Coordinate Transformer: Achieving Single-stage Multi-person Mesh Recovery from Videos

Haoyuan Li[1*]   Haoye Dong[2*]   Hanchao Jia[3]   Dong Huang[2]   Michael C. Kampffmeyer[4]
Liang Lin[5†]   Xiaodan Liang[1,6†]

[1]Shenzhen campus of Sun Yat-sen University   [2]Carnegie Mellon University
[3]Samsung Research China – Beijing (SRC-B)   [4]UiT The Arctic University of Norway
[5]Sun Yat-sen University   [6]Mohamed bin Zayed University of AI

lihy285@mail2.sysu.edu.cn, donghaoye@cmu.edu, hanchao.jia@samsung.com, donghuang@cmu.edu

michael.c.kampffmeyer@uit.no, linliang@ieee.org, xdliang328@gmail.com

Abstract

*Multi-person 3D mesh recovery from videos is a critical first step towards automatic perception of group behavior in virtual reality, physical therapy and beyond. However, existing approaches rely on multi-stage paradigms, where the person detection and tracking stages are performed in a multi-person setting, while temporal dynamics are only modeled for one person at a time. Consequently, their performance is severely limited by the lack of inter-person interactions in the spatial-temporal mesh recovery, as well as by detection and tracking defects. To address these challenges, we propose the **Coord**inate trans**Former** (Coord-Former) that directly models multi-person spatial-temporal relations and simultaneously performs multi-mesh recovery in an end-to-end manner. Instead of partitioning the feature map into coarse-scale patch-wise tokens, Coord-Former leverages a novel Coordinate-Aware Attention to preserve pixel-level spatial-temporal coordinate information. Additionally, we propose a simple, yet effective Body Center Attention mechanism to fuse position information. Extensive experiments on the 3DPW dataset demonstrate that CoordFormer significantly improves the state-of-the-art, outperforming the previously best results by 4.2%, 8.8% and 4.7% according to the MPJPE, PAMPJPE, and PVE metrics, respectively, while being 40% faster than recent video-based approaches. The released code can be found at https://github.com/Li-Hao-yuan/CoordFormer.*

1. Introduction

Considerable progress has been made on monocular 3D human pose and shape estimation from images [5, 35, 17, 22, 39] due to extensive efforts of computer graph-
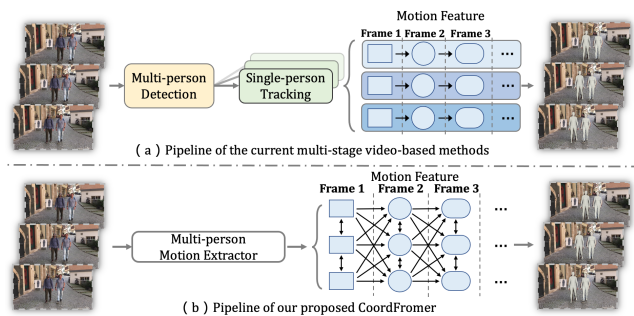

Figure 1. Comparison of video-based multi-person mesh recovery pipelines. (a) **Multi-stage** pipelines [20, 8, 44, 38, 40] explicitly generate tracklets and model single-person temporal mesh sequences independently. (b) Our **single-stage** CoordFormer implicitly matches persons across frames and simultaneously models multi-person mesh sequences in an end-to-end manner.

ics and augmented/virtual reality researchers. However, while frame-wise body mesh detection is feasible, many applications require direct video-based pipelines to avoid spatial-temporal incoherence and missing frame-based detections [20, 8, 38].

Existing video-based methods follow a multi-stage design that involves using a 2D person detector and tracker to obtain the image sequences of a single-person for pose and shape estimation [18, 20, 38, 41, 40]. More specifically, these methods first detect and crop image patches that contain persons, then track these individuals across frames, and associate each cropped image sequence with a person. The frame-level or sequence-level features are then extracted and used to regress 3D human mesh sequences under spatial and temporal constraints. However, the accuracy of the detection and tracking stage greatly affects the performance of these multi-stage approaches, making them particularly sensitive to false, overlapping, and missing detections. Moreover, these multi-stage approaches have a considerable computation cost and lack real-time perspec-

---

*Both authors contributed equally to this work as co-first authors.
†Corresponding author.

tives since the single-person meshes can only be recovered sequence-by-sequence after detection and tracking.

