GenLayNeRF: Generalizable Layered Representations with 3D Model Alignment for Human View Synthesis

Youssef Abdelkareem[1], Shady Shehata[2], and Fakhri Karray[1,2]

[1] University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada. `yafathi,karray@uwaterloo.ca`

[2] Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, United Arab Emirates. `shady.shehata,fakhri.karray@mbzuai.ac.ae`

Abstract. Novel view synthesis (NVS) of multi-human scenes imposes challenges due to the complex inter-human occlusions. Layered representations handle the complexities by dividing the scene into multi-layered radiance fields, however, they are mainly constrained to per-scene optimization making them inefficient. Generalizable human view synthesis methods combine the pre-fitted 3D human meshes with image features to reach generalization, yet they are mainly designed to operate on single-human scenes. Another drawback is the reliance on multi-step optimization techniques for parametric pre-fitting of the 3D body models that suffer from misalignment with the images in sparse view settings causing hallucinations in synthesized views. In this work, we propose, GenLayNeRF, a generalizable layered scene representation for free-viewpoint rendering of multiple human subjects which requires no per-scene optimization and very sparse views as input. We divide the scene into multi-human layers anchored by the 3D body meshes. We then ensure pixel-level alignment of the body models with the input views through a novel end-to-end trainable module that carries out iterative parametric correction coupled with multi-view feature fusion to produce aligned 3D models. For NVS, we extract point-wise image-aligned and human-anchored features which are correlated and fused using self-attention and cross-attention modules. We augment low-level RGB values into the features with an attention-based RGB fusion module. To evaluate our approach, we construct two multi-human view synthesis datasets; DeepMultiSyn and ZJU-MultiHuman. The results indicate that our proposed approach outperforms generalizable and non-human per-scene NeRF methods while performing at par with layered per-scene methods without test time optimization.

Keywords: Novel View Synthesis · NeRF · Multi-Human · Mesh Alignment

1 Introduction

Novel view synthesis (NVS) of scenes with human subjects has numerous applications in telepresence, virtual reality, etc. The extensions [6, 15, 26, 37] of the well-known NeRF [21] architecture achieved competitive synthesis results using sparse views, yet suffered with human subjects due to their complex motions. NeuralBody [25] anchored NeRF with pre-fitted 3D human models to regularize the training producing more photo-realistic output. A main constraint was the inefficient per-scene optimization requirement. Recently, state-of-the-art human-based synthesis methods [3, 12, 20, 44] merged

the concepts of the human model anchors and the image features to generalize to unseen poses and human identities. However, they were only designed to operate on scenes with single human subjects. Multi-human scenes introduce additional challenges due to how humans occlude each other and the complexity of their close interactions. Layered scene representations [42] are a possible solution to operate in the complex multi-person setting. Shuai et al. [30] utilized a layered architecture by representing the human entities using NeuralBody [25] and weakly supervising the human instance segmentation. Nevertheless, the method suffers from the per-scene optimization problem which hinders its applicability to wider real-world domains. Another issue with existing Human NVS methods [12, 30, 44] is the reliance on multi-step optimization methods [2, 29, 43] for the estimation of pre-fitted 3D body models. Such methods hinder the ability of end-to-end learning and suffer from error accumulation throughout the fitting steps which lead to inaccurate parameter fitting and misaligned body models and consequently hurts the synthesis quality of the novel views.

In this paper, we propose generalizable layered neural radiance fields to achieve free-viewpoint rendering of multi-human subjects, while requiring no test-time optimization for novel subjects or poses. We fuse the concepts of implicit feature aggregation and layered scene representations to synthesize novel views of complex human interactions from very sparse input streams. Specifically, we divide the scene into a set of human layers anchored by the 3D human body meshes. We then introduce a novel end-to-end trainable human-image alignment module that utilizes an iterative feedback loop [41] to correct parametric errors in the pre-fitted human models and produces pixel-aligned human layers for better synthesis quality. For view synthesis, we extract a set of point-wise image-aligned and human-anchored features for all views and effectively aggregate them using self-attention and cross-attention modules. We also include an RGB fusion module that embeds the fused features with low-level pixel information from the images for retaining high-frequency details.

Our main contributions are summarized as follows:

- We propose a generalizable layered representation with a novel combination of three attention-based feature fusion modules for free-viewpoint rendering of multi-human scenes from sparse input views while operating on novel human subjects and poses.
- We present a novel human-image alignment module that corrects misalignment errors in the pre-fitted human models through an end-to-end trainable iterative feedback loop coupled with multi-view self-attention feature fusion.
- We surpass state-of-the-art generalizable and non-human per-scene NeRF methods while performing at par with the multi-human per-scene methods without requiring long per-scene training procedures.

