# HPOF: 3D Human Pose Recovery from Monocular Video with Optical Flow

Bin Ji[1,2], Chen Yang[2], Shunyu Yao[2], Ye Pan[1,2*]
whitneypanye@sjtu.edu.cn
[1]John Hopcroft Center for Computer Science, Shanghai Jiao Tong University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
Shanghai, China

**Figure 1: Given challenging in-the-wild videos with low resolutions, our model(HPOF) can reconstruct accurate and realistic 3D human pose from high-speed movement like skating game, where self-occlusion and motion blurs are common.**

## ABSTRACT

This paper introduces HPOF, a novel deep neural network to reconstruct the 3D human motion from a monocular video. Recently, model-based methods have been proposed to simplify the reconstruction task by estimating several parameters that control a deformable surface model to fit the person in the image. However, learning the parameters from a single image is a highly ill-posed problem, and the process is ultimately data-hungry. Existing 3D datasets are not sufficient, and the usage of 2D in-the-wild datasets is often susceptible to the inadequate precision of manual annotations. To address the above issues, our method yields substantial improvements in two domains. First, we leverage optical flow to supervise the 2D rendered images of predicted SMPL models to learn short-term temporal features. Besides, taking long-term temporal consistency into account, we define a novel temporal encoder based on a dilated convolutional network. The encoder decomposes the learning process of human shape and pose, first guarantees the invariance of the body shape, and then simulates a more reasonable forward kinematics process on this basis to achieve more accurate pose estimation. In addition, an adversarial learning framework is applied to supervise the reconstruction progress in a coarse-grained way. We show that HPOF not only improves the accuracy of 3D poses but ensures the realistic body structure throughout the video. We perform extensive experimentation to demonstrate the superiority of our method and analyze the effectiveness of our model, surpassing other state-of-the-arts.

*Corresponding author


*ICMR '21, August 21–24, 2021, Taipei, Taiwan*

ACM ISBN 978-1-4503-8463-6/21/08...$15.00
https://doi.org/10.1145/3460426.3463605

## CCS CONCEPTS

• **Computing methodologies** → **Motion processing**; **Motion capture**; • **Computer systems organization** → **Neural networks**.

## KEYWORDS

pose estimation, motion capture, monocular video, optical flow

**ACM Reference Format:**
Bin Ji[1,2], Chen Yang[2], Shunyu Yao[2], Ye Pan[1,2]. 2021. HPOF: 3D Human Pose Recovery from Monocular Video with Optical Flow. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3460426.3463605

# 1 INTRODUCTION

With the rising interest in personalized VR and immersive experiences comes the need to create high-quality motion capture systems. To ensure high-fidelity recovery, complicated marker-based systems are preferred for professional conditions. Specialized hardware like magnetic trackers, optical cameras, inertial sensors, etc., is involved in these systems. Not only are these systems challenging to deploy and costly, but they come with a large of pre-processing, which hinders their further popularization. On the other end of the spectrum, recent studies develop data-driven learning-based approaches that are efficient and low-cost to perform 3D human pose estimation from the monocular RGB video[37, 50]. For this reason, recovering the 3D human motion from a single RGB camera is taking center stage in this field.

Most of the learning-based methods can be categorized into two classes: skeleton-based approach and model-based approach. Skeleton-based methods learn a sequence of 3D skeleton features directly from video clips. They take into account the hierarchy information of articulated kinematics[41, 42] or anthropometric priors like the symmetry and invariance of the skeleton's bone lengths[54] to infer the 3D joint positions in the camera coordinate and further model the dynamics of the skeleton kinematic tree. However, learning the abstract skeleton features is a highly nonlinear process. These algorithms do not contain enough information to reconstruct a realistic body structure or drive a skinned virtual 3D character. Other issues shift their spotlight towards model-based approaches. With a parametric human model like Skinned Multi-Person Linear (SMPL) model or Adam model[23, 34], model-based approaches[24, 25, 28, 29, 48, 73] can encode more anthropometric knowledge. In this way, the trained neural network can regress more realistic model parameters to fit the human model to the object in the RGB video. However, existing model-based methods inherently encounter the following two problems: (1) regressing rotation matrices is challenging and suffers from insufficient 3D in-the-wild ground truth, and (2) the phenomena of model-image misalignment are widespread because of the erroneous bone length estimation and 3D keypoint estimation.

To tackle these challenges, we look into the optical flow to take full advantage of the parametric model that subsumes more constraints like body shape and proportions of limb size. Instead of settling for indoor 3D datasets combined with diverse in-the-wild videos containing 2D manual keypoint annotations [21, 28, 29], some methods attempt to leverage optical flow to compute the discrepancy between the adjacent frames [62, 69]. It inspires us to exploit optical flow as a valuable feature to supervise the reconstructed 3D human mesh. On the one hand, optical flow can ensure the temporal consistency of the predicted parameters. On the other hand, it can regularize not only joints and bone lengths but shape of the mesh model. And the model with a well-predicted shape can, in turn, promote pose estimation. In this way, we close the loop between pose and shape learning.

In this work, we introduce HPOF, a novel temporal network trained to perform single-person 3D motion reconstruction from the monocular RGB video. Instead of directly extracting skeleton information, our network learns to regress model parameters about shape and pose. The core idea of HPOF is to propose a differentiable forward kinematics(FK) solution via the pose and shape decomposition. First, a temporal encoder is used to learn the bone length invariance in shape parameters and position continuity in pose parameters. Then, on account of the guarantee of bone length invariance, FK is naturally embedded into our network. Given the corresponding 3D joint position ground truth, HPOF can realize the inverse kinematics(IK) process in its backpropagation to learn the regression of pose parameters.

On the other hand, following the work of [21, 28, 29], we utilize 3D datasets combined with in-the-wild 2D datasets to enhance the diversity and realism of training videos. Considering the sizeable temporal continuity error caused by manually annotated 2D datasets, we use the optical flow as an extra 2D cue of motion trajectories, improving the robustness and generality of HPOF. Besides, with the poses sampled from the large-scale 3D motion-capture dataset[36], we implement a motion discriminator to evaluate the motion sequences as a whole. Our model is supervised by regression losses along with an adversarial loss to minimize the reconstruction error between predicted and ground-truth 3D keypoints, 2D keypoints, control parameters, and motion trajectories.

The main contributions of this paper are summarized below:

- We introduce HPOF, a novel end-to-end baseline for 3D human motion reconstruction in video based on optical flow.
- We propose an effective optical-flow-based method to generate rich descriptive 2D supervision information to constrain the shape and pose of the parametric model.
- Our method establishes a positive correlation between pose and shape prediction and improves their prediction substantially at the same time. It mitigates the problems of rotation parameter regression and model-image alignment.
- Our method surpasses other state-of-art models in terms of accuracy and smoothness.

