NeuHMR: Neural Rendering-Guided Human Motion Reconstruction

Tiange Xiang, Kuan-Chieh Wang, Jaewoo Heo,
Ehsan Adeli, Serena Yeung, Scott Delp, Li Fei-Fei
Stanford University
xtiange@stanford.edu

Abstract

Reconstructing 3D human movements from video sequences is an important task in the fields of computer vision, graphics, and biomechanics. Although much progress has been made to infer 3D human mesh based on visual contexts provided in video sequences, generalization to in-the-wild videos still remains challenging for existing human mesh recovery (HMR) methods. To overcome inaccurate prediction, they can perform a second step optimization that refines the inaccurate estimations continuously at test time. Most optimization methods seek fitting of the body joints in the image space with respect to pseudo ground truth predicted by an off-the-shelf key point detector. However, state-of-the-art detectors still introduce errors, especially for challenging poses. In this work, we rethink the dependency on the 2D key point fitting paradigm and present ***NeuHMR****, an optimization-based mesh recovery framework based on recent advances in neural rendering. Our method builds on Human Neural Radiance Fields that allow the refinement of human meshes through animatable 2D renderings. We evaluated our method on two common benchmarks and validated its effectiveness.*

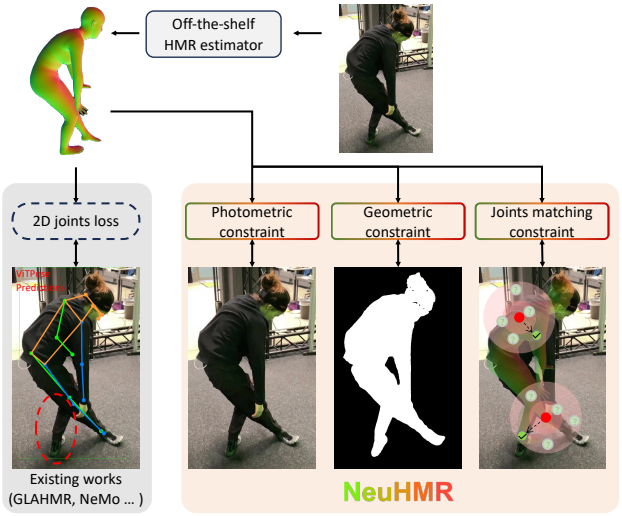

Figure 1. Most existing optimization-based HMR methods rely on pseudo-2D keypoint ground truth estimated by off-the-shelf detectors, but they are bound to make mistakes. See the example in the left part of the figure. In this work, we remove this reliance and demonstrate that using mostly low-level visual constraints based on neural rendering is able to achieve even better pose refinements.

1. Introduction

Reliable 3D reconstruction of a dynamic human requires *synchronized multi-view* captures, which is commonly done prospectively in heavily-equipped laboratories. This fundamental constraint in 3D geometry restricts our ability to curate datasets with ground-truth 3D human meshes. Consequently, it affects not only the size and coverage of such datasets but also hinders the generalization of 3D Human Mesh Reconstruction (HMR) models trained on these datasets. There are two typical types of HMR methods. *Prediction-based* methods infer human meshes from visual contents directly but usually suffer from a lack of high-quality data. *Optimization-based* methods, which refine 3D mesh estimates using additional features from the original video, offer a remedy to this limitation. They rely on minimizing the 2D projection loss with estimated 2D keypoints from off-the-shelf models like OpenPose [4] and ViTPose [50]. Due to the orders of magnitude more datasets with 2D keypoint annotations, covering a wider range of motions, 2D keypoint estimates are often more accurate than reprojected 3D estimates and are easier to generalize to in-the-wild videos. Hence, refining the 3D estimates to match the 2D estimates can improve reconstruction.

However, the inherent flaw with existing optimization-based approaches is that, when the 2D keypoint estimates are incorrect, optimization degrades the original 3D estimates. 2D keypoint estimators are bound to make mistakes (see Figure 1). This paper aims to develop an optimization-based HMR method that relies less on 2D keypoint estimates and achieves better generalization. The ideal optimization-based method should not be bottlenecked by

other off-the-shelf models. Therefore, we seek a data-driven optimization method that obtains its guidance signal from the input target video.

We lean on recent advances in 3D human avatar reconstruction from RGB videos [12] and images [1]. The topic of avatar reconstruction focuses on building an articulated digital human with a matched *appearance*. Intuitively, if the groundtruth avatar of a given video is available, then a pose optimization algorithm can be guided purely by a photometric loss between the rendered avatar and the target human in the input frames [52], removing the dependency on off-the-shelf model predictions.

In this paper, we propose NeuHMR, an optimization-based HMR method based on recent advances in neural rendering. We focus on the setting where a monocular RGB video is given as the input. Using only 2 frames from the input video, NeuHMR first infers a rough estimate of the avatar using our own generalizable human NeRF model. In the second step, we determine a set of anchor frames in the video to be used for registering reliable joint features. In the last step, we use the human avatar, and our proposed optimization strategy to refine the estimated human mesh by comparing the rendered avatar to the input video from different perspectives. Our proposed optimization strategy mostly benefits from the design of the joint matching constraint that enables joint-wise optimization by looking for references from a photometric approach. Empirically, we show that NeuHMR can improve state-of-the-art HMR methods, like Hybrik [23] and HMR2.0 [9]. We evaluated NeuHMR on the recent EMDB dataset [18] which contains multiple challenging in-the-wild videos.

