Generalizable Neural Performer: Learning Robust Radiance Fields for Human Novel View Synthesis

Wei Cheng¹  Su Xu¹  Jingtan Piao¹,²  Chen Qian¹
Wayne Wu¹,³  Kwan-Yee Lin¹,²  Hongsheng Li²
¹ SenseTime Research  ² CUHK-SenseTime Joint Laboratory, CUHK  ³ Shanghai AI Laboratory
{chengwei,xusu,piaojingtan,qianchen}@sensetime.com
{wuwenyan0503,linjunyi9335}@gmail.com  hsli@ee.cuhk.edu.hk

Sparse View Input  Generalizable Neural Performer  Novel View and Geometry

Figure 1. Overview. First row presents some snapshots in our dataset, GeneBody—1.0. It is a multi-view dataset which captures human motion in a variety of pose actions, body shape, clothing and accessories. Second row shows the simplification of our framework.

Abstract

This work targets at using a general deep learning framework to synthesize free-viewpoint images of arbitrary human performers, only requiring a sparse number of camera views as inputs and skipping per-case fine-tuning. The large variation of geometry and appearance, caused by articulated body poses, shapes and clothing types, are the key bottlenecks of this task. To overcome these challenges, we present a simple yet powerful framework, named Generalizable Neural Performer (GNR), that learns a generalizable and robust neural body representation over various geometry and appearance. Specifically, we compress the light fields for novel view human rendering as conditional implicit neural radiance fields with several designs from both geometry and appearance aspects. We first introduce an Implicit Geometric Body Embedding strategy to enhance the robustness based on both parametric 3D human body model prior and multi-view source images’ hints. On the top of this, we further propose a Screen-Space Occlusion-Aware Appearance Blending technique to preserve the high-quality appearance, through interpolating source view ap-
pearance to the radiance fields with a relax but approximate geometric guidance. To evaluate our method, we present our on-going effort of constructing a dataset with remarkable complexity and diversity. The first dataset version, GeneBody-1.0, includes over 2.95M frames of 100 subjects under multi-view cameras capturing, performing a large variety of pose actions, along with diverse body shapes, clothing, accessories and hairdos. Experiments on GeneBody-1.0 and ZJU-Mocap show better robustness of our methods than recent state-of-the-art generalizable methods among all cross-dataset, unseen subjects and unseen poses settings. We also demonstrate the competitiveness of our model compared with cutting-edge case-specific ones. Dataset, code and model will be made publicly available.

1. Introduction

Synthesizing free-viewpoint images of human performers has a wide range of applications in film production, 3D immersive communication and AR/VR gaming. However, the most impressive industry-level rendering results currently require both specialized studio environments [35] and cumbersome artistic design [1] on each human subject. What if such stirring effect could be automatically realized with just using images captured under a casual setting (e.g., very limited number of cameras)? Such capability would dramatically accelerate film production and improves the accessibility of 3D immersive experience in daily life.

In this work, we focus on improving the generalization and robustness in free-viewpoint synthesis for arbitrary human performers from a sparse set of multi-view images. To achieve this, two key challenges need to be solved. First, for the generalization of unseen subjects and unseen actions during inference, the model is asked to represent arbitrary shape and appearance variation caused by different human posing and clothing. Second, rendering high-quality results requires detail preserving of appearance as well as multi-view consistency.

Many previous pioneer works fall short of these goals. There has been attempts either rely on deploying depth sensors for high-quality geometry reconstruction before rendering [10, 12], or building dense camera arrays to capture the changing appearances from different viewing angles [11]. The requirement of professional equipment limits the application of such technologies in personal and daily usage scenarios. Recent works adopt neural networks to learn 3D geometry and appearance from data [21, 33, 36, 54]. Among these methods, the neural implicit representations have shown efficient improvements over image fidelity and reconstruction accuracy. Despite the notable technical revolution they bring in, these methods still suffer from case-specific optimization [33], scanty pose generalization [36, 37], or unrealistic rendering [21, 36, 54]. Some current cutting-edge approaches rely on temporal coherence of the same subject [21, 33] which requires geometry fitting or motion tracking across canonical models. Although lavish geometry and appearance cues lie in 3D body motion across time, there are more instantaneous snapshots sources rather videos in daily life.

To address aforementioned challenges, we present Generalizable Neural Performer, or GNR for short, a novel framework that learns a generalizable and robust neural body representation on various human pose, shape and appearance. We achieve this with several effective designs on tailoring body structural prior and source-view appearance conditions into neural radiance fields learning. Specifically, to be able to represent arbitrary human performers in one single model without per-case finetuning, an Implicit Geometric Body Embedding strategy is firstly introduced. It extracts the geometric information and body semantics from both parametric 3D human body model and multi-view source image feature fusion, which turns the neural radiance field into a conditional implicit field. This strategy has the following properties: 1) The existing parametric models mostly represent “minimally clothed” bodies (e.g., SMPL [25]), which could provide basic geometric prior in pose and constrain the minimal shape. 2) The multi-view input image features encode the firsthand geometric hints of hairdo, cloth and accessories that form the unabridged shape. Thus, the predicted human geometry could be conditionally anchored by the parametric body prior and further bordered by multi-view observation. In short, such an embedding could facilitate network learning a reasonable body pose and shape parameterization via a conditional implicit neural representation for articulated clothed human and thus skipping per-scene optimization. A proper geometry prediction could also in turn help the generalization of appearance.

To robustly producing faithful appearance, we further propose an Screen-Space Occlusion-Aware Appearance Blending (SSOA-AB) technique. It combines lumigraph blending techniques [11, 19] with screen space occlusion estimation. The key aspect of this design is to disentangle blending coefficient from source views into a visibility map, guided by parametric model prior and screen-space self-occlusion information, and a view attention coefficient between source and query views learnt from large-scale training data prior. As the texture from source views provides authentic details, the pixel from these views could be treated as the firsthand base reference color to rectify the unreliable color prediction of the neural radiance field. Whereas not all points could be observed through input views due to sparse view observation as well as occlusions of body parts, directly applying classic lumigraph blending will lead to ghosting texture in fully unobserved part. We thus introduce SSOA-AB, which blends from both source views and

---

1 Project page: https://generalizable-neural-performer.github.io/
Code and model: https://github.com/generalizable-neural-performer/gnr

2 We use the word ‘anchor’ to indicate constraining the inferred human shape to be close to the proper body surface.
Table 1. Summary of key related methodology. Generalizable: Ability to model multiple objects/scenes in same model. Render: Whether has a learnable render. Latent: Latent code types used to condition the network. Prior: Type of prior used to bound the structure of objects/scenes, eg., SMPL [25] as human body prior. Occlusion: Whether consider occlusion during rendering. Supervision: Type of losses. * indicates the type of the supervision is optional.

| Methods     | Generalizable | Render | Latent  | Prior            | Occlusion | Supervision |
|-------------|---------------|--------|---------|------------------|-----------|-------------|
| NeRF [28]   | X             | ✓      | -       | -                | -         | 2D          |
| pixelNerf [49] | ✓             | ✓      | 2D      | -                | 2D        |            |
| IBRNet [42] | ✓             | ✓      | 2D      | -                | -         | 2D          |
| PIFu [36]   | ✓             | X      | 2D      | SMPL (Depth)     | -         | 3D          |
| PaMIR [54]  | ✓             | X      | 2D+3D   | SMPL (Vertices)  | -         | 2D          |
| NB [33]     | ✓             | ✓      | 2D+3D   | SMPL (SDF&Depth) | ✓         | 2D(3D*)     |

To summarize, our work contributes as follows

- We propose a novel approach, named Generalizable Neural Performer, that achieves free-viewpoint synthesis of arbitrary human performers with learning a generalizable and robust implicit neural body representation.