# 2 RELATED WORK

With the boom in the development of deep neural networks, numerous research has been devoted to 3D human motion reconstruction in the last few years. Prior advances mainly focused on 2D pose recognition[7, 40, 57, 68], and improved 2D pose recognition has, in turn, facilitated the more challenging task of 3D human pose estimation[8, 44, 49, 51, 60].

**Skeleton-based 3D pose estimation:** Early paradigms in this field cast 3D human pose estimation as a task of locating the 3D joints on the kinematic tree. Accurate depth map and pose estimation algorithms [7, 55] are proposed to estimate the position of human joints, which provides new inspiration for the research of motion recognition based on human joints. The methods of 3D skeleton estimation can be mainly divided into two categories: one-step methods and two-step methods. One-step methods focus on directly estimating the 3D skeleton pose from the input image.

In comparison, two-step methods estimate 2D skeleton locations first and then upgrade the 2D joints to 3D locations by a learned dictionary of 3D skeleton [1, 63, 75] or regression [12, 37, 45, 58]. 3D skeleton representations vary from 3D Heatmap [50], location map [38] to 2D Heatmap with depth region [74]. Recently, Motion capture (Mocap) and other technologies have been used to collect accurate data and corresponding ground truth, contributing to the impressive performance of these methods. However, one of the skeleton-based 3D pose estimation challenges is that semantically similar actions may not necessarily be numerically similar. The human structural information implicitly estimated by a model may not be realistic.

**Model-based 3D pose and shape estimation:** The parametric human body model contains abundant prior knowledge of the human body. Many pioneers have been committed to predicting the natural 3D pose and shape through a parametric human body model [3, 35, 46]. Compared with the direct regression of 3D human shape and posture [30, 39, 65], adopting a parametric model can reduce the prediction difficulty and provide more convenience for downstream applications since the resulting model is controllable and reasonable. Bogo et al. [5, 32] propose the first method to automatically estimate the 3D pose and shape of the human body from a single unconstrained image. Experiments show that 2D joints alone carry a large amount of information about body shape. This method later gets further developed and extended [33, 43, 46, 52, 64, 65, 71]. To solve the depth ambiguity [5, 32] caused by the input RGB image, many algorithms try to introduce various intermediate variables to improve their performance, such as 2D heatmap input [64], keypoints, silhouettes [52] and semantic part segmentation [44]. Choutas et al. [10] propose ExPose with body-driven attention to reinforce regression on motion as well as hands. Furthermore, some studies exploit temporal context to acquire better performance in video tasks [4, 28].

**Optical Flow in Pose Estimation:** A key advantage of our approach is to constrain the trajectory of a surface model through the synthetic optical flow between successive frames. Several works have been presented exploiting optical flow for pose estimation. Brox et al. [6] use optical flow for 3D pose tracking of rigid objects. At the same time, Fragkiadaki et al. first [13] compute an articulated optical flow field to deal with large part rotations. Tung et al. [62] differentiably match the 3D motion vector projections against their estimated 2D optical flow vectors to realize end-to-end self-supervised learning of motion reconstruction. To enforce photometric consistency in the model textures, Xiang et al. [69] extract the projection of fitted mesh models on the input images and use optical flow to compute the discrepancy between these textures. Previous optical-flow-based methods are either limited to the coarse application of sparse optical flow or require sophisticated calculations like vertex visibility estimation and texture extraction. Our approach applies optical flow to the rasterization of the mesh model in a simple yet effective and feed-forward way, which realizes pixel-wise fine-grain supervised learning.

# 3 METHOD

In this section, we present the solution for 3D human pose reconstruction of video sequences. Fig. 2 shows an instantiation of the proposed HPOF. First, in §3.1, we briefly introduce the pre-knowledge of forwarding kinematics(FK) and its combination with the SMPL model. In §3.2, we present the overall architecture of HPOF. Then, in §3.3, we elaborate on our proposed solution for applying optical flow to supervised learning. Finally, we provide the practical implementation details in §3.4.

## 3.1 Preliminary

**Forward Kinematics:** Given the relative rotation matrix sequence $\mathcal{R} = \{R_{parent(k),k}\}_{k=1}^{K}$, $R \in \mathbb{R}^{3\times 3}$ and initial pose set $\mathcal{T} = \{t_k\}_{k=1}^{K}$, $t \in \mathbb{R}^3$, forward kinematics refers to the process of calculating the joint positions $p_k \in \mathbb{R}^3$ from the joint rotations.

$$p_k = R_k(t_k - t_{parent(k)}) + p_{parent(k)} \tag{1}$$

where K is the number of joints, $R_{parent(k),k}$ means the rotation matrix of leaf joint k relative to its parent joint parent(k), $R_k$ is the global rotation matrix of joint k, and can be computed in a recursive manner: $R_k = R_{parent(k)} R_{parent(k),k}$.

**SMPL Model:** In this paper, we try to fit a SMPL model to the human silhouette in the target image. SMPL has been extensively applied in pose estimation tasks [14, 22, 47]. The 3D mesh model can be controlled by parameters $\Theta = (\theta, \beta) \in \mathbb{R}^{3K+10}$, where $\theta \in \mathbb{R}^{K\times 3}$ are the pose parameters representing the relative rotations of K-1 joints concerning their parent joints and global body rotation of the root joint in the form of axis-angle, $\beta \in \mathbb{R}^{10}$ are the shape parameters that consist of the first ten orthogonal bases of PCA feature space. SMPL first transforms the axis angle $\vec{\theta}_k$ into rotation matrix $R_{parent(k),k}$ for each joint k using the Rodrigues' rotation formula:

$$R_{parent(k),k} = \mathcal{I} + sin(||\theta_j||)[\hat{\vec{\theta}}_j]_\times + (1 - cos(||\theta_j||)[\hat{\vec{\theta}}_j]_\times^2 \tag{2}$$

where $\mathcal{I}$ is the identity matrix, $\hat{\vec{\theta}} = \frac{\vec{\theta}}{||\vec{\theta}||}$ is the unit vector and $[\vec{\theta}]_\times$ is the skew symmetric matrix of $\vec{\theta}$. Then, we compute the forward process in homogeneous coordinates like:

$$H_k = \prod_{j\in A(k)} \left[\begin{array}{c|c} R_{parent(j),j} & t_j - t_{parent(j)} \\ \hline 0 & 1 \end{array}\right] \tag{3}$$

$$p_k = H_k[:3, 3] \tag{4}$$

So it is convenient to perform the FK progress $\mathcal{P} = FK(\mathcal{T}, \mathcal{R})$ in a one-shot recursion and returns the 3D positions of K joints $\mathcal{P} = \{p_k\}_{k=1}^{K}$, $p \in \mathbb{R}^3$. In this way, the differentiable FK function converts shape and pose parameters of SMPL into 3D joint positions.