2   Related Work

2.1   Neural View Synthesis

Recent progress has been made in utilizing neural networks along with differentiable rendering for novel view synthesis [1, 5, 13, 32, 33, 36, 38]. NeRF [21] encapsulated the

full continuous 5D radiance field of scenes inside a Multi-Layer Perceptron (MLP). They achieved photo-realistic results but failed to work on highly deformable scenes with non-static subjects. Deformable NeRF methods [23, 26] modeled the dynamic subjects by training a deformation network that transforms 3D points to a canonical space before querying the MLP. Yet, they show poor synthesis quality for human subjects with complex deformations. NeuralBody [25] anchored NeRF with a deformable human model [18] to provide a prior over the human body shape and correctly render self-occluded regions. However, they lacked generalization capabilities for novel scenes. Per-scene optimization NeRF methods [21,25,25,26,30] need to be trained from scratch on each scene which is often impractical due to the large time and computational costs. Generalizable NeRF methods [34, 35, 39] offer a solution by conditioning NeRF on pixel-aligned features generated from the input images which enhanced the results for unseen scenes with sparse input views. Recently, NHP [12] combined the 3D human mesh with image features to accurately represent complex body dynamics and generalize to novel human subjects and poses. HumanNeRF [44] enhanced the quality through efficient fine-tuning procedures and neural appearance blending techniques. However, the blending module operates on pre-scanned synthetic data with accurate depth maps and cannot be extended to real-world data. One limitation of state-of-the-art generalizable human methods [3, 12, 44] lies in the inability to be extended to multi-human scenes which are challenging due to the inter-human occlusions and interactions.

Layered scene representations [19] were proposed to handle complex scenes with multiple human subjects. ST-NeRF [42] modeled each human layer using a deformable model similar to D-NeRF [26] to achieve editable free-viewpoint rendering. Recently, Shuai et al. [30] extended ST-NeRF by modeling the human subjects using Neural-Body [25] and predicted human segmentation masks as part of the network training. The restriction of both methods is requiring per-scene training procedures for learning, yielding them inefficient to use. We tackle the existing research gap by proposing a generalizable layered scene representation for synthesizing novel views of multi-human subjects through a combination of image features and layered neural radiance fields. We achieve free-viewpoint rendering for scenes with an arbitrary number of humans from very sparse input views, while generalizing to novel subjects and poses at test time without extra optimization.

2.2   Human Mesh Recovery

Mesh Recovery of human subjects has grabbed significant research attention due to its adoption in 3D geometry reconstruction and novel view synthesis. One direction of approaches solves the task through a multi-step optimization process which fits the parametric human models (i.e. SMPL [18]) based on 2D observations such as keypoints or silhouettes [8,31]. Bogo et al. [2] utilized 2D joint predictions from monocular input to guide the SMPL fitting process for single-human scenes. Zhang et al. [43] tackled a more challenging multi-person setting by leveraging triangulated 3D keypoints and a two-step parametric fitting process for enhanced results. The main issues with multi-step methods are breaking the end-to-end learning and the error accumulation throughout the steps, especially in sparse-view datasets. Specifically, 2D keypoints predictions

could suffer from inaccurate joints in certain views which hurts the triangulation process leading to low-quality 3D keypoint predictions. The parametric model fitting is subject to errors due to the abundance of hyperparameters [43] that require meticulous finetuning and the accumulated errors from the previous steps. On the other hand, regression-based approaches aim for better human-image alignment by directly regressing the body models from input images [10, 11, 16, 17, 40, 41]. PyMAF [41] introduced a feedback loop with multi-scale contexts to correct parametric deviations for producing highly aligned meshes from monocular input images for single-humans.

Existing Human NVS approaches [3,12,20,25,44] utilize pre-fitted 3D observations computed using multi-step optimization approaches [29, 43]. However, in sparse-view settings, the pre-fitted predictions suffer from misalignment errors that consequently hurt the quality of the synthesized views. Mihajlovi et al. [20] utilized 3D keypoints instead of body models to avoid parametric fitting errors. L-NeRF [30] introduced a time-synchronization step that accounts for the multi-view image de-synchronization by producing a per-view body model using predicted time offsets. However, they do not account for parametric errors occurring in the multi-step fitting process. In this work, we propose a novel regression-based human-image alignment module that ensures the correction of parametric errors leading to aligned body models with multi-view input.

3   Methodology

3.1   Problem Definition

Given a synchronized set Ω of frames I taken from B sparse input viewpoints of a scene with N arbitrary number of humans, such that $\Omega = \{I_1, .., I_B\}$, our target is to synthesize a novel view frame $\{I_q\}$ of the scene from a query viewing direction q. Each input viewpoint b is represented by the corresponding camera intrinsics K, and camera rotation R and translation t, where $b = \{K_b, [R_b|t_b]\}$. The N pre-fitted 3D human body meshes are given for each input frame. Each human h is represented using the SMPL [18] model which is a deformable skinned model defined in terms of pose and shape parameters Θ_h^0 while also being vertex-based where each model s_h consists of 6,480 vertices, such that $s_h \in \mathbf{R}^{6,480 \times 3}$. For an input view image $I_b \in \mathbf{R}^{H \times W \times 3}$ with height H and width W, we extract a multi-scale feature pyramid $I'_{b,\{0:T-1\}}$ with T levels using a ResNet34 [9] backbone network f, pre-trained on ImageNet, such that $I'_{b,\{0:T-1\}} = f(I_b)$. The operation is carried out for all input views b in $\{1, .., B\}$. A full overview of the proposed architecture is shown in Fig. 1 .

