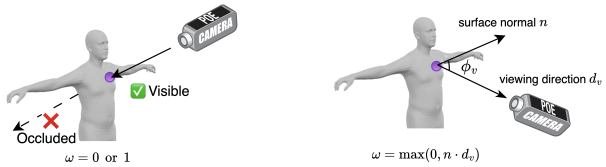
3.6. Cross-View Fusion and Gaussian Prediction

Geometry-aware weighting. A significant challenge in multi-view reconstruction is the aggregation of features from disparate viewpoints. To address this, we implement a visibility-aware fusion mechanism that calculates a contribution weight w_{nv} for each Gaussian-view pair. The final aggregated feature $\bar{\mathbf{f}}_n$ is computed as a weighted sum of the features from all visible views.

The weighting strategy is twofold, incorporating both orientation (Figure 4a) and occlusion (Figure 4b). First, we calculate an orientation-based weight using the dot product between the Gaussian's surface normal $\mathbf{n}_{nv}$ and the viewing direction $\mathbf{d}_v$ from the camera:

$$w_{ori} = \max(0, \mathbf{n}_{nv} \cdot \mathbf{d}_v) \tag{4}$$

This ensures that views facing the Gaussian contribute most significantly, while those viewing it from a rear angle are attenuated. Second, we perform an occlusion test by comparing the depth of the Gaussian center against a pre-computed depth map. A binary visibility mask is assigned, where $w_{occ} = 1$ if the Gaussian is not occluded and $w_{occ} = 0$ otherwise. The final fusion weight is the product of these two components: $w = w_{ori} \cdot w_{occ}$. This mechanism helps the aggregated features to represent the actual visible appearance of the subject.



(a) Orientation-based weight (b) Occlusion-based weight

Figure 4. Geometry-aware weighting

Gaussian Prediction. The fused feature of each Gaussian is finally passed to a shared MLP decoder that predicts its subject-specific attributes defined in subsection 3.4. Since the decoder operates independently on each Gaussian, the prediction process remains parallel and lightweight. To ensure numerical stability, the predicted parameters are mapped to valid ranges through standard output parameterizations, such as normalization for rotation and bounded activations for scale and opacity.

3.7. Training Objectives

Differentiable rendering. The predicted Gaussian representation is rendered into the target views using a differentiable Gaussian splatting renderer gsplat [44]. This enables gradients from image-space supervision to be propagated back to the Gaussian attributes and the upstream feature backbone.

Loss Functions. The model is trained with a composite objective

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{5}$$

where the λ terms control the relative contribution of each loss component.

- **Reconstruction loss.** We supervise the rendered image $\hat{I}$ against the ground-truth image I using an L1 loss $\mathcal{L}_1$ and an SSIM loss $\mathcal{L}_{ssim}$ [42]. The L1 term enforces pixel-level appearance consistency, while SSIM encourages the preservation of local structure and contrast. In addition, we apply a perceptual loss $\mathcal{L}_{perc}$ on image intensity and gradient features to better preserve visual details near the edges. The final reconstruction term is defined as:

$$\mathcal{L}_{\text{rec}} = \lambda_{l1}\mathcal{L}_1 + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{perc}\mathcal{L}_{perc} \tag{6}$$

- **Silhouette loss.** To improve foreground shape alignment, we include a silhouette loss $\mathcal{L}_{sil}$ between the rendered mask and the ground-truth mask. This term helps constrain the spatial extent of the reconstructed avatar and reduces boundary artifacts.
- **Regularization loss.** We apply structural regularization $\mathcal{L}_{reg}$ to keep the predicted Gaussians well-behaved with respect to the mesh-anchored template. In particular, we

regularize the Gaussian scales ($\mathcal{L}_{scale}$), positional offsets ($\mathcal{L}_{\Delta x}$), and opacity ($\mathcal{L}_{\alpha}$). The scale and offset terms are penalized to discourage over-expanded Gaussians and excessive deviation from their mesh anchors. For opacity, we apply a prior that encourages solid foreground coverage. The overall regularization loss is defined as:

$$\mathcal{L}_{\text{reg}} = \lambda_{scale}\mathcal{L}_{scale} + \lambda_{\Delta x}\mathcal{L}_{\Delta x} + \lambda_{\alpha}\mathcal{L}_{\alpha} \tag{7}$$

Specifically, reconstruction loss is the main loss, with the other two serving as the auxiliary losses.

4. Experiments

4.1. Experiment Settings

Baselines. We compare our method with two representative categories of generalizable human avatar reconstruction methods: the implicit NeRF-based baseline SHERF [12] and the explicit Gaussian-based baseline GHG [18]. These baselines allow us to position our method with respect to both implicit and explicit generalizable avatar representations.

Dataset. We evaluate our method on the THuman2.0 dataset, which contains 526 high-quality 3D human scans together with corresponding SMPL-X parameters. Following GHG's experimental protocol, we use the first 426 subjects for training and the remaining 100 subjects for testing.

Evaluation metrics. The experiments evaluate NLF-GS from three aspects: zero-shot reconstruction fidelity, deployment efficiency, and pose drivability. For zero-shot reconstruction fidelity generalization in Section 4.2, we report standard image-quality metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [42], and LPIPS [48]. PSNR and SSIM measure pixel-level and structural reconstruction accuracy, while LPIPS provides a perceptual similarity measure that better reflects perceived visual quality.

