NIKI: Neural Inverse Kinematics with Invertible Neural Networks for 3D Human Pose and Shape Estimation

Jiefeng Li[1*] Siyuan Bian[1*] Qi Liu[1] Jiasheng Tang[3] Fan Wang[3] Cewu Lu[12†]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[3]Alibaba Group

Abstract

*With the progress of 3D human pose and shape estimation, state-of-the-art methods can either be robust to occlusions or obtain pixel-aligned accuracy in non-occlusion cases. However, they cannot obtain robustness and mesh-image alignment at the same time. In this work, we present NIKI (**N**eural **I**nverse **K**inematics with **I**nvertible Neural Network), which models bi-directional errors to improve the robustness to occlusions and obtain pixel-aligned accuracy. NIKI can learn from both the forward and inverse processes with invertible networks. In the inverse process, the model separates the error from the plausible 3D pose manifold for a robust 3D human pose estimation. In the forward process, we enforce the zero-error boundary conditions to improve the sensitivity to reliable joint positions for better mesh-image alignment. Furthermore, NIKI emulates the analytical inverse kinematics algorithms with the twist-and-swing decomposition for better interpretability. Experiments on standard and occlusion-specific benchmarks demonstrate the effectiveness of NIKI, where we exhibit robust and well-aligned results simultaneously. Code is available at https://github.com/Jeff-sjtu/NIKI.*

1. Introduction

Recovering 3D human pose and shape (HPS) from monocular input is a challenging problem. It has many applications [9, 31, 43, 56, 57, 59, 60]. Despite the rapid progress powered by deep neural networks [17, 21, 22, 24, 29, 62], the performance of existing methods is not satisfactory in complex real-world applications where people are often occluded and truncated by themselves, each other, and objects.

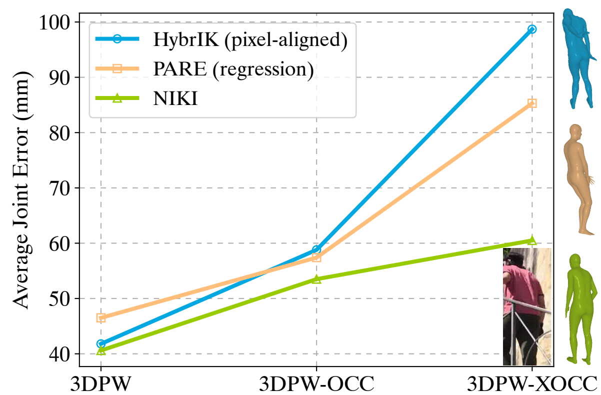

Figure 1. **Trade-off between pixel-aligned accuracy and robustness.** From 3DPW to 3DPW-XOCC, the degree of occlusion increases. The pixel-aligned approach performs well only in non-occlusion cases. The direct regression approach is more robust to occlusions but less accurate in non-occlusion cases. NIKI shows high accuracy and strong robustness simultaneously. Illustrative results on the 3DPW-XOCC dataset are shown on the right.

Existing state-of-the-art approaches rely on pixel-aligned local evidence, *e.g.*, 3D keypoints [14, 29], mesh vertices [37], and mesh-aligned features [62], to perform accurate human pose and shape estimation. Although the local evidence helps obtain high accuracy in standard benchmarks, it fails when the mesh-image correspondences are unavailable due to occlusions and truncations. These pixel-aligned approaches sacrifice robustness to occlusions for high accuracy in non-occlusion scenarios. On the other hand, direct regression approaches are more robust to occlusions. Such approaches directly predict a set of pose and shape parameters with neural networks. By encoding human body priors in the networks, they predict a more physiologically plausible result than the pixel-aligned approaches in severely occluded scenarios. However, direct regression approaches use all pixels to predict human pose and shape, which is a highly non-linear mapping and suf-

[*]Equal contribution.
[†]Cewu Lu is the corresponding author. He is the member of Qing Yuan Research Institute, Qi Zhi Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China.

fers from image-mesh misalignment. Recent work [19, 22] adopts guided attention to leverage local evidence for better alignment. Nevertheless, in non-occlusion scenarios, direct regression approaches are still not as accurate as the pixel-aligned approaches that explicitly model the local evidence. Fig. 1 shows the performance of the state-of-the-art pixel-aligned and regression approaches in scenarios with different levels of occlusions. These two types of approaches cannot achieve mesh-image alignment and robustness at the same time.

In this work, we propose NIKI, a Neural Inverse Kinematics (IK) algorithm with Invertible neural networks, to improve robustness to occlusions while maintaining pixel-aligned accuracy. IK algorithms are widely adopted in pixel-aligned approaches [14, 29] to obtain mesh-image alignment in non-occlusion scenarios. However, existing IK algorithms only focus on estimating the body part rotations that best explain the joint positions but do not consider the plausibility of the estimated poses. Therefore, the output human pose inherits the errors from joint position estimation, which is especially severe in occlusion scenarios. In contrast, NIKI is robust to unreliable joint positions by modeling the bi-directional pose error. We build the bijective mapping between the Euclidean joint position space and the combined space of the 3D joint rotation and the latent error. The latent error indicates how the joint positions deviate from the manifold of plausible human poses. The output rotations are robust to erroneous joint positions since we have explicitly removed the error information by supervising the output marginal distribution in the inverse direction. In the forward direction, we introduce the zero-error boundary conditions, which enforce the solved rotations to explain the reliable joint positions and improve mesh-image alignment. The invertible neural network (INN) is trained in both forward and inverse directions. Since forward kinematics (FK) is deterministic and easy to understand, it aids the INN in learning the complex IK process through inherent bijective mapping. To further improve the interpretability of the IK network, we emulate the analytical IK algorithm by decomposing the complete rotation into the twist rotation and the swing-dependent joint position with two consecutive invertible networks.