**Combination of FK and SMPL:** The existing ground-truth dataset of SMPL parameters is insufficient for the efficient learning of our network. To mitigate the problem, the above FK solution is integrated into HPOF. HPOF supervises the predictions of 3D keypoint locations and further learns accurate pose parameters by the backpropagation of the FK layer. Because errors will accumulate along the kinematic tree during the FK process, first we need to ensure the skeleton consistency. Details will be introduced in §3.2.

Given $\Theta = (\theta, \beta)$, SMPL mesh vertices $M \in \mathbb{R}^{n\times 3}$ are the output of a differentiable function $\mathcal{M}(\theta, \beta)$, with n=6890. Moreover, these vertices have corresponding mesh faces $f \in \mathbb{R}^{N\times 3}$, where N=13775. It is worth noting that SMPL finally predicts 49 joint locations, 24

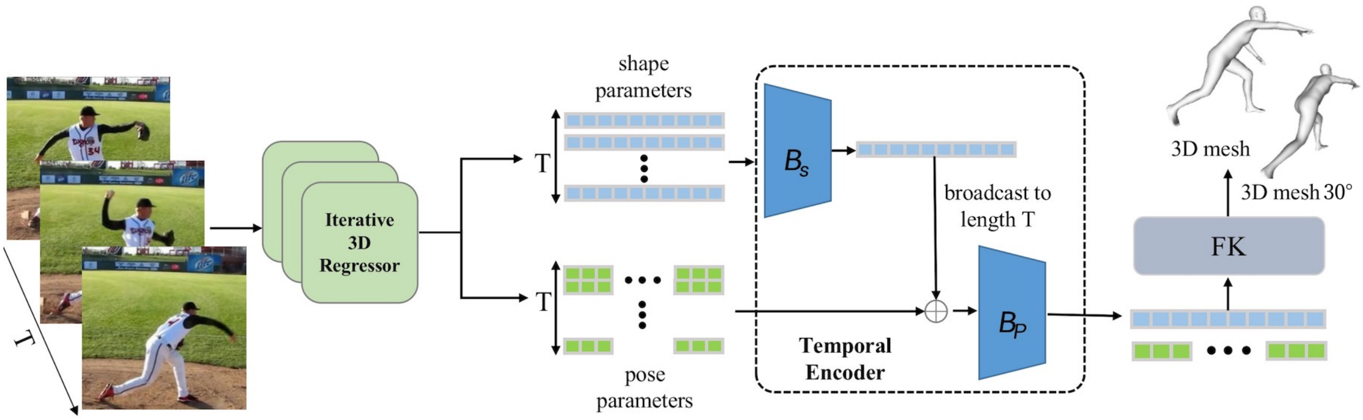


**Figure 2: HPOF architecture. HPOF first uses an iterative regressor to extract per-frame SMPL pose and shape parameters. Then the extracted parameters of past and current frames are fed into a temporal encoder trained to tune the skeleton inconsistency with $B_s$ and pose in-continuity with $B_p$. Finally, an FK layer converts pose parameters to joint locations of SMPL model.**

of them are obtained by FK and the rest are the linear combination results of mesh vertices.

## 3.2 Network Architecture

The overall framework of HPOF is shown in Fig. 2. Given an input video $V = \{I_t\}_{t=1}^{T}$, where $I_t \in \mathbb{R}^{H\times W\times 3}$ can denote each frame containing a single person, HPOF aims to decompose the learning process of pose and shape parameters $\{\Theta_t\}_{t=1}^{T}$ of the SMPL body model and substantially boost the 3D human motion reconstruction. With an iterative regression convolutional neural network, we take each frame $I_t$ as input and output parameters $\Theta_t$. Then we take both past and current frame information into account and exploit a new 2-stage network temporal encoder to learn the skeleton consistency of shape parameters and motion continuity of pose parameters. Following the work of Kanazawa et al. [24] and Kocabas et al. [28], we further employ a sequence-based adversarial network to discriminate between real and fake human motion sequences from a coarse-grained level.

**Iterative regressor:** The intuition behind using an iterative architecture is that pose parameters are tough to learn in a one-shot forward. Given a frame $I_t$, the regressor with a pre-trained ResNet-50 backbone first yields features $f_t \in \mathbb{R}^{1024}$ [15] fed into the iterative module later to infer SMPL parameters recurrently. In particular, given the concatenation of the image feature $f_t$ and the prediction $\Theta_t^i$ of $i$th iteration, the iterative module extract the offset $\Delta\Theta_t^{i+1}$ for the next iteration. Then the parameter set is updated by $\Theta_t^{i+1} = \Theta_t^i + \Delta\Theta_t^{i+1}$. In particular, parameters are first initialized by the mean $\bar{\Theta}$, and the final estimation are denoted by $\{\Theta_1, \Theta_2, ..., \Theta_T\}$. We define the loss function of iterative regressor as:

$$L_{reg} = \sum_{t=1}^{T} ||\Theta_t - \hat{\Theta}_t||_2 \tag{5}$$

**temporal encoder:** Since the single-view task suffers from body occlusion and ambiguity in depth, single-image features are not sufficient enough to yield plausible and accurate pose estimation. We use a temporal encoder consisting of two stages: pose smoother $\boldsymbol{B_p}$ and skeleton controller $\boldsymbol{B_s}$ to make the current frame benefit from past frame information.

During training, $\Theta_t$ is first decomposed into $\theta_t$ and $\beta_t$. In particular, $\theta_t \in \mathbb{R}^{24\times 6}$ is a 6D continuous rotation representation[76] instead of axis angles. The sequence of $\{\beta_1, \beta_2, ..., \beta_T\}$ will be first fed into $\boldsymbol{B_s}$ to mitigate shape inconsistency. Since SMPL mesh model has already been rigged with skeletons, the consistency of skeleton's bone lengths can be guaranteed by that of shape parameters.

$$\beta^* = \boldsymbol{B_s}(\beta_1, \beta_2, ..., \beta_T) \tag{6}$$

Then $\beta^*$ will be broadcast to length T. The combination of set $\{\beta^*\}_1^T$ and $\{\theta_t\}_1^T$ is fed into $\boldsymbol{B_p}$ to generate more temporally coherent results $\{\tilde{\Theta}_1, \tilde{\Theta}_2, ..., \tilde{\Theta}_T\}$. Each $\tilde{\Theta}_i$ benefits from past pose information:

$$\tilde{\Theta}_i = \boldsymbol{B_p}(\Theta_{i-S+1}, \Theta_{i-S+2}, ..., \Theta_i) \tag{7}$$

where $S$ is the receptive field of $B_p$.