For pose drivability in Section 4.3, we mainly provide qualitative results since ground-truth images of the same identities under the target novel poses are not available. The evaluation focuses on whether the reconstructed avatar follows the target motion while preserving identity and appearance consistency.

For deployment efficiency in Section 4.4, we compare avatar generation speed, avatar driving speed, number of required input views, and representation size. These factors are relevant to telepresence-oriented applications because they affect runtime latency, capture complexity, and potential transmission cost.

Finally, the controlled studies in Sections 4.5–4.8 use PSNR, SSIM, and LPIPS under the same evaluation protocol to analyze the effect of input-view number, architecture choices, and auxiliary loss terms. For these experiments, we calculate the metrics using the training views rather than novel views.

4.2. Zero-shot Reconstruction Fidelity

To evaluate zero-shot reconstruction fidelity, we test NLF-GS on unseen THuman2.0 subjects without subject-specific optimization. All avatars are normalized to a height of 1.8m and rendered at 1024×1024 resolution to reduce the effect of scale and framing differences on image-space metrics. We report PSNR, SSIM, and LPIPS, and compare with SHERF and GHG as representative implicit and Gaussian-based generalizable baselines.

Table 3 reports the quantitative generalization results on test subjects for novel views (60°, 180°, and 300°). These views fall exactly halfway between the training views (0°, 120°, and 240°), thoroughly testing the model's ability to render unseen angles [1]. Our method gets much better PSNR and SSIM scores than SHERF. Compared to GHG, our method has a slightly better SSIM (0.950 vs 0.948) but a lower PSNR (27.146 vs 29.267). In terms of LPIPS, our results outperform SHERF while close to GHG, suggesting that the model preserves perceptual similarity and identity-specific details at a similar level. Taken together, these results indicate that our representation strikes a balance between precise reconstruction and identity preservation.

Figure 5 provides qualitative comparisons that further support these observations. Compared with SHERF, our method shows clearer and more consistent visual improvements, which agrees with the quantitative results in Table 3. When compared to GHG, our method produces sharper and more coherent textures, particularly for fine-grained patterns (e.g., the patterns on the first row). A likely reason is that GHG relies on an inpainting-based strategy to estimate unseen regions, which can introduce smoothing artifacts. However, GHG outperforms our method in edge areas like hands and feet. This limitation in our method mainly stems from its reliance on SMPL-X parameters for pose alignment. Since these parameters are estimated and inherently imperfect, small misalignment between the SMPL-X human body template and the actual human geometry can lead to artifacts, especially in high-frequency or articulated regions. In contrast, GHG benefits from its multi-scaffold design, where the human surface is dilated into multiple layers, allowing it to better capture fine geometric details in such areas.

4.3. Drivability Demonstration

To examine pose drivability, we reconstruct avatars from the default sparse-view input (3 views) and animate them

[1] Due to computational constraints, we employ the official SHERF checkpoint pre-trained on the THuman dataset for evaluation on THuman2.0 without further fine-tuning.