We benchmark NIKI on 3DPW [51], AGORA [40], 3DOH [63], 3DPW-OCC [51], and our proposed 3DPW-XOCC datasets. 3DPW-XOCC is augmented from the original 3DPW dataset with extremely challenging occlusions and truncations. NIKI shows robust reconstructions while maintaining pixel-aligned accuracy, demonstrating state-of-the-art performance in both occlusions and non-occlusion benchmarks. The main contributions of this paper are summarized as follows:

- We present a framework with a novel error-aware inverse kinematics algorithm that is robust to occlusions while maintaining pixel-aligned accuracy.
- We propose to decouple the error information from plausible human poses by learning a pose-independent error embedding in the inverse process and enforcing zero-error boundary conditions during the forward process using invertible neural networks.
- Our approach outperforms previous pixel-aligned and direct regression approaches on both occlusions and non-occlusion benchmarks.

2. Related Work

3D Human Pose and Shape Estimation. Prior work estimates 3D human pose and shape by outputting the parameters of statistical human body models [1, 35, 39, 41, 58]. Initial work follows the optimization paradigm [6, 26, 41, 48]. SMPLify [6] is the first automated approach that fits SMPL parameters to 2D keypoint observations. This paradigm is further extended to silhouette [26] and volumetric grids [48].

Recently, learning-based paradigms have gained much attention with the advances in deep neural networks. Existing work can be categorized into two classes: direct regression approaches and pixel-aligned approaches. Direct regression approaches use deep neural networks to regress the pose and shape parameters directly [17, 18, 21–24, 27, 53]. Intermediate representations are used as the weak supervision to improve the regression performance, *e.g.*, 2D keypoints [17] and body/part segmentation [42]. Several studies [16, 24] leverage the optimization paradigm to introduce the pseudo ground truth for better supervision. Pixel-aligned approaches explicitly exploit pixel-aligned local evidence to estimate the pose and shape parameters. Moon *et al*. [37] use the vertex positions to regress the SMPL parameters. Li *et al*. [29] and Iqbal *et al*. [14] propose to map the 3D keypoints to pose parameters. Zhang *et al*. [62] propose the mesh-aligned feedback loop to predict the aligned SMPL parameters. Explicitly modeling local evidence contributes to the state-of-the-art performance of pixel-aligned approaches.

Although pixel-aligned approaches achieve high accuracy in standard benchmarks, they are vulnerable to occlusions and truncations. When the local evidence is not reliable or even does not exist in occluded and truncated cases, such approaches predict physiologically implausible results. Direct regression approaches [15, 19, 22, 46, 63] are more robust to occlusions and truncations but less accurate in non-occlusion scenarios. Zhang *et al*. [63] use the saliency map to infer object-occluded human bodies. Kocabas *et al*. [22] propose part-guided attention to exploit the information about the visibility of body parts. Khirodkar *et al*. [19] use body centermaps to exploit the spatial context. A number of studies [41, 44, 47] propose to use pose prior to improve the plausibility of the estimated poses. Although the local evidence is implicitly used in recent regression approaches, pixel-aligned approaches still dominate non-occlusion benchmarks.

In this work, we combine the merits of pixel-aligned approaches and direct regression approaches. NIKI maintains pixel-aligned accuracy by aligning with the body joints via inverse kinematics while achieving robustness to occlusions and truncations with bi-directional error decoupling.

Inverse Kinematics. The inverse kinematics (IK) process finds the relative rotations to produce the desired positions of body joints. It is an ill-posed problem because of the information loss in the forward process. Traditional numerical approaches [4, 7, 11, 20, 52, 55] are time-consuming due to iterative optimization. The heuristic approaches such as CDC [36], FABRIK [3], and IK-FA [45] are more efficient and have a lower computation cost for each heuristic iteration. Recent work [8, 49] has started using neural networks to solve the IK problem. Zhou *et al*. [64] train a four-layer MLP network to predict the 3D human pose parameterized as 6D vectors. Li *et al*. [29] propose a hybrid analytical-neural solution to accurately predict the body part rotations. Oreshkin *et al*. [38] propose to use prototype encoding to predict rotations from sparse user inputs. Voleti *et al*. [50] extend the same model to arbitrary skeletons. The work of Ardizzone *et al*. [2] is most related to us. They use invertible neural networks (INNs) to solve inverse problems, including the toy inverse kinematics problem in 2D space. However, similar to all the aforementioned approaches, they assume the input body joints are reliable, resulting in vulnerability to occlusions and truncations.

Invertible Neural Network in HPS Estimation. Modeling the conditional posterior of an inverse process is a classical statistical task. Wehrbein *et al*. [54] propose to estimate 3D human poses from 2D poses by capturing lost information with INNs. Several studies [58, 61] leverage INNs to build priors for 3D human pose estimation. The pose priors are learned by normalizing flows that are built with INNs. Biggs *et al*. [5] propose to use the learned prior from normalizing flows to resolve ambiguities. Kolotouros *et al*. [25] propose a conditional distribution with normalizing flows as a function of the input to combine information from different sources. Li *et al*. [28] leverage normalizing flows to capture the underlying residual log-likelihood of the output and propose a novel regression paradigm from the perspective of maximum likelihood estimation. Unlike previous methods, our approach leverages the property of bijective mapping in INNs to decouple joint errors and solve the inverse kinematics problem robustly.