2. Related Work

Human mesh recovery. Recovering human poses and shapes from visual content is a longstanding research problem. Early attempts began with reconstructing human surfaces from full-body scans [2] and configuring structured surface models using body shape coefficients and joint rotations [28]. However, directly modeling humans in the wild from a single RGB image presents considerable challenges. Subsequent works have utilized visual cues and semantic contexts to assist in mesh recovery. HMR [16] pioneered the approach of learning SMPL parameters directly from raw RGB images using an end-to-end architecture. Concurrently, [33] was developed to learn human shape and pose from body silhouettes and joint heat maps. Building on this, Hybrik [23] was proposed to derive 3D skeletons from relative joint rotations and assist in recovering physically realistic 3D meshes. PARE [21] addresses the challenge of mesh recovery under occlusions by first segmenting body parts in images and then using the segmentation to guide part-aware pose regressions. More recently, HMR2.0 [9] introduced an advanced transformer architecture trained on large-scale datasets for more precise recovery of human mesh.

There are also methods that leverage temporal information for better modeling human meshes. Among these, HMMR [17] is one of the earliest works that captures human motions through temporal convolutions. VIBE [20] is another milestone in this line of study that incorporates gated recurrent units to learn SMPL parameters from adjacent frames. More advanced methods utilize transformers to compute self-attention on both spatial and temporal axes [47]. One main component in the training for all of the above methods is the fitting of projected 2D joints to pseudo ground truth. Our methods remove such reliance and find 2D references based on neural rendering.

Optimization methods for human pose. Apart from the aforementioned prediction-based HMR methods, optimization-based approaches strive to iteratively refine human meshes by referring to visual cues implied in the images. SMPLify [3] refines human meshes through iterative fitting based on 2D keypoint priors. Subsequently, more recent methods have advanced human mesh optimization by employing either learned priors [40] or physics-based priors [8, 37, 39, 41]. GLAMR [54] infills inaccurate motion segments in a long sequence and optimizes global motion trajectories. It achieves coherent human motions and stabilized movement predictions. SLAHMR [51] further advances this approach by decoupling the modeling of the camera and the human body, thus enhancing robust optimization of both human motion and global 3D trajectories. Recently, NeMo [46] was introduced to learn the mapping from movement phases to their corresponding SMPL parameters using repetitive motion patterns. Its superior performance has led us to adopt it as the SMPL regressor. TRACE [44] proposes a novel 5D representation comprised of space, time, and identity that enables precise end-to-end reasoning about human motion in both camera and world coordinates. Lastly, WHAM [42] accurately reconstructs 3D human motion in global coordinates while accounting for real-world dynamics by integrating 2D-to-3D keypoint lifting with video features and leveraging camera angular velocity from SLAM methods to estimate global trajectories.

Human rendering. Rendering and animating human avatars has become a trending research direction since the emergence of Neural Radiance Field (NeRF) [30]. Most past works target at photo-realistic rendering of humans from videos captured at multiple camera angles [7, 13, 24, 25, 34, 36, 55, 56]. There is another line of research focus that renders human from videos captured by a single camera [12, 15, 48, 53]. Among which, HumanNeRF [48] is a milestone that for the first time achieves high-quality free-view rendering of dynamic humans from only a monocular video. Subsequent works include OccNeRF [49] and Vid2Avatar

[12] improve human rendering qualities a step further.

One major limitation of neural rendering methods is that the human avatar has to be per-video optimized and can hardly generalize across different videos. NHP [22], TransHuman [32], and GP-NeRF [5] are representatives that achieve generalizable rendering for dynamic humans under multi-view settings. Recently, ActorsNeRF [31] was proposed to extend generalizable rendering on monocular videos. We follow the insights of both GP-NeRF and ActorsNeRF and built a generalizable human NeRF to acquire reliable appearances for any given video.

3. Methods

In this section, we first provide the preliminaries and review basic concepts in human neural radiance fields (section 3.1). Then, we describe our proposed method which consists of two components: a Generalizable Human NeRF (**GHN**) model (section 3.2) and an optimization-based HMR method, **NeuHMR** (section 3.3). Our core idea is to optimize inaccurately estimated human poses with the help of the pretrained GHN model. The overall framework of NeuHMR is outlined in Figure 2. We provide a summary of all defined symbols in the appendix for better reference.

3.1. Preliminaries

Problem formulation. We aim to reconstruct precise 3D human mesh from a monocular video. We follow the paradigm of model-based HMR which uses the SMPL [28] body model. Given an input video, V, with T frames, the target 3D human mesh is represented by a sequence of the 24 major body joints $\mathbf{J}$ and their angles $\theta_{1:T} \in \mathbb{R}^{24\times3\times T}$.

NeuHMR is an optimization-based HMR method. It refines the initial estimate from a HMR method (*e.g.* HMR2.0 [9]). Unlike existing optimization methods that rely on 2D keypoint detectors (*e.g.* ViTPose [50]), we rely on an estimated appearance model, which we detail in section 3.2.

Background. Neural Radiance Field (NeRF) [30] learns mappings from encoded coordinates of 3D points $\{\mathbf{x} \in \mathbb{R}^3\}$ to their radiance $\mathbf{c}$ and density σ':

$$\mathbf{c}, \sigma' = \Phi_{\varphi}^{\text{NeRF}}(\mathbf{x}), \tag{1}$$

where φ denotes the parameters of the NeRF. When casting ray samples $\mathbf{x}$, NeRF utilizes volume rendering [27] to aggregate $\mathbf{c}$ and σ' along the rays:

$$\mathbf{w}_i = \alpha(\mathbf{x}_i) \prod_{j<i} \big(1 - \alpha(\mathbf{x}_j)\big), \tag{2}$$

$$\mathbf{W} = \sum_i \mathbf{w}_i, \qquad \mathbf{C} = \sum_i (\mathbf{w}_i \mathbf{c}_i), \tag{3}$$

where $\alpha(\mathbf{x}_i) = 1 - \exp(-\sigma'_i \delta_i)$ and $\delta_i = z_{i+1} - z_i$ is the z-axis distance between two adjacent ray samples. The final 2D renderings are generated by permuting the per-pixel occupancy $\mathbf{W}$ and rgb color $\mathbf{C}$.