- We present an Implicit Geometric Body Embedding strategy and a Screen-Space Occlusion-Aware Blending technique to facilitate the learning of the representation. We show that a proper design, like these two components, for tailoring prior knowledge of parametric model and source images, can effectively enhance robustness to both human geometry and appearance.

- We contribute a multi-view dataset, GeneBody-1.0, with human samples under various appearances on consideration of ethnicity, age, clothing style and conducting with performing diverse types of actions. We hope this dataset facilitate future research in generalizable human rendering towards real-world scenarios. We also benchmark results of several state-of-the-arts, with hope to offer some insights into current status and future trends in this field.
2. Related Work

2.1. Novel View Rendering

Image-based Rendering. Image-based rendering (IBR) is the fundamental problem of rendering novel views of scenes from sampled views. Plenoptic sampling theorem [7] determines minimum sampling rate for anti-aliased rendering bounded by minimum and maximum scene depths. Methods based on this theorem require dense view sampling to achieve high-quality rendering; many systems devise large-scale camera arrays [44]. Other methods explored to use geometric proxies, like global meshes [5, 13, 45] or local layered depth [27, 55] to alleviate sampling requirement. While the performance of these methods highly relies on the accuracy of geometry estimation and typically fails on low-textured regions.

 Neural Scene Representation. Recent works use neural network to represent a scene [28, 36], although these methods can generate rather satisfying novel views using scene-specific models, the quality of rendered images tends to degrade when the model is trained to fit various objects/scenes. Recent works pixelNerf [49] and IBRNet [42] improves generalization of 3D implicit functions by conditioning query coordinates with image features. These methods are typically unable to generate plausible free-viewpoint video with significant camera movement.

2.2. Human Digitization

Deep Representation with Parametric Models. The emergence of statistical models like SCAPE [3], and SMPL [25] create another cue of modeling humans in a general way, which is also treated as initial guesses in recent works. [34] estimates parametric model from video and renders the full model by per-frame silhouette. [14, 54] extract 3D features from voxelized SMPL to condition implicit function and achieves pose robustness compared to [36]. While the requirement for a precise body scan hinders its robustness due to the biased data.

Modelling and Rendering Human from Images. To model/render human, previous study focus on reconstruction from shape [43], stereo [23] and shading [41] using multi-view systems. While these works are subject-specific and require sophisticated capturing volume setup and system calibration. Recently, [21, 36, 37, 56] have made progress on reconstructing dense human models from a sparse set of images or even a single image in the wild by adopting data-driven techniques. Recently, some concurrent work in preprint [8, 21, 47, 52] generalize human in similar spirit. In contrast to these methods utilizing temporal information to complement the insufficiency of single-frame sparse-view sources, our framework tries to generalize human without extra effort on temporal coherent.

A summarized comparison of proposed method with most relevant related work is shown in Tab. 1.

3. Method

Overview. Recall that given a sparse set of calibrated multi-view images $I = \{I_i\}_{i=1}^{S}$ of a person, our goal is to synthesize high-quality images of that person in arbitrary novel views. To this end, our model is asked to tackle challenges of (1) representing arbitrary human performers in one single model without retraining on target case; (2) producing faithful appearance with detail preserving and multi-view consistency. We alleviate these two dilemmas with introducing an Implicit Geometric Body Embedding strategy (for the former) and a Screen-Space Occlusion-aware Appearance Blending strategy (for the latter) into the learning of a conditional neural radiance field $F^3$.

Specifically, we assume a parametric human body $M$, which fitted to input source images, is given as prior. A deft implicit body shape encoding $G$ is built upon $M$ to anchor performer’s pose and constrain the roughly occupied range in 3D space. In parallel, we extract source images’ features from the encoder $E$ and fuse them via multi-view feature fusion to form the unbridged shape hints complementing with $G$. Since multi-view input image features encode the firsthand geometric cues of off-minimal body part like hairdo, cloth and accessories. Then, for a query point $x$ in 3D space, we look up the corresponding point-aligned features from both $G$ and $E$ to constitute the features of Implicit Geometric Body Embedding (Sec. 3.1) to condition the implicit field for the radiance prediction. Alongside, we fed corresponding pixel color from source views and predicted color from the radiance field into the Screen-Space Occlusion-aware Appearance Blending module (Sec. 3.2) to obtain the rectified color value, which is the final output color of GNR. Details are provided in ensuing subsections.

3.1. Implicit Geometric Body Embedding

Implicit Body Shape Encoding. It is intuitive to use a parametric body model as hands-down geometric prior to anchor human shape to the pose during the learning of implicit field. The SMPLx model $M(\theta, \beta)$ is a mesh consisting of $n_s = 10475$ vertices with fixed topology. $\theta \in \mathbb{R}^{72}$ and $\beta \in \mathbb{R}^{10}$ denote human pose in joints rotation and hyper-parameters of body shape from a statistic model. However, injecting a dense mesh or hyper-parameters of SMPLx is impractical and meaningless, due to its high dimensionality and requirement on shape basis.

We thus propose an alternative to use the signed distance function (SDF) of the mesh to parametrize the Euclidean volume. Specifically, for any 3D location $x$, we first compute its nearest point on the mesh surface $v = \sum_{j=1}^{3} c_j * v_{i,j}$, where $v_{i,j}$ is the $j$-th vertex on the nearest triangular surface $i$, and $c_j$ is the projected coefficients with constraints $\sum_{j=1}^{3} c_j = 1, c_j \geq 0$. Next, the SDF

\begin{equation}
\sigma(\mathbf{d}) = \sqrt{\|\mathbf{x} - \mathbf{y}\|^2 - \delta^2}
\end{equation}

is the basic principle of neural radiance field, or short NeRF [28], is to learn a continuous 5D function by mapping a 3D location $\mathbf{x} = (x, y, z)$ within a bounding volume surrounding the scene and a 2D view direction $\mathbf{d}$ to an emitted color $\mathbf{c} = (r, g, b)$ and a volume density $\sigma$. The pixel is finally rendered via ray integration.
of the query point $x$ can be calculated as $S(x, M) = \text{sign}(x, M) ||x - \bar{v}||_2$, where $\text{sign}(x, M)$ equal to 1 when $x$ is inside $M$ and $-1$ when it is outside $M$.