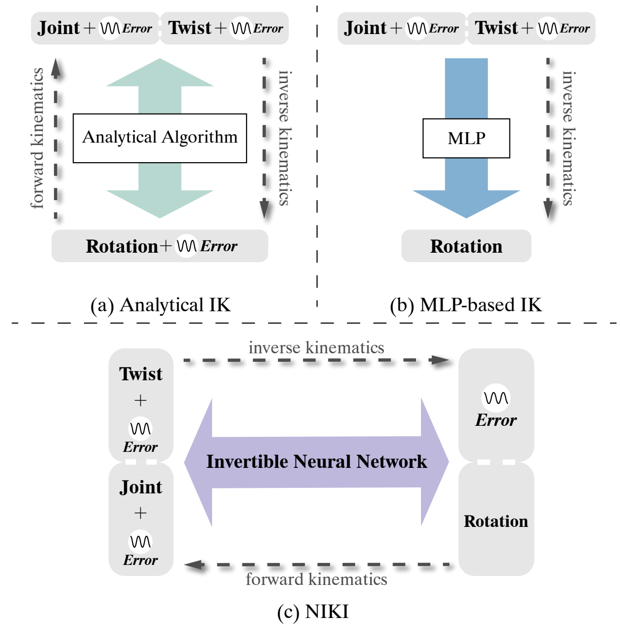

Figure 2. **Illustration** of (a) analytical IK, (b) feedforward MLP-based IK, and (c) NIKI with bi-directional error decoupling.

3. Method

In this section, we present NIKI, a neural inverse kinematics solution for 3D human pose and shape estimation. We first review the formulation of existing analytical and MLP-based IK algorithms in §3.1. In §3.2, we introduce the proposed INN-based IK algorithm with bi-directional error decoupling. In §3.3, we present the overall human pose estimation framework and the learning objective. Then we elaborate on the proposed IK-specific invertible architecture in §3.4. Finally, we provide the necessary implementation details in §3.5

3.1. Preliminaries

The IK process is to find the corresponding body part rotations that explain the input body joint positions, while the forward kinematics (FK) process computes the desired joint positions based on the input rotations. The FK process is well-defined, but the transformation from joint rotations to joint positions incurs an information loss, *i.e*., multiple rotations could correspond to one position, resulting in an ill-posed IK process. Here, we follow HybrIK [29] to consider twist rotations for information integrity.

The conventional IK algorithms only require the output rotations to match the input joint positions but ignore the errors of the joint positions and the plausibility of the body pose. Therefore, the errors of the joint positions will be accumulated in the joint rotations (Fig. 2a). This process

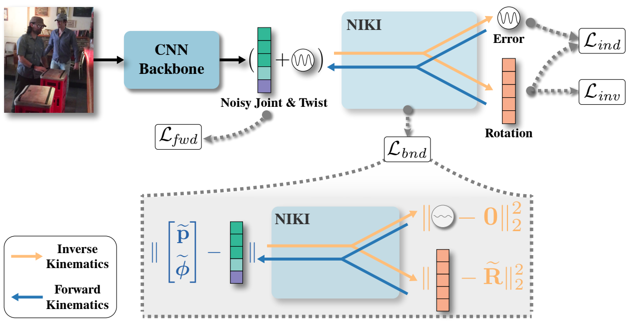

Figure 3. **Overview of the proposed framework.** The input image is fed into the CNN backbone network to estimate the initial joint positions and twist rotations, followed by NIKI to solve the joint rotations.

can be formulated as:

$$\underbrace{\mathbf{R}+\epsilon_r}_{\text{erroneous output}} = \mathrm{IK}_{\text{Analytical}}(\underbrace{\mathbf{p}+\epsilon_p, \boldsymbol{\phi}+\epsilon_\phi}_{\text{erroneous input}} \mid \boldsymbol{\beta}), \tag{1}$$

where $\mathbf{R}$ denotes the underlying plausible rotations, ϵ_r denotes the accumulated error in estimated rotations, $\mathbf{p}$ denotes the underlying plausible joint positions, ϵ_p denotes the position errors, $\boldsymbol{\phi}$ denotes the underlying plausible twist rotations, ϵ_ϕ denotes the twist error, and $\boldsymbol{\beta}$ denotes the body shape parameters.

A straightforward solution to improving the robustness of the IK algorithms is using the standard regression model [8, 63] to approximate the underlying plausible rotations $\mathbf{R}$ given the erroneous input (Fig. 2b):

$$\mathbf{R} \approx \mathrm{IK}_{\text{MLP}}(\mathbf{p}+\epsilon_p, \boldsymbol{\phi}+\epsilon_\phi \mid \boldsymbol{\beta}). \tag{2}$$

Indeed, modeling the IK process with classical neural networks, *e.g.*, MLP, can improve the robustness. However, the output rotations are less sensitive to the change of the joint positions. The errors are highly coupled with the joint positions. Without explicitly decoupling errors from plausible human poses, it is difficult for the network to distinguish between reasonable and abnormal changes in joint positions. Therefore, the output rotations cannot accurately track the movement of the body joints. In practice, we find that the feedforward neural networks could improve performance in occlusion cases but cause performance degradation in non-occlusion cases, where accurate mesh-image alignment is required. Detailed comparisons are provided in Tab. 5.