3.2   Human-Image Alignment Module

Pre-fitted human body models can suffer from misalignment with the input images due to error accumulation throughout the multi-step fitting process [43], especially in sparse view settings, which causes hallucinations in synthesized views. We propose an alignment module that is end-to-end trainable with our NVS architecture and carries out iterative parametric correction with closed feedback [41] to ensure a better alignment of the SMPL models with the multi-view input images. The module takes the pre-fitted SMPL parameters Θ_h^0 as input and returns the aligned and adjusted parameters
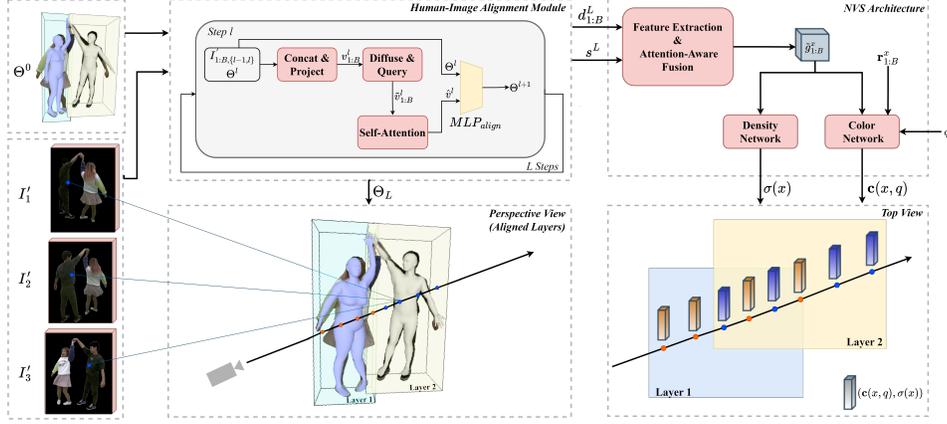
Fig. 1: Overview of the GenLayNeRF approach. We consolidate a layered scene representation where each human subject is modeled using the SMPL model. Regarding our alignment module, at a step l, low and high-resolution feature planes $I'_{1:B,\{l-1,l\}}$ are concatenated and the SMPL vertices are projected on them to produce feature-embedded vertices $v^l_{1:B}$ (Concat & Project). We then diffuse the vertices to continuous spaces and query them at downsampled vertex locations to generate multi-view human features $\tilde{v}^l_{1:B}$ (Diffuse & Query), which are fused using self-attention and passed along with parameters Θ^l to predict the adjusted parameters Θ^{l+1}. In our NVS architecture, we project rays through the aligned scene layers and sample per-layer 3D points within the intersections areas with the layers (shown in the top view). Point-wise features are extracted and fused to output the final fused features $\tilde{g}^x_{1:B}$, which are passed to the density network to predict the volume density $\sigma(x)$, whereas the color network uses the raw RGB values $\mathbf{r}^x_{1:B}$ and q to predict the color $\mathbf{c}(x,q)$.

Θ^L_h. Specifically, we employ an iterative process with L steps, such that, for a step $l > 0$, low-resolution features $I'_{b,l-1}$ from level $l-1$ for view b are upsampled using deconvolution [22] and concatenated with high-resolution feature plane $I'_{b,l}$ at level l resulting in a contextualized and localized feature plane $I''_{b,l}$. Human vertices s^l_h are embedded with image features by projection on the multi-view feature map, such that, $v^l_{h,b} = I''_{b,l}[K_b((R_b s^l_h) + t_b)]$. $v^l_{h,b} \in \mathbf{R}^{6,480 \times C_1}$ represents the features of the vertices projected on feature map $I''_{b,l}$ for human h. The preceding part corresponds to "Concat & Project" in Fig. 1.

Our target is to retrieve a compact and continuous per-human feature representation to be used for parameter adjustment. For that reason, the sparse human vertices $v^l_{h,b}$ need to be diffused into a continuous space that can be queried at any location. We incorporate the SparseConvNet [7,25] architecture which utilizes 3D sparse convolution to diffuse the vertex features into different nearby continuous spaces for every human and view. The diffused vertices are denoted as $d^l_{h,b}$. To obtain the per-human features, we downsample the vertices s^l_h, such that $\tilde{s}^l_h \in \mathbf{R}^{431 \times 3}$, and query the diffused vertex

spaces at the downsampled locations to obtain the multi-view per-human vertex features which are then processed and flattened to obtain a compact version denoted as $\tilde{v}_{h,b}^l \in \mathbf{R}^{1 \times C_2}$. The preceding part corresponds to "Diffuse & Query" in Fig. 1. Afterward, we effectively correlate the multi-view human features using a self-attention module, such that,

$$
\begin{aligned}
mv_h^l &= soft(\frac{1}{\sqrt{d_{k_1}}} query(\tilde{v}_{h,1:B}) \cdot key(\tilde{v}_{h,1:B}^l)^T), \\
\hat{v}_{h,1:B}^l &= mv_h^l \cdot val_1(\tilde{v}_{h,1:B}^l) + val_2(\tilde{v}_{h,1:B}^l), \\
mv_h^l &\in \mathbf{R}^{B \times B}, \ \tilde{v}_{h,1:B}^l \in \mathbf{R}^{B \times C_2},
\end{aligned}
\tag{1}
$$

where key, $query$, and (val_1, val_2) represent the key, query, and value embeddings of the corresponding argument features respectively, and d_{k_1} denotes the dimensionality of the key embedding. $soft$ denotes the softmax operation. We carry out view-wise averaging for multi-view fusion on the view-aware human features such that, $\hat{v}_h^l = \frac{1}{B} \sum_b \hat{v}_{h,b}^l$. Lastly, the fused per-human features are concatenated ($\oplus$) with the current SMPL parameters and passed to a correction MLP that predicts parameter alignment offsets $\Delta\Theta_h^l$ which are added to the current parameters, such that,

$$
\begin{aligned}
\Delta\Theta_h^l &= MLP_{align}([\hat{v}_h^l \oplus \Theta_h^l]), \\
\Theta_h^{l+1} &= \Theta_h^l + \Delta\Theta_h^l,
\end{aligned}
\tag{2}
$$

The updated parameters Θ_h^{l+1} are used to retrieve the adjusted SMPL vertices s_h^{l+1} and are passed to the next step $l + 1$. After L steps, the aligned SMPL parameters Θ_h^L, vertices s_h^L, and diffused spaces $d_{h,1:B}^L$ are passed to our layered NVS architecture.