Note that the $\text{sign}$ function requires $M$ to be a water-tight mesh. This embedding benefits in two ways: 1) The SDF embedding provides information on how close the query point inwards or outwards the minimal human body surface. 2) Converting implicit SDF from mesh, the compact and efficient properties of 3D geometry are preserved, as demonstrated in [26, 30, 53]. Besides, we also incorporate SDF’s derivative with respect to $x$, where we denote it as $S_x'$. The derivative explains in which direction the point is approaching / leaving the mesh surface, or in other word, the normal direction of the surface.

Still and all, the SDF value and its derivative cannot reflect body part semantics, which would lead to local geometry, i.e., body part ambiguity, when the body pose changes. Therefore, to further eliminate such ambiguity, the body semantic knowledge is expected to be incorporated. That is, where on the body does the location of query point $x$ approximately lie. We choose to map the closest point $v$ on the SMPLx model $M(\theta, \beta)$ to its corresponding coordinates $\bar{v}$ in a canonical SMPLx mesh $\bar{M}$ with neutral shape $\bar{\beta}$ and static pose $\bar{\theta}$, and $\bar{M} = M(\bar{\theta}, \bar{\beta})$. Locations of the corresponding points in the canonical space serve as an effective representation of the semantic context regarding the template human body. The complete body shape encoding is written as $G(M, c; \bar{M}) = [S; S_x'; \bar{v}]$, where $[;]$ denotes concatenation. The proposed implicit body shape encoding helps to anchor the implicit radiance field to the proper human body locations in several ways: (1) Geometry generalization. Introducing the dense parametric model information to the implicit field can be interpreted as providing an initial guess of the human pose and shape. The network is thereby trained with a easier task of learning the residual between clothed human and its corresponding minimally clothed body shape. (2) Low cost. To provide meaningful constraints to the implicit field, conventional ways are either expensive in computation and memory as they directly incorporate voxelized body volume/ local 3D patch features [33, 54, 56], or lack geometry prior due to oversimplification of a relative direction to articulate skeleton representation [38]. In contrast, our implicit body shape encoding combines local geometric prior and body semantics from body template, achieving excellent computation and memory efficiency. Such simple representation effectively conditions the implicit field.

Multi-view Feature Fusions and Radiance Prediction. As most parametric models represent ‘minimally clothed’ bodies, simply condition the implicit filed with only implicit body shape encoding is insufficient to infer off-minimal body shape part, e.g., hairdo, cloth and accessories. Thus, we need a complement to offer these cues. As the multi-view input image features intrinsically encode first-hand geometric information of these aspects, the implicit filed is also conditioned by the pixel-aligned features from multi-view inputs aside from implicit body shape encoding. Specifically, we first extract the feature using a convolution network, denoted as $E$. The query point $x$ is projected to 2D location $p$ on each source image planes given the camera pose and intrinsic parameter, and the image feature vector is obtain via $f(x) = E(p)$ with bilinearly interpolation.

We separate density and radiance reasoning network $F_\sigma$, $F_{c_0}$ into two stages:

$$F_\sigma, F_{c_0} = F_2 \circ \phi[F_1(x, G, f)],$$

$^*$Note that we adopt the same network architecture with [36] and initialize its parameters from the pretrained model.
where $F_1$ is an MLP that takes $x$, body model embedding $G$ and image feature vector $f$ as input and outputs an intermediate feature embedding for each of source views. Like [36,49], these intermediate features are fused via average pooling $\phi$, and aggregated feature is then passed into the volume density and radiance reasoning MLP $F_2$ to predict occupancy probability and radiance of $x$.

3.2. Screen-Space Occlusion-aware Appearance Blending

Rendering high-quality images of arbitrary human performer from a general implicit function without retraining on that case is difficult, especially at regions with high-frequency textures, e.g., hair, accessory and cloth textures. Rendering at these regions may be blurry or even glitchy. Recent works learn the camera blending weights of implicit scenes [42] or geometry proxies [11] from source view images. However, these methods require dense view references based on the dense assumption of unstructured lumigraph rendering [5], and fail in sparse setting with further triggering ghosting effect especially in source view occluded part. To solve this, our idea is to disentangle the camera blending field $B$ into an occlusion-aware visibility $o$ and a view attention based blending coefficient $\gamma$.

Screen-Space Occlusion. Ideally, we can infer the texture of any point once a precise occlusion of clothed human and dense source views are provided. However, computing precise occlusion requires either casting rays from a certain point and checks for intersection with reconstructed mesh or estimating depth from source view ray integration. To train such network is difficult and computational ineffective. While luckily, the hands-down parametric body model could provide initial geometric information of self-occlusion and source-view image observations offer visual clue of depth value continuity in out-of-body regions. Thus, we can instead learn an approximate occlusion in the observed screen-space with enforcing the network to diffuse the initial minimal occlusion to clothed ones.

Based on this, we propose a network $F_o$ to learn the occlusion (Learned SSO in Fig 4(b)) from given minimal-clothed occlusion map (Occ.Prior in Fig 4(a)) and image features, the occlusion visibility can be written as

$$o(x, d) = F_o(x, d; f_{x,d}, F_1)$$

where we model $c = \psi(z_{ref} - z)$ as the visibility of $x$ from view direction $d$ in context of SMPLx model, and $z_{ref}$ is the view depth of SMPLx model which can be obtained via a rasterization process, and $z$ is depth of $x$, $\psi$ is a sigmoid.