3.2. Inverse Kinematics with INNs

In this work, to improve the robustness of IK to occlusions while maintaining the sensitivity to non-occluded body joints, we propose to use the invertible neural network (INN) to model bi-directional errors explicitly (see Fig. 2c). In contrast to the conventional methodology, we learn the IK model $g(\cdot;\boldsymbol{\beta},\theta)$ jointly with the FK model $f(\cdot;\boldsymbol{\beta},\theta)$:

$$[\mathbf{p}+\epsilon_p, \boldsymbol{\phi}+\epsilon_\phi] = f(\mathbf{R}, \mathbf{z}_r; \boldsymbol{\beta}, \theta), \tag{3}$$

$$[\mathbf{R}, \mathbf{z}_r] = g(\mathbf{p}+\epsilon_p, \boldsymbol{\phi}+\epsilon_\phi; \boldsymbol{\beta}, \theta), \tag{4}$$

where $\mathbf{z}_r$ is the error embedding that denotes how the input joint positions deviate from the manifold of plausible human poses. Notice that f and g share the same parameters θ, and $f = g^{-1}$ is enforced by the invertible network architecture. We expect that simultaneously learning the FK and IK processes can benefit each other.

In the forward process, we can tune the error embedding $\mathbf{z}_r$ to control the error level of the body joint positions. The body part rotations will perfectly align with the joint positions by setting $\mathbf{z}_r$ to 0, which means no deviation from the pose manifold. In the inverse process, the error is only reflected on $\mathbf{z}_r$. The rotation $\mathbf{R}$ keeps stable against the erroneous input.

Decouple Error Information. The input joints and twists to the IK process contain two parts of information: i) the underlying pose that lies on the manifold of plausible 3D human poses; ii) the error information that indicates how the input deviates from the manifold. We can obtain robust pose estimation by separating these two types of information. Due to the bijective mapping enforced by the INN, all the input information is preserved in the output, and no new information is introduced. Therefore, we only need to remove the pose information from the output vector $\mathbf{z}_r$ in the inverse process. The vector $\mathbf{z}_r$ will automatically encode the remaining error information. To this end, we enforce the model to follow the independence constraint, which encourages $\mathbf{R}$ and $\mathbf{z}_r$ to be independent upon convergence, *i.e.*, $p(\mathbf{z}_r|\mathbf{R}) = p(\mathbf{z}_r)$.

After we separate the error information, we can manipulate the error embedding to let the model preserve the sensitivity to error-free body joint positions without compromising robustness. In particular, we constrain the error information in the forward process with the zero-error condition:

$$[\mathbf{p}, \phi] = f(\mathbf{R}, \mathbf{0}; \boldsymbol{\beta}, \theta). \tag{5}$$

In this way, the rotations will track the joint positions and twist rotations accurately in non-occlusion scenarios. Besides, the zero-error condition can also be extended to the inverse process:

$$[\mathbf{R}, \mathbf{0}] = g(\mathbf{p}, \phi; \boldsymbol{\beta}, \theta). \tag{6}$$

With the independence and zero-error constraints, the network is able to model the error information in both the forward and inverse processes, making NIKI robust to occlusions while maintaining pixel-aligned accuracy.

3.3. Decoupled Learning

The overview of our approach is illustrated in Fig. 3. During inference, we first extract the joint positions and twist rotations with the CNN backbone, which are subsequently fed to the invertible network to predict the complete body part rotations. During training, we optimize FK and IK simultaneously in one network. Hereby, we perform FK and IK alternately with the additional independence loss and boundary loss. The gradients from both directions are accumulated before performing a parameter update.

Inverse Training. In the inverse iteration, the network predicts the body part rotations given the joint positions $\hat{\mathbf{p}}$ and twist rotations $\hat{\phi}$ from the CNN backbone. The loss function is defined as:

$$\mathcal{L}_{inv} = \left\|\hat{\mathbf{R}}_{inv} - \widetilde{\mathbf{R}}\right\|_2^2 + \left\|\mathrm{FK}(\hat{\mathbf{R}}_{inv}) - \mathrm{FK}(\widetilde{\mathbf{R}})\right\|_1, \tag{7}$$

with

$$[\hat{\mathbf{R}}_{inv}, \hat{\mathbf{z}}_r] = g(\hat{\mathbf{p}}, \hat{\phi}; \boldsymbol{\beta}, \theta), \tag{8}$$

where $\widetilde{\mathbf{R}}$ represents the ground-truth rotations, and $\mathrm{FK}(\cdot)$ denotes the analytical FK process to supervise the corresponding 3D joint positions of the predicted pose.

Forward Training. In the forward process, the network predicts the joint positions and twist rotations given the body part rotations. The error of the noisy predictions $\hat{\mathbf{p}}$ and $\hat{\phi}$ should only be determined by the error embedding. Therefore, with the ground-truth rotations $\widetilde{\mathbf{R}}$ and error embedding $\hat{\mathbf{z}}_r$ obtained from the inverse iteration, the forward model should predict the same values as the CNN output:

$$\mathcal{L}_{fwd} = \|\hat{\mathbf{p}}_{fwd} - \hat{\mathbf{p}}\|_1 + \|\hat{\phi}_{fwd} - \hat{\phi}\|_2^2, \tag{9}$$

with

$$[\hat{\mathbf{p}}_{fwd}, \hat{\phi}_{fwd}] = f(\widetilde{\mathbf{R}}, \hat{\mathbf{z}}_r; \boldsymbol{\beta}, \theta). \tag{10}$$