NeRF was originally designed for rendering static scenes. In order to rendering moving human with dynamic poses, a static canonical space is first defined [6, 48] and dynamic human poses $\mathbf{p}$ at different frames can be obtained by deforming the canonical pose to the observation spaces at each frame. This deformation also transforms 3D coordinates between the two spaces, which is achieved through a mapping function $\mathcal{T}$. In details, $\mathcal{T}$ performs a weighted sum of a set of K motion bases defined by rotations R_i and translation t_i of the i_{th} bone of the human body [48]:

$$\hat{\mathbf{x}} = \sum_i^K w_i(\mathbf{x})(R_i\mathbf{x} + t_i), \tag{4}$$

where R_i and t_i can be directly computed from $\mathbf{J}$ and θ. w_i serves as the weights in the observation space, which can be derived from canonical weights [48].

3.2. Generalizable Human NeRF

Given the recent advances in human NeRFs, we propose to use a Generalizable Human NeRF (GHN) model for modeling the appearance. Concretely, we seek a GHN model that takes a video V as input and outputs the parameters of an appearance model parameterized by φ:

$$\varphi = \Phi^{\text{GHN}}(V). \tag{5}$$

The key characteristics of our GHN model are: *(1)* it should take a monocular video as input and *(2)* it should generalize to new videos. Namely, in contrast to human NeRF methods that rely on *multi-view* captures, we need a model that works with monocular input. And, we should not optimize the NeRF from scratch given any new videos.

GHN model architecture. Generalizable rendering requires the framework to learn explicit appearance and geometry semantics from visual cues implied in the video. Considering this, we borrow insights from GP-NeRF [5], which takes as inputs multiple images $\{\mathbf{I}\}$ as conditions to render appearance coherent human avatar with any given poses. For a given image $\mathbf{I}$, we regard the estimated SMPL mesh $\{\mathbf{v}\}$ as a coarse geometry descriptor. The first step of our rendering pipeline is to encode 2D image features via a feature extraction backbone $\mathbf{F} = \mathcal{F}(\mathbf{I})$, which is trained from scratch during training. We associate *2D image features* at the corresponding positions of *3D mesh vertices* when projecting them onto the image space through camera projection $\pi(\cdot)$: $\mathbf{F}_{\pi(\mathbf{x})}$.

We then rely on a rendering network ψ consisting of a SparseConvNet [10] and a transformer layer [45] to regress radiance $\mathbf{c}$ and density σ' at $\mathbf{x}$:

$$\mathbf{c}, \sigma' = \psi(\mathbf{F}_{\pi(\mathbf{x})}, \mathbf{x}). \tag{6}$$

We refer readers to [5] for more details. In the monocular setting, no multi-angle images are available which makes

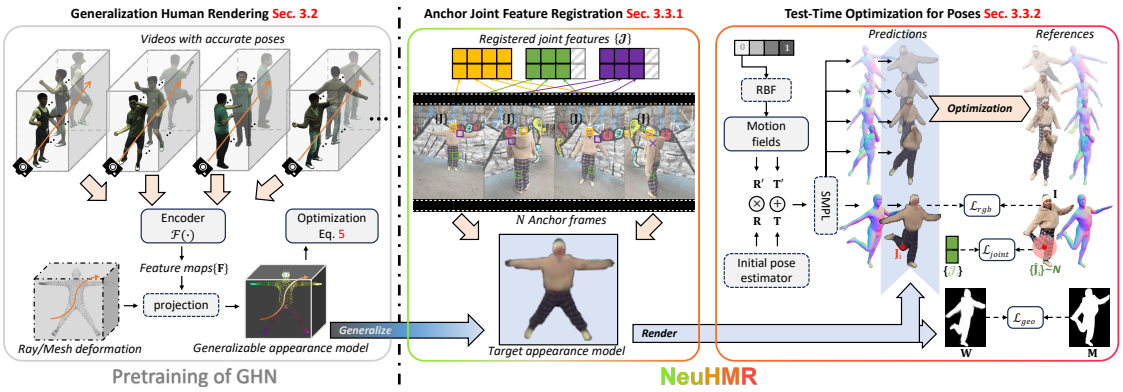

Figure 2. **NeuHMR** is an optimization-based human mesh recovery method that consists of several sequential steps. We first pre-train a Generalizable Human NeRF (GHN) with ground truth poses that learns reliable mappings between 2D images and 3D human appearance/geometry. Then, the learnt GHN is generalized to the target video and is able to render corresponding 2D avatars with the initially estimated poses $\{\mathbf{J}\} \subset \mathbf{p}$. With the human renderings, a set of N anchor frames are first determined as the ones that span a wide range of global orientations and yield the least photometric and geometric losses. 2D features at visible human body joints $\{\mathbf{F_J}\}$ on these anchor frames are collected together and registered as common joint feature descriptors $\{\mathcal{J}\}$. Based on the generalized appearance model and $\{\mathcal{J}\}$, we optimize inaccurate posses through photometric $\mathcal{L}_{rgb}$, geometric $\mathcal{L}_{geo}$, and joints matching constraints $\mathcal{L}_{joint}$.

the above pipeline infeasible. Knowing that moving human shows different body parts to the camera in different frames of a monocular video. Inspired by ActorsNeRF [31], we first pick multiple anchor frames throughout the monocular video and transform the 3D ray samples from the anchor frame spaces to the target observation spaces.