In the training phase, we supervise the pose parameters $\theta$:

$$L_{shape} = \sum_{t=1}^{T} ||\tilde{\theta}_t - \hat{\theta}_t||_2 \tag{8}$$

and the shape parameters $\beta$:

$$L_{pose} = ||\beta^* - \bar{\hat{\beta}}||_2 \tag{9}$$

where $\bar{\hat{\beta}}$ is the average of the ground truth $\{\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_T\}$.

Besidies, we also consider more loss function terms about 2D, 3D joint annotations and acceleration as:

$$L_{2D} = \sum_{t=1}^{T} ||x_t - \hat{x}_t||_2 \tag{10}$$

$$L_{3D} = \sum_{t=1}^{T} ||X_t - \hat{X}_t||_2 \tag{11}$$

$$L_{accel} = \sum_{t=0}^{T-2} ||X_t + X_{t+2} - 2X_{t+1} - \hat{X}_t - \hat{X}_{t+2} + 2\hat{X}_{t+1}||_2 \tag{12}$$

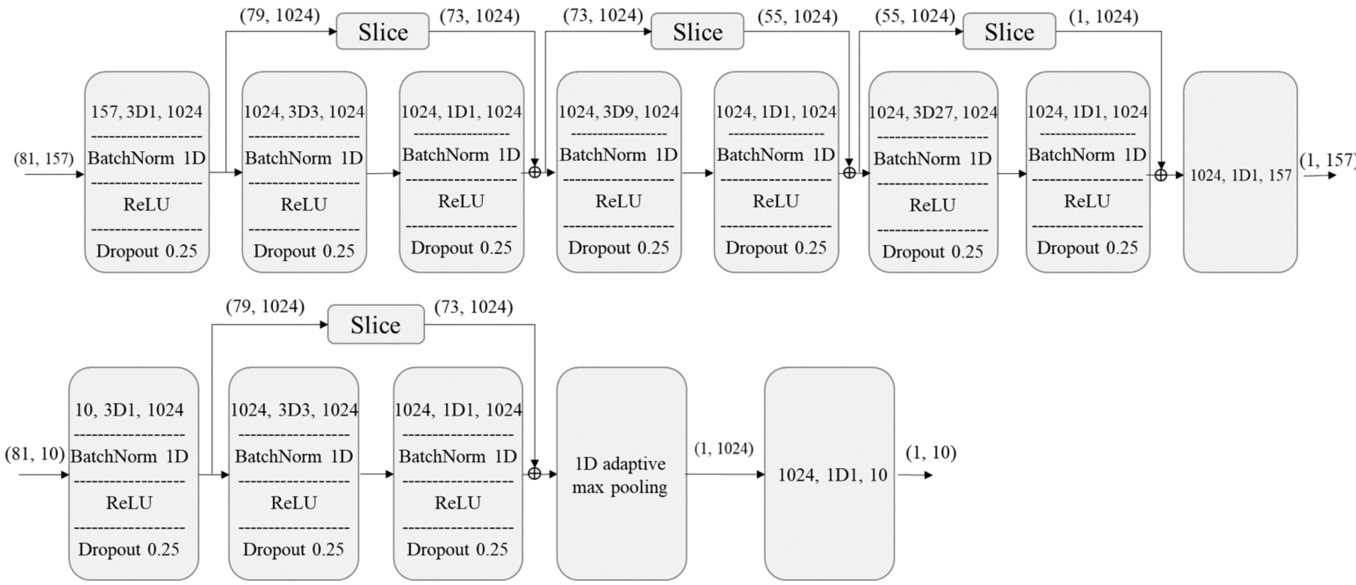


**Figure 3: Instantiation of temporal encoder consisting of $B_p$ and $B_s$. The input of $B_p$ contains pose, shape and cam parameters($157 = 24 * 6 + 10 + 3$) for a recpetive field of 81 frames (B=3 blocks), while the input of $B_s$ are the shape parameters($10$) of each frames. For each module, $157, 3D1, 1024$ denotes input channels $157$, kernels of size $3$ with dilation $1$ and output channels $1024$. In addition, the residuals are sliced from the head to match the output of the subsequent block.**

where $x$ denotes 2D keypoints, $X$ denotes 3D keypoints. The acceleration loss is simple yet effective to provide temporal constraint and assess the quality of temporal encoder in terms of acceleration. Specifically, the total loss function of HPOF is written as :

$$L_{HPOF} = L_{2D} + L_{3D} + L_{shape} + L_{pose} + L_{accel} + L_{GAN} + L_{opt\_flow} \tag{13}$$

where $L_{GAN}$, $L_{opt\_flow}$ will be explained below.

In practice, HPOF utilizes an one-dimensional convolutional network as temporal encoder. An adaptive pooling layer firstly functions as $\boldsymbol{B_s}$ to collapse the temporal axis so as to keep the shape parameters constant in the time domains. Then 1D convolution blocks with residual connection will be applied as $\boldsymbol{B_p}$ to yield smooth predictions over the temporal dimension. Our temporal encoder realizes parallel processing of multiple frames input, which is not possible with classic seq2seq recurrent models [9, 16]. Moreover, convolutional layers are dilated to expand temporal receptive field $S$. Its architecture is shown in Fig. 3.

In addition, Pavllo et al. [53] used 1D temporal convolution to directly lift 2D joint positions into 3D, while we consider additional skeleton consistency. On the other hand, Shi et al. [54] applied a model to directly generate the consistant skeleton from 2D joint positions, which is prone to overfitting issue. We solve the problem by taking the intermediate result from iterative regressor as input and feeding it to $\boldsymbol{B_s}$ further.

**sequence-based adversarial training:** In order to further supervise the generated human motions at the sequence level, HPOF adopts a adversarial training strategy to discriminate whether the predicted motion trajectories embedded on the manifold of plausible human motions. The discriminator $\boldsymbol{D}(\cdot)$ takes as input the sequence of pose parameters $\{\theta_1, \theta_2, ..., \theta_T\}$ (either from groundtruth or prediction) and outputs a value $\in [0, 1]$ to judge whether the sequence is rational. First, we need to train $\boldsymbol{D}(\cdot)$ with the objective:

$$L_D = \mathbb{E}_{\theta \sim p_{data}}[(\boldsymbol{D}(\theta_1, \theta_2, ..., \theta_T) - 1)^2] + \mathbb{E}_{\theta \sim p_g}[\boldsymbol{D}(\theta_1, \theta_2, ..., \theta_T)^2] \tag{14}$$

where $p_{data}$ is the empirical distribution of real motion and $p_g$ is the distribution of generated motion from HPOF.