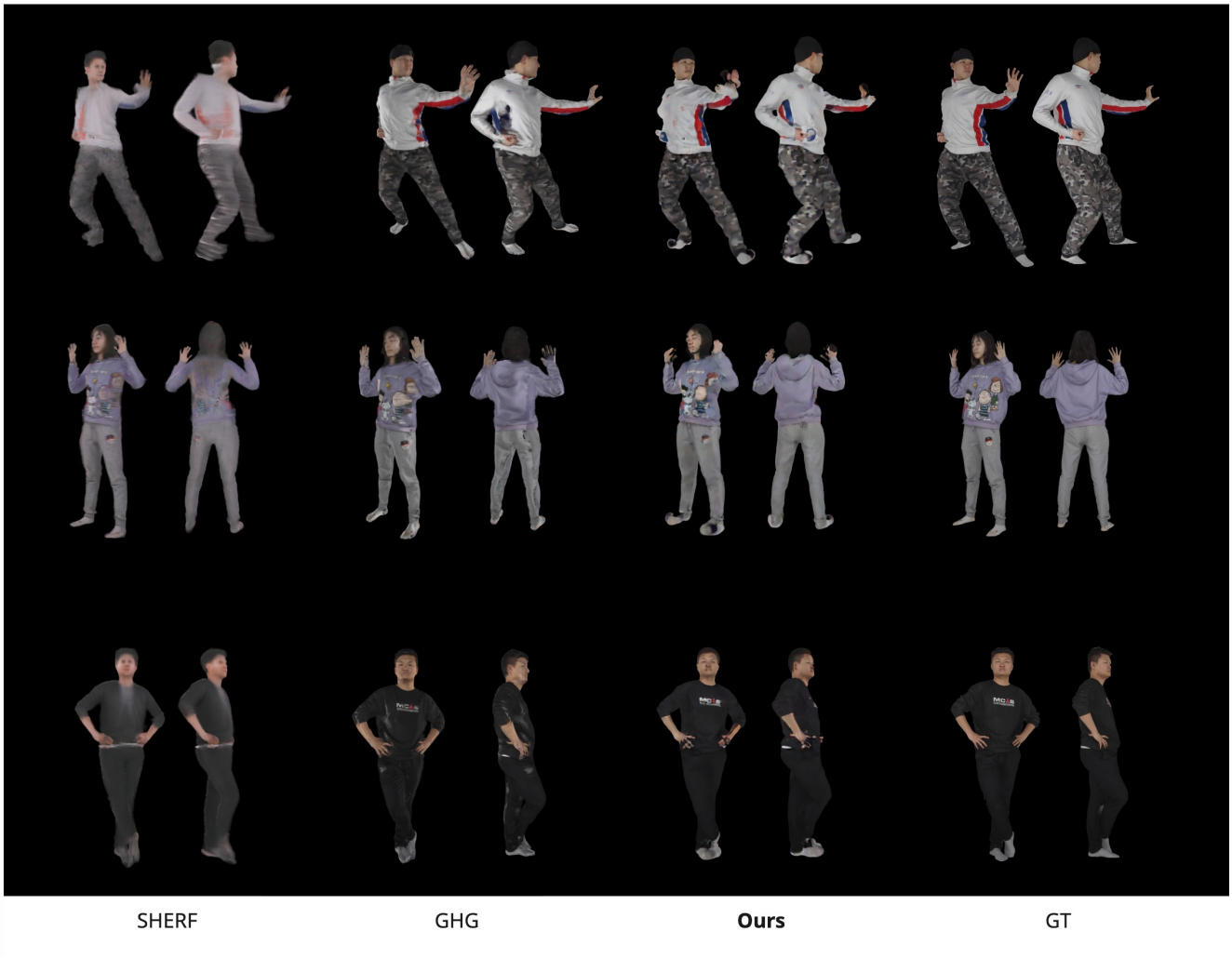


Figure 5. **Qualitative generalization results on reconstruction quality.** We compare our method against SHERF and GHG on the THuman2.0 dataset [45]. SHERF uses 24 views for training but only requires single view for testing, GHG and our method use 3 views for training and testing.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
SHERF	21.603	0.826	0.230
GHG	**29.267**	0.948	**0.060**
NLF-GS (ours)	27.146	**0.950**	0.071

Table 3. **Quantitative comparison on reconstruction quality.** We evaluate generalization to unseen identities, where SHERF uses 24 views for training but only requires a single view for testing, GHG and our method use 3 views for training and testing. See Figure 5 for qualitative results.

using novel SMPL-X pose sequences. Since ground-truth images of the same identities under these target poses are not available, this part is evaluated qualitatively. The focus is whether the reconstructed avatar follows the target motion while preserving identity and appearance consistency. We provide qualitative drivability results in Figure 6 with sequential SMPL-X parameters provided by [10]. These results support our design motivation of using a mixed representation to balance controllable body structure and appearance details, as this combination allows the avatar to be reposed without requiring a separate optimization process for each target motion. At the same time, the visible artifacts (see the second row) show that stronger deformation modeling, temporal consistency, and pose-dependent appearance prediction are necessary for future telepresence-ready avatars.