3.3   Layered Scene Representation

Scenes with multiple humans suffer from inter-human occlusions that become evident when subjects closely interact together. A practical solution to handle complex multi-human scenarios is dividing the scene into distinct layers where each layer models an entity using a neural radiance field [19, 42]. Entities can be humans, objects, or background. Our proposed approach focuses mainly on human layers represented using the SMPL [18] model which is responsible for preserving the local geometry and appearance of humans making it possible to model their complex deformations and occluded areas.

Our aim is to render the full novel view image I_q from a query viewpoint q. To achieve that, we first use the camera-to-world projection matrix, defined as $\mathbf{P}^{-1} = [R_q|t_q]^{-1}K_q^{-1}$, to march 3D rays across the multi-layered scene. In practice, we have a ray for each pixel p in the final image, where the ray origin $r_0 \in \mathbf{R}^3$ is the camera center and the ray direction is given as $i = \frac{\mathbf{P}^{-1}p - r_0}{||\mathbf{P}^{-1}p - r_0||}$. 3D points x are sampled across the rays at specific depth values z, where $x = r(z) = r_0 + zi$. Since we have several human layers in the scene, we determine the intersection areas of the rays with the humans using the 3D bounding box around each layer defined by the minimum and maximum vertex points of the aligned SMPL meshes $s_{1:N}^L$. We then sample depth values within the n_p intersecting areas only such that $z \in [[z_{near_1}, z_{far_1}], .., [z_{near_{n_p}}, z_{far_{n_p}}]]$. This

guarantees that the sampled points lie within the areas of the relevant human subjects as clear in the top view shown in Fig. 1.

3.4 Feature Extraction and Attention-Aware Fusion

In our proposed approach, we extract multi-view image features for each query point x and effectively merge them using attention-based fusion modules to derive the needed spatially-aligned feature vectors. This enables us to extrapolate to novel human subjects and poses by learning implicit correlations between the independent human layers.

Image-aligned And Human-anchored Features Image-aligned point-wise features are extracted by projecting the point x on all the feature maps $I_{b,L}''$ to collect the corresponding image-aligned features for each view b denoted as p_b^x. In addition, human-anchored features are beneficial for maintaining the complex geometric structure of the human body by anchoring the network on the available SMPL body priors. Existing layered scene representations [30] follow the approach of NeuralBody [25] by encoding the vertices of human layers using learnable embeddings that are unique to each layer in each training scene. In our approach, we utilize the vertices $v_{1:N,1:B}^L$ embedded with image features from the alignment module to enable a generalizable approach conditioned on the input images. The radiance field predictor is queried using continuous 3D sampled points. For that reason, we utilize the diffused vertex spaces $d_{h,1:B}^L$ for each human h and transform x to the SMPL coordinate space of its corresponding human layer. Trilinear interpolation is then utilized to retrieve the corresponding human-anchored features g_b^x from the diffused spaces of each view b.

Attention-Aware Feature Fusion To fuse the point-wise feature representations $g_{1:B}^x$, $p_{1:B}^x$ for point x, one strategy is a basic averaging approach [27, 28]. This leads to smoother output and ineffective utilization of the information seen from distinct views. To learn effective cross-view correlations, we employ a self-attention module that attends between all the multi-view human-anchored features $g_{1:B}^x$ where each feature in one view is augmented with the extra features seen from the other views. Each view feature is first concatenated with its corresponding viewing direction d_b'. The formulation is the same as the one shown in Eq. (1). The produced view-aware human-anchored features are denoted as $\hat{g}_{1:B}^x$.

We additionally make use of the rich spatial information in the image-aligned features by carrying out cross-attention from the view-aware human-anchored features to the image-aligned features. The similarity between the multi-view image features and the per-view vertex features is used to re-weigh the image features and embed them with the vertex features. The fused features $\tilde{g}_{1:B}^x$ are calculated with a formulation similar to Eq. (1). The detailed formulation of our cross-attention and self-attention modules is shown in the supplementary material. Afterward, we carry out view-wise averaging, such that $\tilde{g}^x = \frac{1}{B} \sum_b \tilde{g}_b^x$, to generate the final fused feature representation for x.

