A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose

Anonymous Author(s)
Affiliation
Address
email

Abstract

While deep learning reshaped the classical motion capture pipeline with feed-forward networks, generative models are required to recover fine alignment via iterative refinement. Unfortunately, the existing models are usually hand-crafted or learned in controlled conditions, only applicable to limited domains. We propose a method to learn a generative neural body model from unlabelled monocular videos by extending Neural Radiance Fields (NeRFs). We equip them with a skeleton to apply to time-varying and articulated motion. A key insight is that implicit models require the inverse of the forward kinematics used in explicit surface models. Our reparameterization defines spatial latent variables relative to the pose of body parts and thereby overcomes ill-posed inverse operations with an overparameterization. This enables learning volumetric body shape and appearance from scratch while jointly refining the articulated pose; all without ground truth labels for appearance, pose, or 3D shape on the input videos. When used for novel-view-synthesis and motion capture, our neural model improves accuracy on diverse datasets.

1 Introduction

Generative models have evolved from Generative Adversarial Networks (GANs) recreating images [12, 19] to neural scene representations [35, 51, 52] providing control and image understanding for downstream tasks via structured latent variables. However, most 3D models require 3D labels that cannot be crowd-sourced on natural images and require dedicated depth sensors. It is hence an important research problem to learn 3D representations from 2D observations, which is particularly challenging for humans with diverse body shapes and appearances and their non-rigid motion.

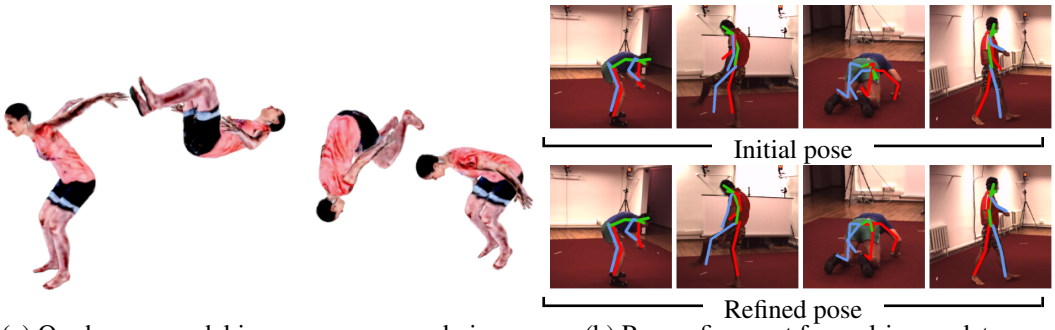
Modern human motion capture techniques typically combine the advantages of discriminative and generative approaches. A feed-forward 3D human pose estimation approach provides a rough initial estimate of human pose. Afterward, a generative approach based on either a high-quality 3D scan of the person [15], or a parametric human body model learned from laser scans [4] refines the estimate iteratively based on the image evidence. Although achieving unprecedented accuracy, existing models require a low-dimensional, restrictive, shape body model or a personalized 3D scan of the user.

We introduce *Articulated Neural Radiance Fields* (A-NeRF) for learning a user-specific neural 3D body model and underlying skeleton pose from unlabelled videos (see Figure 1). When applied to motion capture, it alleviates the need for template models while maintaining the advantages and accuracy of current generative approaches. A-NeRF extends Neural Radiance Fields (NeRF) [35] to work with single videos and articulated motion. NeRF parameterizes the scene implicitly as

$$F_\phi(\Gamma(\mathbf{q}), \Gamma(\mathbf{d})) \mapsto (\sigma, \mathbf{c}), \quad \text{with } \sigma \in \mathbb{R}, \mathbf{c} \in \mathbb{R}^3, \mathbf{q} \in \mathbb{R}^3, \text{ and } \mathbf{d} \in \mathbb{R}^3, \tag{1}$$

by chaining F_ϕ, a Multi-layer Perceptron (MLP), with Γ, the Positional Encoding (PE) [61]. First, the PE maps the input scene point $\mathbf{q}$ and view direction $\mathbf{d}$ to a higher dimensional space that enables the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.



(a) Our human model in unseen poses and views (b) Pose refinement for real-image data

Figure 1: Our A-NeRF optimization for monocular 3D human pose estimation jointly learns a neural body model of the user and works with diverse body poses (left) while also refining the initial 3D articulated skeleton pose estimate from a single or, if available, multiple views without tedious camera calibration (right). Underlying is a surface-free neural representation and skeleton-based embedding coupled with volumetric rendering.

MLP to learn a meaningful scene representation function F_ϕ that subsequently outputs the radiance $\mathbf{c}$ and opacity σ at every point in space. Second, the implicitly described scene (via conditioning on query locations) is rendered via classical ray-marching from computer graphics. The advantage of the MLP representation is that it avoids the complexity of volumetric grids [27], circumvents the artifacts induced by the implicit bias of screen-space convolution [36, 49], and, unlike surface meshes, can have flexible topology. However, the original NeRF only works for static scenes captured from dozens of calibrated cameras such that each 3D point is seen from multiple views.

Our conceptual contribution lies in learning a neural latent representation relative to an articulated skeleton. While explicit models such as the popular SMPL body model [28] deform a surface via forwards kinematics, the implicit form of A-NeRF makes us re-think how skeletons can be integrated—implicit networks require the inverse transformation from 3D world coordinates to the reference skeleton, a significantly harder task that has not been fully explored.