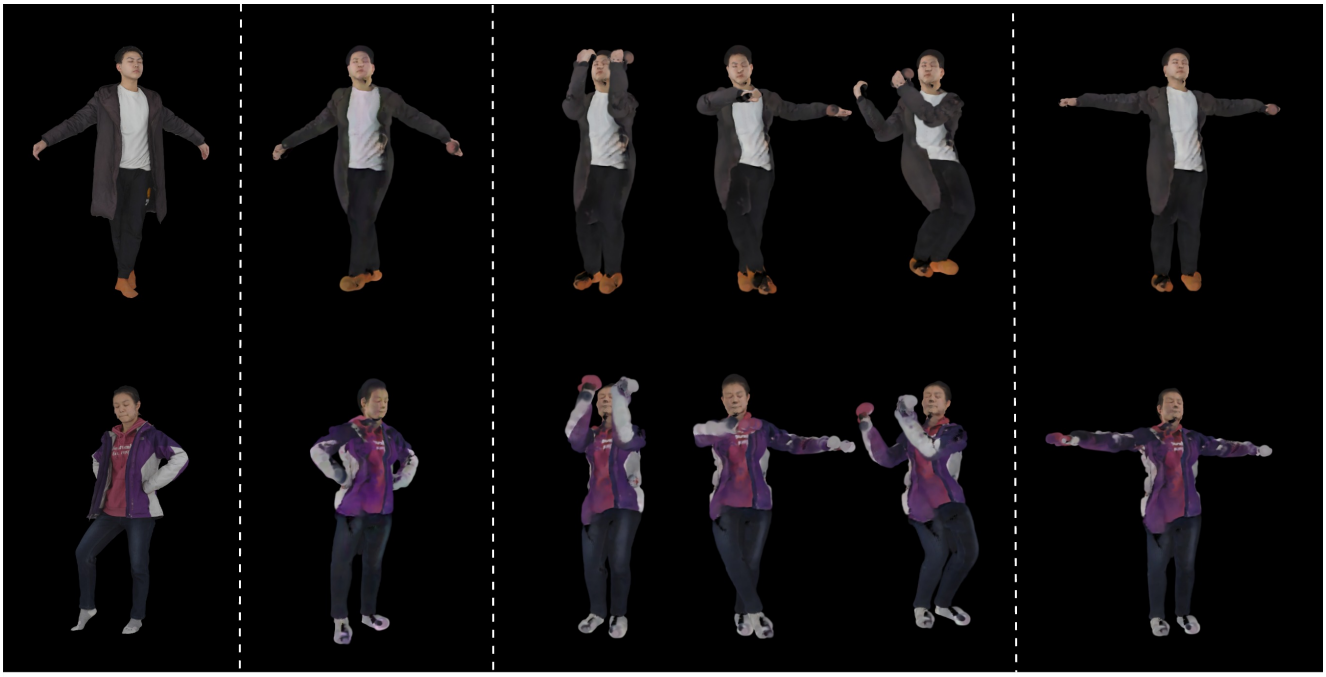


Figure 6. **Qualitative drivability results.** From left to right, we show the ground-truth image, a rendered novel view, several pose-driven animation results under unseen target poses, and a canonical T-pose rendering.

4.4. Deployment Efficiency

Beyond reconstruction quality, telepresence-oriented avatar representations must remain practical under runtime and transmission constraints. Therefore, we compare deployment-related characteristics, including avatar generation speed, avatar driving speed, input-view requirement, and representation compactness. Avatar generation speed measures how fast a method reconstructs a new avatar for an unseen subject from sparse observations, while avatar driving speed measures how fast an already generated avatar can be animated under novel SMPL-X poses. The former reflects the cost of onboarding a new user, whereas the latter is more directly related to real-time telepresence interaction.

As shown in Table 4, NLF-GS achieves 9.57 FPS on a RTX 4070 GPU for generating a new avatar, outperforming SHERF and GHG in feed-forward reconstruction speed. More importantly, once the avatar has been generated, NLF-GS reaches 196.21 FPS for pose-driven animation. This is substantially faster than SHERF, which supports driving but still performs rendering through a neural-field-based pipeline. This result also highlights the benefit of separating avatar generation from avatar driving: the relatively more expensive reconstruction stage only needs to be performed when creating a new avatar, while subsequent pose-driven use can be executed at much higher speed. In terms of capture requirements, our method requires three input views, matching GHG and requiring more views than SHERF. This reflects a trade-off between capture simplicity and reconstruction quality. However, NLF-GS uses substantially fewer Gaussian primitives per avatar than GHG, with approximately 80k Gaussians compared with 1–5M. This compact representation reduces potential storage and transmission overhead, which is important for practical telepresence deployment.

Method	Gen. FPS ↑	Drive FPS ↑	# Views ↓	# Gaussians ↓
SHERF	2.02	2.02	**1**	N/A
GHG	3.26	N/A	3	1–5M
NLF-GS	**9.57**	**196.21**	3	**≈80k**

Table 4. **Comparison of deployment-related efficiency.** Gen. FPS measures the speed of reconstructing a new avatar from sparse input views. Drive FPS measures the speed of driving an existing avatar under novel SMPL-X poses. This value is not reported for GHG because pose-driven animation is not supported. # Views denotes the number of input images required for avatar generation. # Gaussians denotes the number of Gaussian primitives per avatar. This value is not applicable to SHERF because it is an implicit method.

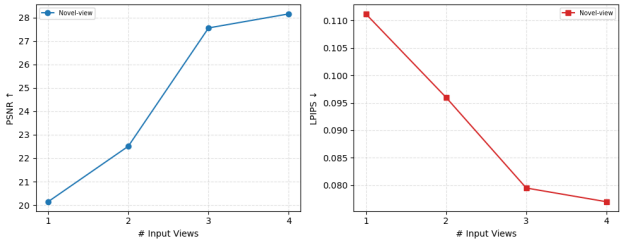


Figure 7. **Experiment on the number of input views for training.** We train models with 1, 2, 3, and 4 input views and evaluate them on novel views. Increasing the number of input views improves reconstruction quality, but the improvement from 3 to 4 views is relatively small, indicating that three views provide a practical trade-off between quality and capture cost.

4.5. Input-view Requirement for Free-view Reconstruction

A practical telepresence avatar should be reconstructed from as few input views as possible while the results remain reliable from any viewpoint. We therefore study how the number of input views affects free-view reconstruction quality. Each model is evaluated on novel views that are not used as input. We choose evenly distributed input views: for one view, we use the front view; for two views, we use the front and back; and for three views, we use cameras spaced 120 degrees apart; for four views, we use cameras spaced 90 degrees apart. As shown in Figure 7, increasing the number of input views consistently improves novel-view reconstruction quality. The one-view and two-view settings suffer from insufficient spatial coverage, leading to stronger ambiguity in occluded or unseen body regions. In contrast, using three input views provides substantially better coverage of the full body and captures most of the available geometric and appearance information. Adding a fourth view further improves the metrics, but the gain is relatively small compared with the additional capture cost. Based on this trade-off, we use three input views as the default setting in the main experiments. This choice provides a practical balance between reconstruction quality and input-view requirement, which is important for deployment-oriented avatar reconstruction.

4.6. Effect of Feature Encoder

We first analyze the effect of the feature encoder while keeping the rest fixed. This experiment asks whether multi-scale image features are important for predicting Gaussian attributes. All variants use the same evaluation protocol as before, and the only difference lies in how per-view image features are extracted.