Independence Loss. The latent error vector is learned in an unsupervised manner by making $\mathbf{R}_{inv}$ and $\mathbf{z}_r$ independent of each other. The pose information in $\mathbf{R}_{inv}$ is supervised by Eq. 7. We then enforce the independence by penalizing the mismatch between the joint distribution of the rotations and error embedding $q\big(\hat{\mathbf{R}}_{inv}, \mathbf{z}_r\big)$ and the product of marginal distributions $p\big(\widetilde{\mathbf{R}}\big)p(\mathbf{z})$:

$$\mathcal{L}_{ind} = \mathcal{D}\big(q\big(\mathbf{R}_{inv}, \mathbf{z}_r\big), p\big(\widetilde{\mathbf{R}}\big)p(\mathbf{z})\big), \tag{11}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ follows the standard normal distribution, $\mathbf{I}$ is the identity matrix, and $\mathcal{D}(\cdot)$ denotes the Maximum Mean Discrepancy [12], which allows us to compare two probability distributions through samples. In addition to the independence constraint, $\mathcal{L}_{ind}$ encourages the error embedding $\mathbf{z}_r$ to follow the standard normal distribution $p(\mathbf{z})$, serving as a regularization.

Boundary Condition Loss. To enforce the solved rotations to explain the reliable joint positions, we supervise the boundary cases where no error occurs. In the inverse process, the output error should be zero when the network is fed with the ground truth:

$$\mathcal{L}^i_{bnd} = \|\hat{\epsilon}_r\|_2^2 + \|\hat{\mathbf{R}}_{bnd} - \widetilde{\mathbf{R}}\|_2^2, \tag{12}$$

with

$$[\hat{\mathbf{R}}_{bnd}, \hat{\epsilon}_r] = g(\widetilde{\mathbf{p}}, \widetilde{\phi}; \boldsymbol{\beta}, \theta), \tag{13}$$

where $\widetilde{\mathbf{p}}$ and $\widetilde{\phi}$ denote the ground-truth joint positions and twist rotations, respectively.

In the forward process, the joint positions and twist rotations should map to the ground truth when the input error vector $\mathbf{z}_r$ is $\mathbf{0}$:

$$\mathcal{L}^f_{bnd} = \|\hat{\mathbf{p}}_{bnd} - \widetilde{\mathbf{p}}\|_1 + \|\hat{\phi}_{bnd} - \widetilde{\phi}\|_2^2, \tag{14}$$

with

$$[\hat{\mathbf{p}}_{bnd}, \hat{\phi}_{bnd}] = f(\widetilde{\mathbf{R}}, \mathbf{0}; \boldsymbol{\beta}, \theta). \tag{15}$$

Overall, the total loss of NIKI is:

$$\begin{aligned}\mathcal{L} =& \lambda_{inv}\mathcal{L}_{inv} + \lambda_{fwd}\mathcal{L}_{fwd} \\ &+ \lambda_{ind}\mathcal{L}_{ind} + \lambda^i_{bnd}\mathcal{L}^i_{bnd} + \lambda^f_{bnd}\mathcal{L}^f_{bnd},\end{aligned} \tag{16}$$

where $\lambda_{inv}, \lambda_{fwd}, \lambda_{ind}, \lambda^i_{bnd}, \lambda^f_{bnd}$ are the scalar coefficients to balance the loss terms.

3.4. Invertible Architecture

One-Stage Mapping. To build a fully invertible neural network for inverse kinematics, we build the one-stage mapping model using RealNVP [10]. Since the IK and FK processes require the skeleton template, we extend the INN to incorporate the conditional shape parameters input. The basic block of the network contains two reversible coupling layers conditioned on the shape parameters. The overall network consists of multiple blocks connected in series to increase capacity. Besides, since the invertible network requires the input and output vectors to have the same dimension, we follow previous work [2] and pad zeros to the input.

Twist-and-Swing Mapping. Although treating the invertible neural network as a black box can let us model both the FK and IK processes at the same time, we further emulate the analytical IK algorithm to improve the performance and interpretability. Specifically, we follow the twist-and-swing decomposition [29] and divide the IK process into two steps: i) from joint positions to swing rotations; ii) from twist and swing rotations to complete rotations. The two-step mapping is implemented by two separate invertible networks:

$$[\mathbf{R}_{sw}, \mathbf{z}_{sw}] = g_1(\mathbf{p} + \epsilon_p; \boldsymbol{\beta}, \theta_1), \tag{17}$$

$$[\mathbf{R}, \mathbf{z}_r] = g_2(\mathbf{R}_{sw}, \boldsymbol{\phi} + \epsilon_\phi; \theta_2), \tag{18}$$

where $\mathbf{R}_{sw}$ is the swing rotations, and $\mathbf{z}_{sw}$ indicates the deviation from the plausible swing rotation manifold.

Since the mappings are bijective, the FK process also follows the twist-and-swing procedure but inversely. We have $f = f_1 \circ f_2 = g_1^{-1} \circ g_2^{-1} = g^{-1}$. In the FK process, the body part rotations are first decomposed into twist and swing rotations. Then the swing rotations are transformed into the joint positions. The intermediate supervision of swing rotations is used in both the forward and inverse training.

Temporal Extension. The invertible framework is flexible. It can be easily extended to solving the IK problem with temporal inputs. The model with static inputs can only identify the errors related to physiological implausibility. In contrast, the temporal model further improves motion smoothness by decoupling errors of implausible human body movements. More details are provided in the supplementary material.