Our core technical novelty is to come up with and evaluate different parameterizations of $\Gamma(\mathbf{q}), \Gamma(\mathbf{d})$ in Eq. 1 to create local coordinates relative to the articulated skeleton. Since a point in 3D world coordinates cannot uniquely be associated with a body part, we resolve the mentioned ill-posed inverse problem by overparameterizing with one embedding per bone. This embeds domain knowledge of how humans move and provides a common frame for the neural network to combine body shape and appearance constraints across the entire captured sequence (see Figure 2).

We demonstrate that all our contributions together enable learning of a neural body model from monocular video, requiring only rough 3D pose estimates for initialization, that reaches a level of detail previously only attained with parametric surface models or multi-view approaches [46].

Scope. We apply the model to motion capture, character animation, and appearance and motion transfer and demonstrate that the pose refinement improves on existing monocular skeleton reconstruction. We learn in the transductive setting, for a specific target video that is known at training time but has no ground truth.

General impact. Building such a self-supervised approach for personalized human body modelling promises to be more inclusive to people and activities that are not well represented in current motion capture datasets. We will make the source code available and easy to use such that this tool becomes available for the general public. It bears the risk that 3D models of people are created from videos without their consent. We urge users to only apply the algorithm on datasets that have been collected for developing and validating motion capture algorithms, as we do here.

2 Related Work

Our approach builds upon and is related to the following work on human pose and shape estimation, human modeling, and neural scene representations [57].

Discriminative Human Pose Estimation. While feed-forward estimation of the 3D joint positions [24, 29–31, 39, 40, 47, 54, 59, 64, 66] or joint angles and bone lengths [48, 67] of the skeleton is highly accurate, such discriminative estimates are prone to misalignment when overlayed onto the input image due to the generalization gap. The skeleton pose can be refined to better match the 2D pose estimates, but this usually leads to larger errors in 3D [32, 33]. We use [21] for initializing skeleton pose and combine it with a neural body model.

Surface-based Generative Body Models. These are obtained by either constraining template meshes via deformation energies [15, 65] or learning parametric human body models from a large collection of laser scans [6, 9, 28]. Their low-dimensional parameters constrain the space of plausible human shapes and motions. This enables real-time reconstructions from single images [7, 13], detailed texturing and displacement mapping [3, 5], and alleviates manual rigging [2]. It also enables optimization within the bounds of the learned prior [11, 14, 23] and weak-supervision when integrated in a differentiable form [25] into neural training processes [4, 18, 21, 38, 41, 60]. Closest to our approach in this category are the model fitting methods by [69] that textures and geometrically refines an untextured parametric quadruped model to zebra images and to [63] that uses optical flow to refine human pose. Although in a similar setting, our surface-free neural body model and volumetric rendering is fundamentally different to their textured triangle mesh that is rigged with forward kinematics and needs to be obtained a-priori.

Implicit Body Models. A-NeRF bears close similarities with body models defined implicitely in terms of level-sets [53] and density of a sum of Gaussians [16, 43, 44] that are used for refining human pose, shape, and appearance via differentiable ray-tracing. However, sum of Gaussians and other primitives only provide rough approximations.

Neural Scene Representations. Recent neural scene representations learn low-dimensional non-linear representations of meshes [26, 58], point clouds [1, 34, 62], sphere sets [22], and dense volumetric grids [27, 50]. Their respective geometric output representations enable rendering with classical rendering techniques but have limited expressiveness, e.g., due to the fixed connectivity of a surface mesh and large memory footprint of discretized volumes. This limitation is overcome by using unconstrained MLPs [35] paired with positional encoding [55] to characterize an arbitrary point in 3D space. Common are surface definitions via level-sets of the MLP that are rendered with sphere tracing [51] and density representations rendered with ray-marching [35]. The rendering step is required for maximum likelihood estimation, to define a likelihood over observable variables—real images—while learning a 3D model. The rendering can be learned too [36, 45, 46] but usually leads to inconsistencies, particularly when training data is scarce.

Closely related to ours is the NASA surface body model [10] that also defines an implicit function as the minimum of individual implicit functions that are rigidly attached to the bones of a skeleton, each conditioned on the entire human pose to model dependencies and learned from 3D scans. By contrast, we learn a volumetric model instead of a surface model and include appearance and rendering. Even more similar is the recent NeuralBody [42] representation, which combines a NeRF with a surface body model and underlying skeleton. By contrast to both approaches, we do not require surface supervision or initialization, condition pose differently, and refine pose, which enables us to learn from single videos in unconstrained environments.

3 Formulation

Objective. Given a sequence $[\mathbf{I}_k]_{k=1}^N$ of N images $\mathbf{I}_k \in \mathbb{R}^{H \times W \times 3}$ stemming from one or several videos of the same person, our goal is to simultaneously estimate the time-varying skeleton poses $[\theta_k]_{k=1}^N$ and learn a detailed body model. Our A-NeRF body model C_ϕ is parametrized by neural network parameters ϕ that define volumetric shape and color while the skeleton captures motion over time. Figure 2 gives an overview of this generative model. It enables a rendering of the virtual body

