Generalizable Human Gaussians for Sparse View Synthesis

Youngjoong Kwon[1], Baole Fang[1*], Yixing Lu[1*], Haoye Dong[1], Cheng Zhang[1], Francisco Vicente Carrasco[1], Albert Mosella-Montoro[1], Jianjin Xu[1], Shingo Takagi[2], Daeil Kim[2], Aayush Prakash[2], and Fernando De la Torre[1]

[1]Carnegie Mellon University [2]Meta Reality Labs

https://humansensinglab.github.io/Generalizable-Human-Gaussians/

Abstract. Recent progress in neural rendering has brought forth pioneering methods, such as NeRF and Gaussian Splatting, which revolutionize view rendering across various domains like AR/VR, gaming, and content creation. While these methods excel at interpolating *within the training data*, the challenge of generalizing to new scenes and objects from very sparse views persists. Specifically, modeling 3D humans from sparse views presents formidable hurdles due to the inherent complexity of human geometry, resulting in inaccurate reconstructions of geometry and textures. To tackle this challenge, this paper leverages recent advancements in Gaussian Splatting and introduces a new method to learn generalizable human Gaussians that allows photorealistic and accurate view-rendering of a new human subject from a limited set of sparse views in a feed-forward manner. A pivotal innovation of our approach involves reformulating the learning of 3D Gaussian parameters into a regression process defined on the 2D UV space of a human template, which allows leveraging the strong geometry prior and the advantages of 2D convolutions. In addition, a multi-scaffold is proposed to effectively represent the offset details. Our method outperforms recent methods on both within-dataset generalization as well as cross-dataset generalization settings.

1 Introduction

Recent advancements in neural rendering techniques, such as Neural Radiance Fields (NeRF) [30], 3D Gaussian Splatting [17], and point-based graphics [2], have unveiled a multitude of captivating applications spanning virtual avatars, asset content creation, or cinematic production. While these methods excel at interpolating a single scene/object from many input views, it is very challenging to generalize to new scenes and objects with few samples, and extrapolating outside the captured views. This limitation is particularly pronounced in the task of photorealistic rendering of humans, a subject of widespread interest and applications. The task of modeling 3D humans from sparse viewpoints is complicated by the inherent complexities of human geometry, including articulations,

* Equal contribution

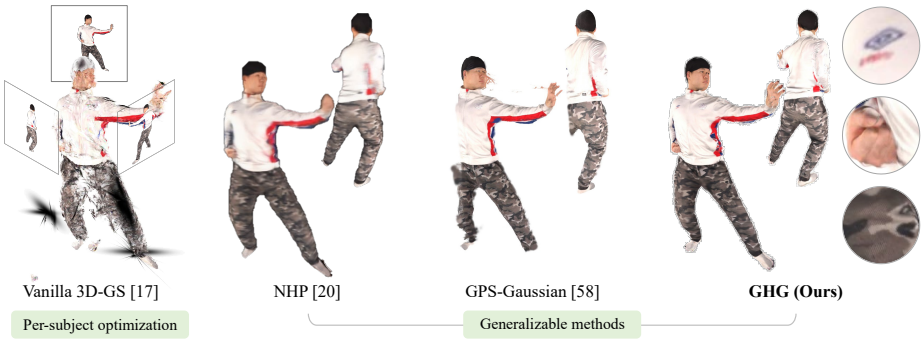

Fig. 1: Generalizable Human Gaussian (GHG). Our method can perform accurate and photorealistic novel view renderings of a new human subject given very sparse inputs (e.g., 3 views) without involving any test-time optimization or fine-tuning. In the sparse-view setup, our GHG approach exhibits superior rendering quality compared to other generalizable methods such as NHP [20] and GPS-Gaussian [58].

self-occlusions, and complex surface geometries like hair. These factors often lead to significant inaccuracies in the reconstruction of both geometry and textures, posing substantial hurdles to generating photorealistic digital humans.

Recent advances in human rendering incorporate implicit neural representations (*e.g.* NeRF) with human template models to facilitate generalizable and robust synthesis under sparse view settings [7,8,20,21,33,57]. While NeRF-based methods have made sigificant progress in generalizable human rendering, they are limited by their slow runtime, mainly due to their computationally intesive per-pixel volume rendering process. Additionally, in sparse-view setting, leveraging recent advances in inpainting models holds promiss for capturing details absent in the input views. However, integrating these modules presents practical challenges as it would further burden the already heavy NeRF-based system.

Recently, explicit representation methods such as 3D Gaussians have gained popularity for their efficient rasterization-based rendering speed. This fast rendering capability enables seamless integration with other models, such as generative [44] or depth estimation models [58], to achieve high-quality novel view rendering. However, these methods encounter difficulties when dealing with human subjects particularly when only sparse input views (e.g., 2-3 views) are available. The inherent challenges in rendering human subjects, such as articulations, self-occlusions, and complex surface geometries, worsen the difficulties in such sparse-view setting (see Figure 1).

To this end, we propose **Generalizable Human Gaussians** (GHG), a method for accurate and photorealistic novel-view renderings of human subjects. GHG enables rendering of a novel human subject given very sparse input views, without requiring any test-time optimization or fine-tuning. To improve performance in the sparse-view setting, our key insight is to leverage human geometry prior by reformulating the optimization of 3D Gaussian parameters within the 2D UV space derived from a human template model. By anchoring the Gaussian parameters onto the surface of the 3D human template model, each

location in the template space can be mapped to each foreground pixel in the corresponding 2D UV map space. Our UV map-based Gaussian representation significantly improves the reconstruction of complex human geometries. Operating on the 2D UV map space enables us to utilize 2D CNNs for the Gaussian optimization which can incorporate information from neighboring pixels unlike MLPs. Additionally, this approach makes our model compatible with inpainting models [53, 54], facilitating seamless integration.

While our human UV map-based representation brings significant advancement in generalization and robustness in rendering, we aim to improve its effectiveness further. Given the inherent disparity between the template body model and real human geometry (*e.g.* clothing or hair), we present a method to bridge this gap. To achieve this, we propose to generate multiple offset meshes through dilating the human template mesh, both at the input and output spaces. These meshes serve as scaffolds, enabling effective encoding of input geometry information, as well as facilitating a richer representation of displacements beyond that can be captured by a single template mesh at the output. Leveraging these multi-scaffold meshes enables more faithful capturing of real human geometries, which often cannot be accurately represented by a single template mesh surface.

We evaluate the efficacy of our Generalizable Human Gaussians on two multi-view human capture datasets: THuman 2.0 [55] and RenderPeople [38]. Existing generalizable human rendering approaches that allow sparse-view input (3 views) are primarily NeRF-based methods [7, 20]. We compare with these methods and demonstrate superior rendering quality in both in-domain and cross-dataset evaluation settings. Additionally, we compare our approach with existing 3D Gaussians-based methods, which either necessitates more input views [58] or per-subject optimization [17], showcasing distinct benefits of our approach.

In summary, our main contributions are as follows:

- We propose a new feed-forward method for accurate and photorealistic novel-view renderings of new humans from very sparse input views. This is achieved by integrating human geometry prior with 3D Gaussians. Specifically, we reformulate the optimization of 3D Gaussian parameters into a task of generating a Gaussian parameter map within the 2D human UV space derived from the human template model.
- We propose a multi-scaffold representation aimed at minimizing the disparity between the template model and real human geometry. This approach enables the Gaussian parameters to be learned across multiple scaffold spaces, allowing for a more comprehensive representation of displacements that surpasses the capacity of a single template mesh space.

2 Related Work

Generalizable NeRF for Human Rendering. Neural Radiance Fields has demonstrated its powerful capability to render 3D scenes with photorealistic

quality. However, they can be only optimized on a single scene, and require images taken from densely sampled cameras to train. To generalize to new scenes without optimization at inference time, some works condition the generation on the pixel-aligned features [41,45,52], cost-volumes [5,47], or image-based rendering [26,46]. Although they have demonstrated high-quality generalization ability on general objects and scenes, directly applying those methods to human subjects is non-trivial due to the complicated human geometries (i.e., articulations and self-occlusions). To effectively address the generalization to humans, a line of works utilize 3D human prior. Specifically, SMPL surface [27] is leveraged as the tool for aggregating the relevant features while preserving its geometric structure [6–10,20,21,57]. Skeletal keypoints are also utilized [29]. Despite their detailed output, their rendering speed is very slow due to the volume rendering process which requires heavy computations to render a single pixel. This deters them from combining with other modules to further improve the performance (*e.g.* inpainting). In our work, thanks to the fast rasterization-based rendering, our model can be combined with 2D-based inpainting module [53,54] to compensate for the unobserved regions inevitable under sparse view settings.

