TransHuman: A Transformer-based Human Representation for Generalizable Neural Human Rendering

Xiao Pan[1,2,*], Zongxin Yang[1], Jianxin Ma[2], Chang Zhou[2], Yi Yang[1,†]
[1]ReLER Lab, CCAI, Zhejiang University, China
[2]Alibaba DAMO Academy, China
{xiaopan, yangzongxin, yangyics}@zju.edu.cn {majx13fromthu, ericzhou.zc}@alibaba-inc.com

Abstract

In this paper, we focus on the task of generalizable neural human rendering which trains conditional Neural Radiance Fields (NeRF) from multi-view videos of different characters. To handle the dynamic human motion, previous methods have primarily used a SparseConvNet (SPC)-based human representation to process the painted SMPL. However, such SPC-based representation i) optimizes under the volatile observation space which leads to the pose-misalignment between training and inference stages, and ii) lacks the global relationships among human parts that is critical for handling the incomplete painted SMPL. Tackling these issues, we present a brand-new framework named TransHuman, which learns the painted SMPL under the canonical space and captures the global relationships between human parts with transformers. Specifically, TransHuman is mainly composed of Transformer-based Human Encoding (TransHE), Deformable Partial Radiance Fields (DPaRF), and Fine-grained Detail Integration (FDI). TransHE first processes the painted SMPL under the canonical space via transformers for capturing the global relationships between human parts. Then, DPaRF binds each output token with a deformable radiance field for encoding the query point under the observation space. Finally, the FDI is employed to further integrate fine-grained information from reference images. Extensive experiments on ZJU-MoCap and H36M show that our TransHuman achieves a significantly new state-of-the-art performance with high efficiency. Project page: https://pansanity666.github.io/TransHuman/

1. Introduction

Rendering free-viewpoint videos of dynamic human performers in high fidelity is vital for many applications such as mixed reality, gaming, and telepresence. Recent
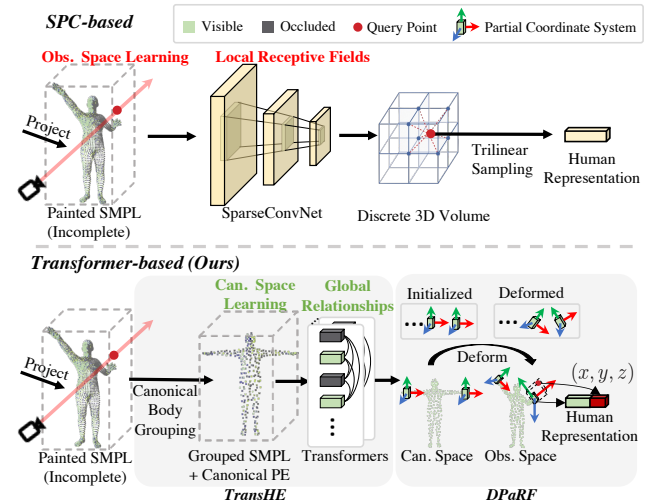

Figure 1. **Comparisons between existing SPC-based and our transformer-based human representations.** Given the incomplete painted SMPL, the SPC-based one optimizes under the varying observation space with limited receptive fields from 3D convolution. Instead, our transformer-based one optimizes under the canonical space with global relationships between human parts.

works [29, 28, 39, 33] integrate the Neural Radiance Fields (NeRF) [27] technology with parametric human prior models (*e.g.*, SMPL [24]) for handling the dynamic human body and achieve fair novel view synthesis results. However, the tedious per-subject optimization and the requirement of dense training views largely hinder the application of such methods. Targeting these issues and inspired by the recent success of generalizable NeRF [42, 4, 37] on static scenes, the task of generalizable neural human rendering is proposed [18], which trains conditional NeRF across multi-view human videos, and can generalize to a new subject in a single feed-forward manner given sparse reference views as input.

Previous methods for generalizable neural human rendering [5, 18] mainly employ the SparseConvNet (SPC) [21]-based human representation (upper row of Fig.

---

1) which first project deep features from reference images onto the vertices of fitted SMPL and then diffuse them to nearby regions via SPC. The final representation is achieved via the trilinear sampling in the discrete 3D feature volume. Such SPC-based representation mainly suffers from the following two aspects: (i) *Volatile observation learning.* The SPC-based one optimizes under the observation space that contains varying poses. This leads to the pose misalignment during training and inference stages, and therefore limits the generalization ability. (ii) *Limited local receptive fields.* As shown in Fig. 1, due to the heavy self-occlusion of dynamic human bodies, the painted SMPL templates are usually incomplete. While, as a 3D convolution network, the limited local receptive fields of SPC make it sensitive to the incomplete input, especially when the occluded regions are large.

To address the aforementioned issues, we propose to first process the painted SMPL with transformers under the *static canonical space* to remove the pose misalignment between training and inference stages and capture the *global relationships* between human parts. Then, a deformation from the canonical to the observation space is required to fetch the human representation of a query point (sampling points on rays) under the observation space. Finally, the fine-grained information directly achieved from the observation space should be further included to the coarse human representation to complement the details.

Motivated by this, we present the TransHuman, a brand-new framework that shows superior generalization ability with high efficiency. TransHuman is mainly composed of Transformer-based Human Encoding (TransHE), Deformable Partial Radiance Fields (DPaRF), and Fine-grained Detail Integration (FDI). (i) *TransHE.* TransHE is a pipeline that processes the painted SMPL under the canonical space with transformers [9]. The core of this pipeline includes a canonical body grouping strategy for the avoidance of semantic ambiguity, and a canonical learning scheme to ease the learning of global relationships. (ii) *DPaRF.* DPaRF deforms the output tokens of TransHE from the canonical space to the observation space and gets a robust human representation for a query point from marched rays. As shown in Fig. 1, the main idea is to bind each token (representing a certain human part) with a radiance field whose partial coordinate system deforms as the pose changes, and the query point is encoded via the coordinates under the deformed partial coordinate systems. (iii) *FDI.* With TransHE and DPaRF, the human representation contains coarse information with human priors yet limited fine-grained details directly captured from the observation space. Therefore, similar to [18], we propose to further integrate the detailed information from the pixel-aligned features at the guidance of the human representation.

Extensive experiments on ZJU-MoCap [29] and H36M [15] demonstrate the superior generalization abil-

ity and high efficiency of TransHuman which attains a new state-of-the-art performance and outperforms previous methods by significant margins, *e.g.*, +2.20 PSNR and −45% LPIPS on ZJU-MoCap [29] under the pose generalization setting.

Our contributions are summarized as follows:

- We propose a brand-new framework TransHuman for the challenging generalizable neural human rendering task which attains a significantly new state-of-the-art performance with high efficiency.

- We propose to process the painted SMPL under the canonical space to remove the pose misalignment during training and inference stages and deform it back to the observation space via DPaRF for robust query point encoding.

- To the best of our knowledge, we make the first attempt to explore the transformers technology around the painted SMPL for capturing the global relationships between human parts.

2. Related Work

2.1. Human Performance Capture

Synthesizing novel views for human performer is a long-standing topic in computer vision and graphics. Traditional methods [10, 6, 12, 7] typically require expensive hardware like depth sensors for getting reasonable results. With the recent success of Neural Radiance Fields (NeRF) [27, 2], many works [29, 28, 39, 33] have attempted to learn the 3D human representation from image inputs via differentiable rendering. However, they require tedious per-subject optimization on dense training images, and can not generalize to unseen subjects, which largely confines the real-world applications.