View Attention based Camera Blending. We also propose a view attention block to learn the appropriate view attention coefficient from angular and visual similarity under large-scale data prior. Specifically, we deploy an attention module originated from transformer block [40], to model the camera blending field $T$ according to the self- and cross-similarities of learned features. $Q$ analogises to the query feature in the context of transformer, and $K$ analogises to the key. In our scenario, they are computed from query point $x$, view directions $d_i$, and view feature embedding $F_{1,i}$ through two separate MLPs. Our final color prediction process can be written as the attention based camera blending:

$$c = softmax\left(\frac{QK^T}{\sqrt{d_k}} \cdot OT\right)V,$$

where $V$ stacks all observed colors, $O$ stacks all the visibilities. Directly blending textures from source view will cause ghosting artifacts in novel view regions occluded from all source images, leading the problem ill-posed in sparse setting, as illustrated in Fig. 4. To alleviate this, we propose to regard the prediction of radiance $c_0$ of Eqn. 1 as an observation of a non-occluded virtual camera that is in identical direction with query view direction $d$. Thus, we could extend blending procedure in Eqn. 3 to associated $N+1$ views. The associated camera blending can apply learnable blending weights in different regions across $N+1$ views of the human performer which enables ghosting-free appearance estimation. The final color prediction can be obtained from $F_c = T \circ F_{c_0}$. We call the whole process of this section as screen-space occlusion-aware appearance blending.
3.3. Volume Rendering with Conditional Neural Radiance Field

Following NeRF [28], we render the color of the rays passing through our implicit neural radiance filed via projective integration to obtain 2D images. Values of each rendered pixel is estimated via integrating density and colors along its corresponding camera ray $r_p = o + t_x$. For a pixel $p$, its corresponding camera ray $r$ is calculated with known camera poses. We then uniformly sample $N$ points $\{\chi_n\}_{n=1}^N$ along the ray within near and far bounds $t_n, t_r$, which are determined according to SMPLx regressions. The expected color $C(r)$ is calculated as:

$$C(r) = \int_{t_n}^{t_r} \exp \left( - \int_{t_n}^{t_r} \sigma(r(s))ds \right) \sigma c \, dt \quad (4)$$

3.4. Loss Function

Photometric Loss. We render the color of each ray using both $c_0$ and blending results $c$, and define photometric loss as the mean squared error (MSE) between the rendered colors and groundtruth pixel colors $c$ for training:

$$L_{\text{photo}} = \sum_{r \in \mathcal{R}} \left[ ||c_0(r) - \hat{c}(r)||_2^2 + ||c(r) - \hat{c}(r)||_2^2 \right] / |\mathcal{R}|$$

where $\mathcal{R}$ is the set of rays in the batch, and $|\mathcal{R}|$ represents the number of rays.

Occupancy and Occlusion Loss. When 3D groundtruth is available, we first supervise the occupancy probability for each point in the casting rays. We model occupancy probability as $\tanh(\sigma)$. Besides, we also guide the occlusion-aware network to learn reasonable 'diffused' occlusion $o$ from data, using groundtruth screen-space occlusion $\psi(z_{gt} - z)$, where $z_{gt}$ is groundtruth rasterized depth from source view. Our occupancy and occlusion loss can be written as:

$$L_{\text{geo}} = \sum_{r \in \mathcal{R}} \sum_{x \in r} \left[ |\tanh(\sigma(x)) - sgn(x)|^2 / |\mathcal{X}| \right]$$

$$+ \sum_{r \in \mathcal{R}} \sum_{x \in r} \left| o - \psi(z_{gt} - z) \right|^2 / |\mathcal{X}|$$

where $x$ is sample point from ray $r$, and $|\mathcal{X}|$ is the number of all sample points.

Overall, our method is optimized by both photo-metric and geometric loss:

$$L = L_{\text{photo}} + \lambda_{\text{geo}} L_{\text{geo}}$$

where $\lambda_{\text{geo}} \in \{0, 1\}$, and set to 1 when the 3D scan is available.

4. Experiments

4.1. Datasets

RenderPeople [2]. A commercial dataset that contains high-fidelity 3D colored static scan meshes with large variety in appearance and poses. We acquire 260 static human scans, and split 240 into a train set and the rest as test set.

ZJUMocap [33]. It contains 10 sequences of human performances, each captured by 23 uniformly distributed synchronized cameras. We split the first 300 frames from three sequences as training set and rest as test set.

GeneBody-1.0. GeneBody for short, is a new dataset that we collected to evaluate the generalization and robustness of human novel view synthesis. It consists of over total 2.95M frames of 100 subjects performing 370 sequences under 48 multi-view cameras capturing, with a variety of pose actions, in different types of body shapes, clothing, accessories and hairdos, ranging the geometry and appearance varies from everyday life to professional occasions. The SMPLx estimation and foreground segmentation are provided for each frame. Although several groups have proposed datasets which have provided essential avenues for research into human bodies rendering, e.g., ZJUMoCap [33] and CMU Plenoptic [16], HUMBI [50]. They suffer from some drawbacks as a benchmark for the generalization capability of human rendering methods. Human samples in these datasets either have limited cloth or accessory types [15, 16, 33, 48], small pose variance [15, 33] or without daily human-object interaction [50]. GeneBody, on the other hand, have a broad distribution across different human races and ages as shown in Fig. 2. Moreover, it gathers a numerous numbers of clothing styles and poses, not only under daily life but also professional occasions such as traditional opera costume, with various outer body accessories like guitars and keyboards, as well as performing actions with large joint rotation like yoga, locking dances and etc, as shown in Fig. 1, 5, 8, 6. We select 40 sequences as training set, and another 10 sequences as test set in this paper.

Unless specified, all synthesis data in RenderPeople is used by rendering around the human in a circle. For all datasets, 4 uniformly distributed view inputs are given for all methods and metrics are evaluated in all camera views.

4.2. Evaluation on Real-World Human Performers

Baselines and Setting. We evaluate two categories of baseline methods: (1) generalization methods, pixelNerf [49] and IBRNet [42]; (2) case-specific methods, NeuralBody (NB) [33], NeuralTexture (NT) [39], NHR [46] and NeuralVolumes (NV) [24]. To evaluate both categories of methods on the same data, we adopt the following experimental setting. For generalization methods, we use the RenderPeople pretrained model (describe in Sec. 4.3) and train 5 epochs on training set of ZJUMoCap and GeneBody separately for each dataset evaluation; and per-case models of case-specific methods are trained on first 300 frames of ZJUMoCap and first 100 frames of GeneBody in test set. All the rest frames are benchmarks for both categories of methods for fair comparison. Such setting leads to unseen ID unseen pose for generalization methods and seen ID unseen pose for case-specific method. Moreover, we use SMPLx UV map as input for NT [39], and densely samples textured
Figure 5. Novel view synthesis on GeneBody test set under unseen ID and poses. We compare qualitative results of novel view of unseen ID and unseen poses on generalization methods (Orange) and case-specific methods (Blue). Subjects in each row come from Normal Case (Kicking, Samba), Hard Shape (Cosplay, Hanfu) and Hard Pose (Yoga) subgroups respectively. Please enlarge for details.

SMPLx for NHR’s [46] point cloud input.