As shown in Table 5, the full model with a finetuned ResNet-FPN encoder achieves the best performance. Compared with the frozen ResNet-FPN variant, joint finetuning improves all metrics, indicating that the encoder benefits from adapting to the avatar reconstruction objective rather than relying only on generic image features. Compared with the single-scale ResNet encoder, the FPN-based model also performs better, showing that multi-scale features are useful for Gaussian attribute prediction. This is expected because human appearance contains both local details, such as hands or faces, and larger body-level context, such as clothing and global appearance.

Encoder	PSNR ↑	SSIM ↑	LPIPS ↓
Single-scale ResNet	25.326	0.941	0.084
Frozen ResNet-FPN	27.745	0.953	0.072
Finetuned ResNet-FPN	**29.245**	**0.963**	**0.064**

Table 5. **Effect of feature encoder.** We compare different feature extraction designs while keeping the fusion module, decoder, losses, and training protocol fixed.

4.7. Effect of Cross-View Fusion

We then study the effect of the cross-view fusion mechanism while keeping the feature encoder fixed as the finetuned ResNet-FPN encoder. This experiment shows how different fusion strategies affect sparse-view avatar reconstruction. Specifically, we compare average fusion, occlusion-only fusion, orientation-only fusion, and the full geometry-aware fusion that combines both orientation and occlusion cues.

As shown in Table 6, average fusion performs the worst across all metrics, confirming that treating all input views equally is suboptimal in sparse multi-view reconstruction. Occlusion-only fusion significantly improves PSNR and SSIM over average fusion, suggesting that removing invisible or blocked observations helps pixel-level reconstruction. However, its LPIPS remains unchanged, indicating that occlusion reasoning alone is insufficient to improve perceptual quality.

Orientation-only fusion achieves the highest PSNR and matches the best SSIM, showing that viewing direction is the most influential cue for selecting useful local evidence. This is reasonable because a Gaussian is usually best reconstructed from views where its associated surface is directly facing the camera. However, orientation alone cannot handle cases where a front-facing region is still occluded by another body part or clothing surface.

The full fusion strategy combines both cues and achieves the best LPIPS, while maintaining similar PSNR and SSIM to orientation-only fusion. Although its PSNR is slightly lower than orientation-only fusion, the improved LPIPS suggests that occlusion reasoning provides complementary perceptual benefits once reliable orientation-based view selection is already established. These results indicate that

orientation is the dominant factor for cross-view feature aggregation, while occlusion acts as a useful complementary cue for improving perceptual consistency.

Fusion Strategy	PSNR ↑	SSIM ↑	LPIPS ↓
Average fusion	25.786	0.945	0.079
Occlusion-only fusion	28.236	0.956	0.079
Orientation-only fusion	**29.481**	0.963	0.068
Full fusion	29.245	**0.963**	**0.064**

Table 6. **Effect of cross-view fusion.** We compare different view aggregation strategies while keeping the feature encoder, decoder, losses, and training protocol fixed.

4.8. Ablation Studies

We evaluate the auxiliary supervision terms by removing one loss component at a time from the full objective. Since the reconstruction loss is essential for training, we keep this part fixed and ablate only the silhouette and regularization terms. All variants use the same architecture, three-view input setting, training split, and 40-epoch training schedule.

Table 7 shows that removing any single loss term has only a minor impact on PSNR and SSIM, but consistently worsens LPIPS. This suggests that while these auxiliary losses do not significantly affect pixel-level accuracy, they play a more important role in preserving perceptual quality.

One reason for this behavior may be the relatively small weights assigned to these terms (e.g., $w_{\text{opacity}} = 0.005$, $w_{\text{scale}} = 0.01$, $w_{\text{offset}} = 0.01$, $w_{\text{sihouette}} = 0.1$, compared to $w_{\text{L1}} = 0.5$), which limits their influence on overall optimization. In addition, some parameters, such as opacity, are already indirectly supervised by the photometric reconstruction loss, as they directly affect the rendered image through differentiable Gaussian splatting. As a result, explicitly constraining these parameters provides limited additional benefit in terms of reconstruction accuracy.

Nevertheless, the consistent increase in LPIPS when these terms are removed indicates that these regularization terms contribute to better perceptual consistency and help stabilize the learning of Gaussian attributes.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Opacity Loss	**29.338**	**0.964**	0.068
w/o Offset Loss	29.245	0.963	0.068
w/o Scale Loss	29.321	0.963	0.068
w/o Silhouette Loss	29.253	0.963	0.069
Full loss	29.245	0.963	**0.064**

Table 7. **Ablation study of different regularization components.** All models are trained for 40 epochs.

5. Discussion

5.1. Summary

This thesis investigates how a human avatar representation can maintain visual fidelity, generalize to unseen identities, realize pose drivability, and achieve deployment efficiency for telepresence-oriented applications. The proposed NLF-GS framework addresses this question by combining an explicit SMPL-X scaffold with Gaussian primitives, where the scaffold provides a consistent human structure and the Gaussian primitives provide efficient rendering and local appearance flexibility.

The experimental results suggest that this mixed representation is a practical direction within the current scope. Quantitative reconstruction results show that the model can generalize to unseen THuman2.0 subjects without subject-specific optimization. The drivability demonstrations further show that the reconstructed avatars can be animated using novel SMPL-X poses. Finally, the compact representation size and real-time driving speed indicate a path toward deployment-oriented avatar representations, although the current system is not yet a full telepresence solution.