3D Gaussian Splatting. 3D Gaussian Splatting is a method to represent a scene with a set of 3D Gaussians [17,39]. By utilizing GPU-parallelized rasterization, they achieve fast rendering speed and have presented impressive ability in novel view synthesis tasks. Some concurrent works utilize human template as the 3D prior and combine it with 3D Gaussians to create animatable representations [14,16,19,24,31,34,51,59,60]. However, they are not generalizable and require new training process for every new subject. Zheng et al. [58] achieve generalization to novel humans by incorporating a stereo-depth estimation module, which serves as a partial geometry prior. However, they suffer when given sparse views with few overlappings and thus depth could not be estimated. Therefore, they can only interpolate between very close views. In this work, we aim for a feed-forward generalizable human rendering method that can work when given very sparse inputs with few or no correspondences by leveraging 3D human prior.

Multi-surface representations. While utilizing 3D human prior has proven its effectiveness in the human rendering task [11,25,35,36,43], representing the geometry gap between human template and the real geometry (*e.g.* loose clothing, hair) is still challenging. Some recent literature [1,22,32,49] utilize multi-surface (shell) [37] to represent the geometry displacement. However, the idea from these works is not directly applicable to our generalization task because they either can be only optimized on a single subject or cannot be conditioned on the input subject information (i.e., unconditional generation from noise) [1]. In this paper, we propose multi-scaffold, a multi-surface-based representation that can effectively represent the human geometric details in the *generalization setting*.

3 Generalizable Human Gaussians (GHG)

Given a set of multi-view images of a subject (that is not in the training set), along with their camera position and fitted human template (i.e., SMPL [27]),

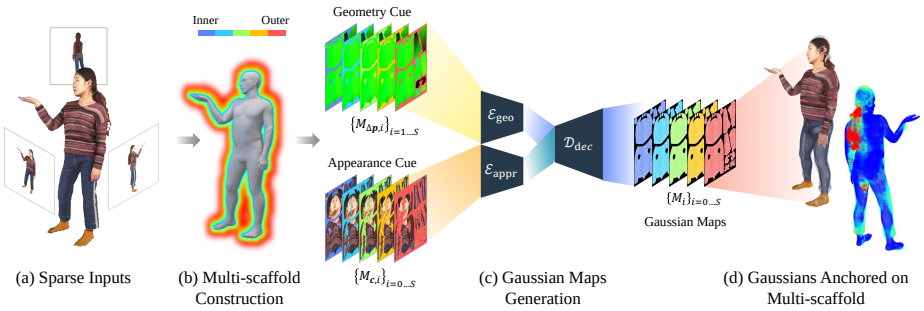

Fig. 2: Overview of GHG. (a) We focus on generalizable human rendering under very sparse view setting. (b) We first construct the multi-scaffolds by dilating the human template surface. The 2D UV space of each scaffold serves to collect the geometry and appearance information from the corresponding 3D locations. (c) The aggregated multi-scaffold input is fed into the network, which generates multi-Gaussian parameter maps. (d) Finally, Gaussians are anchored on the corresponding surface of each scaffold, and rasterized into novel views.

our goal is to render photo-realistic novel views. To address this challenge, we propose Generalizable Human Gaussian (GHG), a feed-forward architecture that does not require any fine-tuning. See Figure 2 for an illustration of GHC. In this section, we first review the 3D Gaussian Splatting and discuss the motivation of GHG (Section 3.1). In Section 3.2, we introduce the main idea of GHG, which reformulates the Gaussian splat fitting as a regression problem in the 2D UV space of the human template. Next, we present our multi-scaffold representation that allows encoding and modeling of the complicated geometric details (Section 3.3). Finally, we describe the end-to-end training objective in Section 3.4.

3.1 Background and Motivation

Notation. Functions (e.g., neural network mapping) are denoted with uppercase calligraphic letters (e.g., $\mathcal{F}$). Vectors are denoted with bold lowercase letters (e.g., $\boldsymbol{p}$). Matrices are denoted with uppercase letters (e.g., M). Sets are denoted with bold uppercase letters (e.g., $\boldsymbol{\Theta}$).

3D Gaussian Splatting (3D-GS). The key idea of 3D-GS [17] is to represent a scene with a set of 3D Gaussians, each of which is characterized by a 3D covariance matrix Σ and a center (mean) position $\mathbf{p}$:

$$\mathbf{G}(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mathbf{p})^T \Sigma^{-1}(\mathbf{x}-\mathbf{p})}. \tag{1}$$

The means of the 3D Gaussians can be initialized by a point cloud using Structure from Motion [42] computed from C images. Each Gaussian $\mathbf{G}$ is parameterized by $\boldsymbol{\Theta} = \{\mathbf{p}, \mathbf{q}, \mathbf{s}, \alpha, \boldsymbol{\eta}\}$ where $\mathbf{p} \in \mathbb{R}^3$ is the center position, $\mathbf{q} \in \mathbb{R}^4$ is the rotation quaternion, $\mathbf{s} \in \mathbb{R}^3$ is the scaling factor, $\alpha \in \mathbb{R}^1$ is the opacity, and $\boldsymbol{\eta} \in \mathbb{R}^{(l+1)^2}$ represents the coefficients of the spherical harmonics (SH) of order l. The covariance matrix is decomposed as $\Sigma = RSS^\top R^\top$, where

$S = diag(\mathbf{s}) \in \mathbb{R}^{3\times3}$ is the scaling matrix and $R \in \mathbb{R}^{3\times3}$ is a rotation matrix derived from the quaternion $\mathbf{q}$.

The rendering of a Gaussian set into an image plane is done by approximating the projection of a 3D Gaussian into pixel coordinates along the depth dimension [17]. Specifically, for each pixel, the final rendered color $\mathbf{c}_{\text{pixel}}$ is obtained by the α-blending of K overlapping Gaussians that are depth-ordered:

$$\mathbf{c}_{\text{pixel}} = \sum_{i\in K} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{2}$$

where α_i is the opacity and $\mathbf{c}_i$ is the RGB color extracted from SH coefficients.

Motivation. Applying GS to our task (i.e., generalizable human rendering from sparse views) is not trivial for the following reasons: First, the original GS was designed for single-scene optimization, making it difficult to adapt to generalization tasks — specifically to reconstruct unseen human subjects without model fine-tuning. Second, accurate point cloud initialization requires a substantial number of input images for Structure from Motion. With the number of input images reduced to as few as 3, vanilla 3D-GS struggles to accurately reconstruct the complex geometry and texture of the human body. See Fig. 1-Vanilla 3D-GS as an example of the view-reconstruction achieved with the original GS. Therefore, in this paper, we focus on adapting the 3D Gaussian Splatting for generalizable human rendering from sparse inputs.

3.2 Learning 3D Gaussians in 2D Human UV Space

UV space of human template. Our goal is to model a generalizable funtion $\mathcal{F}(\{I_c\}_C) = \{\boldsymbol{\Theta}_n\}_{N_\mathbf{G}}$ that estimates the parameters of $N_\mathbf{G}$ Gaussians conditioned on the input images $\{I_c\}_C$. However, due to the complex nature of human geometry that involves articulations and occlusions, it is challenging to regress the parameters only given few sparse observations. Therefore, we propose to incorporate a 3D geometry prior (i.e., a human template model such as SMPL [27]) by attaching the Gaussians on the template surface and regressing their parameters in the 2D UV space of the human template. Specifically, for every foreground pixel of the UV map, we attach a Gaussian on the corresponding 3D human surface point defined by the UV mapping. Then, we regress and store its parameters in the set of 2D UV maps $\boldsymbol{M} = \{M_\mathbf{p}, M_\mathbf{q}, M_\mathbf{s}, M_\alpha, M_\mathbf{c}\}$. $M_\mathbf{p}, M_\mathbf{q}, M_\mathbf{s}, M_\alpha, M_\mathbf{c}$ denotes the map for the position, rotation, scaling, opacity, and RGB color, respectively. Each parameter map has the resolution of $H \times W \times D$, where D is the dimension of each parameter.

2D CNN-based parameter regression. To model the function $\mathcal{F}$, we adopt a 2D Convolutional Network that provides several benefits. First of all, 2D CNN naturally aggregates the information from neighboring pixels. This helps our system to consider the local context and thus maintain the consistency between the adjacent Gaussian parameters, which both contribute to better reconstruction accuracy. In addition, it facilitates integrating other image-based enhancement

models, in our case the inpainting module to hallucinate unobserved regions. In practice, $\mathcal{F}$ is modeled with a U-Net as follows:

$$\mathcal{F} = \mathcal{D}_{\text{dec}}\big(\mathcal{E}(\{I_c\}_C)\big) = \boldsymbol{M}, \tag{3}$$

where $\mathcal{E}$, $\mathcal{D}_{\text{dec}}$ is the U-Net-based encoder and decoder, respectively. $\{I_c\}_C$ denotes the input images, and M is the set of Gaussian parameter maps.