The loss function that back propagated to HPOF architecture is:

$$L_{GAN} = \mathbb{E}_{\theta \sim p_g}[(\boldsymbol{D}(\theta_1, \theta_2, ..., \theta_T) - 1)^2] \tag{15}$$

## 3.3 Supervised-learning through Optical Flow

The intuition behind using optical flow is that fitting a full 3D mesh model to 2D keypoint annotations suffers from manual labeling noise. For two consecutive frames, optical flow refers to a 2D vector field that matches the displacement of a point from the current frame to the next. In this way, we exploit optical flow as the 2D cues of motion flow in our training framework.

Recent progress in optical flow estimation can achieve good performance [17–19, 56, 70]. We define $\boldsymbol{O_t} = (u, v) \in \mathbb{R}^{H \times W \times 2}$ as the dense optical flow field of each frame estimated by the state-of-the-art deep learning method RAFT [61]. Instead of directly matching the sparse 2D projections of visible mesh vertex motions between adjacent frames to optical flow vectors [62], we synthesize the raster images of SMPL models without the background and use the dense optical flow extracted from input images to regularize the motion between them.

Given SMPL mesh vertices $M = \mathcal{M}(\Theta) \in \mathbb{R}^{n \times 3}$, we first project vertices onto the screen with a weak-perspective camera model as:

$$q = s \prod (R\mathcal{M}(\Theta)) + t \tag{16}$$

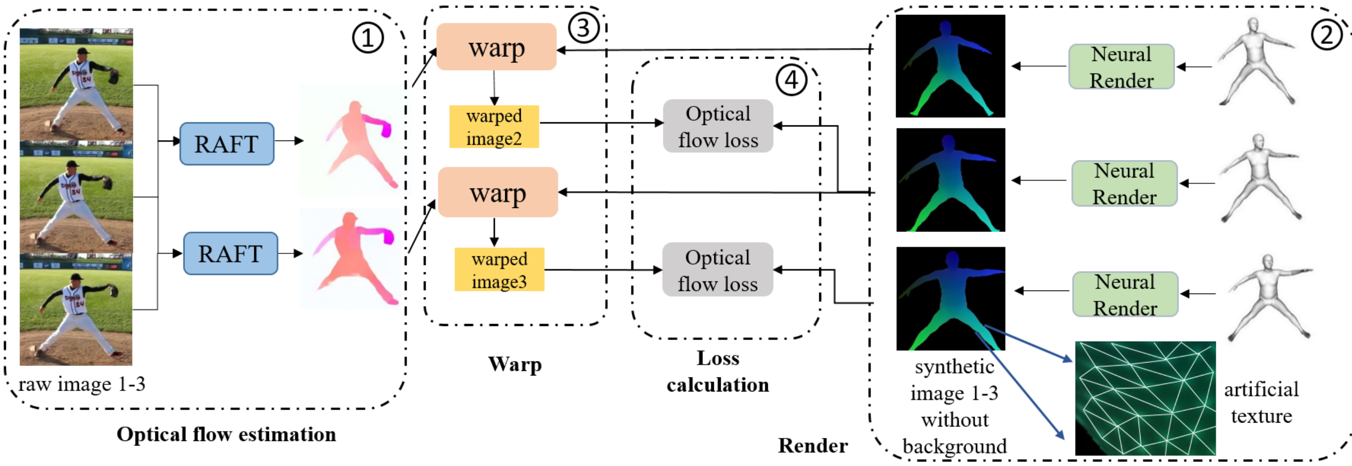


**Figure 4: Supervised-learning through optical flow. There are 4 steps to apply optical flow. First we estimate optical flow with RAFT, then we generate images of mesh model without background and warp them with optical flow to get the warped image of next frame, finally we calculate the loss and propagate it to the training process.**

where $q \in \mathbb{R}^{n\times 2}$ are 2D projections of vertex locations, $t \in \mathbb{R}^2$ and $s \in \mathbb{R}$ represent translation and scale parameters of camera parameters $c = (t, s)$ that are learned by HPOF and $R \in \mathbb{R}^{3\times 3}$ is a global rotation matrix and $\prod$ presents orthographic projection.

To draw the image of SMPL on the screen, we adopt a neural renderer network $\mathcal{R}(\cdot)$ [27] as a differentiable rasterizer. The neural renderer takes as input $q_t$, mesh faces $f$ and texture $T$ and generates image $m_t \in \mathbb{R}^{H\times W\times 3}$ via rendering from the 3D world as:

$$m_t = \mathcal{R}(q_t, f, T) \tag{17}$$

Note that we use an 'artificial texture' with gradient color to identify different parts of model mesh in a simple yet effective manner, rather than costly extract the texture map from the input image $I_t$. In this way, $m_t$ filters out the background noise and contains only the projections of 3D motion.

Then $m_t$ will be warped under the guidance of $O_t$ and output $\hat{m}_{t+1} \in \mathbb{R}^{H\times W\times 3}$, namely:

$$V(x, y, t) = V(x + u, y + v, t + 1) \tag{18}$$

where $V(x, y, t)$ is the intensity of light at pixel $(x, y)$ of $m_t$ and $V(x+u, y+v, t+1)$ is that of $\hat{m}_{t+1}$, which is treated as the groundtruth of frame $t + 1$. The loss function $L_{opt\_flow}$ is defined as:

$$L_{opt\_flow} = \sum_{t=2}^{T} \mathcal{L}_2(m_t, \hat{m}_t) \tag{19}$$

where $\mathcal{L}_2$ demotes the Mean Square Error loss and all these operations are differentiable. In this way, optical flow transfers pose knowledge of the preceding frame to provide short-term guidance for the current frame.

## 3.4 Implementation Details

In this subsection, we elaborate more details about the training and inference process of HPOF. Specifically, HPOF decomposes the training procedure into two phases. In terms of the iterative regressor, we use a ResNet-50 network to extract image features $f_t \in \mathbb{R}^{1024}$ followed by an iterative module with 3 stages to infer SMPL parameters. For the temporal encoder, we set the 1D convolutional module with 3 blocks as is shown in Fig. 3, resulting in the receptive field $S = 81$. We also use Adam optimizer with the learning rate of $1 \times 10^{-5}$ and $5 \times 10^{-5}$ for 3D regressor and temporal encoder respectively. And they will multiply a factor of 0.6 if the estimation does not improve for more than 5 epochs. The weighting coefficients are set as $\lambda_{2D} = 300$, $\lambda_{3D} = 300$, $\lambda_{pose} = 60$, $\lambda_{shape} = 0.06$, $\lambda_{accel} = 60$, $\lambda_{GAN} = 0.5$, $\lambda_{opt\_flow} = 0.0004$

During inference, the branch of optical flow estimation is removed. Given a video, HPOF first utilizes the iterative regressor to estimate the initial SMPL parameters $\Theta_1$, then $\Theta_1$ will be padded to a sequence of length T and pass through the temporal encoder composed of 1D dilated convolutional blocks. For subsequent frames, until the sequence is long enough, we will push the new frame at the end and pop the oldest one at the head.