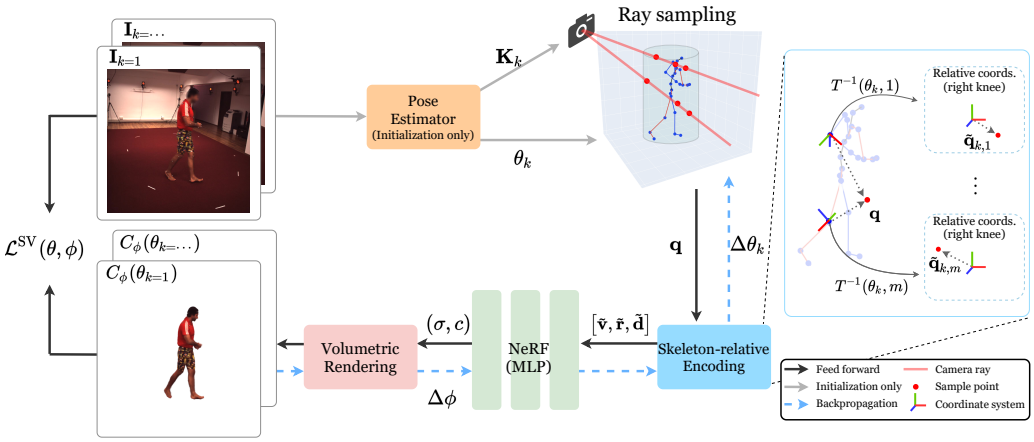


Figure 2: **Overview.** A-NeRF is a generative model that can be rendered and optimized on a photometric loss $\mathcal{L}^{\text{SV}}$ (white). First, the skeleton pose is initialized with an off-the-shelf estimator (orange). Second, this pose is refined via a skeleton-relative embedding (blue) that, when fed to NeRF (green), drives the implicit body model that is rendered by ray-marching (red). A key property of the skeleton-relative embedding is that a single 3D query location maps to an overcomplete reparametrization, with the same point represented relative to each skeleton bone (right).

(a) NeRF (b) Rotate 24° (c) Ours (d) Rotate 24° (e) w/o $\tilde{\mathbf{v}}, \tilde{\mathbf{r}}, \tilde{\mathbf{d}}$ (f) w/o $\tilde{\mathbf{v}}$ (g) Ours

Figure 3: **Importance of our skeleton-relative encodings.** The original NeRF breaks (a) when training on a diverse set of poses and (b) further degrades when the poses are rotated. Even if (e) conditioned directly on θ_k, the NeRF still suffers from artifacts due to the complexity and ambiguity of human articulation. With our skeleton-relative encoding (f, g), the geometry for the subject is consistent under rotation, and the quality is greatly improved, with the full model working best.

model in unseen poses and optimizes its parameters θ and ϕ on the image reconstruction objective

$$\mathcal{L}^{\text{SV}}(\theta, \phi) = \sum_k \underbrace{\|C_\phi(\theta_k) - \mathbf{I}_k\|_1}_{\text{data term}} + \underbrace{\lambda_\theta d(\theta_k - \hat{\theta}_k)}_{\text{pose regularizer}} + \underbrace{\lambda_t \left\| \frac{\partial^2 \theta_k}{\partial t^2} \right\|_2^2}_{\text{smoothness prior}}. \tag{2}$$

with the influence of all three terms balanced by hyperparameters λ_t and λ_θ. The data term measures the distance between the images generated by C_ϕ and the input image with the L1 distance. The pose regularizer encourages the solution to stay close to an initial pose estimate $\hat{\theta}$ obtained from an off-the-shelf predictor [21], tolerating small shifts up to $\epsilon = 0.01$ with $d(x) = \min(\|x\|_2^2 - \epsilon, 0)$. Lastly, the smoothness prior penalizes acceleration $\frac{\partial^2 \theta_k}{\partial t^2}$ between poses of consecutive frames. Minimizing Eq. 2 can be seen as maximizing a corresponding probabilistic model, with the quadratic energy terms being the log-likelihoods of Gaussian distributions. Our focus is on formalizing the neural body model. For simplicity, we continue to write equations in terms of the objective functions used during inference with stochastic gradient descent.

3.1 NeRF and A-NeRF Image Formation Model

Instead of modeling the scene as a collection of triangles or other primitives, we define the human implicitly by a neural network as a function (Eq. 1) defined over all possible 3D points and view directions in space [35]. Likewise, the generated image $C_\phi(u, v; \theta_k)$ has implicit form on the pixel

coordinate (u, v), with the pixel color computed as a sum over queries $\mathbf{q}$ of the scene representation along with view rays $\mathbf{d}$. The ray direction is $\mathbf{d} = \mathbf{K}_k^{-1}(u, v)$, with $\mathbf{K}_k$ the camera intrinsics modeling camera geometry such as the focal length; we estimate it using [21] in our experiments. We define the scene relative to the camera. Hence, no extrinsics are needed. Formally, we write this ray-marching operation as

$$C_\phi(u, v; \theta_k) = \sum_{i=1}^{Q} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i, \quad T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right), \tag{3}$$

with i referring to one 3D query positions $\mathbf{q}_i$ sampled along $\mathbf{d}$ via importance sampling and δ_i the distance to neighboring samples—a constant if samples would be taken at regular intervals. The T_i is the accumulated transmittance for the ray traveling from the near plane to $\mathbf{q}_i$—the fraction of light reaching the sensor from sample point i. The c_i is the light color and intensity emitted or reflected at i. The final pixel color is the sum over all Q samples, with the last sample taking the special role of the background. The background color is easily inferred as the median pixel color over the entire video for static camera setups. We refer to the original paper [35] for additional details.