3.5   Radiance Field Predictor

Color Network. To predict the color $\mathbf{c}$ of point x, we use the query viewing direction q to model the view-dependent effects [21]. In addition, we explicitly augment the high-level features with low-level pixel-wise information to leverage the high-frequency details in the images. This has been achieved with an RGB fusion module which concatenates the high-level features with the encoded raw RGB values $\mathbf{r}_b^x$ for each view b. RGB values from closer input views are assigned higher weights by cross-attending q with the input viewing directions $d'_{1:B}$ such that,

$$
\begin{aligned}
\tilde{c}^x &= MLP_{c_1}(\tilde{g}_{1:B}^x; \gamma(q); p_{1:B}^x), \\
\hat{c}_{1:B}^x &= \{[\tilde{c}^x \oplus \gamma(\mathbf{r}_1^x)], ..., [\tilde{c}^x \oplus \gamma(\mathbf{r}_B^x)]\}, \\
rgb_{att}^x &= soft(\frac{1}{\sqrt{d_{k_2}}} query(q) \, . \, key(d'_{1:B})^T), \\
\mathbf{c}(x,q) &= MLP_{c_2}(rgb_{att}^x \, . \, val_1(\hat{c}_{1:B}^x)), \\
rgb_{att}^x &\in \mathbf{R}^{1 \times B},
\end{aligned}
\tag{3}
$$

Density Network. We predict volume density $\sigma(x)$ for point x using the fused feature $\tilde{g}^x$, such that, $\sigma(x) = MLP_\sigma(\tilde{g}^x)$.

MLP_σ, MLP_{c_1}, and MLP_{c_2} consist of fully connected layers described in the supplementary material. $\gamma : \mathbf{R}^3 \to \mathbf{R}^{(6 \times l)+3}$ denotes a positional encoding [21] with $2 \times l$ basis functions and d_{k_2} is set to 16.

3.6   Layered Volumteric Rendering and Loss Functions

Layered volumetric rendering is used to accumulate the predicted RGB and density for all points across human layers. The points in intersecting areas n_p the layers are sorted based on their depth value z before accumulation. The detailed formulation is shown in the supplementary material. Given a ground truth novel view image I_q^{gt}, all network weights are supervised using the L2 Norm ($||.||$) photo-metric loss. In addition, we include two losses to explicitly supervise the training of our alignment module weights. Given a set of pseudo ground truth 2D keypoints J^{gt}, we derive the predicted 2D keypoints $\tilde{J}$ from the adjusted vertices s^L following PyMAF [41] and minimize the keypoint difference weighted by the ground truth confidence of each body joint. We also include a regularization term on the SMPL parameters to avoid large parametric deviations. The final loss function for our network is written as,

$$
\mathbf{L} = \lambda_{ph}||I_q^{gt} - I_q|| + \lambda_{kpts}||J^{gt} - \tilde{J}|| + \lambda_{reg}||\Theta^L||,
\tag{4}
$$

4   Experiments

In this section, we introduce the datasets, baselines, experimental results, and ablation studies. Details about our training procedure are in the supplementary material.

Table 1: Comparison with generalizable and per-scene NeRF methods on the DeepMultiSyn and ZJU-MultiHuman Datasets. "G" and "S" denote generalizable and per-scene methods, respectively. "*" refers to human-based methods. PSNR and SSIM metric values are the greater the better. "ft" refers to finetuning.

	Method	DeepMultiSyn		ZJUMultiHuman			Method	DeepMultiSyn		ZJUMultiHuman	
		PSNR	SSIM	PSNR	SSIM			PSNR	SSIM	PSNR	SSIM
(a) Seen Models, Seen Poses							*(b) Seen Models, Unseen Poses*				
S	NeRF	15.49	0.497	16.42	0.525	S	L-NeRF*	22.12	0.825	23.02	0.871
	D-NeRF	17.08	0.702	18.53	0.748		PixelNeRF	14.14	0.520	16.88	0.560
	L-NeRF*	24.04	0.858	25.10	0.903		SRF	18.07	0.663	17.93	0.680
	Ours$_{ft}$	**25.05**	**0.889**	**25.21**	**0.916**	G	IBRNet	18.01	0.710	19.84	0.772
	PixelNeRF	14.81	0.534	19.74	0.629		NHP*	20.26	0.677	20.64	0.791
	SRF	20.39	0.724	17.87	0.657		**Ours**	**23.45**	**0.862**	**23.76**	**0.882**
G	IBRNet	19.45	0.741	20.03	0.766	*(c) Unseen Models, Unseen Poses*					
	NHP*	20.91	0.698	21.75	0.813		PixelNeRF	13.12	0.457		
	Ours	**24.01**	**0.859**	**25.02**	**0.901**		SRF	13.95	0.548		
						G	IBRNet	18.80	0.672	Not Applicable	
							NHP*	19.51	0.678		
							Ours	**21.03**	**0.802**		

4.1   Datasets

The existence of readily-available open-source multi-human view synthesis datasets is limited. To solve this challenge, we construct two new datasets, ZJU-MultiHuman and DeepMultiSyn. Both datasets will be published to be used by multi-human view synthesis methods. We also include a subset of the single-human ZJU-MoCap dataset for diversity. Extra details on the datasets are included in the supplementary material.

DeepMultiSyn. The DeepMultiSyn dataset is an adaptation of the 3D reconstruction dataset published by DeepMultiCap [45]. We take the raw real-world multi-view sequences and process them for novel view synthesis. There exist 3 video sequences of scenes containing 2 to 3 human subjects captured from 6 synchronized cameras. Following NeuralBody [25], we use EasyMoCap [29] to fit the SMPL human models for all the subjects in the available frames. Additionally, we predict the human segmentation masks following [14] to separate the humans from the background. This dataset is considered challenging due to the existence of close interactions and complex human actions such as boxing, and dancing activities.