# 4 EXPERIMENTS

In this section, we first describe the experimental setup in detail. Then we conduct ablation experiments and compare our model with some state-of-the-art approaches.

## 4.1 Experimental setup

**Dataset:** To compare with the previous research [24, 25, 29], we evaluate our model on their widely used benchmarks. For 2D datasets, there are in-the-wild datasets, including PoseTrack [2], PennAction [72] annotated with manual 2D joint labels. We unify their annotation types and filter images with less than six visible keypoints. Besides, InstaVariety [26] will be exploited to generate optical flow predictions. For 3D datasets, we employ 3D joint annotations from Human3.6M [20] and ground-truth SPML parameters from 3DPW [67] for training. Unlike typical 3D datasets that are labelled inside a studio, 3DPW has 3D ground truth collected outdoor. We leave aside part of 3DPW as the validation set.

**Table 1: Ablation experiments with structure selection**

| Models | PA-MPJPE↓ | MPJPE↓ | PVE↓ | Accel↓ | Inference Speed↓ |
|---|---|---|---|---|---|
| HPOF(transformers) | 51.9 | 79.3 | 94.0 | 25.3 | 4.3 |
| HPOF(GRU)-2 layers | 50.7 | 77.1 | 92.0 | 23.8 | 4.0 |
| HPOF(GRU)-3 layers | 50.6 | 77.2 | 92.3 | 23.1 | 6.4 |
| HPOF(tempConv)-w/o-OptFlow | 51.3 | 77.2 | 93.7 | 16.5 | - |
| HPOF(tempConv)-w/o-$B_s$ | 50.2 | 76.6 | 91.4 | 16.5 | - |
| HPOF(tempConv) | **49.4** | **73.9** | **88.2** | **16.3** | **1.9** |

**Data Preprocessing:** Each frame is cropped around the person and scaled to a uniform size 224×224 by an affine transformation. The affine transformation matrix needs to be preserved to further rasterize SMPL mesh vertices to the pixel space of original images to calculate the optical-flow-based loss. Moreover, we perform regular data augmentation, including random scaling and flipping.

**Evaluation Metrics:** We evaluate the performance of HPOF with several error metrics: Procrustes-aligned mean per joint position error (PA-MPJPE), mean per joint position error (MPJPE), Per Vertex Error (PVE), and acceleration error (mm/$s^2$).

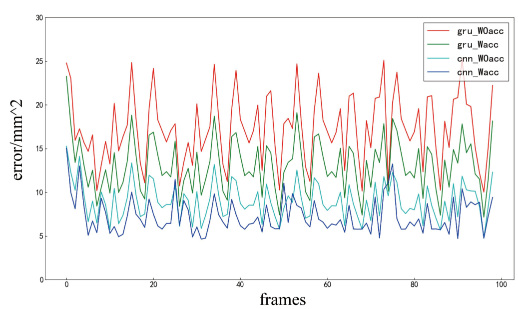


**Figure 6: Efficiency analysis of acceleration loss. Monitor the acceleration error of different structures during the test with or without acceleration loss function**

## 4.2 Ablation Analysis

In this subsection, we conduct ablation studies on 3DPW to analyze the efficacy of core modules. We fix the backbone of the iterative regressor as ResNet50 and vary the configurations of other modules.

Our primary concern is about the performance of the temporal encoder, which will be discussed from the following aspects: (1). structure selection of temporal encoder; (2). the temporal receptive field of $B_s$; (3). acceleration loss.

**Table 2: Ablation experiments with temporal receptive fields**

| Models | PA-MPJPE↓ | MPJPE↓ | PVE↓ | Accel↓ |
|---|---|---|---|---|
| HPOF(tempConv)-1 B | 52.5 | 80.5 | 94.5 | 15.3 |
| HPOF(tempConv)-2 B | **51.5** | **79.8** | **93.6** | **14.2** |
| HPOF(tempConv)-3 B | 55.0 | 82.3 | 96.6 | 17.1 |

To prove the rationality of our proposed temporal module. We replace it with other structures like Gated Recurrent Units(GRU) or transformers [66]. We use HPOF(tempConv/GRU/transformer) to denote HPOF with different temporal encoders. In this experiment, the transformer encoder consists of a stack of $N = 2$ identical layers. Each layer has two blocks: a self-attention block with 8 heads and a position-wise fully connected feed-forward block with 1024 hidden units. The multi-layer GRU we used has a hidden size of 1024.

Results of evaluation metrics and inference time(ms) per forward propagation are shown in Tab. 1. It can be seen that HPOF(tempConv) achieves the best accuracy-speed trade-off. However, experiments with transformer and multi-layer GRU model bring little effect but high computational cost. One intuitive explanation for this is that the task of fine-tuning $\{\theta_1, \theta_2, ..., \theta_T\}$ is simple enough to give more consideration to inference speed. On the other hand, We compare the estimation results from HPOF with and without optical flow module in Tab. 1. We can see that synthetic optical flows do indeed improve the performance of HPOF.

The above experiment provides definitive evidence that optical-flow-based modules can facilitate HPOF for extracting short-term temporal features. At the same time, receptive field $S$ controlled by the number of blocks in $B_p$ is the main factor that affects the ability to capture long-term temporal contexts. We further focus on the receptive field of $B_p$ to make a reasonable trade-off between the capture of short-term and long-term temporal information. We test on increasing numbers of blocks to find appropriate receptive fields, which is denoted by HPOF(tempConv)-xB, where $x \in \{1, 2, 3\}$ corresponding to the receptive field size 9, 27 and 81. From Tab. 2, we can observe that HPOF(tempConv)-2B yields the best results. However, when the receptive field is larger, too much past information will affect the final performance. Note that the experiment of Tab. 2 does not take the optical flow module into account.

Moreover, we compare the results with and without $B_s$ in Tab. 1, we can find that the application of $B_s$ significantly reduces the error on the joint positions and mesh vertices.

In terms of acceleration loss, we analyze the variation trend of acceleration loss during testing and intercept some frames to visualize. As shown in Fig. 6, we can see a generalized decrease in error when the acceleration loss function is utilized across all frames, proving that acceleration loss is a simple yet effective way to regulate our predictions.