Results. As shown in Tab. 2 and Fig. 5, GNR achieves the leading performance rendering novel views without any further unseen human finetuning on both human performance dataset. NV [24] takes multi-view images as the input of autoencoders and estimate mixture of affine transformation from first canonical frame, it fails to estimate the correct warping when unseen body pose differs largely from seen ones, see Samba and Cosplay in Fig. 5. NHR [46] and NT [39] both uses human geometry proxies as input and render the images with a convolution network, they tend to fail when warp fields are improperly estimated which frequently happens in stretch cases, see Hanfu and Yoga. The most related NB [33] extracts latent code from body model vertices, it tends to learn in-precise latent code in non-rigid region with large displacement over SMPLx model, see Hanfu and Cosplay. Note that pixelNerf can be regard as an degradation of our method with image feature condition only, the estimated radiance usually chromatically different from real image.
Table 2. Quantitative Results on GeneBody testset. We evaluate novel view synthesis on human performance with state-of-the-art methods on unseen ID unseen pose. Case-specific methods are also compared on seen ID unseen pose.

| Method        | NV [24]   | NHR [46]  | NT [39]   | NB [33] |
|---------------|-----------|-----------|-----------|---------|
| PSNR ↑        | 19.86     | 20.05     | 21.68     | 20.73   |
| SSIM ↑        | 0.774     | 0.800     | 0.881     | 0.878   |
| LPIPS ↓       | 0.267     | 0.155     | 0.152     | 0.231   |

Table 3. Quantitative Results on RenderPeople testset. We evaluate geometry reconstruction and novel view synthesis with state-of-the-art methods.

| Method       | PIFu [36] | PIFuHD [37] | PaMIR [54] | pixelNerf [49] | IBRNet [42] |
|--------------|-----------|-------------|------------|----------------|--------------|
| Chamfer ↓    | 6.432     | 2.765       | 1.565      | 1.250          | 24.864       |
| Normal ↓     | 0.606     | 0.421       | 0.125      | 0.583          | 0.853        |
| UHD ↓        | 22.254    | 12.621      | 3.585      | 8.842          | 74.960       |
| F-score ↑    | 0.445     | 0.670       | 0.809      | 0.610          | 0.062        |

Figure 6. GeneBody Statistics of Challenging Cases.

Figure 7. Qualitative Results on RenderPeople. Novel view images (top row), geometry (pink in bottom row) and projection error (blue in bottom row) are displayed.

4.3. Reconstruction and Rendering Evaluation on Synthetic Data

Baselines and Setting. We evaluate static human generalization in two scopes, reconstruction and rendering. State-of-the-art generalizable human reconstruction methods PIFu [36], PIFuHD [37] and PaMIR [54], and volume rendering methods pixelNerf [49] and IBRNet [42] are selected as comparison baselines. We re-train all the networks on RenderPeople training set until convergence and evaluate both quantitative results on test set in both 2D and 3D metrics. Specifically, chamfer distance, chamfer normal distance, universal Hausdorff distance (UHD), and F-score are measured between reconstructed mesh and groundtruth scan. Image quality is also estimated in PSNR, SSIM and LPIPS [51], we obtain PIFu, PIFuHD, PaMIR’s image from test views via rasterizing the output mesh.

Results. As shown in Tab. 3 and Fig. 7, GNR achieves outstanding geometry reconstruction performance, and outperforms all methods in rendering quality. Compared with with PaMIR [54] who also utilizes parametric model conditioning, GNR demonstrates its better overall geometry alignment ability (chamfer distance) for body and outer-body accessory because of geometry consistency across different views. On the other hand, despite other radiance based methods usually produce relative noisy object surface, GNR suffers much less due to contribution of body shape embedding. An interesting found is despite IBRNet [42] achieves high-fidelity rendering especially in close views, it provides extremely poor geometry reconstruction, even if a visual hull assumption is made.

4.4. Ablation Study

We examine the impact of important designs on both RenderPeople and GeneBody. Settings are same as previous experiment. To examine the effectiveness of our body shape embedding, attention-based appearance blending and Screen-space Occlusion-aware appearance blend-
Table 4. Ablation Study. Impact of our implicit body embedding, attention-based appearance blending, occlusion-aware blending techniques are studied.

Table 4: Ablation Study.

|                      | Render People |                      |                      |                      |                      |
|----------------------|---------------|----------------------|----------------------|----------------------|----------------------|
|                      | 3D Geometry Reconstruction | Novel View Synthesis |                      |                      |                      |
|                      | Chamfer ↓ Normal ↓ UHD ↓ F-score ↑ | PSNR ↑ SSIM ↑ LPIPS ↓ |                      |                      |                      |
| Ours w/o body        | 1.043 0.281 9.692 0.725 | 27.167 0.946 0.060 |                      |                      |                      |
| Ours w/o att.        | 0.694 0.201 4.848 0.800 | 27.491 0.945 0.056 |                      |                      |                      |
| Ours w/o occ.        | 0.600 0.205 4.949 0.804 | 28.769 0.956 0.045 |                      |                      |                      |
| Full model           | 0.550 0.198 4.879 0.807 | 29.201 0.964 0.043 |                      |                      |                      |

Figure 8. Qualitative Ablation study on GeneBody Dataset. We show the impact of proposed implicit body embedding, attention-based appearance blending, occlusion-aware blending on Peking opera, Street dance sequence of GeneBody. Better zoom in.

5. Conclusion

Creating free-viewpoint videos using a general model is a highly ill-posed problem due to the large shape, dressing, and pose variance of human performer. This paper contributes a dataset with remarkable volume, which covers all aspect in general performance rendering. We also propose a novel approach GNR. Facilitated with proposed Implicit Geometric Body Embedding strategy and a Screen-Space Occlusion-Aware Blending technique, GNR successfully tailors prior knowledge of parametric model and source images, enables robust and effective estimation on human geometry and appearance. Qualitative and quantitative results demonstrate its state-of-the-art performance on multiple human datasets.

Acknowledgements. This work is supported in part by Centre for Perceptual and Interactive Intelligence Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants (Nos. 14204021, 14207319, 14203118, 14208619), in part by Research Impact Fund Grant No. R5001-18, in part by CUHK Strategic Fund.