To address the above issues, we introduce CoordFormer, the first single-stage approach for multi-person 3D mesh recovery from videos that can be trained in an end-to-end manner. As shown in Fig. 1, our method differs from current state-of-the-art approaches [20, 8, 44, 38, 40] by being a single-stage pipeline that implicitly performs detection and tracking through the interaction of feature representations, producing multiple mesh sequences simultaneously.

In particular, CoordFormer leverages a multi-head framework to predict a body center heatmap, which is encoded using our proposed Body Center Attention (BCA). BCA serves as a weak/intermediate person detector that focuses the framework-wide feature representations on potential body centers. Many-to-many temporal-spatial relations among people and across frames are then derived from the BCA-focused features and directly mapped to mesh sequences using our novel Coordinate-Aware Attention (CAA). CAA is integrated into a Spatial-Temporal Transformer (ST-Trans) [44, 26, 24] to capture non-local context relations at the pixel level. See Fig. 2 for an illustration of CAAs motivation. Facilitated by BCA and CAA, CoordFormer advances existing video mesh recovery solutions beyond explicit detection, tracking and sequence modeling. Under various experimental settings on the 3DPW dataset, CoordFormer significantly outperforms the best results of state-of-the-art by 4.2%, 8.8% and 4.7% on MPJPE, PAMPJPE and PVE metrics, respectively. CoordFormer also improves inference speed by 40% compared to the state-of-the-art video-based approaches [20, 38]. Moreover, we demonstrate that enhancing and capturing pixel-level coordinate information significantly benefits the performance under multi-person scenarios.

The main contributions of this work are as follows:

- We propose the first single-stage multi-person video mesh recovery approach, where our BCA mechanism fuses position information and our CAA module enables end-to-end multi-person model training.

- We demonstrate that the pixel-level coordinate correspondence is the most critical factor for performance.

- Extensive experiments on challenging 3D pose datasets demonstrate that the proposed method achieves significant improvements, outperforming the state-of-the-art methods.

2. Related Work

Single image-based 3D human pose and shape estimation. Single image-based methods typically train models to estimate pose, shape, and camera parameters from images, and then output a 3D human mesh using parametric human body models, for example, the SMPL [27]. Significant progress has been made in this area by leveraging
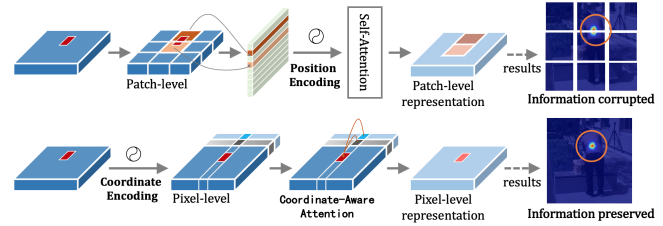

Figure 2. The motivation of our Coordinate-Aware Attention (CAA) module in CoordFormer. (Top) The standard Transformer based modules (such as ST-Trans [44, 26, 24]) model patch-level dependency, which results in corruption of pixel-level features. (Bottom) CAA encodes pixel-level spatial-temporal coordinates and preserves pixel-level dependencies in features.

inherent properties of the 3D human, supervising the models using 2D keypoints [22], semantic segmentation [13], texture consistency [30], interpenetration and depth [13], body shape [9] and IUV maps [16]. However, they primarily use a multi-stage paradigm that is limited by the first stage. BMP [42] improves upon this by proposing a single-stage model that is more robust to occlusions through inter-instance ordinal relation supervision and taking into account body structure. Concurrently, ROMP [32] adopts a multi-head design which predicts a Body Center heatmap and a Mesh Parameter map. Via parsing the Body Center heatmap and sampling from the Mesh Parameters map, ROMP is able to extract and predict 3D human meshes for multi-person scenarios. BEV [33] extends upon this by further leveraging relative depth information to effectively avoid mesh collision in the single-stage design, as well as age information. Despite these advances in estimating human pose and shape from single images, these above methods are restricted to single images and poorly capture motion relations of spatial interaction.