This volumetric model can render opaque and semi-transparent objects of virtually arbitrary size and enables optimization of its parameters by fitting to example images (data term). Even though humans are largely opaque its smoothness is beneficial for gradient-based optimization of complex shapes. The difficulty lies in optimizing a network that outputs density and radiance without having the supervision of either one. In the following, we introduce our skeleton parametrization θ_k and how to introduce it to model dynamic articulated human motion effectively.

3.2 Articulated Skeleton Pose Model

We use a skeleton representation that encodes the connectivity and static bone lengths via a rest pose of the 3D joint locations $\{\mathbf{a}_m\}_{m=1}^{24}$, with the root at $\mathbf{0}$ that are connected via B bones. Dynamics are modeled with per-frame skeleton poses $\theta_k = [\omega_{k,0}, \cdots, \omega_{k,24}]$ that include the relative rotation of 24 joints, $\omega_{k,m}$, to their parents as well as the global position $\omega_{k,0}$ as a special case. For rotation, we experiment with the recently proposed overparametrized representation of [68] ($\omega_{k,m} \in \mathbb{R}^6$) and the traditional axis-angle representation ($\omega_{k,m} \in \mathbb{R}^3$). The subscript $_{k,m}$ indicates that a variable is related to the m-th joint of image $\mathbf{I}_k$. We will exploit that every bone defines a local coordinate system. We write the mapping of a 3D position $\mathbf{p}_{k,m} \in \mathbb{R}^3$ in the m-th local bone coordinates to world coordinates $\mathbf{q} \in \mathbb{R}^3$ via forward kinematics using homogeneous coordinates,

$$\begin{bmatrix}\mathbf{q}\\1\end{bmatrix} = T(\theta_k, m)\begin{bmatrix}\mathbf{p}_{k,m}\\1\end{bmatrix}, \text{ with transformation } T(\theta_k, m) = \prod_{l \in A(m)} \begin{bmatrix}\mathbf{R}(\omega_{k,l}) & \mathbf{a}_l\\ \mathbf{0} & 1\end{bmatrix} \in \mathbb{R}^{4\times 4}, \tag{4}$$

$A(m)$ the ordered set of the joint ancestors of m, and $\mathbf{a}_m \in \mathbb{R}^3$ the joint location of m at the rest pose. Conversely, $T(\theta_k, m)^{-1}$ maps world to local bone coordinates. The rotation matrix $\mathbf{R}(\omega_{k,l})$ is inferred as in [68]. Note that our skeleton is equivalent to SMPL [28] and others, but without their parametric surface model, and can therefore be initialized with any skeleton pose estimator.

3.3 NeRF Positional Encoding and A-NeRF Pose-Relative Encoding

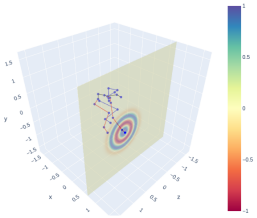
NeRF uses positional encoding, Γ, to augment the input $\mathbf{q}$ and $\mathbf{d}$ with a projection to a $(2L-1)$-dimensional space [35] spanned by L sine and cosine waves of varying frequency,

$$\Gamma(\mathbf{q}) = \Big[\mathbf{q}, \sin\left(2^0\pi\mathbf{q}\right), \cos\left(2^0\pi\mathbf{q}\right), \cdots, \sin\left(2^{L-1}\pi\mathbf{q}\right), \cos\left(2^{L-1}\pi\mathbf{q}\right)\Big]. \tag{5}$$

Each component of the three-dimensional vectors $\mathbf{q}$ and $\mathbf{d}$ are encoded independently. This frequency representation is instrumental for NeRF's success in learning high-frequency details of static scenes [35, 52, 55]. We explored the following variants to generalize NeRF to the dynamic articulated motion of humans, which necessitates a new, time-varying representation.

4D Space-time Encoding. A straightforward extension would be to add time as an additional dimension. However, this led to blurry results in our experiments and is non-interpretable. Note that the number of frames can be hundreds to thousands, which requires some form of aggregation.

Conditioning on pose. Feeding a time-dependent θ_k to the NeRF in addition to the query location is an alternative that would allow the network to learn redundancies better. However, this makes the network learn the complex human body articulation from scratch and leads to artifacts in our ablation.



Skeleton-relative encoding. Our core contribution is to transform the query locations $\mathbf{q}$ and view direction $\mathbf{d}$ relative to the skeleton before determining the color and opacity at that transformed point via NeRF. It is a form of reparameterization that explicitly incorporates domain knowledge of how the human body parts are linked and transformed relative to each other. Intuitively, our implicit formulation turns explicit models, such as SMPL [28], on its head. Instead of deforming the output surface via skinning, the query location is mapped in the inverse direction to local bone-relative coordinate systems before processing through the NeRF network. Our final model uses a combined encoding $\mathbf{e}_k = [h(\tilde{\mathbf{v}}_k)\Gamma(\tilde{\mathbf{v}}_k), \tilde{\mathbf{r}}_k, h(\tilde{\mathbf{v}}_k)\Gamma(\mathbf{d}_k)]$ as input to the NeRF function F_ϕ. Note that the skeleton embedding introduces the desired time dependency, denoted by subscript k. The inlet shows our most crucial contribution, the relative distance encoding $\tilde{\mathbf{v}}_k$ followed by PE with **Cutoff** to reduce the influence of irrelevant bones. We derive the components of our encoding $\mathbf{e}_k$ and the other alternatives below.