Reformulation. In our formulation, $\mathcal{F}$ regresses the set of 2D parameter maps $\boldsymbol{M} = \{M_{\mathbf{p}}, M_{\mathbf{q}}, M_{\mathbf{s}}, M_{\alpha}, M_{\mathbf{c}}\}$. Since the Gaussian positions are fixed on the human template surface, the position map $M_{\mathbf{p}}$ is computed by rasterizing the vertex position of the human template on the 2D UV space. The RGB map $M_{\mathbf{c}}$ is computed as the weighted average of corresponding pixels from all observed views. Specifically, for each pixel in $M_{\mathbf{c}}$, we find their projections to all visible source images and average the source RGB values weighted by visibility:

$$M_{\mathbf{c}} = \sum_{c=1}^{C} \mathrm{W}_{\mathrm{c}}(\boldsymbol{P}) \cdot \Pi\big(I_c, \mathrm{Proj}_{\mathrm{c}}(\boldsymbol{P})\big). \tag{4}$$

$\boldsymbol{P}$ is the Gaussian center positions (i.e., foreground pixel values of the position map M_p). $\mathrm{W}_{\mathrm{c}}(\boldsymbol{P})$ is the normalized visibility of 3D positions $\boldsymbol{P}$ for the c-th camera. C is the number of total input views. I_c is the c-th input view image. Π denotes the bilinear sampling operator. $\mathrm{Proj}_{\mathrm{c}}$ denotes the 3D to 2D projection with respect to the c-th camera. $\Pi\big(I_c, \mathrm{Proj}_{\mathrm{c}}(\boldsymbol{P})\big)$ returns an image that is the result of the of interpolating I_c in the projected coordinates of $\boldsymbol{P}$.

Since $M_{\mathbf{p}}$ and $M_{\mathbf{c}}$ are already computed, the regression of the parameter map $\boldsymbol{M}$ is reduced to $\boldsymbol{M} = \{M_{\mathbf{q}}, M_{\mathbf{s}}, M_{\alpha}\}$. To effectively regress the Gaussian parameter maps, we provide $\mathcal{F}$ with the geometry and appearance cue which have complementary attributes.

The appearance cue provides information of geometric details and how they should look like. The geometry cue facilitates the optimization of the Gaussian parameters to match the appearance details. To provide the geometry cue, we encode the position map $M_{\mathbf{p}}$ using the geometry encoder $\mathcal{E}_{\text{geo}}$. The appearance cue is obtained by encoding the RGB map $M_{\mathbf{c}}$ using the appearance encoder $\mathcal{E}_{\text{appr}}$. Now, we adapt the Equation 3 to condition the parameter map generation on the geometry and appearance cue, as:

$$\mathcal{D}_{\text{dec}}\big(\mathcal{E}_{\text{geo}}(M_{\mathbf{p}}), \mathcal{E}_{\text{appr}}(M_{\mathbf{c}})\big) = \boldsymbol{M}. \tag{5}$$

Inpainting. It is inevitable to have unobserved regions under very sparse view settings. This results in blurriness or missing texture in some areas. To address this issue we incorporate into our architecture a 2D inpainting method. In particular, we create a set of pseudo ground truth texture maps by transferring the ground truth mesh texture map into the human template UV space (see Appx-Fig.10). On this dataset, we train an attention-based generative model $\mathcal{G}_{\text{inpaint}}$ [53, 54] to inpaint the missing regions present in the human template UV space RGB map. At the inference time, we inpaint the RGB map $M_{\mathbf{c}}$ with $\mathcal{G}_{\text{inpaint}}$. We would like to note that this is possible because our 2D CNN-based

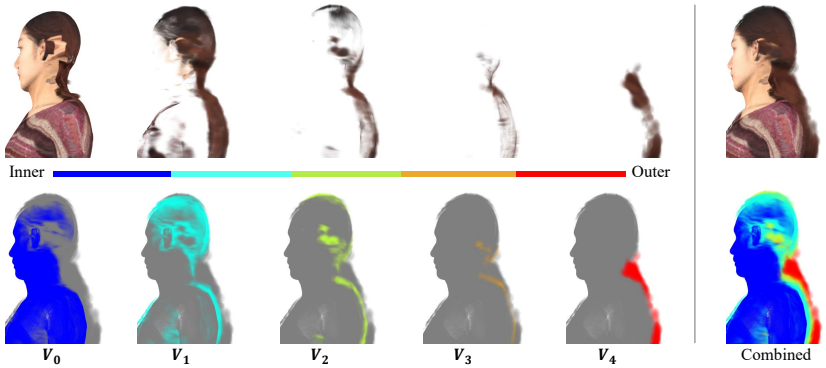

Fig. 3: Illustration of multi-scaffold representation. Each column shows different scaffold levels, with the last column illustrating their combined effect. The top part shows the RGB representation, while the bottom part highlights affected regions, with grey indicating unaffected areas.

system facilitates the combination with a 2D-based inpainting module. Our parameter map regression is again adapted to:

$$\mathcal{D}_{\text{dec}}\Big(\mathcal{E}_{\text{geo}}(M_{\mathbf{p}}), \mathcal{E}_{\text{appr}}\big(\mathcal{G}_{\text{inpaint}}(M_{\mathbf{c}})\big)\Big) = \boldsymbol{M}. \tag{6}$$

Please refer to the supplementary materials for details of the inpainting network.

3.3 Modeling Geometric Details with Multi-scaffolds

The utilization of a human template model helps to reconstruct the shape and appearance with sparse views. However, this is not enough to effectively represent accurately details that are offseted from the human surface such as hair or loose clothing due to the following reasons: (1) The appearance details deviating from the template surface (e.g., ponytail) cannot be accurately represented with the input appearance cue (i.e., RGB map M_c in Eq. (6)).

Therefore, to narrow this geometry gap, we propose to utilize multiple scaffolds constructed through dilation of the human template mesh. These multi-scaffold representation facilitates the effective encoding of the geometry gap information into the input, and allows for more versatile output Gaussians to represent the displacement details more accurately. Specifically, we create the multiple scaffolds $\{\boldsymbol{V}_i\}_{i=1...S}$ by offsetting the human template vertices $V_0 = \{\mathbf{v}_{0,\text{j}}\}$ along its outward vertex normal direction:

$$\boldsymbol{V}_i = \{\mathbf{v}_{\text{i,j}} | \mathbf{v}_{\text{i,j}} = \mathbf{v}_{0,\text{j}} + i \cdot d \cdot \hat{\mathbf{n}}_{\text{j}}\}, \tag{7}$$

where $\mathbf{v}_{i,j}$ is the j-th vertex of i-th outer scaffold, $\hat{\mathbf{n}}_{\text{j}}$ is the j-th vertex normal, d defines the offset between scaffolds, S is the number of outer scaffolds. We use $d = 1cm$ and $S = 4$ in the experiments.

Input. We adapt the geometry and appearance cues to aggregate information from the entire level of scaffolds. We redefine the geometry cue as the feature

extracted from the concatenation of offset maps which record the displacement between each scaffold:

$$\mathcal{E}_{\text{geo}}(M_{\Delta\mathbf{p},1} \oplus \ldots \oplus M_{\Delta\mathbf{p},S}), \text{ where } M_{\Delta\mathbf{p},i} = M_{\mathbf{p},i} - M_{\mathbf{p},i-1}. \tag{8}$$

$M_{\Delta\mathbf{p},i}$, $M_{\mathbf{p},i}$ is the offset map and position map of the i-th scaffold, respectively. $\oplus$ is the concatenation operation. The appearance cue is redefined as:

$$\mathcal{E}_{\text{appr}}(M_{\mathbf{c},0} \oplus \ldots \oplus M_{\mathbf{c},S}), \text{ where } M_{\mathbf{c},i} = \sum_{c=1}^{C} \mathrm{W}_{\mathrm{c}}(\boldsymbol{P}_i) \cdot \Pi(I_{\mathrm{c}}, \mathrm{Proj}_{\mathrm{c}}(\boldsymbol{P}_i)). \tag{9}$$

Here $M_{\mathbf{c},i}$, $\boldsymbol{P}_i$ is the RGB map and the Gaussian center positions corresponding to the i-th level scaffold, respectively. Note that inpainting is done only to the RGB map of 0-th level scaffold (i.e., $M_{\mathbf{c},0} = \mathcal{G}_{\text{inpaint}}(M_{\mathbf{c},0})$)

Output. There are numerous possible design choices to model the displacement details such as hair. For example, the scaling can be enlarged to cover the gap, or we could directly learn Gaussian mean offsets. However, we empirically found out that these lead to unstable training and hinder the system from converging because of the high degree-of-freedom (see Fig. 6). Therefore, we attach Gaussians on each scaffold, and regress their parameters *within each scaffold*. This is realized by confining the maximum scaling of the Gaussians as the offset between scaffolds. Our final formulation is defined as:

$$\mathcal{D}_{\text{dec}}\big(\mathcal{E}_{\text{geo}}(M_{\Delta\mathbf{p},1} \oplus \ldots \oplus M_{\Delta\mathbf{p},S}), \mathcal{E}_{\text{appr}}(M_{\mathbf{c},0} \oplus \ldots \oplus M_{\mathbf{c},S})\big) = \{\boldsymbol{M}_i\}_{i=0\ldots S}. \tag{10}$$

where $\{\boldsymbol{M}_i\}$ is the set of parameter maps corresponding to the i-th level scaffold.