ZJU-MultiHuman. The ZJU-MultiHuman dataset consists of one video sequence with 600 frames taken from 8 uniformly distributed synchronized cameras. The video sequence was published online [29] with the calibration files. The captured scene contains 4 different human subjects. Similar to DeepMultiSyn, we predict the SMPL models and segmentation masks utilizing [14, 29].

4.2   Baselines

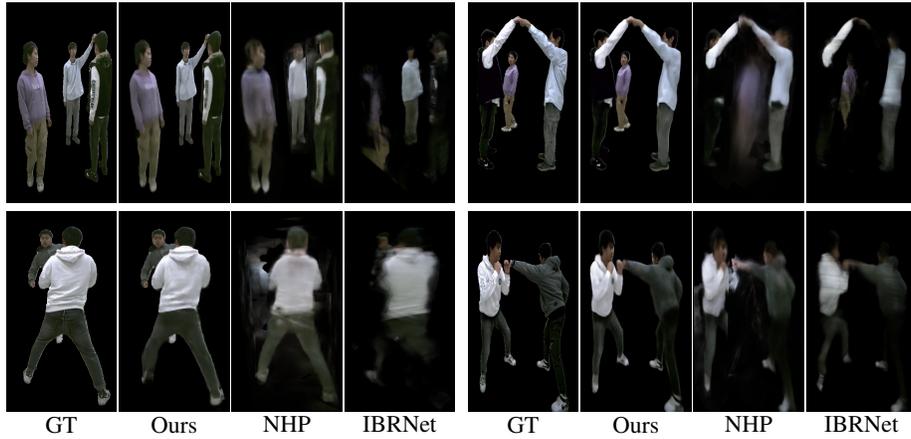We compare our proposed approach with generalizable and per-scene NeRF methods.

Fig. 2: Comparison with generalizable methods on **seen models/unseen poses** [top row] and **unseen models/unseen poses** [bottom row] for the DeepMultiSyn Dataset.
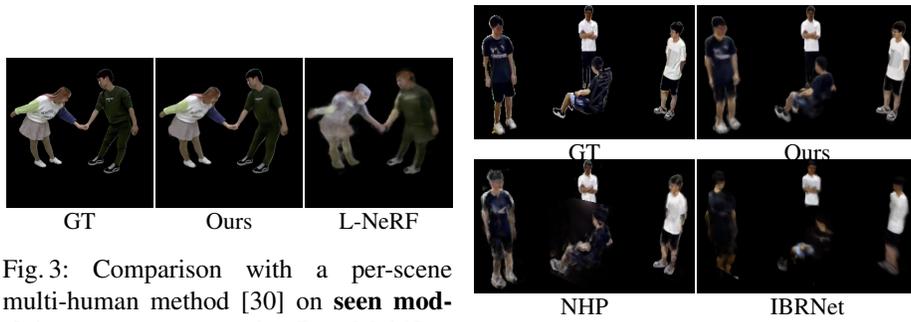

Fig. 3: Comparison with a per-scene multi-human method [30] on **seen models/unseen poses** on the DeepMultiSyn Dataset.
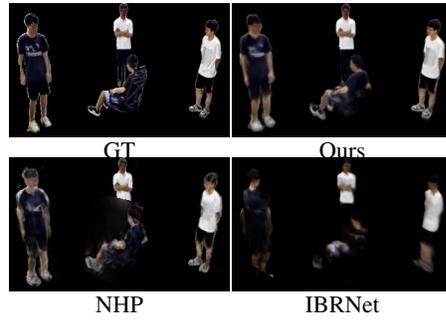

Fig. 4: Comparison with generalizable methods on **seen models/unseen poses** for the ZJU-MultiHuman Dataset.

Comparison with generalizable NeRF methods. Generalizable human-based NeRF methods [3, 12, 20, 44] operate only on scenes with single humans. We choose to compare against NHP [12] after adjusting it to work on multi-human scenes by using the segmentation masks to render a separate image for each individual in the scene. We then superimpose the human images based on their depth to render the novel view image. Regarding non-human methods, PixelNeRF [39] is the first to condition NeRF on pixel-aligned features for generalization. IBRNet [35] and SRF [4] additionally utilize image-based rendering and stereo correspondences, respectively, to achieve generalizable properties. All methods were trained on all human scenes simultaneously.

Comparison with per-scene methods. We evaluate our performance compared to the multi-human layered scene representation approach [30], denoted as L-NeRF. We also compare against D-NeRF [26] and the original NeRF [21] method. All of the mentioned approaches are trained on each scene separately with the same train-test splits.
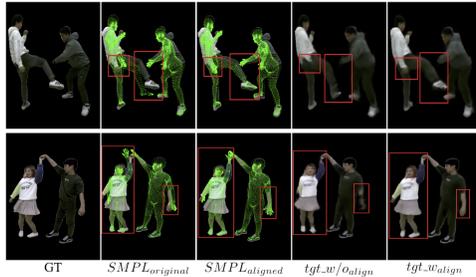
GT $SMPL_{original}$ $SMPL_{aligned}$ tgt_w/o_{align} tgt_w_{align}

Fig. 5: Visualization of the output of our human-image alignment module ($SMPL_{aligned}$) given the misaligned pre-fitted model ($SMPL_{original}$). "tgt_w_{align}" and "tgt_w/o_{align}" denote the rendered image with/without our alignment module.