## 4.3 Comparison to state-of-the-art results

Tab. 3 shows the comparisons of our method with state-of-the-arts on the 3DPW dataset. In particular, HPOF(+) trained on the dataset


**Figure 5: Qualitative comparison between HPOF (*white*) and VIBE [28](*green*) both in door and outdoor. The models are tested on NVIDIA GTX1080 GPU, and HPOF is significantly faster than VIBE. As shown in the figure, VIBE performs worse than our approach in estimating the pose of the extremities. The phenomena of model-image misalignment are obvious.**

similar to [24, 28, 59], while HPOF(†) and VIBE(†) also use 3DPW for training. We evaluate the performance of models with all the metrics mentioned above. Since both of state-of-the-arts VIBE[28] and SPIN[29] use the same regression module, we experiment HPOF with pre-trained HMR from SPIN[29] as its iterative regressor. From Tab. 3, we observe significant improvements in the MPJPE and PVE and acceleration metrics. In particular, when compared with VIBE, HPOF reduces MPJPE and PVE by more than 10 per cent (73.9 vs 82.9) and (88.2 vs 99.1) and achieves 3x processing speed in terms of temporal module mentioned in the ablation study (1.9ms vs 4.0ms per image), demonstrating its outperforming efficiency.

**Table 3: Quantitative comparison with other methods on 3DPW dataset**

| Models | PA-MPJPE↓ | MPJPE↓ | PVE↓ | Accel↓ |
|---|---|---|---|---|
| Arnab *et al.* [4] | 72.2 | - | - | - |
| Kolotouros *et al.* [31] | 70.2 | - | - | - |
| Kolotouros *et al.* [29] | 59.2 | 96.9 | 116.4 | 29.8 |
| Kanazawa *et al.* [25](+) | 72.6 | 116.5 | 139.3 | **15.2** |
| Doersch *et al.* [11] | 74.7 | - | - | - |
| Sun *et al.* [59](+) | 69.5 | - | - | - |
| VIBE *et al.* [28](+) | 56.5 | 93.5 | 113.4 | 27.1 |
| HPOF(+) | 53.1 | 84.7 | 97.5 | 25.6 |
| VIBE *et al.* [28](†) | 51.9 | 82.9 | 99.1 | 23.4 |
| HPOF(†) | **49.4** | **73.9** | **88.2** | 16.3 |

Furthermore, we conduct a visualization experiment to compare the results of HPOF and VIBE. As is shown in Fig. 5, VIBE fails to track the details of limbs, such as the hands and feet. The leading cause for this phenomenon is that VIBE only utilizes a motion discriminator to tell realistic motion from an overall perspective. Sometimes, to keep the rationality of actions, VIBE tends to be more conservative with significant range movement and ignore some details. While for HPOF, we use a large temporal receptive field to guarantee long-term motion tendency and optical flow loss to capture short-term saltation information.

## 5 DISCUSSION

In this paper, we present an end-to-end approach HPOF to realize 3D human reconstruction from monocular video. Considering the temporal consistency between consecutive frames, HPOF adopts a temporal encoder to learn the skeleton invariance and pose continuity across all frames. With a large receptive field, the temporal encoder can realize long-term motion perception. On the other hand, HPOF employs the synthetic optical flow as extra 2D cues of motion trajectories to facilitate HPOF for capturing short-term temporal information. Our method is fully differentiable and allows simultaneously training of 3D pose and shape in an end-to-end manner. We carefully set up the experiments to prove that HPOF can surpass state-of-the-art methods.

In addition to addressing the problems mentioned above, future works include: (1). Using physics-based trajectory optimization to predict the whole body reasonably from a limited visual range; (2). Realizing real-time inference of HPOF for deploying the model to 3D development platforms like Unity; (3). Learning the pose prior from different motion styles.

## ACKNOWLEDGMENTS

We gratefully acknowledge support from Shanghai Sailing Program (20YF1421200), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and VokaTech.

## REFERENCES

[1] Ijaz Akhter and Michael J. Black. 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 1446–1455. https://doi.org/10.1109/CVPR.2015.7298751

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. 2018. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 5167–5176. https://doi.org/10.1109/CVPR.2018.00542

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

[4] Anurag Arnab, Carl Doersch, and Andrew Zisserman. 2019. Exploiting Temporal Context for 3D Human Pose Estimation in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 3395–3404. https://doi.org/10.1109/CVPR.2019.00351

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. 2016. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 9909)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). 561–578. https://doi.org/10.1007/978-3-319-46454-1_34

[6] Thomas Brox, Bodo Rosenhahn, Daniel Cremers, and Hans-Peter Seidel. 2006. High Accuracy Optical Flow Serves 3-D Pose Tracking: Exploiting Contour and Flow Based Constraints. In *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 3952)*, Ales Leonardis, Horst Bischof, and Axel Pinz (Eds.). 98–111. https://doi.org/10.1007/11744047_8

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 1 (2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

[8] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby T. Tan. 2019. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 723–732. https://doi.org/10.1109/ICCV.2019.00081

[9] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). 1724–1734. https://doi.org/10.3115/v1/d14-1179

[10] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. 2020. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*. Springer, 20–40.

[11] Carl Doersch and Andrew Zisserman. 2019. Sim2real transfer learning for 3D human pose estimation: motion to the rescue. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 12929–12941. http://papers.nips.cc/paper/9454-sim2real-transfer-learning-for-3d-human-pose-estimation-motion-to-the-rescue

[12] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. 2018. Learning Pose Grammar to Encode Human Body Configuration for 3D Pose Estimation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). 6821–6828. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16471

[13] Katerina Fragkiadaki, Han Hu, and Jianbo Shi. 2013. Pose from Flow and Flow from Pose. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. 2059–2066. https://doi.org/10.1109/CVPR.2013.268

[14] Peng Guan, Alexander Weiss, Alexandru O. Balan, and Michael J. Black. 2009. Estimating human shape and pose from a single image. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. 1381–1388. https://doi.org/10.1109/ICCV.2009.5459300

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[17] Markus Hofinger, Samuel Rota Bulò, Lorenzo Porzi, Arno Knapitsch, Thomas Pock, and Peter Kontschieder. 2020. Improving Optical Flow on a Pyramid Level. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII (Lecture Notes in Computer Science, Vol. 12373)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). 770–786. https://doi.org/10.1007/978-3-030-58604-1_46

[18] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. 2018. LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 8981–8989. https://doi.org/10.1109/CVPR.2018.00936

[19] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 1647–1655. https://doi.org/10.1109/CVPR.2017.179

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 7 (2014), 1325–1339. https://doi.org/10.1109/TPAMI.2013.248

[21] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. 2020. Exemplar Fine-Tuning for 3D Human Pose Fitting Towards In-the-Wild 3D Human Pose Estimation. *CoRR* abs/2004.03686 (2020). arXiv:2004.03686 https://arxiv.org/abs/2004.03686