3.4 Training and Optimization

Gaussian parameter map regressor. To train our Gaussian parameter map regressor $\mathcal{F}$ (i.e., $\mathcal{E}_{\text{geo}}$, $\mathcal{E}_{\text{appr}}$, $\mathcal{D}_{\text{dec}}$), we employ multi-view RGB and mask supervision. Specifically, we sample N target views from the positions in between the input views and generate the RGB and mask predictions. The predictions are supervised by minimizing the loss objective $\mathcal{L} = \frac{1}{N}(\lambda_1 \cdot \mathcal{L}_1 + \lambda_{\text{ssim}} \cdot \mathcal{L}_{\text{ssim}} + \lambda_{\text{mask}} \cdot \mathcal{L}_{\text{mask}})$, where $\mathcal{L}_1$, $\mathcal{L}_{\text{ssim}}$ are L_1 and SSIM loss [48] computed between the ground truth and predicted RGB images, respectively. $\mathcal{L}_{\text{mask}}$ is the Binary Cross Entropy loss computed between the ground truth and predicted foreground mask. We use $N = 3$, $\lambda_1 = 0.8$, $\lambda_{\text{ssim}} = 0.2$, $\lambda_{\text{mask}} = 0.02$ in our experiments. We used a single GPU with 20G memory during training. AdamW optimizer [28] with an initial learning rate of $2e^{-4}$ was used. We train the parameter regressor for $100k$ iterations with a single batch size, which takes around 10 hours.

Inpainting network. When training the inpainting network $\mathcal{G}_{\text{inpaint}}$, $\mathcal{L}_G = \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} + \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{adv}}$ is minimized, where $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{adv}}$ are L_1 loss and adversarial loss computed between the inpainted results and pseudo ground truth. $\lambda_{\text{rec}} = 10$ and $\lambda_{\text{adv}} = 1$ are used in training. The loss objective for the inpainting discriminator $\mathcal{D}_{\text{inpaint}}$ is defined the discriminator loss between the inpainted image and pseudo

Table 1: Comparison with NeRF-based methods for (a) in-domain and (b) cross-domain sparse view synthesis. For all the methods, we use 3 views during both training and testing. GHG achieves competitive results for both settings. TH: THuman [55]. RP: RenderPeople [38]. See Figure 4 and 5 for qualitative results.

	(a) In-domain: TH → TH			**(b) Cross-domain: TH → RP**		
Method	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓
NHP [20]	**23.32**	184.69	136.56	22.34	172.56	137.23
NIA [21]	23.20	181.82	127.30	**22.45**	168.15	124.80
GHG (ours)	21.90	**133.41**	**61.67**	21.02	**137.73**	**60.85**

ground truth. $\mathcal{G}_{\text{inpaint}}$ and $\mathcal{D}_{\text{inpaint}}$ are trained alternatively for 40 epochs. We use Adam optimizer [18] with an initial learning rate of $1e^{-4}$ and decay the learning rate by half every 10 epoch. It takes around 4 hours to train the inpainting module on a single GPU with a batch size of 1. Note that $\mathcal{G}_{\text{inpaint}}$ is trained only once separately from the Gaussian map regressor, and it is a general model that works for different new subjects at inference.

4 Experiments

4.1 Baselines, Datasets, and Metrics

Baselines. We benchmark our method against state-of-the-art generalizable human rendering techniques from two categories: 3D human template-conditioned NeRF methods NHP [20] and NIA [21], and depth-based 3D Gaussian method GPS-Gaussian [58]. Additionally, we compared with the original vanilla 3D Gaussians [17], which are optimized per subject.

Datasets. We conducted experiments on two datasets: the THuman [55] and RenderPeople [38] dataset. The THuman dataset comprises 526 high-quality 3D scans, texture maps, and corresponding SMPL-X parameters. 100 subjects were reserved for the evaluation, following GPS-Gaussian [58]. The RenderPeople dataset encompasses 3D human scans representing diverse clothing styles, races, and ages, totaling 956 subjects split into 756 train and 200 test subjects. SMPL-X parameters were estimated using off-the-shelf methods [3, 4].

Metrics. We generated images at a resolution of 1024×1024 for evaluating our results. To assess the quality of our results, we employed several metrics. Initially, we utilized the peak signal-to-noise ratio (PSNR), a standard metric. However, PSNR may not fully reflect human perception, as it can assign a low error to very blurry and unrealistic results [56]. Therefore, we also incorporated the learned perceptual image patch similarity (LPIPS) [56] and the Fréchet inception distance (FID) [12], which better align with human perceptions.

4.2 Comparison with NeRF-based methods

We compare with NHP [20] and NIA [21], which are the two competitors that are most similar to our settings in that (1) they focus on generalizable human

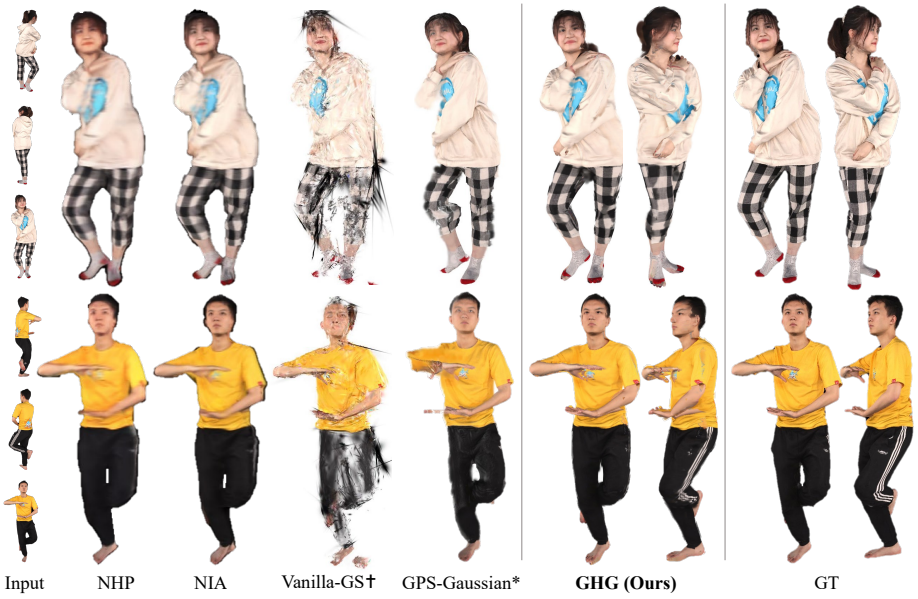

Fig. 4: Qualitative comparisons. All methods are trained and tested on THuman dataset [55]. †Unlike the other methods, Vanilla-GS [17] is per-subject optimized on the testing subjects. *GPS-Gaussian [58] is trained and tested with 5 input views, whereas NHP [20], NIA [21] and our method are trained and tested with 3 input views.

rendering from very sparse (i.e., 3) input views and (2) use the human template as 3D prior. Ours, NHP, and NIA are trained / evaluated on the THuman dataset with the same NHP protocol, where three randomly chosen input views are used during training and the same three canonical views are used during evaluation.

In-domain generalization. Table 1-(a) shows the in-domain generalization result where we evaluate on test subjects from THuman dataset. We achieve the best performance on the perception-based metrics LPIPS and FID, and comparable PSNR. As shown in Figure 4, the single-layer representation of NHP and NIA where the features are aggregated on a single surface of a naked body leads to the mixture of visual details and produce blurry results. On the other hand, our method collects visual information from the multi-scaffold and thus recovers sharp and high-frequency details including hair, wrinkles, and logos.