5.2. Limitations and Future Directions

Feature alignment and SMPL-X dependency. Despite the proposed progress toward generalizable and drivable avatar reconstruction, several limitations remain. First, the visual fidelity of the reconstructed avatars still lags behind recent Gaussian-avatar methods that optimize a representation for a single subject. This is partly a consequence of our generalizable design: NLF-GS predicts Gaussian attributes from locally sampled image features in latent space. The reconstruction quality is sensitive to the accuracy of the estimated SMPL-X parameters, since errors in body localization directly affect where features are sampled. In addition, projecting 3D Gaussian locations onto 2D feature maps introduces potential discretization and alignment errors between image coordinates and feature coordinates. This could be improved by replacing the current image-feature sampling strategy with a more stable texture-based representation. For example, mapping Gaussian primitives to a UV texture space may provide more accurate and consistent feature localization than directly projecting Gaussian centers onto image feature maps. Such a design could reduce errors caused by image-feature misalignment, but texture-space mapping is also more computationally expensive and may require additional foreground extraction or post-processing to avoid allocating representation capacity to background regions. Therefore, its efficiency and deployment costs must be considered for telepresence applications.

In addition, the current pipeline still depends on SMPL-X parameters as input. In a practical telepresence scenario,

these parameters must be estimated from incoming images or video streams, introducing additional computational cost and another possible source of error. Since many existing SMPL-X estimators, such as PyMAF-X [47], already infer body parameters from image features, a natural extension is to integrate SMPL-X estimation directly into the avatar reconstruction pipeline. This would enable an end-to-end system that maps captured images or videos to a structured Gaussian avatar without relying on externally pre-computed body parameters. At the same time, joint optimization of body estimation and Gaussian reconstruction could reduce error propagation between the localization and anchoring.

Uniform Gaussian allocation. Currently, the use of a relatively compact set of Gaussian primitives further limits the model's ability to capture high-frequency details. This uniform allocation is simple and structured, but it does not reflect the uneven distribution of visual complexity across the human body. Future work could adopt an adaptive Gaussian allocation strategy, assigning more primitives to detail-sensitive regions such as the face, hands, hair, and clothing boundaries, while maintaining fewer primitives on smoother or less deformable regions such as limbs and large clothing surfaces. Such a design could improve fine-detail reconstruction without unnecessarily increasing the total number of Gaussians, thereby limiting additional storage and transmission cost.

Occlusion and unobserved regions. The method remains limited in its ability to infer severely occluded or unobserved body regions. As shown in the second row of Figure 6, when the hands are hidden inside the pockets, the model cannot reliably reconstruct plausible hand geometry or appearance. Although we expect the model to learn a generic prior that hands should resemble skin-like regions, the current local sampling mechanism mainly relies on nearby visible evidence. This limitation directly affects the perceived realism of articulated avatars, especially for hands and faces. This remains an open challenge. One possible direction is to train on larger-scale datasets so that the model can learn stronger human priors, as discussed below.

Identity Metadata and Temporal consistency. Finally, the current training data does not include identity metadata or video sequences. The current dataset has different scans of the same subject, but without the metadata information, they are regarded as different subjects in the model's view. Thus, identity metadata would provide stronger priors for idenitity preservation. Moreover, the model is trained from static multi-view observations, so it does not explicitly learn temporal consistency across poses. This limits its ability to produce animated results. Training with video sequences could help the model learn motion-dependent appearance changes and support more stable reconstruction under continuous driving. This would move the current static reconstruction framework closer to a full 4D avatar representation suitable for immersive telepresence.

5.3. Research Reflection

From telepresence to the representation layer. Although immersive telepresence remains the long-term motivation of this thesis, a complete telepresence system involves many interconnected components, including human capture, pose estimation, avatar reconstruction, compression, data streaming, and real-time rendering. This thesis focuses on the representation layer as a technically bounded but essential step toward the larger goal. Rather than attempting to build a full communication system, the work investigates how an avatar representation can support visual fidelity, generalization to unseen identities, pose drivability, and deployment efficiency within a unified framework.

From hybrid ideas to mixed representation. As discussed in section 2, different avatar representation families provide different strengths. This motivates the use of mixed representations that combine complementary properties from different families. There are several possible ways to construct such mixed representations. One direction is a part-wise hybrid design, where different body regions are represented by different primitives, for example using an implicit representation for detailed regions such as the head and explicit primitives for coarser regions such as clothing or the torso. Such designs may offer strong local quality, but integrating different rendering and deformation mechanisms into a single coherent pipeline remains challenging. This thesis instead explores a more unified mixed representation, where the parametric body model provides spatial organization and pose-driven structure, while the Gaussian primitives provide efficient rendering and local appearance flexibility. The hybrid part-wise representation remains a valuable direction to explore in the future.

The Shift Toward Scalable Explicit Representations. Placing this work within the broader context of computer vision, the field is currently experiencing a major shift from implicit neural representations (like NeRFs) to explicit primitives (like 3D Gaussian Splatting). While explicit methods have solved the rendering speed bottleneck, the field still struggles with scalability. Most current high-fidelity Gaussian avatars require long, per-subject optimization, which contradicts the real-time demands of telepresence. By framing avatar generation as a feed-forward prediction task on a shared SMPL-X scaffold, NLF-GS connects to the broader Computer Science trend of building generalizable models. It demonstrates that we do not have

to sacrifice pose drivability to achieve zero-shot generalization, offering a scalable path forward for interactive 3D vision systems.