Training Losses. After obtaining per pixel color and density via Equation 3, the framework is mainly optimized towards the input video itself through pixel-wise photometric loss [30]. Moreover, we noticed that due to the huge cross-subject diversity in the dataset, the rendered human body is usually incomplete in geometry. We propose to regularize the accumulation of ray densities to be close to the binary human segmentation mask. Overall, the training of our GHN is supervised by the combination of losses:

$$\sum_i (||\mathbf{C}_i - \mathbf{C}'_i||_2^2 + \frac{1}{2}||\mathbf{W}_i - \mathbf{M}_i||_2^2), \tag{7}$$

where $\mathbf{C}'_i$ is the RGB value from $\mathbf{I}$ at the pixel position of i, $\mathbf{M}$ is the binary human segmentation mask of $\mathbf{I}$ that can be pre-computed by an off-the-shelf model (*e.g.* SAM [19]).

3.3. NeuHMR

Given the trained GHN and a new target video, NeuHMR starts with the estimated appearance model and performs optimization in two steps. In the first step, we preprocess the target video to determine N anchor frames based on rendering errors and register anchor features of each body joints by sampling on the 2D feature maps (section 3.3.1). In the second step, we present a way to generalize the trained appearance model onto arbitrary humans and rely on the 2D renderings and pre-processed anchor joint features to optimize inaccurate poses at test-time (section 3.3.2).

3.3.1 Anchor Joint Feature Registration

Optimization-based methods usually rely on an assumption that in a few video frames, the human poses are easier to capture and the pose estimations are rather accurate [51, 54]. Our method follows the same intuition and condition the GHN on n of the N anchor frames.

Anchor frame selection. We determine the anchor frames through two criteria. First, the anchor frames must be the ones that have relatively accurate initial pose estimations. We approximate pose accuracy according to photometric and geometric losses computed between the humans in the frames and their corresponding renderings (later in section 3.3.2 Another criteria is to include as diverse global orientations of the poses as possible. Since a body joint may have very different semantics when being viewed at different angles (e.g. front and back of head), it is important to ensure the poses in the anchor frames are from a complete coverage at multiple angles. We ensure a wide range coverage of body orientations by first projecting the global rotation matrix $\in \mathbb{R}^3$ onto the surface of a unit sphere. We then run Farthest Point Sampling (FPS) [38] on the surface vectors to sift $\hat{N} > N$ candidates. The N anchor frames are determined as the ones with the least loss.

Joint feature registration. After defining N relatively well-aligned anchor frames, we can extract reliable 2D joint features. The registration is achieved via collecting 2D features for each of the joints $\{\mathbf{J}\}$ on the N anchor frames through grid sampling. However, it is common that some joints are invisible to the camera due to self-occlusion. Considering this, for each frame t, we perform Z-buffer tests using the rendered human depth map $\mathcal{D}^t$ and the z-axis distances of the body joints $\{z_i^t\} \in \mathbf{J}_i^t = \{x_i^t, y_i^t, z_i^t\}$ in the

camera space. Each pixel of $\mathcal{D}^t$ indicates the distance between the camera and a closest point on the human surface. We determine the visibility $\mathcal{V}_i^t$ of a joint $\mathbf{J}_i^t$ via:

$$\mathcal{V}_i^t = \mathbb{1}(|\mathcal{D}^t_{\mathbf{J}_i^t} - z_i^t| < \epsilon), \tag{8}$$

where $\mathbb{1}(\cdot)$ is an indicator that returns 1 for true condition and 0 otherwise, $\epsilon = 0.2$ is a tolerance threshold. $\mathcal{V}_i^t = 1$ indicates joint $\mathbf{J}_i^t$ locates closest to the camera and not been occluded by other parts of the body. We then register the joint features as the average of all visible joints throughout all N anchor frames:

$$\mathcal{J}_i = \frac{1}{|\mathcal{V}_i|} \sum_{t<N} (\mathcal{V}_i^t \mathbf{F}_{\mathbf{J}_i^t}). \tag{9}$$

The registered joint features $\{\mathcal{J}\}$ will be used in the joints matching constraint for optimization (later in section 3.3.2).

3.3.2 Test-Time Optimization for Poses

Motivation. Given the estimated appearance model, the goal of this phase is therefore utilizing such 2D renderings to guide the refinement of inaccurate human poses. Specifically, we account for three constraints for optimization, which try to align photometric color, body geometry, and joint key points between renderings and video frames in the 2D space. In this step, the appearance model is frozen from being updated.

Motion fields. To refine frame-wise human poses, we build a neural motion field (NeMo) [46] to map every time stamp in a video to the corresponding pose parameters. Specifically, it is parameterized as a MLP $\Phi_\eta^{\text{MLP}}(\cdot)$ with trainable weights η that takes as input an embedded scalar value $t \in [0, 1]$ indexing the progress of the action in a given video. It outputs residuals to be appended on all the joint angles, global orientation as well as translation:

$$\begin{aligned} \{\mathbf{R}'_{joints}, \mathbf{R}'_{global}, \mathbf{T}'\} &= \Phi_\eta^{\text{MLP}}(\texttt{RBF}(t)), \\ \Rightarrow \mathbf{R}_{joints} &:= \mathbf{R}_{joints} \cdot \mathbf{R}'_{joints}, \\ \Rightarrow \mathbf{R}_{global} &:= \mathbf{R}_{global} \cdot \mathbf{R}'_{global}, \\ \Rightarrow \mathbf{T} &:= \mathbf{T} + \mathbf{T}', \end{aligned} \tag{10}$$

where $\mathbf{R}_{joints}/\mathbf{R}'_{joints}/\mathbf{R}_{global}/\mathbf{R}'_{global} \in \mathbb{R}^{K\times 3\times 3}$, $\mathbf{T}/\mathbf{T}' \in \mathbb{R}^3$, and $\texttt{RBF}(\cdot)$ is the radial basis function.