Cross-domain generalization. To confirm the cross-dataset generalizability of our approach, we train a model on the THuman dataset and evaluate it on the challenging RenderPeople dataset without any test-time optimization. The RenderPeople dataset exhibits a more diverse data distribution compared to the training dataset (THuman), encompassing variations in race, age, and apparel. Yet, our GHG significantly outperforms NHP and NIA on the perception-based metrics LPIPS and FID. In Figure 5, our method recovers fine details such as facial expressions, clothing patterns, and textures. However, since PSNR is pixel-wise computed, a slight deviation from the ground truth can lead to a lower score.

Table 2: Comparison with Gaussian Splatting-based methods on the THuman dataset [55]. Due to GPS-Gaussian [58] requiring at least 5 input views for reasonable results, we train and test all methods with 5 views.

Method	PSNR↑	LPIPS↓	FID↓
Vanilla-GS (per-subject) [17]	17.62	220.30	210.03
GPS-Gaussian [58]	20.69	**123.30**	46.26
GHG (ours)	**22.06**	132.42	**37.10**

Fig. 5: Qualitative results on cross-domain generalization. We train the models on THuman dataset [55] and test on Renderpeople dataset [38] without model finetuning. GHG can render high-frequent details and accurate geometry of the novel subject.

This can explain our lower performance in terms of PSNR in Table 1-(a),(b), while NHP with blurry results achieves the highest performance.

4.3 Comparison with Gaussian Splatting-based methods

We show the comparison with GPS-Gaussian [58] and the original vanilla 3D Gaussians [17]. Although the original GPS-Gaussian does not focus on sparse view synthesis as ours, we include comparison with them because we are both 3D Gaussian-based methods and explore generalization onto unseen human subjects. In our exploration of GPS Gaussian, we observed that GPS-Gaussian requires a substantial overlap between inputs for stereo-depth computation and cannot perform adequately with fewer than five views. Therefore, for comparison purposes in Table 2, we employ five uniformly distributed input views for training and evaluation of GPS-Gaussian, vanilla 3D Gaussian, and our method. Unlike the other methods, vanilla Gaussian is per-subject optimized and thus trained on the testing human subjects. As presented in Table 2, our approach achieves comparable results to GPS-Gaussian [58] with better PSNR and FID scores, while significantly outperforms the vanilla Gaussian method. Visual comparisons in Figure 4 reveal that our model, trained and tested with 3 input views, exhibits more accurate geometries and finer detail reconstruction compared to GPS-Gaussian, trained and tested with 5 input views. Particularly, GPS-Gaussian suffers from inaccurate geometry and missing contents, possibly due to its inherent high demand for larger overlap between input views.

Table 3: Ablation study on the multi-scaffold representation. S: single scaffold. **S⋆:** single scaffold with a large scale. **S†:** single scaffold with a learnable offset. **✓:** multiple scaffolds.

	Input Scaffold		Output Scaffold			
	Geometry	Appearance	Gaussian Map	PSNR↑	LPIPS↓	FID↓
a	S	S	S	22.30	145.74	84.38
b	S	S	✓	22.44	142.84	73.91
c	S	✓	✓	22.96	136.81	72.43
d	✓	S	✓	22.60	144.27	81.31
e	✓	✓	S⋆	23.11	145.03	90.16
f	✓	✓	S†	**23.39**	145.55	87.49
g	✓	✓	✓	21.90	**133.41**	**61.67**

4.4 Ablation Studies and Analyses

We conducted ablation studies on the THuman dataset, evaluating variants of our GHG model on unseen subjects in Figure 6 and Table 3, 4, and 5.

Effect of multi-scaffold representation. We examine the impact of the multi-scaffold representation for different configurations of inputs/outputs, see Table 3. First, we trained three input variants where the multi-scaffold input is either partially (c, d) or not used at all (b). We retained the multi-scaffold output (i.e., multi-Gaussian map generation). The lack of multi-layer geometry and appearance information leads to perceptual performance degradation. Second, we designed two output variants where only a single Gaussian map is generated, where we maintained the multi-scaffold input. In our original model, scale of each Gaussian map is confined up to the distance between its next scaffold (i.e., $1cm$). To represent the displacement between the template model and the real geometry, either scale is allowed to grow up to $4cm$ (Table 3-(e)) or scale is still confined to $1cm$ but we additionally learn a Gaussian center offset that can move up to $4cm$ (Table 3-(f)). However, as shown in Figure 6-(2),(3), the lack of regularization generates blurry results. The lowest performance of output variants without multi-scaffold representation in Table 3-(e),(f) again validates our design choice where we build 3D Gaussians on multiple scaffolds.

Importance of geometry and appearance cue. To study the effect of integration of the geometry and appearance cue, we train a variant with the geometry cue completely removed (Table 4-first row, Figure 6-4) and a variant with appearance cue removed (Table 4-second row, Figure 6-5). Removing one of them leads to visual artifacts and failure of recovering geometry gap such as hair.

Effect of inpainting. Under our sparse-view setting, it is inevitable to have insufficient observations (Figure 6-6). Our 2D architecture allows us to easily combine with the 2D-based inpainting module and hallucinate the unobserved regions, thus leads to improved quality (Figure 6-7, Table 5).

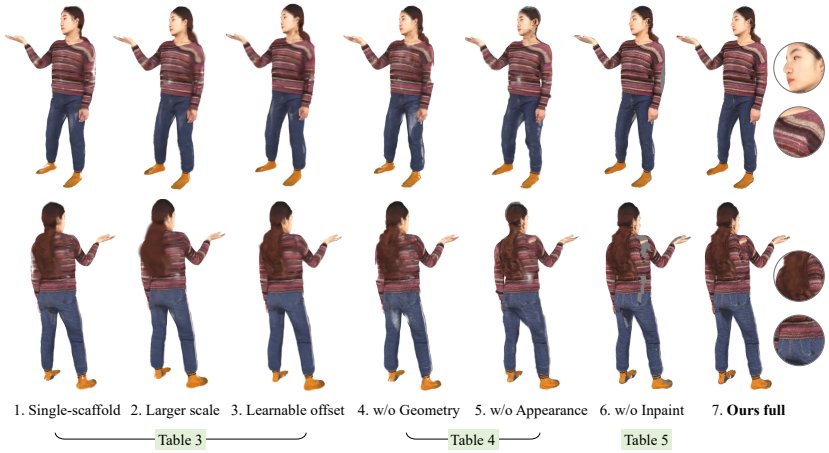

Fig. 6: Ablation studies. 1) Result only using the template mesh. 2) Illustrates the result of using a larger scale. 3) Depicts the result of learning the Gaussian offset. 4) Shows the model devoid of geometry information. 5) Illustrates the model without appearance cues. 6) Shows the model without inpainting. 7) Presents our model.

Table 4: Ablation study on the geometry cue and appearance cue. ✗/✓indicates completely remove/keep the encoding branch. See Figure 2 for an illustration.

Geo.	App.	PSNR↑	LPIPS↓	FID↓
✗	✓	**22.74**	138.35	69.12
✓	✗	21.83	146.05	78.56
✓	✓	21.90	**133.41**	**61.67**

Table 5: Ablation study on the texture inpainting network. ✗/✓indicates without/with the inpainting network on the 2D UV space, respectively. Please see Figure 6 for a comparison result.

Inpainting	PSNR↑	LPIPS↓	FID↓
✗	21.65	135.42	67.15
✓	**21.90**	**133.41**	**61.67**

5 Conclusion

We present Generalizable Human Gaussians (GHG), a feed-forward architecture capable of synthesizing novel views of new humans using sparse input views, without the need for test-time optimization. Our key insight is the reformulation of 3D Gaussian parameter optimization into the generation of parameter maps within the 2D human UV space. This allows us to leverage the human geometry prior, addressing challenges such as articulations and self-occlusions. Additionally, by framing the task as a 2D problem, we can exploit local neighboring information and integrate 2D-based inpainting modules to hallucinate unobserved regions. Finally, we propose a multi-scaffold approach to effectively represent and bridge the geometry gap between the human template and real human geometry. Experimental results show that our method can generate high-quality renderings surpassing state-of-the-art approaches.