3.5. Implementation Details.

We adopt HybrIK [29] as the CNN backbone to predict the noisy body joint positions and the twist rotations. The input of the IK model includes the joint positions $\mathbf{p} \in \mathbb{R}^{3K}$, twist rotations parameterized in 2-dimensional vectors, *i.e.*, $\boldsymbol{\phi} \in \mathbb{R}^{2(K-1)}$, and the confidence scores of each joint $\mathbf{s} \in \mathbb{R}^{K}$, where K denotes the total number of human body joints. The output of the model consists of the body part rotations parameterized as a 6D vector for each part, *i.e.*, $\mathbf{R} \in \mathbb{R}^{6K}$, and the error embedding $\mathbf{z}_r \in \mathbb{R}^{D_z}$. The IK model is conditioned on the shape parameters $\boldsymbol{\beta}$, which is also predicted by the CNN backbone. We pad the input with a zero vector with the dimension $M = D_z + 2$ to satisfy the dimension constraint of the invertible neural network. The networks are trained with the Adam solver for 50 epochs with a mini-batch size of 64. The learning rate is set to 1×10^{-3} at first and reduced by a factor of 10 at the 30th and 40th epochs. Implementation is in PyTorch. Detailed architectures are provided in the supplementary material.

4. Experiments

Datasets. We employ the following datasets in our experiments: (1) 3DPW [51], an outdoor benchmark for 3D human pose estimation. (2) AGORA [40], a synthetic dataset with challenging environmental occlusions. (3) 3DOH [63], a 3D human dataset where human activities are occluded by objects. (4) 3DPW-OCC [51], a different split of the original 3DPW dataset with a occluded test set. (5) 3DPW-XOCC, a new benchmark for 3D human pose estimation with extremely challenging occlusions and truncations. We simulate occlusions and truncations by randomly pasting occlusion patches and cropping frames with truncated windows.

	Method	3DPW MPJPE ↓	PA-MPJPE ↓	PVE ↓
Regression	HMR [17]	130.0	81.3	-
	SPIN [24]	96.9	59.2	116.4
	ROMP [46]	85.5	53.3	103.1
	METRO [32]	77.1	47.9	88.2
	PARE [22]	74.5	46.5	88.6
Pixel-aligned	PyMAF [62]	92.8	58.9	110.1
	I2L [37]	93.2	58.6	-
	KAMA [14]	-	51.1	97.0
	Mesh Graphormer [33]	74.7	45.6	87.7
	HybrIK (ResNet) [29]	76.2	45.1	89.1
	HybrIK (HRNet) [29]	72.9	41.8	88.6
	NIKI (One-Stage)	71.7	41.0	86.9
	NIKI (Twist-and-Swing)	**71.3**	**40.6**	**86.6**

Table 1. **Quantitative comparisons with state-of-the-art methods on the 3DPW dataset.** Symbol "-" means results are not available.

Method	3DPW-XOCC MPJPE ↓	PA-MPJPE ↓	PVE ↓
HybrIK [29]	148.3	98.7	164.5
PARE [22]	139.4	85.3	151.6
PARE* [22]	114.2	67.7	133.0
NIKI (One-Stage)	117.0	64.4	135.6
NIKI (Twist-and-Swing)	**110.7**	**60.5**	**128.6**

Table 2. **Quantitative comparisons with state-of-the-art methods on the 3DPW-XOCC dataset.** Symbol $*$ means finetuning on the 3DPW-XOCC train set.

Training and Evaluation. NIKI is trained on 3DPW [51] and Human3.6M [13] and evaluated on 3DPW [51] and 3DPW-XOCC [51] to benchmark the performance on both occlusions and non-occlusion scenarios. We use the AGORA [40] train set only when conducting experiments on its test set. For evaluations on the 3DPW-OCC [51] and 3DOH [63] datasets, we train NIKI on COCO [34], Human3.6M [13], and 3DOH [63] for a fair comparison. Procrustes-aligned mean per joint position error (PA-MPJPE) and mean per joint position error (MPJPE) are reported to assess the 3D pose accuracy. Per vertex error (PVE) is also reported to evaluate the estimated body mesh.

4.1. Comparison to State-of-the-art Methods

We evaluate NIKI on both standard and occlusion-specific benchmarks. Tab. 1 compares NIKI with previous state-of-the-art HPS methods on the standard 3DPW dataset. We report the results of NIKI with one-stage mapping and twist-and-swing mapping. We can observe that in the standard benchmark, pixel-aligned approaches obtain better performance than direct regression approaches. NIKI

Method	AGORA			
	NMVE ↓	NMJE ↓	MVE ↓	MPJPE ↓
HMR [17]	217.0	226.0	173.6	180.5
SPIN [24]	216.3	223.1	168.7	175.1
EFT [16]	196.3	203.6	159.0	165.4
ROMP [46]	227.3	236.6	161.4	168.0
PyMAF [62]	200.2	207.4	168.2	174.2
PARE [22]	167.7	174.0	140.9	146.2
SPEC [23]	126.8	133.7	106.5	112.3
CLIFF [30]	83.5	89.0	76.0	81.0
HybrIK [29]	81.2	84.6	73.9	77.0
NIKI (One-Stage)	72.2	75.9	65.7	69.1
NIKI (Twist-and-Swing)	**70.2**	**74.0**	**63.9**	**67.3**

Table 3. **Quantitative comparisons with state-of-the-art methods on the AGORA dataset.**