To tackle this issue and inspired by the recent advances of generalizable NeRF methods [42, 4, 37], the generalizable neural human rendering task is explored [18, 11, 5, 44], At the core of this task is to properly exploit the human prior from the pre-fitted parametric human model. One line of works [44, 11] take the parametric human model as the medium of the deformation between observation and canonical spaces using blend skinning technology [14, 19, 22], and optimize conditional NeRF under a canonical pose. Instead, another line of works [18, 5] directly diffuse the painted parametric human model under the observation space via SparseConveNet (SPC) [21] for a human representation with approximate priors, and the final condition feature for a query point is the hybrid of human representation and pixel-aligned features. Obviously, a high-quality human representation is critical in this paradigm, yet the SPC-based one optimizes under the varying observation
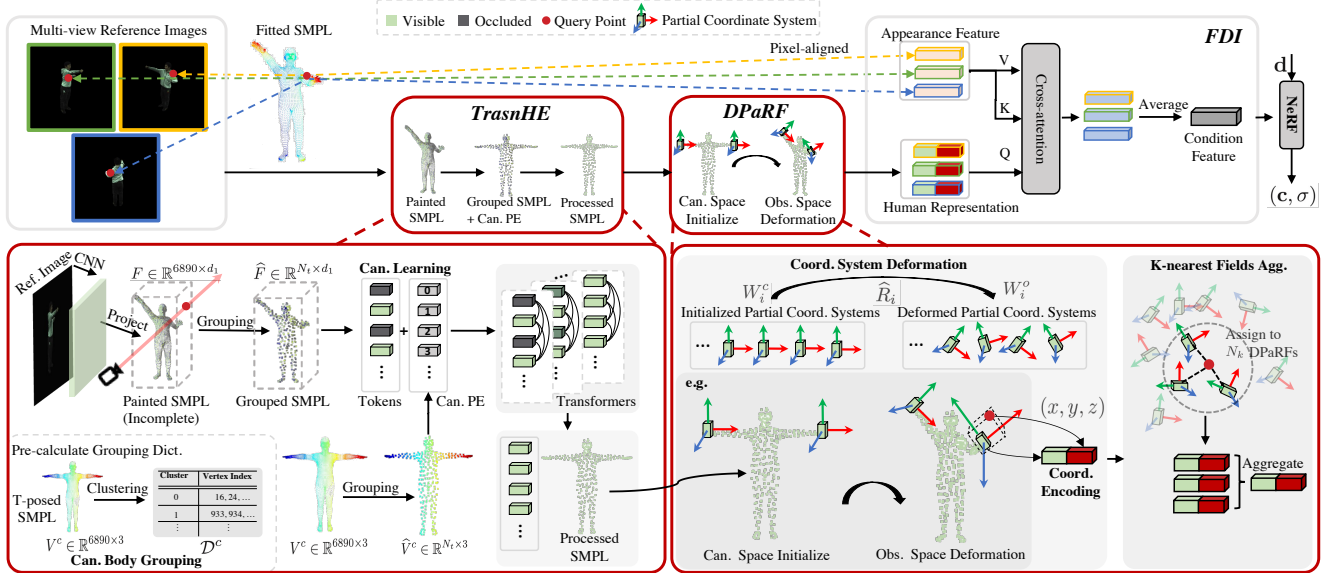
Figure 2. **Overview of TransHuman.** TransHE first builds a pipeline for capturing the global relationships between human parts via transformers under the canonical space. Then, DPaRF deforms the coordinate system from the canonical back to the observation space and encodes a query point as an aggregation of coordinates and condition features. Finally, FDI further gathers the fine-grained information of the observation space from the pixel-aligned appearance feature under the guidance of human representation.

space, lacks the global perspective, and is restricted by the trilinear sampling in discrete 3D volumes.

Targeting these issues, we present TransHuman with an advanced human representation based on transformers [36, 35, 9], and outperforms the previous state-of-the-art methods by significant margins.

2.2. Transformers with Neural Radiance Fields

With the significant advances of the transformer architecture [8, 9, 3, 30], several works [20, 17, 32, 37, 16, 41] have attempted to introduce it with NeRF technology. Specifically, [20] combines transformers with CNN [13] as a stronger feature extractor for reference images, [17, 32, 37] use transformers as the aggregator of source view features, and [16, 41] introduce the pre-trained transformers [30, 3] as a semantic prior to relieve the dense requirement of training views.

Differently, in this paper, we make the first attempt to apply the transformer technology around the surface of painted SMPL for a stronger human representation that captures the global relationship between human parts.

3. Method

Overview. The task of generalizable neural human rendering targets on learning conditional NeRF across multi-view videos of different subjects, which can generalize to unseen subjects in a single feed-forward pass given sparse reference views. At the core of the task is to get a high-quality condition feature that contains accurate subject information for each query point sampled on rays. To this end, we pro-

pose a novel framework named TransHuman which shows superior generalization ability. As shown in Fig. 2, TransHuman is mainly composed of three aspects: Transformer-based Human Encoding (TransHE), Deformable Partial Radiance Fields (DPaRF), and Fine-grained Detail Integration (FDI). § 3.1 introduces the TransHE which builds a pipeline for capturing the global relationships between human parts via transformers under the canonical space. § 3.2 demonstrates the DPaRF which deforms the processed SMPL back to the observation space and fetch a robust human representation. § 3.3 presents the FDI module that further gathers the fine-grained information directly from the observation space with the guidance of human representation. After that, we introduce the volume rendering in § 3.4, and the training and inference pipelines in § 3.5.

3.1. Transformer-based Human Encoding

For simplicity, we start by introducing the process of a single reference image that is applicable for all other views, and the multi-view aggregation will be detailed in § 3.3. Given a reference images I for a certain time step and its corresponding pre-fitted SMPL model $V^o \in \mathbb{R}^{6890 \times 3}$ under the observation pose [†], we first project the d_1-dimensional deep features of I extracted by CNN to the vertices of V^o based on the camera information, and get the painted SMPL $F \in \mathbb{R}^{6890 \times d_1}$. Previous methods [18, 5] have mainly employed the SPC [21] to diffuse the painted SMPL to nearby space (Fig. 1). However, they optimize under the varying observation space which leads to the pose misalign-

---

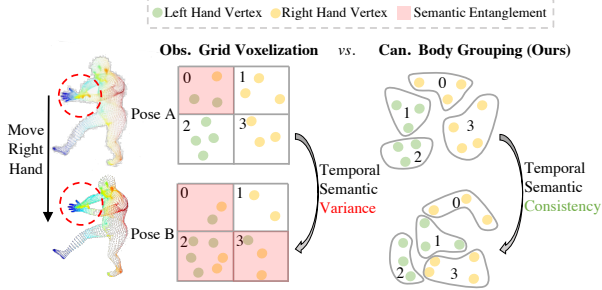[†]We use the SMPL coordinate system unless otherwise specified.