References

1. Abdal, R., Yifan, W., Shi, Z., Xu, Y., Po, R., Kuang, Z., Chen, Q., Yeung, D.Y., Wetzstein, G.: Gaussian shell maps for efficient 3d human generation. arXiv preprint arXiv:2311.17857 (2023) 4
2. Aliev, K.A., Sevastopolsky, A., Kolos, M., Ulyanov, D., Lempitsky, V.: Neural point-based graphics. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 696–712. Springer (2020) 1
3. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: European Conference on Computer Vision (ECCV). Springer (aug 2020) 10
4. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In: Advances in Neural Information Processing Systems (NeurIPS) (December 2020) 10
5. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021) 4
6. Chen, J., Yi, W., Ma, L., Jia, X., Lu, H.: Gm-nerf: Learning generalizable model-based neural radiance fields from multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20648–20658 (2023) 4
7. Chen, M., Zhang, J., Xu, X., Liu, L., Cai, Y., Feng, J., Yan, S.: Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In: European Conference on Computer Vision. pp. 222–239. Springer (2022) 2, 3, 4
8. Cheng, W., Xu, S., Piao, J., Qian, C., Wu, W., Lin, K.Y., Li, H.: Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. arXiv preprint arXiv:2204.11798 (2022) 2, 4
9. Gao, Q., Wang, Y., Liu, L., Liu, L., Theobalt, C., Chen, B.: Neural novel actor: Learning a generalized animatable neural representation for human actors. IEEE Transactions on Visualization and Computer Graphics (2023) 4
10. Gao, X., Yang, J., Kim, J., Peng, S., Liu, Z., Tong, X.: Mps-nerf: Generalizable 3d human rendering from multiview images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 4
11. Habermann, M., Liu, L., Xu, W., Pons-Moll, G., Zollhoefer, M., Theobalt, C.: Hdhumans: A hybrid approach for high-fidelity digital humans. Proceedings of the ACM on Computer Graphics and Interactive Techniques **6**(3), 1–23 (2023) 4
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 10
13. Ho, H.I., Song, J., Hilliges, O.: Sith: Single-view textured human reconstruction with image-conditioned diffusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 20
14. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. arXiv preprint arXiv:2312.02134 (2023) 4
15. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: Tech: Text-guided reconstruction of lifelike clothed humans. arXiv preprint arXiv:2308.08545 (2023) 20

16. Jena, R., Iyer, G.S., Choudhary, S., Smith, B., Chaudhari, P., Gee, J.: Splatarmor: Articulated gaussian splatting for animatable humans from monocular rgb videos. arXiv preprint arXiv:2311.10812 (2023) 4
17. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023) 1, 3, 4, 5, 6, 10, 11, 12
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
19. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: Hugs: Human gaussian splats. arXiv preprint arXiv:2311.17910 (2023) 4
20. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. Advances in Neural Information Processing Systems **34**, 24741–24752 (2021) 2, 3, 4, 10, 11, 20, 21
21. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural image-based avatars: Generalizable radiance fields for human avatar modeling. In: International Conference on Learning Representations (2023) 2, 4, 10, 11, 20, 21
22. Kwon, Y., Liu, L., Fuchs, H., Habermann, M., Theobalt, C.: Deliffas: Deformable light fields for fast avatar synthesis. arXiv preprint arXiv:2310.11449 (2023) 4
23. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: 2019 International Conference on 3D Vision (3DV). pp. 643–653. IEEE (2019) 24
24. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. arXiv preprint arXiv:2311.16096 (2023) 4
25. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM transactions on graphics (TOG) **40**(6), 1–16 (2021) 4
26. Liu, Y., Peng, S., Liu, L., Wang, Q., Wang, P., Christian, T., Zhou, X., Wang, W.: Neural rays for occlusion-aware image-based rendering. In: CVPR (2022) 4
27. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023) 4, 6
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7 9
29. Mihajlovic, M., Bansal, A., Zollhoefer, M., Tang, S., Saito, S.: Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In: European conference on computer vision. pp. 179–197. Springer (2022) 4
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 1
31. Moreau, A., Song, J., Dhamo, H., Shaw, R., Zhou, Y., Pérez-Pellitero, E.: Human gaussian splatting: Real-time rendering of animatable avatars. arXiv preprint arXiv:2311.17113 (2023) 4
32. Ouyang, H., Zhang, B., Zhang, P., Yang, H., Yang, J., Chen, D., Chen, Q., Wen, F.: Real-time neural character rendering with pose-guided multiplane images. In: European Conference on Computer Vision. pp. 192–209. Springer (2022) 4
33. Pan, X., Yang, Z., Ma, J., Zhou, C., Yang, Y.: Transhuman: A transformer-based human representation for generalizable neural human rendering. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 3544–3555 (2023) 2

34. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: Ash: Animatable gaussian splats for efficient and photoreal human rendering (2023) 4
35. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14314–14323 (2021) 4
36. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021) 4
37. Porumbescu, S.D., Budge, B., Feng, L., Joy, K.I.: Shell maps. ACM Transactions on Graphics (TOG) **24**(3), 626–633 (2005) 4
38. RenderPeople. http://renderpeople.com (2018) 3, 10, 12, 19
39. Robertini, N., Casas, D., Rhodin, H., Seidel, H.P., Theobalt, C.: Model-based outdoor performance capture. In: Proceedings of the 2016 International Conference on 3D Vision (3DV 2016) (2016), http://gvv.mpi-inf.mpg.de/projects/OutdoorPerfcap/ 4
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 19
41. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2304–2314 (2019) 4
42. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016) 5
43. Su, S.Y., Yu, F., Zollhöfer, M., Rhodin, H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. Advances in Neural Information Processing Systems **34**, 12278–12291 (2021) 4
44. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023) 2
45. Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z., et al.: Is attention all nerf needs? arXiv preprint arXiv:2207.13298 (2022) 4
46. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021) 4
47. Wang, S., Wang, Z., Schmelzle, R., Zheng, L., Kwon, Y., Sengupta, R., Fuchs, H.: Learning view synthesis for desktop telepresence with few rgbd cameras. IEEE Transactions on Visualization and Computer Graphics (2024) 4
48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004) 9
49. Wang, Z., Shen, T., Nimier-David, M., Sharp, N., Gao, J., Keller, A., Fidler, S., Müller, T., Gojcic, Z.: Adaptive shells for efficient neural radiance field rendering. arXiv preprint arXiv:2311.10091 (2023) 4
50. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 512–523 (2023) 20

51. Ye, K., Shao, T., Zhou, K.: Animatable 3d gaussians for high-fidelity synthesis of human motions. arXiv preprint arXiv:2311.13404 (2023) 4
52. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021) 4
53. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. arXiv preprint arXiv:1806.03589 (2018) 3, 4, 7, 24
54. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. arXiv preprint arXiv:1801.07892 (2018) 3, 4, 7
55. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021) 3, 10, 11, 12, 19
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 10
57. Zhao, F., Yang, W., Zhang, J., Lin, P., Zhang, Y., Yu, J., Xu, L.: Humannerf: Efficiently generated human radiance field from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7743–7753 (2022) 2, 4
58. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 2, 3, 4, 10, 11, 12, 20
59. Zhou, Z., Ma, F., Fan, H., Yang, Y.: Headstudio: Text to animatable head avatars with 3d gaussian splatting. arXiv preprint arXiv:2402.06149 (2024) 4
60. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars. arXiv preprint arXiv:2311.08581 (2023) 4

A Appendix - Overview

This appendix is organized as follows: Sec. B discusses the limitations and future works; Sec. C presents the societal impacts our work can have; Sec. D shows additional results including video results, comparison with single-view methods, ablations on number of outer scaffolds, ablation study with different loss supervision, ablations on the number of input views at inference time, and runtime at inference. Sec. E provides information regarding reproducibility, which includes implementation details.

B Limitations and Future Works

Although our method achieves state-of-the-art results in terms of visual quality and runtime, it is not free from limitations. (1) While our method effectively compensates for minor inaccuracies in SMPL-X estimations through the use of multi-scaffolds, significant deviations in SMPL-X from the input images could compromise the quality of our results, as our Gaussians are anchored to the SMPL-X surface. (2) Currently, the number of scaffolds is determined empirically. It would be an interesting direction to explore adaptive scaffolds based on subject attributes (e.g., loose or tight clothing). (3) The performance of our inpainting network is constrained by the small number of ground truth texture maps available during training, which in turn limits its ability to generate detailed hallucinations when given a single-view input. Therefore, integrating and fine-tuning generative models trained on extensive datasets (e.g., Stable Diffusion model [40]) could substantially improve our network's hallucination capabilities and generalizability, which is a promising direction for future work.

C Societal Impacts

Our proposed method can push immersive entertainment and communication to a more affordable setting. For example, our work has the potential to enhance the accessibility of telepresence experiences by facilitating the creation of avatars from minimal RGB images. Moreover, the technology presents benefits to film and game production by enabling efficient synthesis of large-scale 3D human avatars with low costs.