Table 2: Ablation study results on **seen models** and **unseen poses** for the Deep-MultiSyn dataset. "# V." denotes the number of views.

crs	slf	rgb	align	V.	PSNR ↑	SSIM ↑
				3	20.92	0.7860
✓				3	21.45	0.8005
✓	✓			3	21.98	0.8093
✓	✓	✓		3	22.19	0.8260
✓	✓	✓	✓	3	**23.45**	**0.8620**
✓	✓	✓	✓	1	21.98	0.8091
✓	✓	✓	✓	2	22.32	0.8379
✓	✓	✓	✓	4	23.72	0.8711

4.3   Experimental Results

Our evaluation spans three generalization settings as follows:

Seen Models, Seen Poses. In this setting, we test on the same human subjects and poses that the model is trained on. Tab. 1a indicates the results in terms of the per-scene and generalizable baselines. Regarding the generalizable approaches, our method exhibits the best overall performance on both datasets on all metrics. For the per-scene approaches, our proposed method performs at par with the state-of-the-art per-scene baseline (L-NeRF), while effectively saving computational and time resources by taking 50 hours to converge on all the scenes simultaneously compared to 144 hours for per-scene training. After per-scene finetuning, our method surpasses L-NeRF on both datasets. Qualitative comparisons for the per-scene methods are included in the supplementary material.

Pose Generalization. We additionally test all approaches on the same human subjects seen during training, but with novel poses. On both datasets, Tab. 1b shows that our approach outperforms all the generalizable NeRF methods on all metrics. L-NeRF lags behind our method on the DeepMultiSyn dataset due to the complex novel poses which validates the pose generalization ability of our method on challenging motions. In Fig. 2 and Fig. 4, IBRNet fails to model the full body of the human subjects properly, while NHP fails to represent areas of occlusions where subjects highly overlap. However, our method successfully models the body shapes and can handle overlapping areas which validates the effectiveness of the layered scene representation in the generalizable multi-human setting. Fig. 3 shows how L-NeRF fails to properly render the appearance of subjects when presented with complex unseen poses.

Human Generalization. A challenging setting would be testing on human subjects and poses not seen during training. This was done on the DeepMultiSyn dataset by leaving out two different scenes for testing. Tab. 1c validates that our method has the best generalization capability as it outperforms all other methods by a large margin. The bottom row of Fig. 2 shows that our method better represents the main body features of the

novel human subjects. IBRNet fails to render some body parts like the legs, while NHP suffers from more blur artifacts, especially in overlapping areas. In the supplementary material, we show that our method surpasses NHP by a large margin even on single humans in the ZJU-MoCap dataset for both pose and human generalization settings.

4.4 Ablation Studies

Effect of Human-Image Alignment. We evaluate the impact of the proposed human-image alignment module on the synthesis quality. Quantitatively, Tab. 2 shows the superior enhancement offered by the alignment module (align) on both metrics. In Fig. 5, we demonstrate the large misalignment between the pre-fitted SMPL model and the image which caused severe hallucinations in the synthesized image (areas with red boxes). Our module successfully aligns the SMPL model with the images leading to higher-quality synthesis results. We include additional results of our module in the supplementary material.

Effect of Fusion modules. We assess the effect of different fusion modules on the synthesis results. From Tab. 2, the second row uses the cross-attention module (crs) in Sec. 3.4 and it shows a noticeable improvement over doing basic average pooling in the first row. This indicates the effectiveness of the correlation learned between the vertex and image features. The addition of the self-attention module (slf) in Sec. 3.4 in the third row led to the incorporation of multi-view aware features and achieved a slight enhancement on both metrics. The last row adds the raw RGB fusion module (rgb) in the Color Network presented in Sec. 3.5. It enhances the performance, especially on the SSIM metric, validating the importance of utilizing low-level information.

Effect of Number of Views. We evaluate the performance of our proposed approach when given a different number of input views at test time. Tab. 2 indicates that using 4 views leads to an enhancement in both metrics due to the extra information available. Decreasing the number of views gradually degrades the performance. However, using only one input view, our method outperforms all the generalizable NeRF methods in Tab. 1 that use 3 input views.

5 Limitations & Future Work

Several enhancements to our proposed method could be investigated further. As our two proposed datasets were sufficient to show the generalization capability of our method, there is room for improvement by elevating the diversity in terms of the number of scenes, camera views, distinct humans, and complex actions. This would lead to better generalization capabilities on broader challenging scenarios. Furthermore, our method suffers from blur artifacts representing human clothing details such as skirts as seen in Fig. 3. One could experiment with integrating a deformation model [26] to represent small deformations such as textured clothing. In addition, adjustments could be made to allow for human-image alignment for more complex body models such as SMPL-X [24]. Lastly, a research direction could explore the optimization of the body model parameters from scratch with multi-view time synchronization taken into consideration.