Figure 3. **2D illustration of the semantic ambiguity issue.** Naive grid voxelization under the observation space leads to spatial semantic entanglement and temporal semantic variance issues, while the semantics with our canonical body grouping strategy is temporally consistent and spatially disentangled.

ment between training and inference stages, and the limited receptive fields of 3D convolution blocks make it sensitive to the incomplete painted SMPL input caused by the heavy self-occlusions of human bodies. Tackling these issues, we present a pipeline named Transformer-based Human Encoding (TransHE) that captures the global relationships between human parts under the canonical space. The key of TranHE includes a canonical body grouping strategy for avoiding the semantic ambiguity and a canonical learning scheme to ease the optimization and improve the generalization ability.

Canonical Body Grouping. Directly taking all the vertex features of F as input tokens of transformers is neither effective considering the misalignment between fitted SMPL and the ground truth body, nor efficient due to the large vertex number, *i.e.*, 6890. A possible solution is to directly perform the grid voxelization [25] on F under the observation pose. However, due to the complex human poses, this will lead to the semantic ambiguity issue. More concretely, the gathered vertices in each voxel are highly different as the pose changes (*i.e.*, temporal semantic variance), and a voxel might include vertices from dispersed semantic parts (*i.e.*, spatial semantic entanglement), as illustrated in Fig. 3.

To tackle this issue, we propose that grouping the vertices under the canonical space and then applying this canonical grouping to all the observation poses is a better choice. Compared with the varying observation poses, the canonical pose is both *static* and more *stretched*, which can largely relieve the semantic ambiguity issue via the consistent split among different poses (*i.e.*, temporal semantic consistency) and more disentangled semantics in each voxel (*i.e.*, spatial semantic disentanglement), as shown by the right part of Fig. 3.

Formally, we first process the canonically posed (T-posed) SMPL $V^c \in \mathbb{R}^{6890 \times 3}$ with a clustering algorithm (*e.g.*, k-means [1]) based on the 3D coordinates, and get a grouping dictionary $\mathcal{D}^c$ caching the indexes of the SMPL vertices that belong to the same cluster, as illustrated in

Fig. 2. Notice that we only need to calculate $\mathcal{D}^c$ once before training. Then, for each iteration, the features from the same cluster are aggregated via average pooling:

$$\widehat{F} = \mathcal{G}_{\mathcal{D}^c}(F), \quad \widehat{F} \in \mathbb{R}^{N_t \times d_1}, \tag{1}$$

where N_t is the number of clusters (tokens), and $\mathcal{G}_{\mathcal{D}^c}(\cdot)$ indicates indexing based on $\mathcal{D}^c$ and then performing average pooling in each cluster.

Canonical Learning. After grouping, we now have a decent number of input tokens, and the next question is about the choice of position embedding for each token. Since we need the condition feature of a query point under the observation space, a possible choice is to directly learn under the observation space (same as SPC-based methods [18, 5]) and use the 3D coordinates of each token under the observation pose as the position information, *i.e.*, $\widehat{V}^o = \mathcal{G}_{\mathcal{D}^c}(V^o) \in \mathbb{R}^{N_t \times 3}$. However, except for the pose misalignment issue mentioned previously, $\widehat{V}^o$ is also varying for different time steps, which leads to the unfixed patterns of position embeddings that make it harder to capture the global relationships between human parts.

Hence, to address these issues, we propose to learn the global relationships under the static canonical space for removing the pose-misalignment and easing the learning of global relationships:

$$\widehat{F}' = \mathcal{T}(\widehat{F}, \gamma_1(\widehat{V}^c)), \tag{2}$$

where $\widehat{V}^c = \mathcal{G}_{\mathcal{D}^c}(V^c)$ is the token positions under the canonical space, $\gamma_1(\cdot) : \mathbb{R}^{3 \rightarrow d_1}$ represents the positional encoding used in the original NeRF [27], $\mathcal{T}(\cdot) : \mathbb{R}^{d_1 \rightarrow d_1}$ indicates the transformers, and $\widehat{F}' \in \mathbb{R}^{N_t \times d_1}$ is the output tokens with learned global relationships between each other.

3.2. Deformable Partial Radiance Fields

For deforming the processed SMPL back to the observation space and get a robust human representation, we present the Deformable Partial Radiance Fields (DPaRF). The main idea of DPaRF is to bind each output token of TransHE with a conditional partial radiance field for a certain semantic part whose coordinate system deforms as the pose changes under the observation space, and the query points from rays are encoded as the coordinates under the deformed coordinate system, as shown in Fig. 2.

Coordinate System Deformation. Given the i-th token $\widehat{F}'_i \in \mathbb{R}^{d_1}$ from the TransHE output, a coordinate system $W_i^c \in \mathbb{R}^{3 \times 3}$ is initialized under the canonical space which takes $\widehat{V}_i^c \in \mathbb{R}^3$ as the origin [†]. Then, as the pose changes under the observation space, we rotate W_i^c with the rotation matrix $\widehat{R}_i \in \mathbb{R}^{3 \times 3}$ of token i:

$$W_i^o = \widehat{R}_i W_i^c, \tag{3}$$

[†] Without loss of generality, we set W_i as the identity matrix for all the tokens for simplicity.

where $\widehat{R}_i$ is the averaged rotation matrix for vertices belonging to the i-th token, *i.e.*, $\widehat{R} = \mathcal{G}_{\mathcal{D}^c}(R) \in \mathbb{R}^{N_t \times 3 \times 3}$, and $R \in \mathbb{R}^{6890 \times 3 \times 3}$ can be calculated via blending the rotation matrices of 24 joints with the blend weights provided by SMPL [24].

Coordinate Encoding. After that, for a query point $\mathbf{p}$ sampled from the rays under the observation space, we get its coordinate $\overline{\mathbf{p}}_i$ under the DPaRF of the i-th token with:

$$\overline{\mathbf{p}}_i = W_i^o(\mathbf{p} - \widehat{V}_i^o). \tag{4}$$

And the final fetched human representation from the DPaRF of the i-th token is:

$$\mathbf{h}_i = [\widehat{F}'_i; \gamma_2(\overline{\mathbf{p}}_i)], \quad \mathbf{h}_i \in \mathbb{R}^{d_2}, \tag{5}$$

where $[;]$ indicates the concatenation, and $\widehat{F}'_i$ is the condition feature for the i-th DPaRF.

K-nearest Fields Aggregation. Finally, for a more robust representation, we assign a query point $\mathbf{p}$ to its N_k nearest DPaRFs, and aggregate them based on the distances:

$$\mathbf{h} = \sum_{i=1}^{N_k} softmax(-\frac{\|\mathbf{p} - \widehat{V}_i^o\|_2}{\sum_i \|\mathbf{p} - \widehat{V}_i^o\|_2})\mathbf{h}_i, \quad \mathbf{h} \in \mathbb{R}^{d_2}. \tag{6}$$

3.3. Fine-grained Detail Integration

With TransHE and DPaRF, for a query point $\mathbf{p}$, we can actually achieve a set of human representations from N_v reference views $\mathbf{h}^{1:N_v} = \{\mathbf{h}^j\}_{j=1}^{N_v} \in \mathbb{R}^{N_v \times d_2}$ following the same procedure. $\mathbf{h}^{1:N_v}$ contains coarse information with human priors (*e.g.*, geometry constraints and certain color information) yet lacks the fine-grained information (*e.g.*, lighting, textures) for high-fidelity novel view synthesis. Therefore, inspired by [18], we further integrate the fine-grained information from the pixel-aligned appearance feature $\mathbf{a}^{1:N_v} = \{\mathbf{a}^j\}_{j=1}^{N_v} \in \mathbb{R}^{Nv \times d_2}$ at the guidance of human representation $\mathbf{h}^{1:N_v}$.