References

[1] Metahuman, 2017. 2
[2] Renderpeople, 2017. 3, 7
[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In ACM SIGGRAPH 2005 Papers, pages 408–416. 2005. 4
[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In European conference on computer vision, pages 561–578. Springer, 2016. 13
[5] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pages 425–432. 2001. 4, 6
[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. 13, 14
[7] Jin-Xiang Chai, Xin Tong, Shing-Chow Chan, and Heung-Yeung Shum. Plenoptic sampling. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pages 307–318. 2000. 4
[8] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Jiasheng Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. arXiv preprint arXiv:2112.04312, 2021. 4
[9] Wei Cheng, Lan Xu, Lei Han, Yuanfang Guo, and Lu Fang. ihuman3d: Intelligent human body 3d reconstruction using a single flying camera. In Proceedings of the 26th ACM international conference on Multimedia, pages 1733–1741. 2018. 15
[10] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts
Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016. 2

[11] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, 1996. 2, 6

[12] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (TOG)*, 36(4):1, 2017. 2

[13] Peter Hedman and Johannes Kopf. Instant 3d photography. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 4

[14] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 4

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 3, 7

[16] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohi Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3, 7

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 13, 14

[18] Hiroharu Kato, Yoshihata Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 13

[19] Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. *arXiv preprint arXiv:2103.11571*, 2021. 2

[20] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human shape and pose via model-fitting in the loop. In *ICCV*, 2019. 13

[21] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 4

[22] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 16(2):150–162, 1994. 15

[23] Yebin Liu, Qionghai Dai, and Wenli Xu. High-resolution 3d human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2, 3, 4, 5, 6, 9, 15

[24] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 5

[25] Justus Thies, Michael Zollhoefer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 7, 8, 9, 16, 17
[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017. 6

[41] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In ACM SIGGRAPH Asia 2009 papers, pages 1–11. 2009. 4

[42] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. arXiv preprint arXiv:2102.13090, 2021. 3, 4, 6, 7, 9, 16

[43] Michael Waschbäusch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus Gross. Scalable 3d video of dynamic scenes. The Visual Computer, 21(8):629–638, 2005. 4

[44] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. In ACM SIGGRAPH 2005 Papers, pages 765–776. 2005. 4

[45] Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and Werner Stuetzle. Surface light fields for 3d photography. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pages 287–296, 2000. 4

[46] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1682–1691, 2020. 7, 8, 9, 16, 17

[47] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion. arXiv e-prints, page arXiv:2110.13746, Oct. 2021. 4

[48] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. ACM Trans. Graph., 37(2):27:1–27:15, May 2018. 7

[49] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. Pixelnerf: Neural radiance fields from one or few images. arXiv preprint arXiv:2012.02190, 2020. 3, 4, 6, 7, 9, 16

[50] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2990–3000, 2020. 3, 7

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018. 9

[52] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Generalizable neural human radiance field from sparse inputs. arXiv preprint arXiv:2112.02789, 2021. 4

[53] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. arXiv preprint arXiv:2011.14565, 2020. 5

[54] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. 2, 3, 4, 5, 9, 14

[55] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fryffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817, 2018. 4

[56] Pierre Zins, Yuanlu Xu, Edmond Boyer, Stefanie Wuhrer, and Tony Tung. Learning implicit 3d representations of dressed humans from sparse views. arXiv preprint arXiv:2104.08013, 2021. 4, 5
Appendix

A. Introduction

In this appendix, we provide detailed discussion on the design of proposed method Generalizable Neural Performer (GNR) and additional qualitative and quantitative experiments.

Specifically, this material is organized as follows: (1) the elaborate description of Implicit Geometric Body Embedding (Sec. B); (2) the detailed description of network architecture designs, training and inference strategies for GNR (Sec. C); (3) more quantitative and qualitative results on both synthetic and real datasets (Sec. D).

B. Implicit Body Shape Embedding

Recall that GNR propose to anchor the network with a parametric dense body model, which is achieved by the implicit body shape embedding with local geometric and semantics. In this section, we provide elaborate descriptions over the implicit shape embedding stage. Specifically, we first provide details of the body model fitting algorithm from multi-view images in Sec. B.1. The closest point searching algorithm and gradient computation of signed distance function of our implicit body shape embedding is then discussed in Sec. B.2 and in Sec. B.3 respectively. Lastly, in Sec. B.4 we explain our grid-based acceleration method and its implementation in modern GPU.

B.1. Multi-view Body Fitting

As illustrated in Fig. S1, to estimate the accurate and uniform SMPLx [32] mesh and parameters \( \theta, \beta \) from multiple input views with camera matrices \( R \), we first predict a coarse result as the initialization from a randomly selected input image by using [17]. Then, we extend the SMPLify [4] with tailoring a multi-view joints reprojection error into the optimization:

\[
E_{proj} = \sum_{n=1}^{N} \sum_{k=1}^{K} \omega_{n,k} \rho(sR_nJ_n,k(\theta, \beta) + t) - P_{n,k})
\]

(S1)

where \( P = \{P_k|k = 1, ..., K\} \) is the estimated 2D joints locations from OpenPose [6], and \( \omega_{n,k} \) is the corresponding confidence weight of joint \( k \) in image \( I_n \). \( J_k \) is the 3D joint \( k \) induced by \( \theta, \beta \) in homogeneous coordinate. \( s \) and \( t \) are scale factor and global translation of the SMPLx model, finally \( \rho \) is the robust Geman-McClure penalty function.

Only minimizing the reprojection error may lead to misfit body shape due to the loose body shape constraint from joints location. Thus, we further define the intersection over union (IoU) error from the projected region of SMPLx and the mask \( M = \{M_n|n = 1, ..., N\} \) of image \( I \), to encourage the inferred model fitting tighter to the given shape.

\[
E_{IoU} = \sum_{n=1}^{N} IoU(M_n, \Gamma(M(\theta, \beta); K, R_n))
\]

(S2)

where \( \Gamma \) is the differentiable rendering function [18] given the model mesh and output the 2D mask using perspective projection, and \( IoU(A, B) = |A \cap B|/|A \cup B| \).

The overall energy function can be written as

\[
E_{total} = \lambda_{proj} E_{proj} + \lambda_{IoU} E_{IoU} + \lambda_{\theta} E_{\theta}(\theta) + \lambda_{\beta} E_{\beta}(\beta)
\]

(S3)

\( \{\lambda_{proj}, \lambda_{IoU}, \lambda_{\theta}, \lambda_{\beta}\} \) are trade-off parameters. \( E_{\theta}(\theta), E_{\alpha}, E_{\beta} \) are regularization terms defined in [20] preventing impossible pose and shape. When the groundtruth camera poses are given, we optimize the SMPLx parameters via

\[
\theta^*, \beta^*, s^*, t^* = \text{arg min}_{\theta, \beta, s, t} E_{total}(\theta, \beta, R)
\]

we adopt the Adam gradient descent method to solve the optimal solution until it converges to a predefined threshold.