Video-based 3D human pose and shape estimation. The existing video-based methods are similarly built based on SMPL and extract SMPL parameters from frames [34, 18, 3]. However, in these methods, a greater focus is put on modeling temporal consistency and motion coherence. As their image counterparts, video methods follow a two-stage design where people are first detected and features of the bounding-boxes are extracted. In the second stage, tracking is used to capture the motion sequence and refine the pose and shape estimation. More specifically, Sun et. al [34] disentangle skeleton features for improving the learning of spatial features and develop a self-attention temporal network for modelling temporal relations. Additionally, they propose an unsupervised adversarial training strategy for guiding the representation learning of motion dynamics in the video. HMMR [18] proposes a temporal encoder that learns to capture 3D human dynamics in a semi-supervised manner, while Arnab *et al.* [3] presents a bundle-adjustment-based algorithm for human mesh optimization and a new dataset consisting of in-the-wild videos. Compared to tem-

poral convolutions and optimization across frames, recurrent structures and attention mechanisms provide superior motion information for mesh regression. VIBE [20] first extracts features from each frame and uses a temporal encoder, i.e. bidirectional gated recurrent units (GRU), to model temporal relations and obtain consistent motion sequences. For more realistic mesh results, the discriminator adopts an attention mechanism to weight the contribution of distinct frames. TCMR [42] proposes the PoseForecast approach composed of GRUs, which integrates and refines static features by fusing pose information from past and future frames to ensure motion consistency. MPS-Net [38] further extends the non-local concept to capture motion continuity, as well as temporal similarities and dissimilarities. MPS-Net further develops a hierarchical attentive feature integration to refine temporal features observed from past and future frames. However, these methods only optimize the motion of individual people and ignore the spatial interactions among people, which is crucial in multi-person scenarios. CoordFormer, instead, adopts a single-stage design for multi-person mesh recovery, aiming at modeling spatial-temporal relations and constraints across frames.

3. CoordFormer

Overview. We present the CoordFormer framework (see Fig. 3) to advance multi-person temporal-spatial modelling for video-based 3D mesh recovery. We take inspiration from single-stage image-based approaches for mesh recovery [32] and leverage a multi-head design that predicts a Body Center heatmap as well as a Mesh parameter map. To further capture the spatial-temporal relations, we introduce two novel modules: (1) the BCA mechanism (Sec. 3.1), which focuses spatial-temporal feature extraction on persons for better performance and faster convergence, and (2) the CAA module (Sec. 3.2) incorporated in a Spatial-Temporal Transformer (Sec. 3.3), which preserves pixel-level spatial-temporal coordinate information. CAA avoids the spatial information degradation which usually occurs in the patch-level tokenization of standard vision transformers.

For completeness and notation consistency we briefly present the Body Center heatmap and the Mesh Parameter map which are predicted by the backbone network. They follow [32] and are computed for all the T frames in a video.

Body Center Heatmap $\mathbf{C}_m \in \mathbb{R}^{T \times 1 \times H \times W}$: $\mathbf{C}_m$ (where H=W=64) represents the likelihood of there being a 2D human body center at a given pixel in the image, where each potential body center is characterised by a Gaussian distribution. Following [32], scale information such as body size is encoded in the kernel size of the Gaussian k. More specifically, let d_{bb} be the diagonal length of the person bounding box and W be the width of the Body Center heatmap, then k is computed as:

$$k = k_l + \frac{\sqrt{2}W}{d_{bb}}^2 k_r, \qquad (1)$$

where k_l is the minimum kernel size, k_r is the range of k.

Note, for in-the-wild images of multiple people, $\mathbf{C}_m$ not only contains the scale information of every potential target, but also contains strong location information that can be leveraged to reduce redundancy and focus features. This is further explored in Sec. 3.1.

Mesh Parameter map $\mathbf{P}_m \in \mathbb{R}^{T \times 145 \times H \times W}$: $\mathbf{P}_m$ (where H=W=64) contains the camera parameters $\mathbf{A}_m \in \mathbb{R}^{T \times 3 \times H \times W}$ and SMPL parameters $\mathbf{S}_m \in \mathbb{R}^{T \times 142 \times H \times W}$.