Fine-grained Appearance Features. For the appearance features, instead of directly using projected deep features from CNN, *i.e.*, the one used when painting SMPL, we additionally concatenate the projected RGB-level information from the raw images and then fuse them with a fully connected layer $FC(\cdot) : \mathbb{R}^{3+d_1 \rightarrow d_2}$. The projected RGB features can complement the misaligned and lost details caused by the down-sample operation in CNN.

Coarse-to-fine Integration. Then, we employ a cross-attention module which takes the human representation $\mathbf{h}^{1:N_v}$ as the query, and the appearance feature $\mathbf{a}^{1:N_v}$ as the key and value, and get the integrated feature $\mathbf{f}^{1:N_v} \in \mathbb{R}^{N_v \times d_2}$. The final condition feature $\mathbf{f} \in \mathbb{R}^{d_2}$ of query point $\mathbf{p}$ is achieved via the average pooling on the view dimension: $\mathbf{f} = \sum_{j=1}^{N_c} \frac{1}{N_c} \mathbf{f}^j$.

3.4. Volume Rendering

Desnity & Color Prediction. The final density $\sigma(\mathbf{p}) \in \mathbb{R}^1$ and color $\mathbf{c}(\mathbf{p}) \in \mathbb{R}^3$ are predicted as:

$$\begin{aligned}
\sigma(\mathbf{p}) &= MLP_\sigma(\mathbf{f}), \\
\mathbf{c}(\mathbf{p}) &= MPL_\mathbf{c}(\mathbf{f}, \gamma_3(\mathbf{d})),
\end{aligned} \tag{7}$$

where MLP_σ and $MLP_\mathbf{c}$ are NeRF MLPs for density and color predictions, respectively, and $\mathbf{d}$ is the unit view direction of the ray.

Differentiable Rendering. Then, for a marched ray $\mathbf{r}(z) = \mathbf{o} + z\mathbf{d}$, where $\mathbf{o} \in \mathbb{R}^3$ represents the camera center, and $z \in \mathbb{R}^1$ is the depth between a pre-defined bounds $[z_n, z_f]$, its color $\mathbf{C}(\mathbf{r})$ is calculated via the differentiable volume rendering [27]:

$$\mathbf{C}(\mathbf{r}) = \int_{z_n}^{z_f} T(z)\sigma(z)\mathbf{c}(z)dz, \tag{8}$$

where $T(z) = exp(-\int_{z_n}^z \sigma(s)ds)$ represents the probability that the ray travels from z to z_n.

3.5. Training & Inference

Training Losses. We compare the rendered pixel colors with the ground truth ones for supervision. Similar to [39], we employ the MSE loss for pixel-wise and perceptual loss [43] for patch-wise supervision, which is more robust to misalignments. The random patch sampling [39] is employed for supporting perceptual loss training. The overall loss is:

$$\mathcal{L} = \mathcal{L}_{MSE} + \lambda\mathcal{L}_{PER}, \tag{9}$$

where we set $\lambda = 0.1$ by default.

Inference. During the inference stage, for each time step, N_v reference views are provided and the rendered target views are compared with the ground truth ones for calculating the metrics. Notably, GP-NeRF [5] has proposed a fast rendering scheme that leverages the coarse geometry prior from the 3D feature volume to filter out useless points. Similarly, our framework also supports such strategy by simply using the SMPL template as the geometry prior instead (detailed in the appendix).

4. Experimental Results

4.1. Experimental Settings

Datasets. We benchmark on ZJU-MoCap [29] and H36M [15] for verifying the effectiveness of our TransHuman.

(i) ZJU-MoCap [29] provides multi-view videos of 10 human subjects with 23 synchronized cameras, together with the pre-fitted SMPL parameters and human masks. Each video spans between 1000 to 2000 frames and contains complicated motions like "Taichi" and "Twirl". Following [18, 5], 10 subjects are split into 7 source subjects

Method	Dataset		Per-subject training	Unseen		Results		
	Train	Test		Pose	Subject	↑ PSNR	↑ SSIM	↓ LPIPS
Pose Generalization								
NV [TOG19] [23]	ZJU-7	ZJU-7	✓	✓	✗	22.00	0.818	-
NT [TOG19] [34]	ZJU-7	ZJU-7	✓	✓	✗	22.28	0.872	-
NHR [CVPR20] [40]	ZJU-7	ZJU-7	✓	✓	✗	22.31	0.871	-
NB [CVPR21] [29]	ZJU-7	ZJU-7	✓	✓	✗	23.79	0.887	-
NHP [NIPS21] [18]	ZJU-7	ZJU-7	✗	✓	✗	24.60	0.910	0.147
GP-NeRF [ECCV22] [5]	ZJU-7	ZJU-7	✗	✓	✗	25.05	0.909	0.159
Ours	ZJU-7	ZJU-7	✗	✓	✗	**27.25**	**0.936**	**0.087**
Identity Generalization								
NV [TOG19] [23]	ZJU-3	ZJU-3	✓	✓	✗	20.84	0.827	-
NT [TOG19] [34]	ZJU-3	ZJU-3	✓	✓	✗	21.92	0.873	-
NHR [CVPR20] [40]	ZJU-3	ZJU-3	✓	✓	✗	22.03	0.875	-
NB [CVPR21] [29]	ZJU-3	ZJU-3	✓	✓	✗	22.88	0.880	-
PVA [arXiv21] [31]	ZJU-7	ZJU-3	✗	✓	✓	23.15	0.866	-
PixelNeRF [CVPR21] [42]	ZJU-7	ZJU-3	✗	✓	✓	23.17	0.869	-
KeyNeRF [ECCV22] [26]	ZJU-7	ZJU-3	✗	✓	✓	25.03	0.897	-
GP-NeRF [ECCV22] [5]	ZJU-7	ZJU-3	✗	✓	✓	24.55	0.902	0.157
NHP [NIPS21] [18]	ZJU-7	ZJU-3	✗	✓	✓	24.94	0.905	0.144
Ours	ZJU-7	ZJU-3	✗	✓	✓	**26.15**	**0.918**	**0.098**
GP-NeRF† [ECCV22] [5]	ZJU-7	ZJU-3	✗	✓	✓	26.83	0.924	0.132
Ours†	ZJU-7	ZJU-3	✗	✓	✓	**27.55**	**0.933**	**0.090**
One-shot Generalization								
NHP [NIPS21] [18]	ZJU-7	ZJU-3	✗	✓	✓	23.20	0.877	0.182
Ours	ZJU-7	ZJU-3	✗	✓	✓	**24.11**	**0.891**	**0.142**
Cross-dataset Generalization								
NHP [NIPS21] [18]	ZJU-7	H36M	✗	✓	✓	18.84	0.820	0.222
Ours	ZJU-7	H36M	✗	✓	✓	**20.48**	**0.856**	**0.169**

Table 1. **Comparisons of generalization ability with the state-of-the-art methods.** We achieve a significantly new sate-of-the-art performance compared with both generalizable [31, 42, 5, 18, 26] and per-subject methods [23, 34, 40, 29]. Following [18], the per-subject optimization methods are trained on the training part of each subject since they can not generalize to unseen subjects, which is actually an easier task. "†" means using the officially released human split from GP-NeRF [5] and employing the overfitting trick used in GP-NeRF.