- **Reference Pose Encoding** One could compensate motion by attaching the query $\mathbf{q}$ in world coordinates at frame k to the closest bone m and transforming it with

$$\mathbf{a}_k = T(\theta_0, m)T^{-1}(\theta_k, m)\mathbf{q}. \tag{6}$$

  This puts the query relative to the bone m as in frame k but with the skeleton in rest pose θ_0. NeRF could then learn without change in the 3D space of the rest pose as done before for surfaces [56]. However, this cannot capture non-rigid pose-dependent effects, such as muscle bulging, and has ambiguities when $\mathbf{q}$ is at equal distance to two bones.

- **Bone-relative Position (Rel. Pos.)** To remove these ambiguities and the ill-posed association to a single part, we map $\mathbf{q}$ relative to each bone m with,

$$\tilde{\mathbf{q}}_k = [\tilde{\mathbf{q}}_{k,1}, \cdots, \tilde{\mathbf{q}}_{k,24}] \text{ and } \tilde{\mathbf{q}}_{k,m} = T^{-1}(\theta_k, m)\mathbf{q}. \tag{7}$$

  The resulting individual bone coordinates are well suited to model the overwhelmingly rigid motion of the corresponding body part. Moreover, the overparameterization of position by concatenating all local encodings enables learning when complex interactions are necessary. However, such an embedding for all bones increases the dimensionality by an order of magnitude.

- **Relative Distance (Rel. Dist.)** Much simpler to compute are distances from $\mathbf{q}$ to all bones m,

$$\tilde{\mathbf{v}}_k = [\tilde{\mathbf{v}}_{k,1}, \cdots, \tilde{\mathbf{v}}_{k,24}], \text{ with } \tilde{\mathbf{v}}_{k,m} = \|\tilde{\mathbf{q}}_{k,m}\|_2 \in \mathbb{R}. \tag{8}$$

  This radial encoding is used in our final model in favor of $\tilde{\mathbf{q}}$ because it naturally captures spherically shaped limbs, is lower-dimensional, and thereby improves reconstruction accuracy.

- **Relative Direction (Rel. Dir.)** Since the distance encoding is invariant to direction, we additionally obtain the direction vector to capture the orientation information of $\mathbf{q}$,

$$\tilde{\mathbf{r}}_k = [\tilde{\mathbf{r}}_{k,1}, \cdots, \tilde{\mathbf{r}}_{k,24}], \quad \tilde{\mathbf{r}}_{k,m} = \frac{\tilde{\mathbf{q}}_{k,m}}{\|\tilde{\mathbf{q}}_{k,m}\|_2} \in \mathbb{R}^3. \tag{9}$$

  Note that by contrast to all other embeddings, this direction encoding did not profit from subsequent PE. We therefore pass it directly into $\mathbf{e}_k$.

- **Relative Ray Direction (Rel. Ray.)** NeRF models a radiance field that is a function of the position and view direction. To encode the view direction, we transform $\mathbf{d}$ to obtain the relative view-ray direction $\tilde{\mathbf{d}}$, similar to query position,

$$\tilde{\mathbf{d}}_k = [\tilde{\mathbf{d}}_{k,1}, \cdots, \tilde{\mathbf{d}}_{k,24}], \quad \tilde{\mathbf{d}}_{k,m} = [T^{-1}(\theta_k, m)]_{3\times3}\mathbf{d} \in \mathbb{R}^3, \tag{10}$$

  with $[T^{-1}(\theta_k, m)]_{3\times3}$ the rotational part of the bone-to-world transformation $T^{-1}(\theta_k, m)$.

- **Cutoff.** We desire a local embedding where points should not be influenced by all but only nearby bones. To this end, we introduce a windowed version of positional encoding by multiplying the encoding with respect to bone m by $h(\tilde{\mathbf{v}}_{k,m}) = 1 - S(-\tau(\tilde{\mathbf{v}}_{k,m} - t))$, with S the sigmoid step function, t the cutoff point, τ the sharpness. This leaves queries unaffected by distant bones.

Our embedding choice of $\mathbf{e}_k = [h(\tilde{\mathbf{v}}_k)\Gamma(\tilde{\mathbf{v}}_k), \tilde{\mathbf{r}}_k, h(\tilde{\mathbf{v}}_k)\Gamma(\mathbf{d}_k)]$ has the advantage of being invariant to the global shift and rotation of the person and preserves the piece-wise rigidity of articulated motion while still allowing for pose-dependent deformation. In addition to these embeddings, we also consider other variants. See the supplemental material and Section 4 for a detailed discussion.

4 Evaluation

We performed experiments to validate that A-NeRF learns accurate body models and poses, with fewer assumptions (single view, uncalibrated, and w/o a parametric surface model) than the related works. This makes it applicable to fine-grained pose refinement that improves the estimates of state-of-the-art methods. The supplements provide additional comparisons and ablation studies.

Inference and implementation details Our A-NeRF model is learned without supervision on a single or multiple videos of the same person. Camera intrinsics, bone lengths for setting $\mathbf{a}_m$, and pose θ_k are initialized with [21] for every frame k. These poses are then optimized for 500k iterations on objective Eq.2, together with the neural parameters ϕ, which model shape and appearance. We then stop optimizing pose and continue to train ϕ with only the data term for additional 200k iterations, which further improves visual fidelity. We form a training batch by randomly sampling $N_{\text{batch}} = 3072$ rays from all available images. Therefore, an image k can have no or only very few samples, leading to noisy updates of θ_k. To counteract, we accumulate the gradient update $\Delta\theta$ for $G \geq (30N_{\text{batch}})/N$ iterations, so that each θ_k is expected to be sampled $30 \geq$ times before the gradient update. To speed up the rendering part, we use the same coarse-to-fine importance sampling strategy of [35] and restrict samples to be within a coarse human silhouette estimated with [8] unless the background is monochrome. The training is done on 2 Nvidia Tesla V100 32GB GPUs, which takes approximately 60 hours. It takes around 1-4 seconds to render one $512 \times 512 \times 3$ image using a single Nvidia Tesla V100 GPU. We tuned the hyperparameters on Subject 1 of the Human3.6M training set. The supplemental materials provides additional details, including the network architecture.