Method	3DPW-OCC			3DOH	
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJPE ↓	PA-MPJPE ↓
Zhang *et al*. [63]	-	72.2	-	-	58.5
SPIN [24]	98.4	62.5	135.1	104.3	68.3
HMR-EFT [16]	95.8	62.0	120.5	75.2	53.1
HybrIK [29]	90.8	58.8	111.9	40.4	31.2
PARE [22]	91.4	57.4	115.3	63.3	44.3
NIKI (One-Stage)	88.2	55.3	109.7	38.9	29.2
NIKI (Twist-and-Swing)	**85.5**	**53.5**	**107.6**	**38.8**	**28.7**

Table 4. **Quantitative comparisons with state-of-the-art methods on the 3DPW-OCC and 3DOH datasets.**

significantly outperforms the most accurate direct regression approach by 5.9 mm on PA-MPJPE (12.7% relative improvement). Besides, NIKI obtains comparable performance to pixel-aligned approaches, showing a 1.2 mm improvement on PA-MPJPE.

Tab. 2 demonstrates the robustness of NIKI to extreme occlusions and truncations. We report the results of the most accurate pixel-aligned and direct regression approaches on the 3DPW-XOCC dataset. It shows that direct regression approaches outperform pixel-aligned approaches in challenging scenes, which is in contrast to the results in the standard benchmark. NIKI improves the PA-MPJPE performance by **38.7**% compared to HybrIK and **10.1**% compared to PARE that finetuned on the 3DPW-XOCC train set.

The results of NIKI on other occlusion-specific datasets are reported in Tab. 3 and Tab. 4. NIKI shows consistent improvements on all these datasets, demonstrating that NIKI is robust to challenging occlusions and truncations while maintaining pixel-aligned accuracy. Specifically, NIKI improves the NMJE performance on AGORA by **12.5**% compared to the state-of-the-art methods. We can also observe that the twist-and-swing mapping model is consistently superior to the one-stage mapping model. More discussions of limitations and future work are provided in the supplementary material.

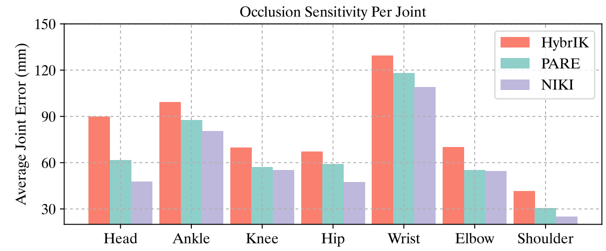

Figure 4. **Per joint occlusion sensitivity analysis** of three different methods: HybrIK [29], PARE [22], and NIKI.

		3DPW		3DPW-XOCC	
		MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
(a)	Analytical IK [29]	74.9	41.8	148.3	98.7
	MLP-based IK [64]	83.2	50.3	121.1	68.5
	Vanilla INN-based IK	73.8	43.1	115.3	64.4
	Ardizzone *et al*. [2]	79.1	45.6	119.5	67.3
(b)	NIKI	**71.3**	**40.6**	**110.7**	**60.5**
	w/o Independence	71.6	40.8	112.6	61.4
	w/o Boundary	73.6	43.0	112.5	62.9
	w/o Bi-directional Train	74.0	43.0	111.9	61.9

Table 5. **Ablation experiments on the 3DPW and 3DPW-XOCC datasets.**

4.2. Ablation Study

INN *vs*. NN. To demonstrate the superiority of the invertible network in solving the IK problem, we compare NIKI with existing IK algorithms, including the analytical [29], MLP-based [64], the INN-based [2], and the vanilla INN baseline. [2] is trained without modeling the error information. The vanilla INN baseline is trained in both the forward and inverse directions without independence and boundary constraints. Quantitative results are reported in Tab. 5a. The MLP-based method shows better performance in occlusions scenarios compared to the analytical method. However, it is less accurate in the standard non-occlusion scenarios since it cannot accurately track the movement of the body joints. The vanilla INN performs better than the MLP model in both standard and occlusion datasets but is still less accurate than the analytical method in the standard dataset. NIKI surpasses all methods in both datasets.

Effectiveness of Error Decoupling. To study the effectiveness of error decoupling, we evaluate models that are trained without enforcing independence or boundary constraints. Quantitative results are summarized in Tab. 5b. It shows that the independence loss for inverse error modeling contributes to a better performance in occlusion scenarios and barely affects the performance in non-occlusion scenarios. Besides, the boundary constraints contribute to a better alignment. Without boundary constraints, the model shows performance degradation in non-occlusion scenarios.

Figure 5. **Quantitative results on COCO (rows 1-3) and 3DPW-XOCC (rows 4-5) datasets.** From left to right: Input image, (a) HybrIK [29] results, (b) PARE [22] results, and (c) NIKI results.

Effectiveness of Bi-directional Training. To further validate the effectiveness of bi-directional training, we report the results of the baseline model that is only trained with the inverse process. Without bi-directional training, we also cannot apply the boundary condition in the forward direction, which means that we only decouple the errors in the inverse process. As shown in Tab. 5b, the IK model cannot maintain the sensitivity to non-occluded body joints in the standard benchmark without forward training.

Sensitivity Analysis. We further follow Kocabas *et al*. [22] to conduct the occlusion sensitivity analysis. Fig. 4 shows the per-joint breakdown of the mean 3D error from the occlusion sensitivity analysis for three different methods on the `3DPW` test split. Although HybrIK [29] obtains high accuracy on the `3DPW` dataset, it is quite sensitive to occlusions. NIKI is more robust to occlusions and improves the robustness of all joints. We also qualitatively compare HybrIK, PARE, and NIKI in Fig. 5. NIKI performs well in challenging occlusion scenarios and predicts well-aligned results. More occlusion analyses and qualitative samples are provided in the supplementary material.