(ZJU-7) and 3 target subjects (ZJU-3), and each subject is further divided into training and testing parts. We strictly follow the officially released human split from [18] for training and testing. We refer the detailed split information to the appendix. To prove that our method can welly handle the incomplete painted SMPL, we additionally report the performance of the one-shot generalization setting, *i.e.*, only 1 reference view is provided during inference.

(ii) H36M [15] records multi-view videos with 4 cameras and includes multiple subjects with complex motions. We use the preprocessed one by [28] which contains representative subjects S1, S5, S6, S7, S8, S9, S11, and their corresponding SMPL parameters and human masks. We verify the cross-dataset generalization ability with H36M, *i.e.*, trained on ZJU-MoCap and then directly inference on H36M. The first 3 views are taken as the reference views, and the last one is used as the target view.

Evaluation Metrics. For novel view synthesis, we report the commonly used Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [38], and Learned Perceptual Image Patch Similarity (LPIPS) [43]

as the evaluation metrics. For 3D reconstruction, following [18, 5], we only report the qualitative results since ground truth meshes are unavailable.

4.2. Implementation Details

In line with [18], we take the ResNet-18 [13] (only the first 3 layers are used) as the CNN for extracting the deep features from reference images and set the multi-view number $N_v = 3$. The number of clusters (tokens) in human body grouping is set as $N_t = 300$, and the light-weight ViT-Tiny [9] is employed as the transformer module. Each query point is assigned with $N_k = 7$ DPaRFs. Following [18, 5], we train on ZJU-MoCap with 512×512 resolutions, and for each ray we sample 64 points by default during both the training and inference stages.

4.3. Comparisons with State-of-the-art

Baselines. Following [18, 5], we compare with both per-subject optimization methods [29, 34, 40, 23] and generalizable methods [31, 42, 26, 18, 5]. For per-subject optimization methods, an individual model is trained on the training part of each subject. Notably, previous state-
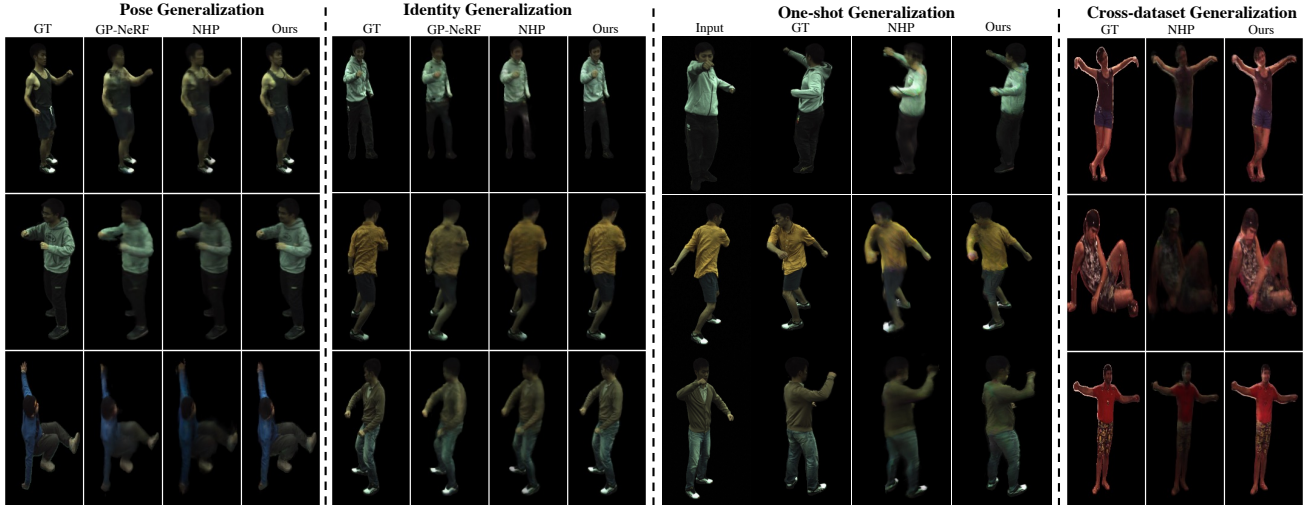
Figure 4. **Visualization comparisons with previous state-of-the-art methods on ZJU-MoCap (pose generalization, identity generalization) and H36M (cross-dataset generalization).** Our method shows significantly better generalization ability with better body geometry and more accurate details like textures and lighting.
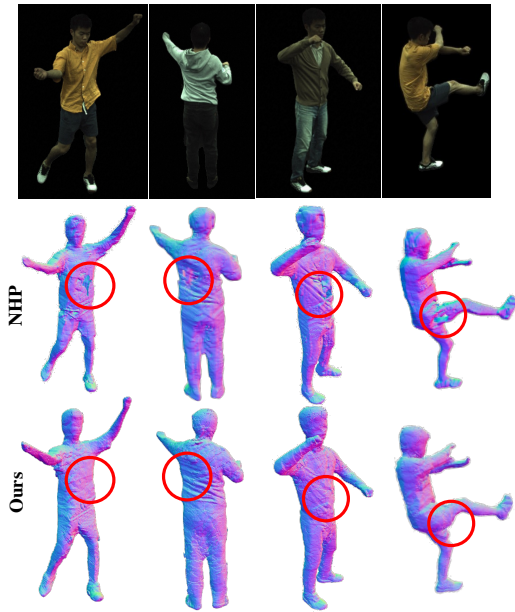

Figure 5. **3D reconstruction under the identity generalization setting.** Our method achieves more complete geometry with details like wrinkles compared with NHP [18] which employs a SPC-based human representation.

of-the-art methods for generalizable neural human rendering [18, 5] actually use different human splits in their officially released code and are not in line with the one used in their papers (performance is not reproducible). Hence, for fair comparisons, we **unify them under the released human split of NHP [18]**. Specifically, we report the performance of NHP [18] using the official checkpoint, and re-run the official code of GP-NeRF [5] under the unified human split. Note that, GP-NeRF has employed an overfitting trick which we think is unreasonable, *i.e.*, they overfit the test

reference views instead of randomly sampling during the training stage. This trick leaks the test information to the training stage, therefore we remove it in our re-running. We also provide the comparisons under the released human split of GP-NeRF with the overfitting trick, where our method outperforms it consistently by large margins.

Novel View Synthesis. We compare the quantitative results with previous state-of-the-art methods in Table 1. Obviously, we outperform them by significant margins under all the settings. Notably, for the identity generalization setting, the per-subject methods are directly trained on the target subjects while our method is only trained on the source subjects, yet we still outperform them by large margins, *i.e.*, +3.27 in PSNR. Compared with the recent SPC-based generalizable methods [18, 5], our method also shows healthy margins, *i.e.*, +2.20 PSNR and −45% LPIPS compared with the second-best under the pose generalization setting. For the more challenging cross-dataset generalization setting, we also outperform the baseline methods remarkably albeit these two datasets [29, 15] have significantly different distributions, which proves the superior generalization ability of our TransHuman.