Photometric constraint $\mathcal{L}_{rgb}$. One of the primary benefits of optimizing human poses over 2D renderings is the access to visual signals that can be used to guide the correction of poses. We simply adopt the L2 norm of per-pixel distance (similar to the training of GHN) to apply the photometric constraint. With the correct pose, this constraint will lead to minimum possible loss values.

Geometric constraint $\mathcal{L}_{geo}$. However, due to the limits of rendering quality, the rendered rgb images on a new video are usually over-smoothed missing lots of local contexts. Without high quality pixel-pixel correspondences, optimizing the poses with photometric constraint solely is sub-par. Here we apply another constraint to enforce matching between the geometric silhouette of the rendered human and the binary human segmentation mask. We borrow the Jaccard index loss from binary segmentation tasks that computes intersections over unions on the accumulation of ray densities $\mathbf{W}$ and the segmentation mask $\mathbf{M}$:

$$\mathcal{L}_{geo} = 1 - \frac{\sum_i (\mathbf{M}_i \cdot \mathbf{W}_i)}{\sum_i \mathbf{M}_i + \sum_i \mathbf{W}_i - \sum_i (\mathbf{M}_i \cdot \mathbf{W}_i)}. \tag{11}$$

Noteworthy, this geometric constraint differs from that in Equation 7. We observed that using Equation 7 for pose optimization can lead to the rendered human being smaller than the target. Equation 11 does not suffer from this issue.

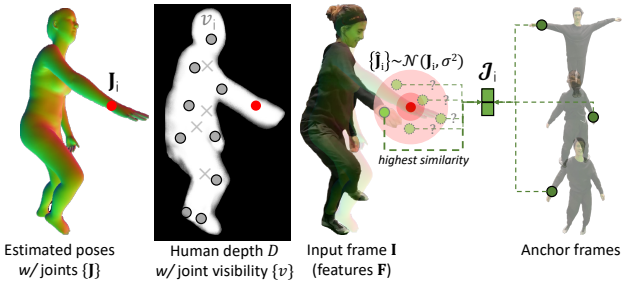

Figure 3. Optimizing poses through joints matching constraint.

Joints matching constraint $\mathcal{L}_{joint}$. Geometric constraint is helpful in terms of aligning the overall human shape, but still insufficient to correct positioning of detailed body joints. Without relying on 2D pseudo ground truth, here we present an alternative approach to find reliable optimization targets for the joints with the help of the registered joint features $\{\mathcal{J}\}$. See Figure 3 for an illustration.

Our method is built on an observation that, state-of-the-art human pose estimators only generate mild errors [9, 23], therefore the correct joints should locate somewhere near to the initial estimations. Following this intuition, we model the distribution of possibly correct joints for $\mathbf{J}$ as Gaussian $\mathcal{N}(\mathbf{J}, \sigma^2)$ centered at $\mathbf{J}$ with a self-defined σ. For each joint $\mathbf{J}_i$ at each video frame $\mathbf{I}$, S samples are randomly drawn from the Gaussian $\{\hat{\mathbf{J}}_i\} \sim \mathcal{N}(\mathbf{J}_i, \sigma^2)$ to be as potential candidates for the correct joint locations $|\{\hat{\mathbf{J}}_i\}| = S$. If the human body is properly rendered, there exists at least one candidate joint $\hat{\mathbf{J}}_i$ on the *input frame* $\mathbf{I}$ that shares similar semantics to the registered joint features $\mathcal{J}_i$. We rely on the same encoder network $\mathcal{F}(\cdot)$ in the GHN to extract features at candidate 2D joint positions $\mathbf{F}_{\{\hat{\mathbf{J}}\}}$. We define the joints matching constraint as:

$$\begin{aligned} \mathcal{L}_{joint} &= \frac{1}{|\{\mathbf{J}\}|} \sum_i ||\hat{\mathbf{J}}_i^{\omega(i)} - \mathbf{J}_i||^2, \\ \omega(i) &= \arg\max_k \cos(\mathbf{F}_{\{\hat{\mathbf{J}}_i^k\}}, \mathcal{J}_i), \end{aligned} \tag{12}$$

ZJU-MoCap	init.	MPJPE↓	PVE↓
***prediction**-based HMR methods*			
VIBE [20]	-	101.3	113.0
HMR2.0 [9]	-	78.0	94.7
Hybrik [23]	-	77.3	86.5
***optimization**-based HMR methods*			
SLAHMR [54]	HMR2.0	75.7	87.1
GLAMR [54]	HMR2.0	93.1	120.0
GLAMR [54]	Hybrik	91.5	111.6
Ours	HMR2.0	74.7	89.6
Ours	Hybrik	**74.2**	**86.9**

Table 1. Quantitative comparisons on ZJU-MoCap 393 [35].