However, our work might also introduce potential challenges, primarily related to the accessible creation of realistic human images. This could lead to deep-fake human avatars on social media, with implications for misinformation and the degradation of trust in digital content. To mitigate such risks, it is urgent to promote ethical guidelines and regulations on synthetic media. We strongly appeal transparent use of such technology as it should align with societal interests and foster trust rather than skepticism.

D Additional results

D.1 Video results

Video results of comparison with the state-of-the-art baselines on the in-domain generalization task (i.e., trained and tested on THuman 2.0 dataset [55]) and cross-dataset generalization task (i.e., trained on THuman 2.0 and tested on RenderPeople [38]) can

Fig. 7: Comparison with single-view reconstruction methods: ECON [50], TeCH [15], and SiTH [13]. Our method outperforms the baselines in terms of faithfulness to the given observation.

be found in the project website*. For the in-domain generalization task, we compare our GHG with (1) human template-conditioned NeRF, generalization from sparse view methods NHP [20] and NIA [21], and (2) generalizable 3D Gaussian Splatting for human rendering method GPS-Gaussian [58]. Note that GPS-Gaussian is trained and tested with 5 input views due to the rectification requirement. NHP, NIA, and ours are trained and tested with 3 input views. For the cross-dataset generalization task, we show comparison with our main baselines NHP and NIA. Our method can recover sharp and fine details compared to human template-conditioned NeRF baselines. Due to the lack of full 3D prior, GPS-Gaussian suffers in maintaining multi-view consistency between the novel views generated using different input views. On the other hand, ours maintains robust and accurate geometry reconstruction utilizing the 3D human template.

D.2 Comparison with single-view methods

Fig. 7 shows comparisons with SOTA single-view reconstruction methods that are based on 3D human prior: ECON [50], TECH [15], and SiTH [13]. We used their officially released implementation for the comparison. Our sparse-view work outperforms in terms of accuracy and faithfulness to the observed data, as can be seen in Fig. 7. Also, the single-view methods either require per-subject optimization (ECON, TeCH) or run at relatively slow speed (e.g., ECON 3 min / TeCH 4 hr / SiTH 2 min). On the other hand, ours is a feed-forward method that runs at $4fps$, which is $\times 480$ faster than SiTH.

D.3 Ablations

Ablation on the number of scaffolds. In Tab. 6, we study the impact of number of outer scaffolds. Variants with different number of outer scaffolds are trained and tested. The performance increase is saturated as more than 5 outer scaffolds are used. Therefore, we use 4 outer scaffolds as our final model. In Fig. 8, we show how the

* https://humansensinglab.github.io/Generalizable-Human-Gaussians

Table 6: Ablation study on the number of outer scaffolds used. We trained and tested variants with different numbers of scaffolds that are outside the original SMPL-X surface. The variant with only the base template is denoted as "0 scaffold". The performance increase is saturated as more than 5 outer scaffolds are used.

# Out scaffolds.	PSNR↑	LPIPS↓	FID↓
0	22.30	145.74	84.38
1	**22.77**	139.16	75.66
2	22.28	137.65	73.54
3	21.87	136.38	65.19
4 **(Ours full)**	21.90	**133.41**	**61.67**
5	22.13	134.73	63.80
6	22.09	135.52	64.81

Table 7: Ablation study on the supervision. ✗/✓indicates completely remove/keep the loss supervision. Our L_1-only supervision result (a) still outperforms the human template-conditioned NeRF methods NHP and NIA, which are also trained with L_1-only supervision. This validates the effectiveness of our proposed multi-scaffold.

	L_1	SSIM	Mask	Multi-view	PSNR↑	LPIPS↓	FID↓
NHP	✓	✗	✗	✗	**23.32**	184.69	136.56
NIA	✓	✗	✗	✗	23.20	181.82	127.30
a	✓	✗	✗	✗	23.05	142.57	71.97
b	✓	✓	✗	✗	22.69	136.44	69.50
c	✓	✓	✗	✓	22.03	134.82	62.04
Ours full	✓	✓	✓	✓	21.90	**133.41**	**61.67**

number of scaffolds affects the reconstruction of offset details such as hair (a,c) and loose clothing (b,c).

Ablation on the supervision. Tab. 7 shows the impact of different loss supervision employed during training. Note that our variant with L_1-only supervision (Tab. 7-a) already outperforms the human template-conditioned generalizable NeRF methods NHP and NIA, which are also trained with L_1-only supervision, in terms of perceptual metrics LPIPS and FID. This validates that our gain is not only from the different supervision but also from our proposed multi-scaffold. Our full model that leverages multi-view supervision with L_1, SSIM, and mask loss achieves the highest performance on the perception-based metrics. Note that multi-view supervision is possible by leveraging the fast 3D Gaussian splatting.

Ablation on the number of input views at inference. We trained our model using 3 input views and tested with different number of input views at inference time in Tab. 8. The performance improves as more observations are available. However, note that our performance when only given two views is still comparable to the 3-view results. This demonstrates the effectiveness of our method under sparse view setting.

Performance on the randomly selected input views. During evaluation, we followed the convention of previous sparse view 3D human reconstruction works [20,21] that use 3 uniformly distributed inputs. However, we additionally ran the evaluations given 3 random views 10 times and computed the mean metrics. We verified that the

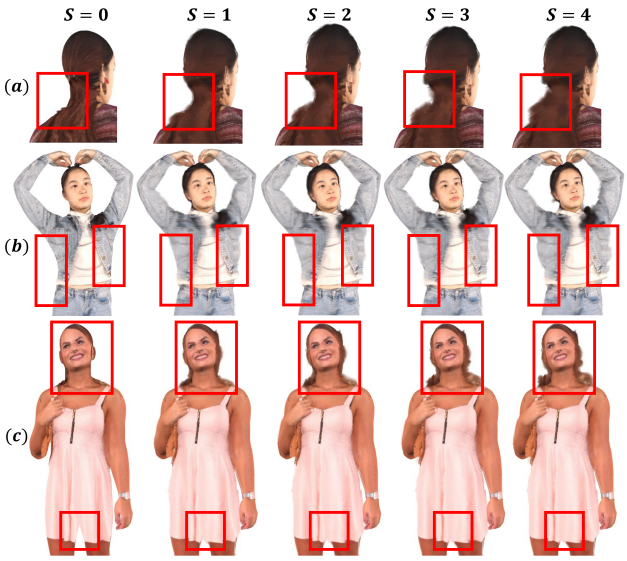

Fig. 8: Multi-scaffold helps reconstruct hair and loose clothing. S denotes the number of outer scaffolds.

Table 8: Ablation study on the number of input views at inference. We trained our model using 3 input views, and tested with different numbers of input views at inference time. The performance improves as more observations are available.

# Inputs.	PSNR↑	LPIPS↓	FID↓
1	20.08	152.54	99.13
2	21.79	132.61	78.56
3	21.90	133.41	61.67
4	22.01	133.68	53.40
5	**22.07**	**131.80**	**35.00**

performance difference between the uniformly and randomly sampled inputs is minimal – PSNR is 1.5%, and LPIPS is 0.3%.

D.4 Runtime at inference

Our GHG runs at $4fps$ for rendering a single $1K$ (1024×1024) image on a single NVIDIA RTX A4500 GPU. However, note that inpainting network takes most of our runtime (74%). Without the inpainting network, ours runs at $15fps$. More efficient inpainting model can be explored to further reduce the runtime.

The detailed breakdown of runtime is as follows. Our pipeline can be divided into three stages: (1) constructing multi-scaffold (2) Gaussian parameter map generation (3) rasterization. **(1) Constructing multi-scaffold:** RGB map for each scaffold is aggregated on the UV space of human template. Our inpainting network inpaints the

missing regions of the innermost scaffold RGB map in 180.89 ms. **(2) Gaussian parameter map generation:** Multi-Gaussian parameter maps are generated in 57.97 ms. **(3) Rasterization:** Rasterization takes 5.78 ms. In total, GHG takes 244.65 ms to render a single $1K$ image.

We would like to highlight that our method runs faster than the sparse-view generalizabl human NeRF methods NHP and NIA ($0.01 fps$ to render a single 1K image) while outperforming their visual quality.