B.2. Closest Point Searching

As mentioned in the main paper, given a mesh of body model \( M \) parametrized by body shape \( \beta \) and pose \( \theta \), we expect to find a points \( v \) on the mesh surface, such that it has the least distance to query point \( x \). This problem is identical to least distance problem to single triangular face, which can be formulated as a quadratic programming (QP) problem with equality and inequality constraints

\[
\min_{c_1,c_2,c_3} \| \sum_{i=1}^{3} c_i v_i - x \|_2
\]

(S5)

\[s.t \quad c_1, c_2, c_3 \geq 0; \quad c_1 + c_2 + c_3 = 1\]

where \( (v_1, v_2, v_3) \) are vertices who construct a triangular face. This QP problem with inequality constraint can be optimized via active set method [29]. Note that this optimization process is numerical efficient, only \( 4 \times 4 \) matrix inversion is required in maximum 5 steps. Depending on optimal active set \( A \), there exist three types of optimal points:

- When \( \text{size}(A) = 0 \) or \( A \) is empty set, the optimal nearest point lies inside the triangular;
- When \( \text{size}(A) = 1 \), the optimal point lies on one of the edges of the triangular;
- When \( \text{size}(A) = 2 \), the optimal point lies on one of the vertices of the triangular;
**B.3. Gradient of Signed Distance Function**

Our implicit body shape embedding is constructed on signed distance function (SDF) and its gradient $S'$. Recall that the SDF is defined as

$$S(x, M) = \text{sign}(x, M) ||x - v||^2, \quad (S6)$$

$$\text{sign}(x, M) = \begin{cases} 
1 & \text{if } x \text{ inside } M, \\
-1 & \text{if } x \text{ outside } M.
\end{cases} \quad (S7)$$

where $v$ indicates the closest point of $x$.

The $S'$ can be written as

$$S'(x, M) = \text{sign}(x, M) \frac{(x - v)}{||x - v||^2} \quad (S8)$$

Note that $S$ is not strictly differentiable when multiple nearest point can be found, in practice we choose the first indexed nearest point to calculate gradient. Eqn. S8 satisfies eikonal equation, where the gradient of SDF is a vector with norm equals to 1. Although the defined SDF’s gradient of a mesh is not a inward normal vector field due to the discretized surface, the gradient keeps smooth property, as illustrated in Fig. S2 (a). It is piece-wise constant in the prism of a triangular face, and piece-wise smooth in the space between two adjacent prisms.

**B.4. Grid-based Hierarchical Searching**

When computing Eqn. S6, naively searching from the full set of triangular faces is computational inefficient. Therefore, we utilize a grid based hierarchical searching method to narrow down the searching space. Given an input mesh $M$ of parametric body model, we first discretize the volume into voxel grids $V = \{V_i | i = 1, \ldots, m \times n \times k\}$. Integers $m, n, k$ here are resolution in three axes. Related faces which lie inside or cross each voxel are memorized as $P(V_i)$, as shown in Fig. S2 (b). Then, we perform basic mesh operations, e.g., nearest point searching, ray casting, etc, in a hierarchical manner. More specifically, in the coarse stage mesh operations are performed on voxel grids, when voxel $V_i$ is visited, we test all faces in $P(V_i)$ in the fine stage. Take closest point searching as an example, the hierarchical searching algorithm can be summarized as:

To further accelerate the algorithm, we execute the process of Alg. 1 in parallel in modern GPUs. More precisely, each CUDA core optimizes the closest point with respect to one face in current voxel $V_i$. We evaluate the computational efficiency of body embedding methods in PaMIR [54] and GNR in term of execution time. PaMIR requires 0.387s to extract embedding from a 3D convolution network for 64k points, while our method only takes 0.0208s for same number of points. Note that the volume resolution is $127^3$, and execution time and memory resources required for 3D convolution increase cubically if a higher volume resolution is needed.

---

**Figure S1. Multi-view Body Fitting Pipeline.** 2D keypoints from multiple views are detected from OpenPose [6], and the initial SMPLx model is estimated by HMR [17] from one keyframe. Our multi-view fitting algorithm adopts two optimization terms, $E_{proj}$ models the skeleton reprojection error, and $E_{IoU}$ force the final SMPLx to fit tightly with shape masks. The final output is optimized in an iterative manner.

**Figure S2. (Left) Gradient of SDF.** Region in blue is the triangular prism, where gradient is piece-wise constant; Region in orange is the space where gradient is piece-wise smooth. (Right) Voxel Grid of Mesh. Volume is discretized into voxels, faces related to each voxel are indexed.
Algorithm 1: Hierarchical Searching of Closest Point

Initialization: Discretize the volume $V$; Compute the related faces $P(V)$; Rank the non-empty voxels with Manhattan distance $d_m; d_{min} = +\infty; i = 0$;
while $d_{min} > \sqrt{2}d_m$ do
   for face in $P(V)$ do
      Solve Eqn. S5, and get distance $d = ||\mathbf{x} - \mathbf{v}||_2$;
      if $d < d_{min}$ then
         $d_{min} = d$;
      end
   end
   $i = i + 1$;
end

C. Implementation Details

C.1. Network Architecture

We describe detailed network architecture of our GNR, which consists of a image encoder $E$, the main implicit function who contains three MLPs $F_1, F_2, F_3$ and predicts volume density $\sigma$ and color $c_0$, and appearance blending network $T$, as illustrated in Fig. S3.

Image Encoder. For image feature encoder $E$, stacked hourglass network is used whose parameters are initialized from the pre-trained geometry model of [36] and outputs a 128-channel feature map with downsampling resolution. To resolve the overly smoothed effect due to low-resolution embedding, we then adopt a transposed convolution network to upsmaple feature map and produce our final image embedding in full resolution, during which the resized input image is concatenated after feature channel as an observation reference before transposed convolution at each scale.

Multiview Feature Fusion As shown in Fig. S3, the main part of implicit function consists of three sub-network $F_1, F_2, F_3$, and two pooling operations $\phi, \psi$. Specifically, for volume density prediction $F_\sigma$, we concatenate the point coordinate, SMPLx embedding (discussed in Sec. B), and image features. The concatenated inputs are fed to the MLP $F_1$ with a depth of 3 and width 256, the multi-view feature is aggregated using average pooling $\phi$. The volume density is predicted via a second 5-layer MLP, and query point coordinate with body embedding are inserted in the intermediate layers like [28]. The color prediction module $F_{c_0}$ is a shallow MLP with a depth of 3, inputs with first 3 order spherical harmonics of view direction $d$ and weighted pooled features by $\psi$.

Appearance Blending. The appearance blending network $T$ consists of two MLPs with width 16, the query MLP takes weighted pooled feature with point coordinate and view direction and outputs $Q$; the key MLP maps $K$ with the intermediate features and outgoing direction. The blending weights are normalized through a softmax operator, for simplicity we denoted it as $\kappa$ in Fig. S3, and final color is predicted by camera blending of $N + 1$ cameras.

C.2. Normalized Human Coordinate and Visual Hull Bounding

Normalized Human Coordinate. To unify human body in different shapes, poses, orientations or even different setups into the same capture volume, we adopt a normalized SMPL coordinate within a range of $[-1,1]^3$, and rotated by SMPL’s global orientation. The query points and view directions are transformed into normalized SMPL coordinates before feeding to the network.