EMDB	init.	MPJPE↓	PVE↓
***prediction**-based HMR methods*			
VIBE [20]	-	125.9	146.8
HMR2.0 [9]	-	115.0	129.7
Hybrik [23]	-	103.7	117.2
***optimization**-based HMR methods*			
NeMo [46]	HMR2.0	158.1	202.9
GLAMR [54]	HMR2.0	120.5	138.8
GLAMR [54]	Hybrik	113.6	133.4
Ours	HMR2.0	105.7	125.2
Ours	Hybrik	**95.4**	**109.6**

Table 2. Quantitative comparisons on EMDB protocol 1 [18].

where cos stands for cosine similarity. As the optimization proceeds, poses are being corrected continuously and getting closer to the correct joint positions. We repeat the above process to re-determine $\{\hat{\mathbf{J}}\}$ at each step.

We optimize NeuHMR through aggregation of the three constraints combined with different weight scales $\lambda_{(\cdot)}$:

$$\arg\min_{\eta}(\lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{geo}\mathcal{L}_{geo} + \lambda_{joint}\mathcal{L}_{joint}). \quad (13)$$

4. Experiments

Please see supplementary materials for more experiments.

4.1. Datasets

ZJU-MoCap [35]. This dataset contains videos captured from 23 angles, recording various human activities in an indoor motion capture setting. Given its high-quality annotations, this dataset is commonly used as a benchmark for human rendering. Following the common training protocol [5, 48], we train a GHN using all camera angles for the 8 subjects (313, 315, 377, 386, 387, 390, 392, 394) and evaluate the pose optimization results on the first camera angle for the left-out subject (393). This dataset is only used to demonstrate a proof of concept.

EMDB [18]. To evaluate our method on more challenging in-the-wild videos, we adopt EMDB, the most recent HMR benchmark, for a comprehensive study. Benchmarks on this dataset follow two protocols, as suggested, we compare methods on protocol 1, which includes videos for the most challenging 17 activities performed by 8 different actors and encompasses both indoor and outdoor scenes. As this dataset is not commonly utilized as a benchmark for human rendering, humans in some videos are only partially visible to the camera, thus posing significant challenges for human rendering and NeuHMR in general.

4.2. Evaluations

Comparison. NeuHMR is mainly compared against three state-of-the-art optimization based methods: GLAMR [54], SLAHMR [51], and NeMo [46]. For comprehensive comparisons, NeuHMR is experimented to optimize upon multiple different pose estimations generated by VIBE [20] — a classic video HMR method, Hybrik — the best performing method on EMBD [18], and HMR2.0 [9] — the most recent state-of-the-art HMR method.

Metrics. We rely on the commonly used Mean Per Joint Position Error (MPJPE) and Per Vertex Error (PVE) in millimeters (mm) to measure the distance between optimized SMPL joints/vertices to the ground truth references in the 3D space. Since the human poses estimated by different methods are in different 3D coordinates, we transform all of SMPL joints/meshes to the camera space by using corresponding camera parameters.

4.3. Implementation Details

Details for generalizable human rendering. We pretrained a GHN on all frames from all camera angles of the 8 ZJU-MoCap subjects to cover as diverse human movements as possible. We adopted ConvNeXt-tiny [26] as the backbone feature extractor $\mathcal{F}(\cdot)$, and used an additional 1×1 convolution to obtain the 2D feature maps $\mathbf{F}$ at $\frac{1}{4}\times$ resolution [5]. We set the number of input frames to the GHN, n, to 2, which is the minimum number that could possibly cover the complete view of a human (front and back). To better decouple the human from the background during training, 512 rays are sampled on both the human body area and the background area. For both training and inference, we cast only 32 samples per ray to enhance rendering efficiency. We adopted the AdamW [29] optimizer with a weight decay of $1e^{-4}$ and an initial learning rate of $1e^{-4}$, which is exponentially decayed to $1e^{-5}$ over 35,000 steps.

Details for NeuHMR. For better efficiency, we resize the ZJU-MoCap videos to $\frac{1}{4}\times$ and the EMDB videos to $\frac{1}{8}\times$ their original resolutions. During the anchor joint feature registration, we employ FPS to select $N = \frac{T}{10}$ anchor frames per video. After registration, we construct a motion field [46] with four simple MLP layers to regress the resid-

uals (Equation 10) from the 100-dimensional RBF features at each frame. For the joint matching constraint, we sample up to $S = 128$ joint candidates from the Gaussian distribution with $\sigma = 4$. We adopt the AdamW optimizer to refine the motion field over 5 epochs. The learning rate is initially set at $1e^{-6}$ and gradually decays to $1e^{-7}$, following a Cosine Annealing schedule. The loss weights are $\lambda_{rgb} = 2.5$, $\lambda_{geo} = 0.001$, and $\lambda_{joint} = 0.003$. Furthermore, we observed that the 24 SMPL joints are densely located when projecting onto images and some of them share very similar semantics due to the limited resolution of $\mathbf{F}$. Consequently, we disregard crowded joints at the body trunk, hands, and feet, resulting in the utilization of only 16 out of the 24 SMPL joints for the joint matching loss. We also add a simple joint regularization term to penalize high-degree rotations at the head, hands, and feet.

4.4. Results

We first report quantitative results on the left-out ZJU-MoCap sequence in Table 1 to demonstrate the feasibility of NeuHMR. Hybrik, without employing optimization-based HMR methods to refine the poses, achieves the best MPVPE of 85.5. GLAMR, a global optimization-based method, prioritizes refining the global trajectory but compromises the metrics for local joints/vertices. Consequently, GLAMR optimized poses yield even lower MPJPE/MPVPE than the initial estimations. Conversely, SLAHMR excels in refining short videos, achieving marginal quantitative improvements. Among all comparison methods, our NeuHMR achieves the best MPJPE and comparable MPVPE.