6   Conclusion

We introduce a generalizable layered scene representation for free-viewpoint rendering of multi-human scenes using very sparse input views while operating on unseen poses and subjects without test time optimization. We additionally present a novel end-to-end human-image alignment module that corrects parametric errors in the pre-fitted body models leading to pixel-level alignment of human layers with the input images. Regarding view synthesis, we divide the scene into a set of multi-human layers. We then generate point-wise image features and human-anchored features and utilize a combination of cross-attention and self-attention modules that effectively fuse the information seen from different viewpoints. In addition, we introduce an RGB fusion module to embed low-level pixel values into the color prediction for higher-quality results. We assess the efficacy of our approach on two newly proposed multi-human datasets. Experimental results show that our method outperforms state-of-the-art generalizable NeRF methods in different generalization settings and performs at par with layered per-scene methods without long per-scene optimization runs. We also validate the effectiveness of our alignment module by showing its significant enhancement on the synthesis quality. Our module could be integrated with existing SMPL-based synthesis methods to elevate the performance by improving the human-image alignment.

References

1. Aliev, K.A., Ulyanov, D., Lempitsky, V.S.: Neural point-based graphics. ArXiv **abs/1906.08240** (2020)
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. ArXiv **abs/1607.08128** (2016)
3. 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 **abs/2204.11798** (2022)
4. Chibane, J., Bansal, A., Lazova, V., Pons-Moll, G.: Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021)
5. Flynn, J., Broxton, M., Debevec, P.E., DuVall, M., Fyffe, G., Overbeck, R.S., Snavely, N., Tucker, R.: Deepview: View synthesis with learned gradient descent. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2362–2371 (2019)
6. Gao, C., Shih, Y., Lai, W.S., Liang, C.K., Huang, J.B.: Portrait neural radiance fields from a single image. ArXiv **abs/2012.05903** (2020)
7. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 9224–9232 (2018)
8. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. 2009 IEEE 12th International Conference on Computer Vision pp. 1381–1388 (2009)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)

10. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 11107–11117 (2021)
11. Kundu, J.N., Rakesh, M., Jampani, V., Venkatesh, R.M., Babu, R.V.: Appearance consensus driven self-supervised human mesh recovery. ArXiv **abs/2008.01341** (2020)
12. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. In: NeurIPS (2021)
13. Li, J., Feng, Z., She, Q., Ding, H., Wang, C., Lee, G.H.: Mine: Towards continuous depth mpi with nerf for novel view synthesis. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12558–12568 (2021)
14. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020). https://doi.org/10.1109/TPAMI.2020.3048039
15. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Lv, Z.: Neural 3d video synthesis. ArXiv **abs/2103.02597** (2021)
16. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1954–1963 (2020)
17. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12919–12928 (2021)
18. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
19. Lu, E., Cole, F., Dekel, T., Xie, W., Zisserman, A., Salesin, D., Freeman, W.T., Rubinstein, M.: Layered neural rendering for retiming people in video. ACM Transactions on Graphics (TOG) **39**, 1 – 14 (2020)
20. Mihajlović, 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 (2022)
21. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
22. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 1520–1528 (2015)
23. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D., Seitz, S., Martin-Brualla, R.: Deformable neural radiance fields. https://arxiv.org/abs/2011.12948 (2020)
24. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
25. 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. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9050–9059 (2021)
26. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural radiance fields for dynamic scenes. https://arxiv.org/abs/2011.13961 (2020)
27. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2304–2314 (2019)
28. Saito, S., Simon, T., Saragih, J.M., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 81–90 (2020)

29. Shuai, Q., Geng, C., Fang, Q., Peng, S., Shen, W., Zhou, X., Bao, H.: Easymocap - make human motion capture easier. Github (2021), https://github.com/zju3dv/EasyMocap
30. Shuai, Q., Geng, C., Fang, Q., Peng, S., Shen, W., Zhou, X., Bao, H.: Novel view synthesis of human interactions from sparse multi-view videos. ACM SIGGRAPH (2022)
31. Sigal, L., Balan, A.O., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NIPS (2007)
32. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhöfer, M.: Deepvoxels: Learning persistent 3d feature embeddings. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2432–2441 (2019)
33. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. arXiv: Computer Vision and Pattern Recognition (2019)
34. Trevithick, A., Yang, B.: Grf: Learning a general radiance field for 3d scene representation and rendering. ArXiv **abs/2010.04595** (2020)
35. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.A.: Ibrnet: Learning multi-view image-based rendering. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4688–4697 (2021)
36. Wu, M., Wang, Y., Hu, Q., Yu, J.: Multi-view neural human rendering. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1679–1688 (2020)
37. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9416–9426 (2021)
38. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. ArXiv **abs/1612.00814** (2016)
39. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4576–4585 (2021)
40. Zanfir, A., Bazavan, E.G., Xu, H., Freeman, B., Sukthankar, R., Sminchisescu, C.: Weakly supervised 3d human pose and shape reconstruction with normalizing flows. ArXiv **abs/2003.10350** (2020)
41. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 11426–11436 (2021)
42. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Transactions on Graphics **40**, 1 – 18 (2021)
43. Zhang, Y., Li, Z., An, L., Li, M., Yu, T., Liu, Y.: Lightweight multi-person total motion capture using sparse multi-view cameras. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 5540–5549 (2021)
44. Zhao, F., Yang, W., Zhang, J., Lin, P.Y., Zhang, Y., Yu, J., Xu, L.: Humannerf: Generalizable neural human radiance field from sparse inputs. ArXiv **abs/2112.02789** (2021)
45. Zheng, Y., Shao, R., Zhang, Y., Yu, T., Zheng, Z., Dai, Q., Liu, Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. International Conference on Computer Vision (ICCV) pp. 6219–6229 (2021)