Datasets. We evaluate on the following benchmarks, and additionally on synthetic data created from **SURREAL** and **Mixamo** characters, which are listed in the supplement.

- **Human 3.6M [17]** The dataset consists of 5 training and 2 testing subjects (S9/S11) with ground truth 3D joint locations. We follow two widely adopted test protocols denoted as Protocol I [18, 20, 21] and Protocol II [37, 54], in which we evaluate 14/17-joint estimation error on every $5^{th}/64^{th}$ frame of the test videos, respectively. See the supplement for details.
- **MPI-INF-3DHP [31]** This dataset is a standard benchmark for human pose estimation. It consists of 4 indoor and 2 outdoor subjects with challenging human poses. The number of frames per subject range from 276 to 603.
- **MonoPerfCap [65]** The dataset consists of human performance video captured with a monocular camera in both indoor and outdoor settings. We use two subjects, Weipeng_outdoor and Nadia_outdoor, for our qualitative experiments. The two subjects have 1151 and 1635 frames, respectively, for training.

Pose Metrics. We report the PA-MPJPE metric, the Euclidean distance between Procrustes-aligned (PA) predictions and ground truth 3D joint position averaged over all frames and joints of the test set. The PA alignment in scale and orientation is essential for comparing approaches that do not assume knowledge of the ground truth calibration and are, hence, ill-posed to the factors that the alignment removes. Following prior work [21, 20, 31], we also report percentange of correct keypoints (PCK) for MPI-INF-3DHP; the percentage of joints that lie within a distance of 150mm to the ground truth.

Novel-View-Synthesis and Character Animation. Our body model is generative, which allows us to train on a single or multiple uncalibrated videos and alter viewpoint and human pose. Figure 4 shows renderings of the same persons from a new camera angle; novel-view-synthesis. Likewise, Figure 5 demonstrates character animation, where the view is fixed, and the underlying skeleton is re-posed by manually changing joint angles or by transferring the motion between characters. A-NeRF is the first model that learns a detailed human body model with such capabilities without needing a 3D surface or multi-view supervision. While NeuralBody [42] can also learn photo-realistic human models from monocular images, their model anchors its representation on the SMPL 3D

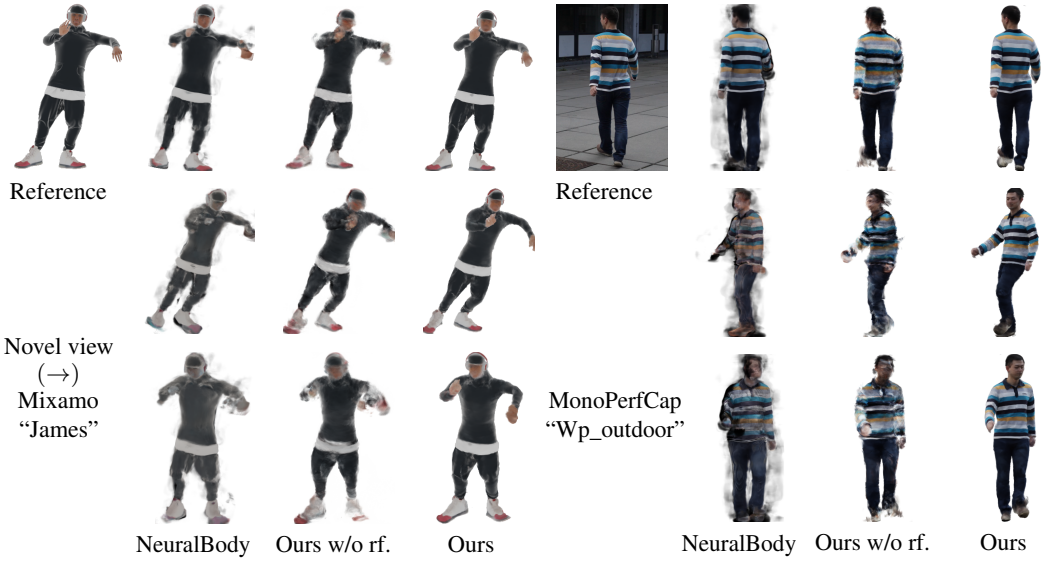


Figure 4: **Novel view synthesis from all models.** All models are trained with SPIN [21] estimated human pose and camera parameters. The renderings from A-NeRF (ours) align better with the reference images, and the rendered novel views show better details (limbs, facial features).

Figure 5: **Motion retargeting and animation. Top rows:** Pose transfer, with the source pose reconstructed by A-NeRF and rendered with different target body models. **Last row:** Animating the A-NeRF model while keeping the lower or upper part of the body fixed.

surface. When trained on our single-view setup with the same noisy estimated poses $\hat{\theta}$ as input, NeuralBody suffers from severe artifacts and less detail when training, as it assumes 3D ground truth, which is only available in controlled conditions. The importance of our joint body and pose optimization is further validated by the ablation with refinement disabled (Ours w/o rf.), which similarly produces ghosting artifacts around extremities.