Following the success on the toy ZJU-MoCap sequence, we conducted comprehensive evaluations on the large-scale EMDB dataset (Table 2). We observed that Hybrik maintains its superior performance as a prediction-based method. Surprisingly, when incorporating NeMo, the human poses become over-smoothed across frames, and local metrics deteriorate further. GLAMR strikes a balance between global and local optimizations without significantly lowering the quantitative performances. NeuHMR outperforms all the counterparts on both MPJPE and MPVPE, validating its effectiveness on in-the-wild videos. We omit SLAHMR results here since it fails on most EMDB videos due to its incapacity to handle long sequences[1].

We present qualitative comparisons on the left-out ZJU-MoCap sequence (against SLAHMR [51]) and two sequences from EMDB (against GLAMR [54]): `outdoor_climb` and `outdoor_big_stairs_down`, in Figure 4. We observe finer articulations at human limbs on our optimized meshes, which again validates that the proposed rendering guidance is able to perform well when pseudo 2D references are limited (e.g. multiple 2D keypoints are crowded together).

[1]An alternative approach is segmenting long videos into smaller clips and optimizing them individually. However, it requires over a thousand SLAHMR runs, which are extremely resource-consuming.

Methods	Time↓ (s per frame)
GLAMR [54]	0.1
SLAHMR [51]	31.6
Ours (joint feature registration)	0.8
Ours (pose optimization)	13.5

Table 3. Comparisons on methodology efficiency.

4.5. Extensive Studies

Analysis of joint matching. As we argued in section 1, existing 2D keypoint detectors are prone to errors, which can adversely affect the performance of optimization when used the key point predictions as pseudo ground truth. With this in mind, NeuHMR employs a photometric approach to optimize finer-grained body joints through the proposed joint matching constraint. Nevertheless, we questioned whether this constraint functions as anticipated and if it indeed outperforms reliance on ViTPose alone. To this end, we conducted extensive experiments, measuring the L1 distances in the 2D space between ground truth joints, ViTPose estimated joints (with confidence score > 0.5), and our matched joints across all EMDB videos. Without losing generality, we calculate the distances on the pelvis —— the most critical joint and the joint that is most likely to be visible in videos. The distance distributions are shown in Figure 5. Although ViTPose did a better job of predicting joints at frames with simple human poses, it occasionally incurs significant mistakes (red arrows) while still remaining highly confident. On average, our matched joints approximate the ground truth with a smaller L1 error of 0.008, which is $\frac{1}{4}$ of the errors made by ViTPose.

Running time comparisons. Each optimization step in NeuHMR requires the complete rendering of a human avatar. Despite our efficient implementation of GHN, we were intrigued to see whether this restriction makes our method significantly slower than other optimization-based HMR methods. We report the average per-frame optimization time in Table 3.

Among these methods, GLAMR is the fastest and SLAHMR has the longest optimization process. Note that, all of NeuHMR, GLAMR, and SLAHMR utilize pretrained models —— GHN for NeuHMR, CVAE motion infiller for GLAMR, and Humor [40] for SLAHMR. We do not consider pretraining as part of the optimization process.

Ablation studies. Ablation results are reported in Table 4. We first disabled the proposed joint feature registration, instead implementing joint matching with features extracted directly from the input frame and its corresponding *rendered image*. This resulted in a significant

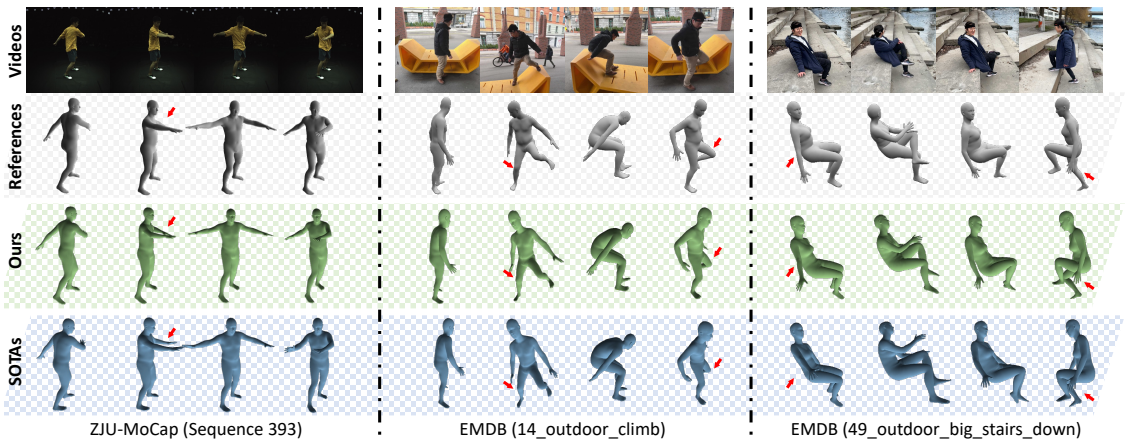

Figure 4. Qualitative comparisons on the refined SMPL meshes for sequences in ZJU-MoCap [35] and EMDB [18]. NeuHMR refined meshes are more aligned to the ground truth at twisted body limbs and crowded joints. Major differences are highlighted.

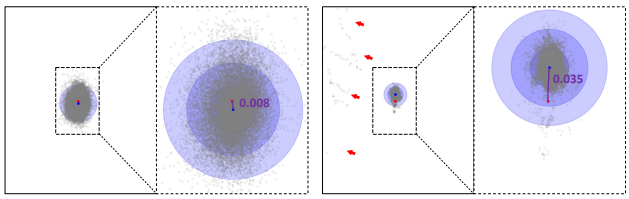

Figure 5. Pelvis distance distributions at two scales. A Gaussian is fit on each distribution. Purple dots are the mean of the Gaussians while red dots are at the coordinate (0,0) indicating no errors.