The qualitative comparisons are illustrated in Fig. 4, where our TransHuman gives significantly better details and body geometry. We attribute this to the careful design of our framework, *i.e.*, the global human representation brings more complete body geometry, the canonical learning scheme gives better generalization ability, and FDI further includes more fine-grained details like textures and lighting.

3D Reconstruction. The 3D reconstruction results are illustrated in Fig. 5. Compared with NHP [18] that uses the SPC-based human representation, our method achieves a more complete and fine-grained geometry with details like

Method	↑ PSNR	↑ SSIM	↓ LPIPS
obs. body grouping	25.28	0.909	0.111
obs. PE	25.80	0.915	0.102
can. body grouping + can. PE	**26.15**	**0.918**	**0.098**

Table 2. **Ablation of TransHE.** Our canonical body grouping together with the canonical learning scheme performs best.

Method	↑ PSNR	↑ SSIM	↓ LPIPS
w/o coordinate	25.80	0.912	0.123
absolute coordinate	25.76	0.912	0.116
w/o k-nearest fields	26.05	0.916	0.099
full model	**26.15**	**0.918**	**0.098**

Table 3. **Ablation of DPaRF.** Coordinate encoding is critical and the k-nearest fields aggregation can further bring improvements.

Method	↑ PSNR	↑ SSIM	↓ LPIPS
w/o **a**	24.58	0.898	0.134
w/o **h**	24.66	0.897	0.143
w/o RGB	26.05	0.917	0.101
full model	**26.15**	**0.918**	**0.098**

Table 4. **Ablation of FDI.** Using the appearance feature and human representation individually leads to the drop of performance, and the raw RGB feature can bring certain improvement.

Method	↑ PSNR	↑ SSIM	↓ LPIPS
SPC + trilinear	25.14	0.907	0.102
TransHE + DPaRF (ours)	**26.15**	**0.918**	**0.098**

Table 5. **Comparision with SPC-based representation.** Our transformer-based representation outperforms the SPC-based one significantly.

Method	Param.	Inference Time	Inference Mem.	Training Mem.	PSNR
NHP [18]	**5.80M**	1h55min	6.4GB	12.2GB	24.94
GP-NeRF [5]	9.52M	**9min**	10.3GB	11.0GB	24.55
Ours-16pts	6.08M	**9min**	**5.7GB**	**7.8GB**	**25.39**
Ours	6.08M	17min	6.2GB	7.8GB	26.15

Table 6. **Efficiency comparisons under the identity generalization setting.** With the same inference time, our method outperforms GP-NeRF [5] significantly in PSNR albeit requiring fewer parameters and training/inference memory. The performance can further be greatly improved at the cost of certain additional inference time and minor inference memory.

wrinkles.

4.4. Ablation Studies

Following [18], we perform ablation studies under the identity generalization setting. Due to the limited space, we refer more detailed ablation studies to the appendix.

Ablation of TransHE. We first study the effectiveness of canonical body grouping and canonical learning scheme in Table 2. When performing the body grouping under the observation space with grid voxelization ("obs. body grouping"), the performance suffers a significant drop from 26.15 to 25.28 in PSNR. As introduced in § 3.1, performing grouping under the observation space leads to the semantic ambiguity issue, therefore leading to worse performance. Then, "obs. PE" changes the position embedding of input tokens from the canonical positions $\hat{V}^c$ to observation positions $\hat{V}^o$, and also observes a significant decrease, e.g., -0.35 in PSNR. The canonical learning scheme eases the optimization and removes the pose misalignment between training and inference stages, therefore leading to better performance.

Ablation of DPaRF. We verify the effectiveness of DPaRF in Table 3. "w/o coordinate" represents removing the coordinate part from the human representation. As expected, the performance drops by significant margins (-0.35 in PSNR). Coordinates contain the accurate position information of query point in each DPaRF, therefore is important. "absolute coordinate" indicates using the absolute coordinate of query point, i.e., $\mathbf{p}$ instead of $\bar{\mathbf{p}}$ in Eq. 5, and the performance does not show significant improvement compared with "w/o coordinate". This further proves the importance of using the coordinate under the deformed coordinate systems. Finally, "w/o k-nearest fields" shows that the k-nearest fields aggregation design can bring certain improvement on all the metrics.

Ablation of FDI. We first perform the ablation of FDI by individually removing the appearance feature part ("w/o **a**") or the human representation part ("w/o **h**"). As illustrated in Table 4, merely using either of them gives an unsatisfactory

performance. Then, "w/o RGB" shows that the raw RGB features can further bring a measure of improvement.

Comparisons with SPC-based representation. To further verify the effectiveness of our proposed transformer-based human representation, we directly replace the TransHE and DPaRF modules with SPC and trilinear sampling in our code. We follow [18] to configure the SPC including the architecture and input resolution. As shown by Table 5, our proposed transformer-based representation outperforms the SPC-based one by significant margins among all the metrics under a fair comparison setting.

4.5. Efficiency Analysis

We compare the efficiency of our method with previous state-of-the-art methods in Table 6 under the identity generalization setting (438 frames). For a fair comparison with the previously fastest method GP-NeRF [5] under the same inference time, we provide a fast version of our method by reducing the sampling points per ray from 64 to 16 during inference ("Ours-16pts"). Obviously, with the same inference time, our method still outperforms GP-NeRF by 0.84 in PSNR albeit using merely 64% parameters, 55% inference memory, and 71% training memory, and the performance can be further significantly improved with acceptable additional cost. This proves that our TransHuman is both effective and efficient.

5. Conclusion

In this paper, we propose a brand-new framework named TransHuman for the generalizable neural human rendering task. At the core of TransHuman is a canonically optimized human representation with global relationships between human parts captured by transformers which shows superior generalization ability compared with previous methods. However, there are remaining challenges to be explored, such as the joint optimization of fitted SMPL and training on unconstrained multi-view capture setups. We hope that our efforts will motivate more researchers in the future.

References

[1] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020. 4

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 1, 2

[5] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *ECCV*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 11

[6] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 2

[7] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 6

[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] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2022. 2

[12] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6

[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. 2

[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. 2, 5, 6, 7, 13, 14

[16] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 3

[17] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 3

[18] 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. 1, 2, 3, 4, 5, 6, 7, 8, 11, 14

[19] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000. 2

[20] Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 3

[21] Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 806–814, 2015. 1, 2, 3

[22] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 2

[23] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019. 6

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1, 5

[25] Jiageng Mao, Yujing Xue, Minzhe Niu, et al. Voxel transformer for 3d object detection. *ICCV*, 2021. 4

[26] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*, 2022. 6

[27] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 4, 5

[28] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 1, 2, 6, 14

[29] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 1, 2, 5, 6, 7, 13, 14

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[31] Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pva: Pixel-aligned volumetric avatars. *arXiv preprint arXiv:2101.02697*, 2021. 6

[32] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 3

[33] Gusi Te, Xiu Li, Xiao Li, Jinglu Wang, Wei Hu, and Yan Lu. Neural capture of animatable 3d human from monocu-lar video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 275–291. Springer, 2022. 1, 2

[34] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 6

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 3

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[37] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 1, 2, 3

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

[39] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 1, 2, 5

[40] 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. 6

[41] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 736–753. Springer, 2022. 3

[42] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1, 2, 6

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 6

[44] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022. 2

Method	↑ PSNR	↑ SSIM	↓ LPIPS	Inference Time
w/o progressive	**26.15**	**0.918**	**0.098**	56min
full model	**26.15**	**0.918**	**0.098**	**17min**

Table 7. **Effectiveness of progressive rendering strategy.** The progressive rendering strategy can reduce the inference time by around 70% while without influencing the performance.