[22] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart C. Nabbe, Iain A. Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2019. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1 (2019), 190–204. https://doi.org/10.1109/TPAMI.2017.2782743

[23] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 8320–8329. https://doi.org/10.1109/CVPR.2018.00868

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-End Recovery of Human Shape and Pose. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 7122–7131. https://doi.org/10.1109/CVPR.2018.00744

[25] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. 2019. Learning 3D Human Dynamics From Video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 5614–5623. https://doi.org/10.1109/CVPR.2019.00576

[26] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. 2019. Learning 3D Human Dynamics From Video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 5614–5623. https://doi.org/10.1109/CVPR.2019.00576

[27] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Neural 3D Mesh Renderer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 3907–3916. https://doi.org/10.1109/CVPR.2018.00411

[28] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. 5252–5262. https://doi.org/10.1109/CVPR42600.2020.00530

[29] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 2252–2261. https://doi.org/10.1109/ICCV.2019.00234

[30] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. 2019. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4501–4510.

[31] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. 2019. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 4501–4510. https://doi.org/10.1109/CVPR.2019.00463

[32] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. 2017. Unite the People: Closing the Loop Between 3D and 2D Human Representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[33] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. 2017. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6050–6059.

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (2015), 248:1–248:16. https://doi.org/10.1145/2816795.2818013

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

[36] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture As Surface Shapes. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 5441–5450. https://doi.org/10.1109/ICCV.2019.00554

[37] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2659–2668. https://doi.org/10.1109/ICCV.2017.288

[38] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*. 506–516. https://doi.org/10.1109/3DV.2017.00064

[39] Gyeongsik Moon and Kyoung Mu Lee. 2020. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. *arXiv preprint arXiv:2008.03713* (2020).

[40] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*

*(Lecture Notes in Computer Science, Vol. 9912)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). 483–499. https://doi.org/10.1007/978-3-319-46484-8_29

[41] Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan. 2018. Pose Partition Networks for Multi-person Pose Estimation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 11209)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). 705–720. https://doi.org/10.1007/978-3-030-01228-1_42

[42] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. 2019. Single-Stage Multi-Person Pose Machines. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 6950–6959. https://doi.org/10.1109/ICCV.2019.00705

[43] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*. IEEE, 484–494.

[44] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. 2018. Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation. In *2018 International Conference on 3D Vision, 3DV 2018, Verona, Italy, September 5-8, 2018*. 484–494. https://doi.org/10.1109/3DV.2018.00062

[45] Sungheon Park, Jihye Hwang, and Nojun Kwak. 2016. 3D Human Pose Estimation Using Convolutional Neural Networks with 2D Pose Information. In *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 9915)*, Gang Hua and Hervé Jégou (Eds.). 156–169. https://doi.org/10.1007/978-3-319-49409-8_15

[46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10975–10985.

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 10975–10985. https://doi.org/10.1109/CVPR.2019.01123

[48] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. 2020. Human Mesh Recovery from Multiple Shots. *CoRR* abs/2012.09843 (2020). arXiv:2012.09843 https://arxiv.org/abs/2012.09843

[49] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. 2018. Ordinal Depth Supervision for 3D Human Pose Estimation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 7307–7316. https://doi.org/10.1109/CVPR.2018.00763

[50] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. 2017. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 1263–1272. https://doi.org/10.1109/CVPR.2017.139

[51] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. Learning to Estimate 3D Human Pose and Shape From a Single Color Image. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 459–468. https://doi.org/10.1109/CVPR.2018.00055

[52] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 459–468.

[53] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 7753–7762. https://doi.org/10.1109/CVPR.2019.00794

[54] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. MotioNet: 3D Human Motion Reconstruction from Monocular Video with Skeleton Consistency. *ACM Trans. Graph.* 40, 1 (2020), 1:1–1:15. https://doi.org/10.1145/3407659

[55] Jamie Shotton, Andrew W. Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. 1297–1304. https://doi.org/10.1109/CVPR.2011.5995316

[56] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 8934–8943. https://doi.org/10.1109/CVPR.2018.00931

[57] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 5693–5703. https://doi.org/10.1109/CVPR.2019.00584

[58] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. 2017. Compositional Human Pose Regression. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2621–2630. https://doi.org/10.1109/ICCV.2017.284

[59] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. 2019. Human Mesh Recovery From Monocular Images via a Skeleton-Disentangled Representation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 5348–5357. https://doi.org/10.1109/ICCV.2019.00545

[60] Vince Tan, Ignas Budvytis, and Roberto Cipolla. 2017. Indirect deep structured learning for 3D human body shape and pose prediction. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. https://www.dropbox.com/s/wrx7dzegq3wim04/0722.pdf?dl=1

[61] Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12347)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). 402–419. https://doi.org/10.1007/978-3-030-58536-5_24

[62] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. 2017. Self-supervised Learning of Motion Capture. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5236–5246. https://proceedings.neurips.cc/paper/2017/hash/ab452534c5ce28c4fbb0e102d4a4fb2e-Abstract.html

[63] Hsiao-Yu Fish Tung, Adam W. Harley, William Seto, and Katerina Fragkiadaki. 2017. Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation from Unpaired Supervision. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 4364–4372. https://doi.org/10.1109/ICCV.2017.467

[64] Hsiao-Yu Fish Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. 2017. Self-supervised learning of motion capture. *arXiv preprint arXiv:1712.01337* (2017).

[65] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 20–36.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[67] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X (Lecture Notes in Computer Science, Vol. 11214)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). 614–631. https://doi.org/10.1007/978-3-030-01249-6_37

[68] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 4724–4732. https://doi.org/10.1109/CVPR.2016.511

[69] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. 2019. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 10965–10974. https://doi.org/10.1109/CVPR.2019.01122

[70] Gengshan Yang and Deva Ramanan. 2019. Volumetric Correspondence Networks for Optical Flow. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 793–803. https://proceedings.neurips.cc/paper/2019/hash/bbf94b34eb32268ada57a3be5062fe7d-Abstract.html

[71] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. 2018. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2148–2157.

[72] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. 2013. From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. 2248–2255. https://doi.org/10.1109/ICCV.2013.280

[73] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. 2019. End-to-End Hand Mesh Recovery From a Monocular RGB Image. In *2019 IEEE/CVF*

*International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 2354–2364. https://doi.org/10.1109/ICCV.2019.00244

[74] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. 2017. Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 398–407. https://doi.org/10.1109/ICCV.2017.51

[75] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. 2016. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 4966–4975. https://doi.org/10.1109/CVPR.2016.537

[76] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the Continuity of Rotation Representations in Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 5745–5753. https://doi.org/10.1109/CVPR.2019.00589