5. Conclusion

In this paper, we propose NIKI, a neural inverse kinematics solution for accurate and robust 3D human pose and shape estimation. NIKI is built with invertible neural networks to model bi-directional error information in the forward and inverse kinematics processes. In the inverse direction, NIKI explicitly decouples the error information from the manifold of the plausible human poses to improve robustness. In the forward direction, NIKI enforces zero-error boundaries to obtain accurate mesh-image alignment. We construct the invertible neural network by emulating the analytical inverse kinematics algorithm with twist-and-swing decomposition to improve interpretability. Comprehensive experiments on standard and occlusion-specific datasets demonstrate the pixel-aligned accuracy and robustness of NIKI. We hope NIKI can serve as a solid baseline for challenging real-world applications.

Acknowledgments. This work was supported by the National Key R&D Program of China (No. 2021ZD0110704), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, and Shanghai Science and Technology Commission (21511101200).

References

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

[2] Lynton Ardizzone, Jakob Kruse, Sebastian Wirkert, Daniel Rahner, Eric W Pellegrini, Ralf S Klessen, Lena Maier-Hein, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. In *ICLR*, 2019. 3, 5, 7

[3] Andreas Aristidou and Joan Lasenby. Fabrik: A fast, iterative solver for the inverse kinematics problem. *Graphical Models*, 2011. 3

[4] A Balestrino, Giuseppe De Maria, and L Sciavicco. Robust control of robotic manipulators. *IFAC Proceedings Volumes*, 1984. 3

[5] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. In *NeurIPS*, 2020. 3

[6] 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 *ECCV*, 2016. 2

[7] Samuel R Buss and Jin-Su Kim. Selectively damped least squares for inverse kinematics. *Journal of Graphics tools*, 2005. 3

[8] Akos Csiszar, Jan Eilers, and Alexander Verl. On solving the inverse kinematics problem using neural networks. In *M2VIP*, 2017. 3, 4

[9] Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *CVPR*, 2023. 1

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *ICLR*, 2017. 5

[11] Michael Girard and Anthony A Maciejewski. Computational modeling for the computer animation of legged figures. In *SIGGRAPH*, 1985. 3

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012. 5

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2013. 6

[14] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. Kama: 3d keypoint aware body mesh articulation. In *3DV*, 2021. 1, 2, 6

[15] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 2

[16] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2021. 2, 7

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 6, 7

[18] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 2

[19] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, 2022. 2

[20] Charles A Klein and Ching-Hsiang Huang. Review of pseudoinverse control for use with kinematically redundant manipulators. *IEEE Transactions on Systems, Man, and Cybernetics*, 1983. 3

[21] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2

[22] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 1, 2, 6, 7, 8

[23] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *ICCV*, 2021. 2, 7

[24] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1, 2, 6, 7

[25] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 3

[26] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 2

[27] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&d: Learning human dynamics from dynamic camera. In *ECCV*, 2022. 2

[28] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021. 3

[29] 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 *CVPR*, 2021. 1, 2, 3, 5, 6, 7, 8

[30] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 7

[31] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *CVPR*, 2023. 1

[32] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 6

[33] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 6

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 2015. 2

[36] David G Luenberger, Yinyu Ye, et al. *Linear and nonlinear programming*. Springer, 1984. 3

[37] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 1, 2, 6

[38] Boris N Oreshkin, Florent Bocquelet, Felix G Harvey, Bay Raitt, and Dominic Laflamme. Protores: Proto-residual network for pose authoring via learned inverse kinematics. In *ICLR*, 2021. 3

[39] Ahmed AA Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Supr: A sparse unified part-based human representation. In *ECCV*, 2022. 2

[40] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. 2, 6

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 2

[42] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, 2018. 2

[43] 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 *CVPR*, 2021. 1

[44] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 2

[45] Nizar Rokbani, Alicia Casals, and Adel M Alimi. Ik-fa, a new heuristic inverse kinematics solver using firefly algorithm. In *Computational intelligence applications in modeling and control*. Springer, 2015. 3

[46] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 2, 6, 7

[47] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *ECCV*, 2022. 2

[48] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, 2018. 2

[49] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *CVPR*, 2018. 3

[50] Vikram Voleti, Boris N Oreshkin, Florent Bocquelet, Félix G Harvey, Louis-Simon Ménard, and Christopher Pal. Smpl-ik: Learned morphology-aware inverse kinematics for ai driven artistic workflows. In *SIGGRAPH Asia*, 2022. 3

[51] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2, 6

[52] Charles W Wampler. Manipulator inverse kinematic solutions based on vector formulations and damped least-squares methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 1986. 3

[53] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *ICCV*, 2021. 2

[54] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *ICCV*, 2021. 3

[55] William A Wolovich and H Elliott. A computational technique for inverse kinematics. In *CDC*, 1984. 3

[56] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans obtained from normals. In *CVPR*, 2023. 1

[57] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: implicit clothed humans obtained from normals. In *CVPR*, 2022. 1

[58] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, 2020. 2, 3

[59] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *CVPR*, 2023. 1

[60] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *CVPR*, 2023. 1

[61] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *ECCV*, 2020. 3

[62] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 1, 2, 6, 7

[63] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 2, 4, 6, 7

[64] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3, 7