E Implementation details

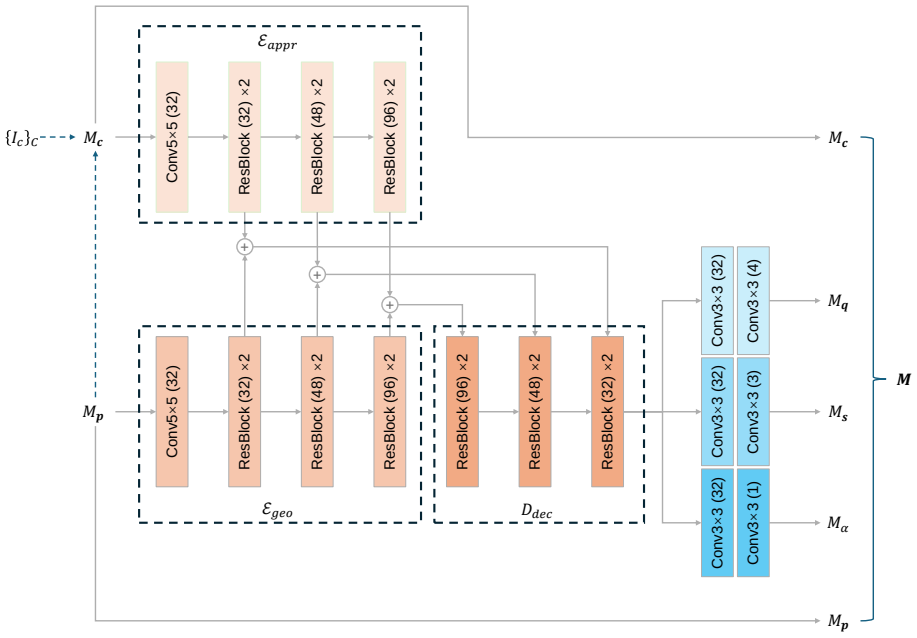

Fig. 9: Network architecture for Gaussian parameter map generation.

E.1 Gaussian parameter map generation

The architecture design of our Gaussian parameter map generation network is presented in Fig. 9. Our network is composed of two encoders $\mathcal{E}_{\text{appr}}$, $\mathcal{E}_{\text{geo}}$ and one decoder $\mathcal{D}_{\text{dec}}$. The feature maps extracted by $\mathcal{E}_{\text{appr}}$ and $\mathcal{E}_{\text{geo}}$ are added together before being fed into $\mathcal{D}_{\text{dec}}$. Moreover, $M_{\mathbf{s}}$ and M_α are sent into *Softplus* and *Sigmoid* activation layers, respectively, after the convolution layers. Note that in the figure, the number following each layer name and sitting in the bracket denotes its output channel size.

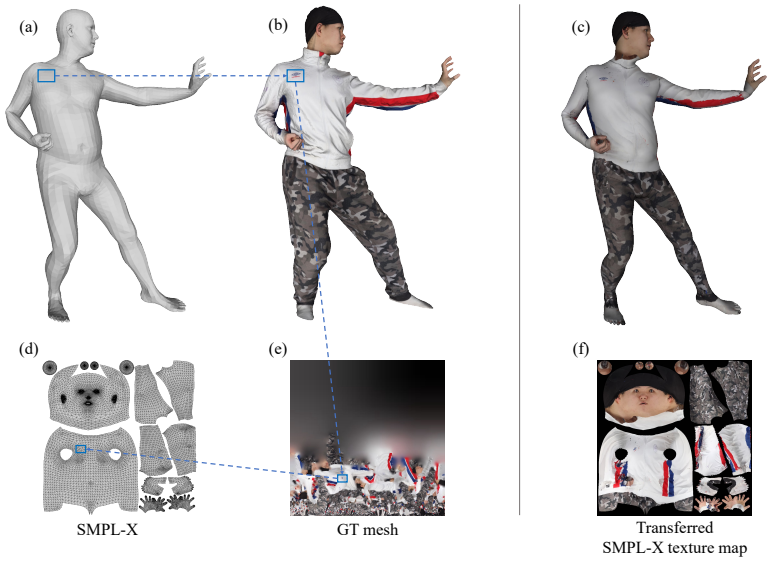

Fig. 10: Illustration of texture transfer on to the SMPL-X UV space. For each point on the SMPL-X model (a), the nearest point on the scanned mesh (b) is found. Then, we get the corresponding position of this point on the scan's UV map (e), which will be mapped to the matching location on the SMPL-X's UV map (d). Resulting on the transferred texture map (f) and the colored mesh (c).

E.2 Inpainting

Pseudo ground truth generation To create the pseudo ground truth texture map on the SMPL-X UV space, we follow the approach proposed in Lazova et al [23]. The process is illustrated in Fig. 10. For each point on the SMPL-X model, we identify the nearest point on the scanned object. Next, we determine the corresponding position of this point on the scan's UV map. We then transfer the color from this position on the scan's UV map to the corresponding location on the SMPL-X's UV map.

Network architecture Fig. 11 shows the inpainting module architecture. The inpainting network follows the DeepFillv2 design [53]. The inpainting network is composed of a generator $\mathcal{G}_{\text{inpaint}}$ and a discriminator $\mathcal{D}_{\text{inpaint}}$. In the generator, all convolutions are gated convolutions with a kernel size of 3×3 if not specified, where *GatedConv*, *DilateGatedConv*, *GatedConvDown*, *GatedConvUp* have a stride of 1, 1, 2, 0.5, respectively. The four *DilateGatedConv* layers in *DilatedBlock* have a dilation of 2, 4, 8, 16, respectively. The *Attention* layer is a self-attention layer. In the discriminator, all convolutions are common 2D convolutions, where *Conv*, *ConvDown* have a stride of 1, 2, respectively. Besides, all convolution layers are followed by ELU activation. Note that in the figure, the number following each layer name and sitting in the bracket denotes its output channel size.

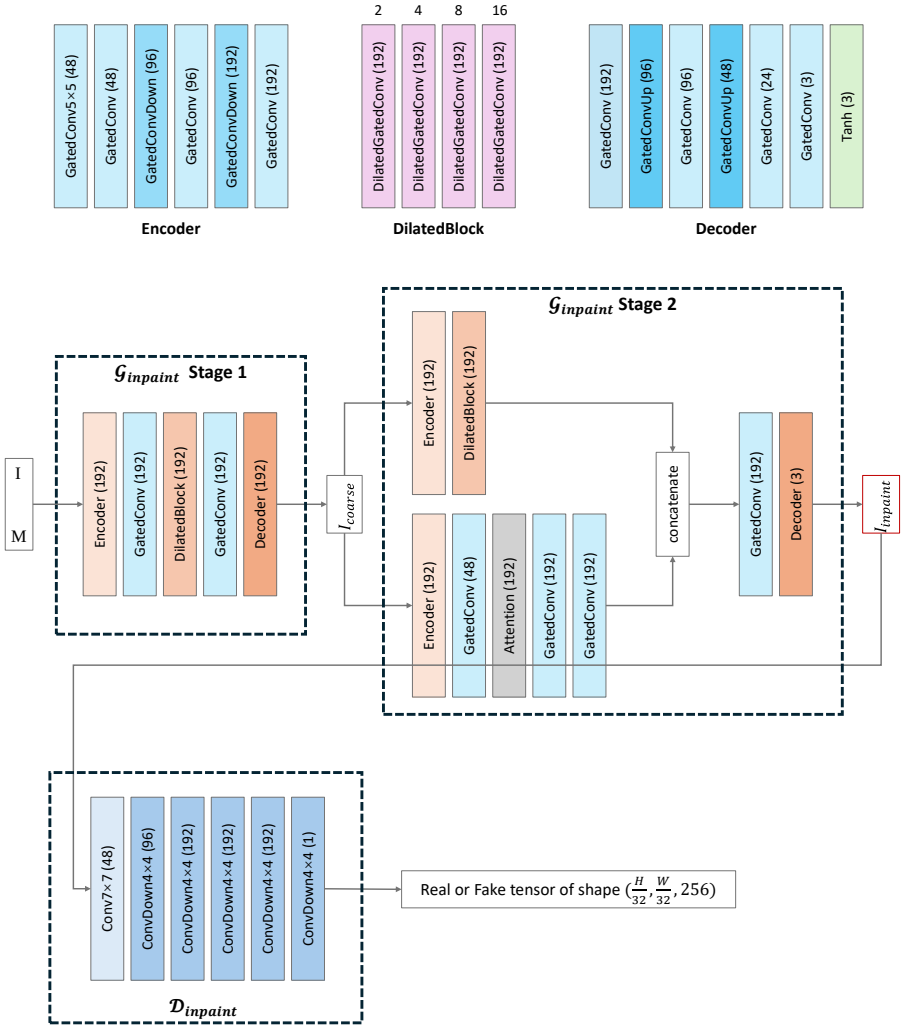

Fig. 11: Inpainting network.