Ablations	MPJPE↓	PVE↓
w/o $\mathcal{L}_{rgb}$	99.3	115.0
w/o $\mathcal{L}_{geo}$	98.8	114.3
w/o $\mathcal{L}_{joint}$	100.7	116.4
w/o registered joint features	101.1	116.5
Full NeuHMR	**95.4**	**109.6**

Table 4. Ablation results on different constraints of NeuHMR.

performance drop on both MPJPE and PVE, validating the importance of this step. We then assessed the impact of each optimization constraint by individually removing them from the complete training objectives. Removing the proposed $\mathcal{L}_{joint}$ prevented NeuHMR from aligning finer-grained body joints, resulting in the second-worst performances. Surprisingly, we observed that the removal of $\mathcal{L}_{rgb}$ did not preclude the model from achieving competitive performances. This may be attributed to the fact that the humans in most EMDB videos exhibit changing appearances, which could adversely impact pose optimization.

5. Conclusion & Discussion

Limitations. While NeuHMR mainly follows data-driven optimizations through neural rendering. The performance of NeuHMR is correlated to the rendering quality of the GHN. NeuHMR is also prone to subpar human renderings due to appearance inconsistencies across videos. Moreover, our method is designed to focus on per-frame optimization without considering valuable temporal information, which may result in unstable estimations across frames. Please see supplementary materials for examples.

Conclusions. In this work, we present NeuHMR, a human mesh recovery method that optimizes inaccurately estimated human poses. Most of the existing optimization-based counterparts rely on pseudo ground truth 2D keypoints, estimated by an off-the-shelf key point detector, to refine human poses by minimizing reprojection errors. However, we argue that even the most advanced keypoint detectors still make mistakes, and optimizing poses towards the inaccurate 2D references leads to inferior results. NeuHMR is designed to optimize human poses with the help of more reliable low-level visual constraints instead of pseudo-2D keypoints. We start by training a generalizable human radiance field with accurate poses and then generalize the appearance model to any target videos. With human appearances, we determine a set of frames and register features at the corresponding joint positions. Human pose estimations are eventually optimized via the combination of photometric, geometric, and joint matching constraints. We evaluated NeuHMR on two common benchmarks with extensive analysis, which verified the effectiveness of this novel approach. Please refer to the appendix for future works as well as limitations in the current approach.

Acknowledgments. This work was partially funded by the NIH Grant R01AG089169 and P41EB027060, the Gordon and Betty Moore Foundation, the Jaswa Innovator Award, Stanford HAI, Stanford HAI graduate fellowship, and Stanford Wu Tsai Human Performance Alliance.

References

[1] Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia*, 2023. 2

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1

[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 *European Conference on Computer Vision*, pages 222–239. Springer, 2022. 3, 6

[6] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[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*, page 145–156, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 2

[8] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13106–13115, 2022. 2

[9] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023. 2, 3, 5, 6

[10] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 3

[11] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 1

[12] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 2, 3

[13] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *Computer Vision – ECCV 2018*. Springer International Publishing, 2018. 2

[14] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. Kama: 3d keypoint aware body mesh articulation. In *2021 International Conference on 3D Vision (3DV)*, pages 689–699. IEEE, 2021. 2

[15] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 2

[16] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[17] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[18] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14632–14643, 2023. 2, 6, 8

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6

[21] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2

[22] 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*, 34:24741–24752, 2021. 3

[23] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 2, 5, 6

[24] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 419–436. Springer, 2022. 2

[25] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 2

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6

[27] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 3

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3, 1

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[30] 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. 2, 3, 4

[31] Jiteng Mu, Shen Sang, Nuno Vasconcelos, and Xiaolong Wang. Actorsnerf: Animatable few-shot human rendering with generalizable nerfs. *arXiv preprint arXiv:2304.14401*, 2023. 3, 4, 1

[32] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. Transhuman: A transformer-based human representation for generalizable neural human rendering. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3544–3555, 2023. 3

[33] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 2

[34] 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. 2

[35] 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. 6, 8

[36] 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, 2021. 2

[37] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018. 2

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4

[39] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 71–87. Springer, 2020. 2

[40] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 2, 7, 1

[41] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (ToG)*, 40(4):1–15, 2021. 2

[42] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion, 2024. 2

[43] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, pages 744–760. Springer, 2020. 1, 2

[44] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[45] 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

[46] Kuan-Chieh Wang, Zhenzhen Weng, Maria Xenochristou, Joao Pedro Araujo, Jeffrey Gu, C Karen Liu, and Serena Yeung. Nemo: 3d neural motion fields from multiple video instances of the same action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 6

[47] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13211–13220, 2022. 2

[48] 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. 2, 3, 6, 1

[49] Tiange Xiang, Adam Sun, Jiajun Wu, Ehsan Adeli, and Li Fei-Fei. Rendering humans from object-occluded monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3239–3250, 2023. 2

[50] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1, 3

[51] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21222–21232, 2023. 2, 4, 6, 7, 1, 3

[52] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 2

[53] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[54] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 2, 4, 6, 7, 1, 3

[55] Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, and Jingyi Yu. Relightable neural human assets from multi-view gradient illuminations, 2023. 2

[56] Tiansong Zhou, Jing Huang, Tao Yu, Ruizhi Shao, and Kun Li. Hdhuman: High-quality human novel-view rendering from sparse views, 2023. 2