Human Pose Estimation. Training A-NeRF includes a form of test-time optimization (see Figure 1, right), only the initialization from [21] is trained supervised on the Human3.6M training set. Table 1

Table 1: **Quantitative evaluation on Human3.6M [17]**. Our pose refinement improves consistently upon the SPIN baseline, with largest improvements for extremities (PA-Wrist).

	Human 3.6M				MPI-INF-3DHP	
	Protocol I	Protocol II	Prot. II Wrist	Prot. II Multi-view		
Method	PA-MPJPE↓	PA-MPJPE↓	PA-Wrist↓	PA-MPJPE↓	PA-MPJPE↓	PCK↑
MotioNet [48]	54.6	-	-	-	-	-
HoloPose [14]	46.5	-	-	-	-	-
VIBE [20]	41.4	n/a	n/a	n/a	**64.6**	**89.3**
SPIN [21]	**41.1**	-	-	-	67.5	76.4
Baseline (SPIN github [21])	45.1⋆	41.9⋆	66.5⋆	34.0	68.2⋆	79.3⋆
Ours (w/o smoothness prior)	41.9	**39.6**	**57.3**	**28.0**	66.9	80.4
Ours	**41.7**	n/a	n/a	n/a	**66.8**	**80.4**

⋆ Reevaluation of publicly-available neural network model because missing Protocol II evaluation or not reproducible.

shows that on Human 3.6M, A-NeRF reaches comparable results with other single-view approaches, and achieves a 7% improvement in PA-MPJPE upon the baseline used for pose initialization on Protocol I ($45.1 \rightarrow 41.9$) and 5.5% on Protocol II ($41.9 \rightarrow 39.6$). Unfortunately, the publicly available baseline model does not reproduce the numbers reported in the paper. Note that these are average numbers across all joints, including easy-to-predict hip, shoulder, and head joints. Our largest gains are on the extremities, e.g., with an improvement of 14% (9.2 mm) for the wrist joint on Protocol II (PA-Wrist). On MPI-INF-3DHP, Table 1 shows the results averaged over the 6 test subjects from MPI-INF-3DHP. Despite having a low number of frames and human poses available for learning the skeleton-relative encoding, A-NeRF still provides moderate improvements over the baseline estimations.

Multi-view extension. A-NeRF can also leverage multi-view refinement (Table 1, Protocol II Multi-view), even without access to ground truth camera calibration (see supplementary).

Ablation study. Our detailed ablation studies are reported in the supplemental document. In summary, they reveal: i) Embedding relative 3D position, $\tilde{\mathbf{q}}_k$, instead of our proposed radial embedding $\tilde{\mathbf{v}}$ yields only half as good pose refinements. ii) Not refining the skeleton pose degrades texture and body shape. iii) We quantify image quality in terms of PSNR on a novel view from a multi-view recording. Our embedding choices keep the dimensionality moderate while improving on or matching the PSNR of higher-dimensional variants. iv) For a fixed number of images, the visual quality is better with one long video rather than multiple shorter multi-view recordings.

Limitations and Failure Cases. Our computation time is the biggest bottleneck in extending A-NeRF to long sequences and multiple actors. Although a single static camera suffices, A-NeRF requires to see the person from all sides in varying poses to learn pose dependencies from data. The inlet shows a rendering of an extreme breakdance pose retargeted to a model trained on normal walking motions. Hence, the source pose is unseen during training and far from the data distribution, which leads to artifacts.

Reference Failure

5 Conclusion

We propose a new way for integrating articulated skeleton models and implicit functions via an overcomplete re-parametrization. It includes learning an interpretable 3D representation from 2D images; a personalized volumetric density field with texture detail and time-varying poses of the actor depicted in the input. The underlying ill-posed problem of mapping a single query location to multiple parts is addressed with an overparametrization over nearby parts. To the best of our knowledge, A-NeRF is the first approach to define NeRF models for extreme and articulated motion on unconstrained video and this new approach scores high on the Human 3.6 Million benchmark. Importantly, it works from a single video and naturally extends to multi-view and does not require camera calibration in either scenario. This is an important step towards making motion capture more accurate and practical. In future work, we will learn a general human model from a database of subjects instead of individuals.

References

[1] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky. Neural point-based graphics. *arXiv preprint arXiv:1906.08240*, 2019.

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-moll. Video Based Reconstruction of 3D People Models. In *CVPR*, 2018.

[3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *3DV*, pages 98–109. IEEE, 2018.

[4] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *CVPR*, pages 1175–1186, 2019.

[5] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019.

[6] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *CVPR*, pages 1–8, 2007.

[7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578. Springer, 2016.

[8] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[9] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. URL `https://expose.is.tue.mpg.de`.

[10] B. Deng, J. Lewis, T. Jeruzalski, G. Pons-Moll, G. Hinton, M. Norouzi, and A. Tagliasacchi. Nasa: neural articulated shape approximation. *arXiv preprint arXiv:1912.03207*, 2019.

[11] J. Dong, Q. Shuai, Y. Zhang, X. Liu, X. Zhou, and H. Bao. Motion capture from internet videos. In *ECCV*, pages 210–227. Springer, 2020.

[12] I. Goodfellow, J. Pouget-abadie, M. Mirza, B. Xu, D. Warde-farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *NeurIPS*, pages 2672–2680, 2014.