- In terms of camera parameters, $\mathbf{A}_m = (\xi, t_x, t_y)$ describes the 2D scale and translation information for every person in each frame, such that the 2D projection $\hat{\mathbf{J}}$ of the 3D body joints $\mathbf{J}$ can be obtained as $\hat{\mathbf{J}}_x = \xi \mathbf{J}_x + t_x, \hat{\mathbf{J}}_y = \xi \mathbf{J}_y + t_y$.

- The SMPL parameters, $\mathbf{S}_m$, describe the 3D pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ of the body mesh at each 2D position. For every potential person, $\boldsymbol{\theta} \in \mathbb{R}^{6 \times 22}$ describes the 3D rotations in the 6D representation [45] of each body joint apart from the hands, and $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are the shape parameters. Combining $\boldsymbol{\theta}$ with $\boldsymbol{\beta}$, SMPL establishes an efficient mapping to a human 3D Mesh $\boldsymbol{M} \in \mathbb{R}^{6890 \times 3}$.

3.1. BCA: Body Center Attention

The Body Center Attention mechanism is at the core of CoordFormer. It aims to fuse position information and acts as a learnable feature indexer by leveraging the representation pattern of the body center heatmap $\mathbf{C}_m$. Each pixel in $\mathbf{C}_m$ represents a potential person and learning relations at this pixel-level through Multi-Head Self-Attention (MHSA) would result in redundant calculations as most pixels do not contain people. Instead, we leverage the fact that $\mathbf{C}_m$ contains effective position information which can be used as a natural additional attention map for locating people in the corresponding frame. We thus use the **Body Center** heatmap as the **Attention** map, i.e. Body Center Attention, to focus and extract features of all persons.

Specifically, given an input video sequence $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^T$ with T frames, we first use the backbone to extract the feature map $\mathbf{F}_m \in \mathbb{R}^{T \times H \times W \times C}$. To enhance the perception of the coordinate system, we extend $\mathbf{F}_m$ with coordinate channels [25] resulting in $\mathbf{F}_{coord}$ and predict $\mathbf{C}_m$ from it. Finally, we compute the focused features as the Hadamard product between $\mathbf{C}_m$ and $\mathbf{F}_m$. Note, here we leverage $\mathbf{F}_m$ instead of $\mathbf{F}_{coord}$, to avoid altering the coordinate features of $\mathbf{F}_{coord}$.

Let $\mathbf{F}_m^t \in \mathbb{R}^{H \times W \times C}$, $\mathbf{F}_{coord}^t \in \mathbb{R}^{H \times W \times (C+2)}$ and $\mathbf{C}_m^t \in \mathbb{R}^{H \times W \times 1}$ be the feature map, coordinate feature map and Body Center heatmap of the tth frame, respectively. The focused feature map of the tth frame $\mathbf{F}_{focus}^t \in$
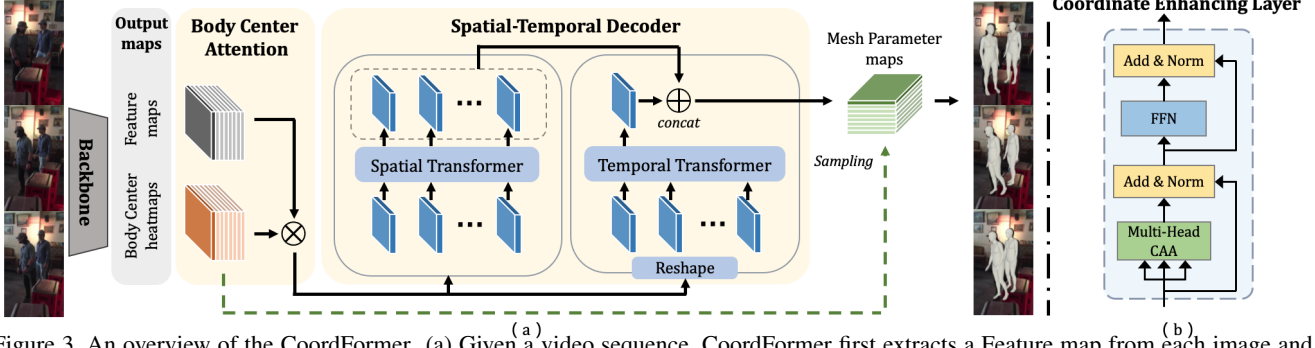
Figure 3. An overview of the CoordFormer. (a) Given a video sequence, CoordFormer first extracts a Feature map from each image and predicts the Body Center heatmap that reflects the probability of each position being a body center. Then CoordFormer leverages our proposed BCA mechanism and Spatial-Temporal Decoder to predict the pixel-level Mesh Parameter map that contains SMPL and camera parameters. Finally, the Body Center heatmap is parsed and the 3D mesh results are sampled. (b) The Coordinate Enhancing Layer that the Spatial Transformer and the Temporal Transformer of CoordFormer are comprised of. Each layer consist of multi-head CAA operations, a feed-forward network (FFN), Layernorm, and skip connections.