Dataset choice and the need for better benchmarks. The dataset choice reflects a trade-off between relevance and feasibility. THuman2.0 provides high-quality clothed scans and pre-fitted SMPL-X parameters, making it suitable for controlled avatar reconstruction. However, it is still limited in evaluating long-term pose drivability and temporal stability. Future telepresence-oriented research would benefit from datasets that combine identity diversity, accurate body parameters, dynamic motion sequences, and multi-view appearance observations.

Evaluation trade-off. The evaluation strategy also reflects the difficulty of assessing avatar representations through purely objective metrics. Reconstruction fidelity can be measured with standard image-based metrics such as PSNR, SSIM, and LPIPS, but other aspects such as pose drivability, identity preservation during animation, and perceived interaction quality are harder to summarize with a single score. These aspects often require qualitative inspection or user-centered evaluation.

Therefore, this thesis combines quantitative reconstruction metrics, deployment-related factors, and qualitative drivability demonstrations. This broader evaluation reflects the motivation of telepresence-oriented avatar reconstruction, where the most useful representation is not necessarily the one with the highest visual score alone, but the one that balances visual fidelity, generalization, pose drivability, and deployment efficiency. Future work would benefit from more objective and reproducible metrics for the aspects discussed above. A user study could further strengthen the evaluation by assessing perceptual realism and interactive quality.

5.4. Societal Context and Stakeholder Implications

This thesis explores a possible path toward telepresence-oriented avatar generation, which carries distinct implications for various stakeholders.

For daily users, the accessibility of digital communication is a primary concern. Because NLF-GS requires only three sparse input views to generate an avatar, it removes the barrier of needing multi-camera capture studios, allowing users to participate in virtual workspaces or remote education using standard consumer cameras.

For industry developers, deployment efficiency is critical. The compact nature of our representation with approximately 80k Gaussians, compared to the millions required by other methods, significantly reduces data storage and transmission costs. Furthermore, driving the avatar at 196 FPS on consumer-grade hardware ensures it can support smooth, real-time interaction. On a larger scale, shifting away from computationally heavy volume rendering to efficient explicit primitives could reduce the energy consumption and carbon footprint.

However, from a societal security perspective, realistic and generalizable avatar generation introduces severe risks. Because NLF-GS can reconstruct unseen identities in a feed-forward manner without subject-specific optimization, the barrier to creating unauthorized deepfake avatars is drastically lowered. Malicious actors could potentially animate a realistic person using just a few photos sourced from the internet. To mitigate this, future telepresence systems must build in technical safeguards, such as cryptographic watermarking of AI-generated content and strict verification protocols before the model is called to generate an avatar.

6. Conclusion

We present NLF-GS, a feed-forward framework that enables the zero-shot reconstruction of unseen identities, while remaining compatible with novel-view rendering, pose-driven animation, and deployment-efficient inference. Experiments show that our method can generate high-quality avatar representations with fast avatar generation and real-time pose-driven animation. Although several issues, including occlusion and temporal consistency, remain, we believe our proposed model provides a practical representation-level step toward telepresence scenarios and could benefit future research.

7. Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Chirag and Pablo, for guiding me through this research journey.

Thank you, Chirag, for inviting me into the creative world of telepresence. We share the same feeling of missing our families on the other side of the world, and that shared longing gave this study a deep, personal meaning for me. I admire your approach to life and work, and I walk away from every one of our talks having learned something new. Along with that, Ojas and Anuj also provided me with brilliant ideas when I'm stuck in a loop; I truly appreciate the discussions.

I am also deeply grateful to Pablo. It is because of your constant encouragement that I found the strength to bounce back from setbacks. I truly enjoyed my visits to CWI, where I was always met with a warm welcome and helpful feedback. Thank you as well for your patient review, which made this thesis so much stronger. Lastly, I want to thank Dr. Petr Kellnhofer for his time and for joining my committee.

Beyond my mentors, I feel very lucky to have shared this journey with a wonderful group of peers from Tapri lab. In

our weekly meetings, we safely shared our confusions and ideas, and we made sure to celebrate every step forward together.

Finally, I want to thank my friends and family for their love and support throughout my work, in particular Madalena for your lasting help and careful proofreading. Another special thanks belongs to Tim. Not only did your steady encouragement keep me going during the hard times, but your brilliant help with debugging, and the use of your computer, played a crucial role in bringing this project to life. I feel incredibly lucky to have you by my side.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[2] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision (ECCV)*, 2022. 4

[3] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[4] Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[5] Alexander Clemm, Maria Torres Vega, Hemanth Ravuri, Tim Wauters, and Filip De Turck. Toward truly immersive holographic-type communication: Challenges and solutions. *IEEE Communications Magazine*, 58(1):1–7, 2019. 2

[6] Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[7] Jingnan Gao, Zhuo Chen, Xiaokang Yang, and Yichao Yan. Anisdf: Fused-granularity neural surfaces with anisotropic encoding for high-fidelity 3d reconstruction. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[8] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016. 4