Method	↑ PSNR	↑ SSIM	↓ LPIPS
NHP	25.65	0.917	0.148
GP-NeRF	26.46	0.918	0.158
Ours	**28.08**	**0.939**	**0.087**

Table 8. **Fitting performance on training frames.** Our method shows the best fitting ability compared with previous methods.

Method	↑ PSNR	↑ SSIM	↓ LPIPS
$N_t = 100$	26.04	0.917	0.100
$N_t = \mathbf{300}$	**26.15**	**0.918**	**0.098**
$N_t = 500$	26.10	0.917	0.100
$N_t = 1000$	26.07	0.917	0.100

Table 9. **Influence of cluster number N_t.** $N_t = 300$ gives the best performance.

A. Progressive Rendering

GP-NeRF [5] has proposed a progressive rendering strategy using the coarse geometry provided by the output 3D feature volume of SPC to reduce the number of rendering points. Although there is no SPC in our framework, we find that simply using the fitted SMPL as the alternative works pretty well. Specifically, after sampling points on marched rays from the target view, we only render the points whose euclidean distance to the SMPL template is smaller than $0.1m$. Then, for these close points, we first get the density values for all of them, and then only send part of them whose density value is larger than 0 for the following color inference, which is in line with [5]. The effectiveness of this strategy is illustrated in Table 7. While without decreasing the performance, the inference time is reduced by around 70%. Notably, even without using such accelerating strategy, the inference is still over 2 times faster than NHP [18] (56min *vs*. 1h55min, Table 6). This strongly proves the efficiency of our method.

B. Performance on Training Frames

Following previous methods [18, 5], we report the fitting performance on the training set in Table 8. We achieve the best fitting performance among the generalizable methods, which shows the superior capacity of our method.

C. Additional Ablation Studies

We provide more detailed ablation studies in this section.

C.1. Influence of Cluster Number N_t

We study the influence of cluster (token) number by varying it sequentially as $\{100, 300, 500, 1000\}$. As shown

Method	↑ PSNR	↑ SSIM	↓ LPIPS
$N_k = 1$	26.05	0.917	0.099
$N_k = 3$	26.11	0.918	0.100
$N_k = 5$	26.13	0.918	0.100
$N_k = \mathbf{7}$	**26.15**	**0.918**	**0.098**
$N_k = 9$	26.10	0.917	0.100

Table 10. **Influence of k-nearest number N_k.** $N_k = 7$ performs best.

Method	↑ PSNR	↑ SSIM	↓ LPIPS
w/o $\mathcal{L}_{PER}$	**26.16**	0.916	0.146
full model	26.15	**0.918**	**0.098**

Table 11. **Influence of perceptual loss.** Perceptual loss mainly improves the LPIPS with less effect on PSNR and SSIM.
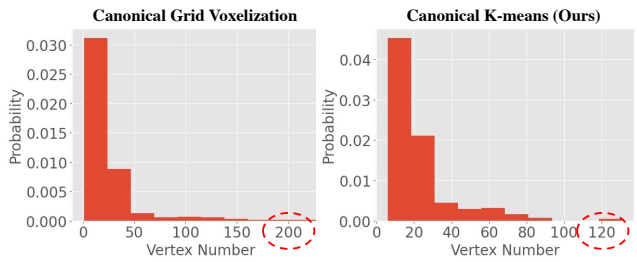

Figure 6. **Comparisons of vertex number distributions between canonical grid voxelization and canonical k-means.** Canonical k-means gives more uniform split with smaller variance.

in Table 9, too large cluster number does not bring further improvement. As mentioned in § 3.1, there exists misalignment between the fitted SMPL and the ground truth body. Larger cluster number may also include more misleading information, and we only intend to take the human representation as the coarse-level guidance, therefore we set $N_t = 300$.

C.2. Influence of K-nearest Number N_k

We show the influence of k-nearest number N_k in Table 10. When using no k-nearest fields aggregation, *i.e.*, $N_k = 1$, the performance suffers a relatively significant drop in PSNR. This shows that using k-nearest fields aggregation can improve the robustness of human representation. When $N_k > 1$, the performance tends to be more stable, and we choose N_k as 7 since it gives the best performance.

C.3. Influence of Perceptual Loss

In Table 11, we demonstrate the influence of perceptual loss. Obviously, perceptual loss can largely improve the LPIPS, *i.e.*, make the results visually pleasing, while shows less effect on PSNR and SSIM. Without perceptual loss, we still outperform previous methods by consistent margins in PSRN and SSIM.

Method	↑ PSNR	↑ SSIM	↓ LPIPS
can. grid voxelization	26.01	0.917	0.100
can. k-means (ours)	**26.15**	**0.918**	**0.098**

Table 12. **Comparisons between using k-means and grid voxelization in canonical body grouping.**
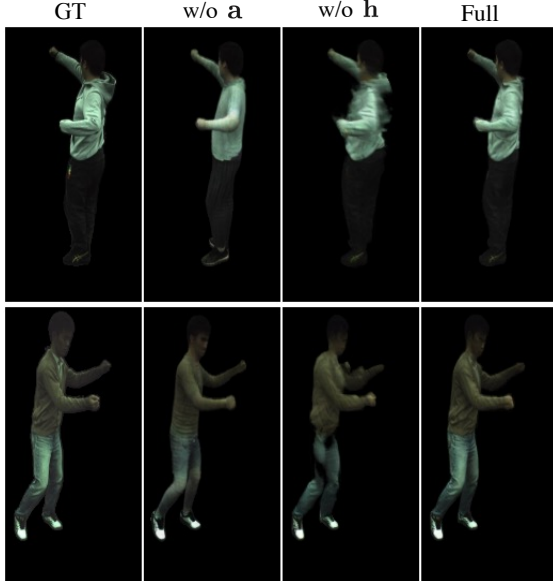

Figure 7. **Ablation of human representation h and appearance feature a in FDI.** Human representation **h** provides geometry constraints from human priors and coarse color information, and further integrates fine-grained information from appearance features **a** with FDI.

C.4. Canonical K-means *vs*. Canonical Grid Voxelization

In canonical body grouping, we employ the k-means clustering to get the grouping dictionary. Actually, using grid voxelization under the canonical space is also feasible. However, the uniform grid leads to the large variance of vertex number in each voxel considering the shape of human body, as shown in Fig. 6. Therefore, we use k-means instead for a more uniform split. As illustrated in Table 12, canonical k-means performs better than canonical grid voxelization.

D. Additional Visualization Examples

We provide more visualization examples in this section.

D.1. Ablation of FDI

To better illustrate the functional difference between human representation **h** and appearance feature **a** in FDI, we provide the ablation in Fig. 7. Obviously, the human representation **h** contains geometry constraints from human priors with coarse color information, while **a** shows more vivid colors with poor geometry. Hence, we propose to take the coarse human representation as the guidance for integrating

proper fine-grained details from the appearance feature.