Visual Hull Bounding. Ray casting strategy in conventional NeRF is highly inefficient because a large amount of the samples fall in the blank space even if the searching space is bounded by a near body cylinder [9]. We use the visual hull [22] to effectively bound the sampling space. First, binary masks are extracted from source images, and the masks are then expanded via a morphological dilation operation. Samples are then back-projected to the image planes; only points that fall within all expanded masks are treated as valid samples. Invalid points are assigned with zero volume density and zero radiance during ray integration. We also randomly sample a small number of points from the free space in each sample batch to regularize the volume density in the exteriors of the visual hull. This sampling strategy dramatically improves the training efficiency and inference speed compared to the conventional method.

C.3. Runtime and Hardware

We train and test GNR on 8 NVIDIA V100 GPU in parallel using PyTorch [31]. In each training iteration, we randomly samples 4096 pixels with 256 sample points on each camera ray. The network is trained in an averaged 7.11 iterations per second rate using multi-process distributed data parallel. The network converges at 20th epoch on RenderPeople, and 3rd epoch on fine-tuning ZJUMoCap and GeneBody. For testing, GNR renders $512 \times 512$ images at a rate of averaging 6.25 fps.

D. Additional Experiments

D.1. More Quantitative Results On Real-World Captured Datasets

Besides the results analyzed in the main paper, we demonstrate and discuss our GNR with additional experimental settings on GeneBody-1.0 and ZJU-Mocap i.e., zero-shot testing and in-domain testing.

Quantitative Results under Zero-shot Testing on GeneBody and ZJU-Mocap. We add an experiment setting additionally on real-world datasets called zero-shot testing, namely GNR and all compared methods are pretrained on RenderPeople and tested on unseen subjects in GeneBody without finetuning. This setting is designed to show the generalization ability under an even harder scenario where the domain gap between RenderPeople and GeneBody would hinder the fidelity of the view synthesis. Tab. S1 shows that
Implicit Geometric Embedding

Query point coordinate $x$, its implicit body shape embedding $G$ and reprojected image feature from $i$th view $f_i$ are input to $F_1$ for intermediate feature embedding; Intermediate features $F_{1,i}$ from all views are aggregated via an average pooling operator $\phi$ are then feed to occupancy prediction network $F_2$; $F_{1,i}$ are concatenated with observed color and sent to color prediction network $F_3$ after weighted pooling $\psi$. $x$, $G$, and $d$ are injected to intermediate layers of $F_2$ and $F_3$ respectively. The final color is fused by learned camera blending from appearance blending network $T$.

Figure S3. **Network Architecture Details and Data Flow.** Query point coordinate $x$, its implicit body shape embedding $G$ and reprojected image feature from $i$th view $f_i$ are input to $F_1$ for intermediate feature embedding; Intermediate features $F_{1,i}$ from all views are aggregated via an average pooling operator $\phi$ are then feed to occupancy prediction network $F_2$; $F_{1,i}$ are concatenated with observed color and sent to color prediction network $F_3$ after weighted pooling $\psi$. $x$, $G$, and $d$ are injected to intermediate layers of $F_2$ and $F_3$ respectively. The final color is fused by learned camera blending from appearance blending network $T$.

Figure S4. **Qualitative Results on ZJU MoCap.** We compare our methods with case-specific methods NV [24], NT [39], NHR [46] and NB [33] on this dataset. Images are from unseen camera poses of sequences in the test-set of ZJU Mocap. Our model is under the setting of training on RenderPeople and fine-tuning on the subset of ZJU MoCap. These two performers are unseen ID and meanwhile unseen pose to our model.

Quantitative Results under In-Domain Testing. We exhibit more results with two settings that under In-Domain Testing, as shown in Tab S2. Specifically, the first setting (denoted as † in Table S2) is case-specific training but testing on unseen poses, where generalizable methods such as IBRNet [42] and ours are trained following same protocol as case-specific methods such as [24, 33, 39, 46]. This setting demonstrates the effectiveness of GNR of modeling the human geometry and appearance under a rather fair protocol comparing to case-specific methods. The second setting (denoted as ‡ in Tab. S2) is in-domain finetuned but testing on unseen ID unseen poses, where both generalizable methods are finetuned on training set of GeneBody-1.0, and tested on the listed 10 test sequences. This setting examines the capability of generalization in the same domain. From Tab. S2, GNR performs better under both settings, and is proved to benefit more than [42] from in domain finetuning.

D.2. More Qualitative Results On ZJU-MoCap

We present qualitative results of our model compared with case-specific methods in the test-set of ZJU MoCap.

| Method        | GenBody PSNR↑ | GenBody SSIM↑ | GenBody LPIPS↓ | ZJU-Mocap PSNR↑ | ZJU-Mocap SSIM↑ | ZJU-Mocap LPIPS↓ |
|---------------|---------------|---------------|---------------|----------------|----------------|-----------------|
| pixelNerf [49] | 20.40         | 0.87          | 0.37          | 24.84          | 0.90          | 0.13            |
| IBRNet [42]   | 19.67         | 0.85          | 0.24          | 23.96          | 0.89          | 0.17            |
| GNR           | 22.44         | 0.89          | 0.14          | 24.96          | 0.90          | 0.11            |

Table S1. **Quantitative Results of Zero-shot Testing on GeneBody and ZJU-Mocap Datasets.** We evaluate generalizable methods on the setting of pretraining on RenderPeople dataset, without any finetuning on either real-world dataset.
Table S2. Result on GeneBody-1.0 in terms of PSNR and SSIM. We compare our method with case-specific in two settings. † denotes \textit{SeenID and UnseenPose} without any pretrain, ‡ means \textit{UnseenID and UnseenPose} trained on GeneBody training test.

|        | PSNR↑ | SSIM↑ | LPIPS↓ |
|--------|-------|-------|--------|
| NV [24] | 24.47 | 0.88  | 0.193  |
| NHR [46]| 23.39 | 0.87  | 0.191  |
| NB [33] | 23.00 | 0.90  | 0.186  |
| GNR     | 28.77 | 0.94  | 0.068  |

Table S3. Quantitative results on V-Sense.

Dataset, as shown in the Fig S4. NV [24] trends to esimate wrong affine warping when human pose differs largely from seen poses. NT [39] and NHR [46] use UV and point cloud as geometry proxy, while such proxy can not model the non-rigid deformation from human body, leading blurry texture. Our method achieve comparable visual result on total un-seen pose with optimized NB [33] model.

D.3. More Quantitative and Qualitative Results On V-Sense

We compare our method to case-specific methods such as [24, 33, 46] on V-Sense dataset, and provide quantitative comparison in this subsection. The experiment protocol follows that of Tab. 2 in the main paper. As shown in Tab. S3, our method perform better than other case-specific methods. The qualitative comparison are shown in Fig. S5. GNR preserves better geometry fidelity comparing to methods such as [33, 46] and high frequency texture than [24].

Figure S5. Qualitative Results on V-Sense. We compare our methods with case-specific methods NV [24], NHR [46], NB [33].