[10] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[11] Hezhen Hu, Zhiwen Fan, Tianhao Wu, Yihan Xi, Seoyoung Lee, Georgios Pavlakos, and Zhangyang Wang. Expressive gaussian human avatars from monocular rgb video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[12] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 7

[13] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 4

[14] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

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

[16] 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 (CVPR)*, 2024. 3

[17] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3, 4

[18] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. Generalizable human gaussians for sparse view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 7

[19] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[20] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, 2000. 4

[21] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[22] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 3

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 4

[24] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6), 2021. 2, 3

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6), 2015. 2, 3

[26] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[27] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[28] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. Transhuman: A transformer-based human representation for generalizable neural human rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 4

[30] Cheng Peng, Jingxiang Sun, Yushuo Chen, Zhaoqi Su, Zhuo Su, and Yebin Liu. Parametric gaussian human model: Generalizable prior for efficient and realistic human avatar modeling. In *International Conference on 3D Vision (3DV)*, 2026. 3

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4

[32] Sen Peng, Weixing Xie, Zilong Wang, Xiaohu Guo, Zhonggui Chen, Baorong Yang, and Xiao Dong. Rmavatar: Photorealistic human avatar reconstruction from monocular video based on rectified mesh-embedded gaussians. *Graphical Models*, 139:101266, 2025. 3

[33] Pablo Pérez, Ester González-Sosa, Jes'us Guti'errez, and Narciso García. Emerging immersive communication systems: Overview, taxonomy, and good practices for qoe assessment. In *Frontiers in Signal Processing*, 2022. 2

[34] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[35] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[36] Renderpeople. 3d people for renderings, 2023. 4

[37] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[38] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[39] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 4, 6

[40] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[41] Hanzhang Tu, Ruizhi Shao, Xue Dong, Shunyuan Zheng, Hao Zhang, Lili Chen, Meili Wang, Wenyu Li, Siyan Ma, Shengping Zhang, et al. Tele-aloha: A low-budget and high-authenticity telepresence system using sparse rgb cameras. In *ACM SIGGRAPH Conference Papers*, 2024. 2

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6, 7

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[44] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research (JMLR)*, 26, 2025. 6

[45] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4, 8

[46] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 3

[47] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 45(10):12287–12303, 2023. 12

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[49] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[50] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4

[51] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. In *International Conference on 3D Vision (3DV)*, 2025. 2, 3

[52] Anton Zubekhin, Heming Zhu, Paulo Gotardo, Thabo Beeler, Marc Habermann, and Christian Theobalt. Giga: Generalizable sparse image-driven gaussian humans. In *International Conference on 3D Vision (3DV)*, 2026. 2, 3

NLF-GS: A Mixed Mesh-Gaussian Representation for Generalizable and Drivable Human Avatars

Supplementary Material

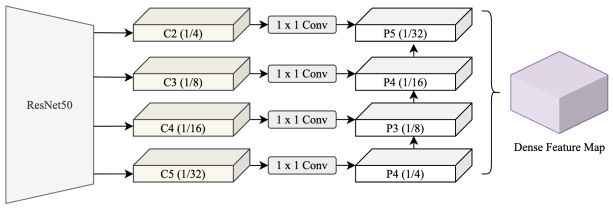


Figure 8. ResNet-FPN feature encoder used to extract multi-scale dense features from each input view.

A. Backbone Architecture

This section describes how the ResNet-based FPN is used as the feature extractor. Specifically, for each input image I_v, we extract multi-scale backbone features from stages $\{\mathbf{C}_2^v, \mathbf{C}_3^v, \mathbf{C}_4^v, \mathbf{C}_5^v\}$ and transform them into pyramid features $\{\mathbf{P}_2^v, \mathbf{P}_3^v, \mathbf{P}_4^v, \mathbf{P}_5^v\}$ through the standard top-down FPN pathway with lateral connections:

$$\{\mathbf{P}_2^v, \mathbf{P}_3^v, \mathbf{P}_4^v, \mathbf{P}_5^v\} = \mathrm{FPN}(\mathbf{C}_2^v, \mathbf{C}_3^v, \mathbf{C}_4^v, \mathbf{C}_5^v). \quad (8)$$

For each view v, we upsample all pyramid levels to a common spatial resolution and concatenate them along the channel dimension to obtain a unified dense feature map F_v, which serves as the per-view appearance representation for Gaussian feature querying.

We use FPN to adapt to different levels of detail. Human appearance contains structures at very different spatial scales, ranging from fine details such as facial regions or hands, to coarser body-level context over the torso and limbs. A multi-scale feature pyramid therefore provides a more suitable representation than a single-resolution feature map: shallow features retain greater spatial detail, while deeper features provide stronger semantic context with larger receptive fields. By combining them, each Gaussian can be conditioned not only on its local projected appearance but also on the surrounding neighborhood at multiple scales.

B. AI Usage Statement

During the preparation of this thesis, artificial intelligence tools, including ChatGPT, Gemini and Claude, were used to support the research and writing process in the following ways:

- **Brainstorming:** AI was used to identify relevant literature and outline complex concepts during the early stages of research. It also assisted in iteratively refining the overall structure of the framework.
- **Coding Assistance:** AI was used to assist with debugging errors, writing analysis scripts, and refining the code implementation.
- **Writing:** AI tools were utilized to correct grammatical errors, improve sentence clarity, and most importantly, ensure a consistent academic style throughout the document.