[13] P. Guan, A. Weiss, A. O. Bãlan, and M. J. Black. Estimating human shape and pose from a single image. In *CVPR*, pages 1381–1388, Sep. 2009. doi: 10.1109/ICCV.2009.5459300.

[14] R. A. Guler and I. Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, pages 10884–10894, 2019.

[15] M. Habermann, W. Xu, , M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM TOG (Proc. SIGGRAPH)*, jul 2019.

[16] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. ARCH: animatable reconstruction of clothed humans. In *CVPR*, 2020. doi: 10.1109/CVPR42600.2020.00316. URL `https://doi.org/10.1109/CVPR42600.2020.00316`.

[17] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014.

[18] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018.

[19] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[20] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[21] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[22] C. Lassner and M. Zollhöfer. Pulsar: Efficient sphere-based neural rendering. In *CVPR*, 2021.

[23] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, July 2017.

[24] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *CVPR*, June 2020.

[25] S. Liu, T. Li, W. Chen, and H. Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, pages 7708–7717, 2019.

[26] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *ACM TOG (Proc. SIGGRAPH)*, 37(4), July 2018. ISSN 0730-0301. doi: 10.1145/3197517.3201401. URL `https://doi.org/10.1145/3197517.3201401`.

[27] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM TOG (Proc. SIGGRAPH)*, 38(4), July 2019. ISSN 0730-0301. doi: 10.1145/3306346.3323020. URL `https://doi.org/10.1145/3306346.3323020`.

[28] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM TOG (Proc. SIGGRAPH)*, 34(6):1–16, 2015.

[29] C. Luo, X. Chu, and A. L. Yuille. Orinet: A fully convolutional network for 3d human pose estimation. In *BMVC*, page 92, 2018. URL `http://bmvc2018.org/contents/papers/0289.pdf`.

[30] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[31] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*. IEEE, 2017. doi: 10.1109/3dv.2017.00064. URL `http://gvv.mpi-inf.mpg.de/3dhp_dataset`.

[32] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In *TOG*, volume 36, July 2017. doi: 10.1145/3072959.3073596. URL `http://gvv.mpi-inf.mpg.de/projects/VNect/`.

[33] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. In *ACM TOG (Proc. SIGGRAPH)*, 2020.

[34] M. Meshry, D. B. Goldman, S. Khamis, H. Hoppe, R. Pandey, N. Snavely, and R. Martin-Brualla. Neural rerendering in the wild. In *CVPR*, pages 6878–6887, 2019.

[35] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[36] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, pages 7588–7597, 2019.

[37] A. Nibali, Z. He, S. Morgan, and L. Prendergast. 3d human pose estimation with 2d marginal heatmaps. In *WACV*, 2019.

[38] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018.

[39] S. Park, J. Hwang, and N. Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *European Conference on Computer Vision (ECCV)*, pages 156–169, 2016. doi: 10.1007/978-3-319-49409-8_15. URL `https://doi.org/10.1007/978-3-319-49409-8_15`.

[40] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017.

[41] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018.

[42] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021.

[43] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *ICCV*, pages 765–773, 2015.

[44] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*, pages 509–526, 2016.

[45] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation. In *ECCV*, 2018.

[46] H. Rhodin, V. Constantin, I. Katircioglu, M. Salzmann, and P. Fua. Neural scene decomposition for multi-person motion capture. In *CVPR*, 2019.

[47] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *CVPR*, pages 1216–1224, Honolulu, United States, July 2017. IEEE. doi: 10.1109/CVPR.2017.134. URL `https://hal.inria.fr/hal-01505085`.

[48] M. Shi, K. Aberman, A. Aristidou, T. Komura, D. Lischinski, D. Cohen-Or, and B. Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM TOG (Proc. SIGGRAPH)*, 40(1):1–15, 2020.

[49] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, et al. Textured neural avatars. In *CVPR*, pages 2387–2397, 2019.

[50] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019.

[51] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, pages 1121–1132, 2019.

[52] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. *NeurIPS*, 33, 2020.

[53] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *ICCV*, pages 951–958, 2011.

[54] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, pages 2602–2611, 2017.

[55] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020.

[56] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *CVPR*, pages 103–110. IEEE, 2012.

[57] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B. Goldman, and M. Zollhöfer. State of the Art on Neural Rendering. *Computer Graphics Forum (EG STAR 2020)*, 2020.

[58] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG (Proc. SIGGRAPH)*, 2019.

[59] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *CVPR*, July 2017.

[60] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NeurIPS*, pages 5242–5252, 2017.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[62] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7467–7477, 2020.

[63] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, pages 10965–10974, 2019.

[64] J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *CVPR*, June 2020.

[65] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *TOG*, 37(2):27, 2018.

[66] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *ICCV*, pages 2344–2353, 2019.

[67] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *ECCV Workshop*, 2016.

[68] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019.

[69] S. Zuffi, A. Kanazawa, T. Berger-Wolf, and M. J. Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *ICCV*, pages 5358–5367. IEEE, Oct. 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] At the end of Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] At the end of Section 1.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We will release the code upon acceptance with detailed explanation.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Details are provided in the main document and further ones given in the supplementary materials.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We experimented on Human 3.6M S9 Protocol II with 3 different random seeds. We trained for 100k iterations each, and the resulting PA-MPJPE difference is negligible (≤ 0.03). We did not report this in the paper.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [No] All datasets in the papers are free for academic use.

   (c) Did you include any new assets either in the supplemental material or as a URL? [No]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]