D.2. Comparisons with State-of-the-art

We provide more comparison examples with previous state-of-the-art methods in Fig. 8.

E. Human Split

We list the detailed human split information in Table 13. We hope that it can serve as a standard split for the following researchers. The code will also be available upon acceptance.
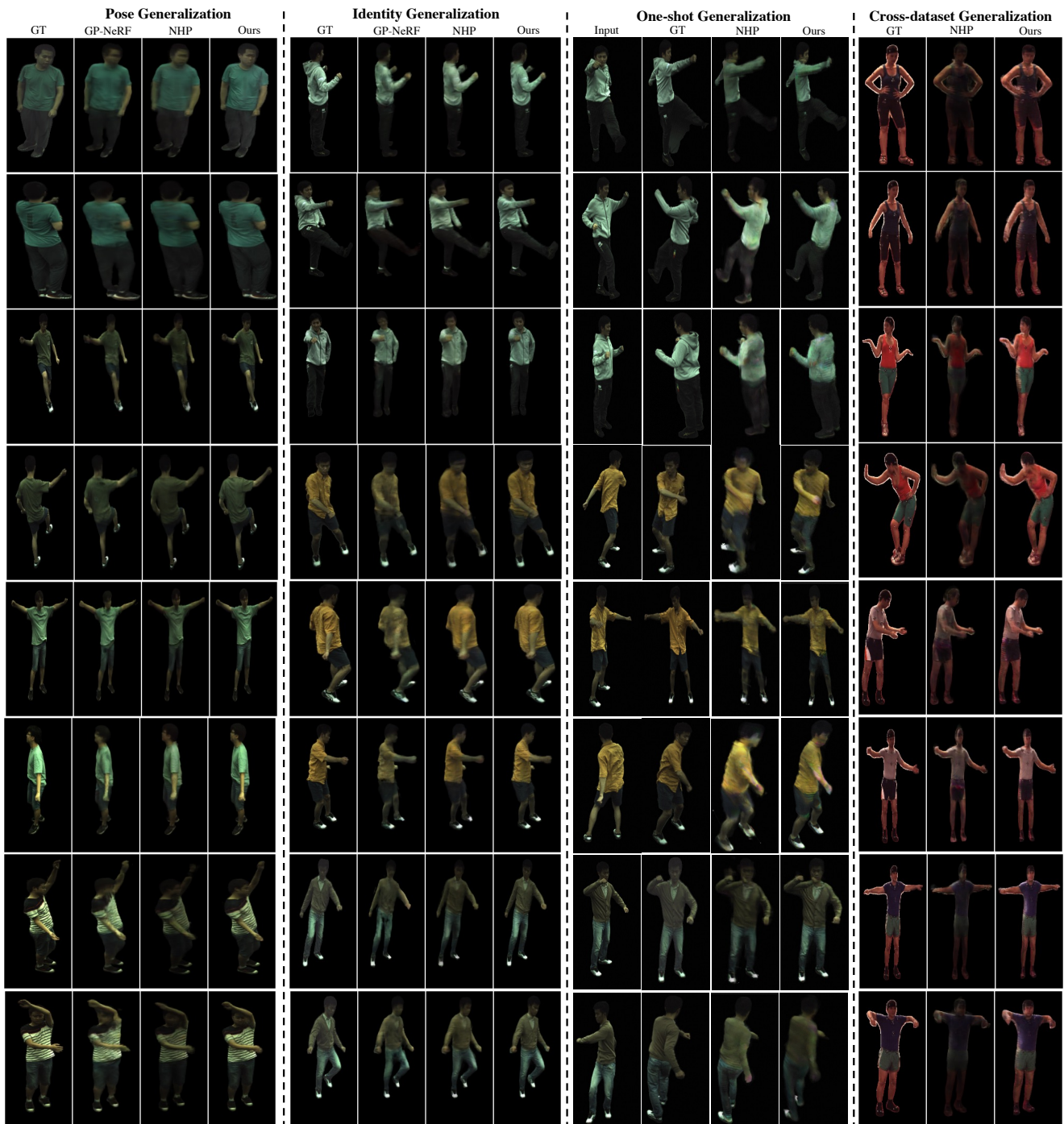
Figure 8. **Supplemented visualization examples on ZJU-MoCap [29] (pose generalization, identity generalization, one-shot gener-alization) and H36M [15] (cross-dataset generalization).**

Human ID	[Start, End)	Interval	Frame Number	Total Frames	Reference View	Target View
			Training Frames			
313	[0, 60)	1	60 × 21		Rand 3	Rand 1
315	[0, 400)	6	67 × 21		Rand 3	Rand 1
377	[0, 300)	30	10 × 23		Rand 3	Rand 1
386	[0, 300)	6	50 × 23	7589	Rand 3	Rand 1
390	[700, 1000)	6	50 × 23		Rand 3	Rand 1
392	[0, 300)	6	50 × 23		Rand 3	Rand 1
396	[810, 1080)	5	54 × 23		Rand 3	Rand 1
			Pose Generalization			
313	[60, 1060)	30	34 × 6		0, 7, 15	3, 5, 10, 12, 18, 20
315	[400, 1400)	30	34 × 6		0, 7, 15	3, 5, 10, 12, 18, 20
377	[300, 617)	30	11 × 6		0, 7, 15	3, 5, 10, 12, 18, 20
386	[300, 646)	30	12 × 6	798	0, 7, 15	3, 5, 10, 12, 18, 20
390	[0, 700)	30	24 × 6		0, 7, 15	3, 5, 10, 12, 18, 20
392	[300, 556)	30	9 × 6		0, 7, 15	3, 5, 10, 12, 18, 20
396	[1080, 1350)	30	9 × 6		0, 7, 15	3, 5, 10, 12, 18, 20
			Identity Generalization			
387	[0, 654)	30	22 × 6		0, 7, 15	3, 5, 10, 12, 18, 20
393	[0, 658)	30	22 × 6	438	0, 7, 15	3, 5, 10, 12, 18, 20
394	[0, 859)	30	29 × 6		0, 7, 15	3, 5, 10, 12, 18, 20
			One-shot Generalization			
387	[0, 654)	30	22 × 6		0	3, 5, 10, 12, 18, 20
393	[0, 658)	30	22 × 6	438	0	3, 5, 10, 12, 18, 20
394	[0, 859)	30	29 × 6		0	3, 5, 10, 12, 18, 20
			Cross-dataset Generalization			
S1	[0, 750)	150	5 × 1		0, 1, 2	3
S5	[0, 1250)	150	9 × 1		0, 1, 2	3
S6	[0, 750)	150	5 × 1		0, 1, 2	3
S7	[0, 1500)	150	10 × 1	54	0, 1, 2	3
S8	[0, 1250)	150	9 × 1		0, 1, 2	3
S9	[0, 1300)	150	9 × 1		0, 1, 2	3
S11	[0, 1000)	150	7 × 1		0, 1, 2	3

Table 13. **Detailed human split.** For ZJU-MoCap [29] (training frames, pose, identity, and one-shot generalization), we follow the human split from the officially released code of NHP [18], while for H36M [15] (cross-dataset generalization), we follow the split from [28]. Note that in ZJU-MoCap, "313" and "315" only contains 21 camera views, while the left ones have 23 camera views.