$\mathbb{R}^{H \times W \times C}$ can then be computed as follows,

$$\mathbf{F}^t_{coord} = ACC(\mathbf{F}^t_m), \tag{2}$$

$$\mathbf{C}^t_m = Conv_c(\mathbf{F}^t_{coord}), \tag{3}$$

$$\mathbf{F}^t_{focus,c} = \odot(\mathbf{C}^t_m, \mathbf{F}^t_{m,c}), \tag{4}$$

where $ACC(\cdot)$ indicates adding the coordinate channels, $\odot(\cdot, \cdot)$ indicates the Hadamard product, $Conv_c(\cdot)$ is the head convolution layers to obtain the Body Center heatmap, and $\mathbf{F}^t_{focus,c}$ and $\mathbf{F}^t_{m,c}$ indicate the cth channel of $\mathbf{F}^t_{focus}$ and $\mathbf{F}^t_m$, respectively.

As obtaining $\mathbf{C}_m$ is arguably the simplest learning task in the multi-head framework, it represents a reliable source to obtain the focused features $\mathbf{F}_{focus}$ and facilitates the effectiveness of BCA.

3.2. CEL : Coordinate Enhancing Layer

After establishing the existence and location of the people in the video, the motion sequence features must be used to determine their temporal relationships. Moreover, in multi-person scenarios, it is imperative to understand the spatial-temporal interactions to facilitate accurate mesh recovery. The spatial-temporal constraints between all known entities must therefore be modeled effectively.

Inspired by the progress on Spatial-Temporal Transformers (ST-Trans) with joint coordinates as input [26, 44, 46], we adopt a powerful ST-Tran as the base model for our Spatial-Temporal Decoder. However, directly applying a ST-Tran on $\mathbf{F}_{focus}$ does not produce the desired results. This is because the patch-level position information captured from the Position-Encoding [36] is not enough to regress the precise joint coordinates required for our single-stage design. Moreover, as illustrated in Fig. 2, vision transformers [11] that split features into patches and extract tokens from them, can lead to a degradation in the pixel-level information, especially for $\mathbf{C}_m$. Empirical evidence for this is provided in the supplementary material.

To add precise coordinate information across frames and maintain the pixel-level representation of $\mathbf{C}_m$ and $\mathbf{P}_m$, we introduce the CAA module that expands the self-attention operation of the Transformer Encoder [36]. Unlike Position-Encoding [36], which provides only rough location information at the patch-level, the CAA module captures the the coordinate relationships between (t, x, y) of each pixel. As depicted in Fig. 4, we extend $\mathbf{C}_m$ with both time and axis coordinates, enabling us to leverage $\mathbf{C}_m$ for detection while also leveraging the coordinate features to capture relations.

Specifically, we set Pixel coordinate $\mathbf{PC} \in \mathbb{R}^{W \times 1} = [1, 2, 3..., W]$ and Time coordinate $\mathbf{TC} \in \mathbb{R}^{T \times 1} = [1, 2, 3..., T]$ and repeat them to $\mathbf{PC}_r \in \mathbb{R}^{T \times W \times 1}$ and Time coordinate $\mathbf{TC}_r \in \mathbb{R}^{T \times H \times 1}$. The input feature $\mathbf{F}^1_{input} \in \mathbb{R}^{T \times H \times (W+2)}$ of the 1th CEL is then the concatenation of $\mathbf{F}_{focus}$, $\mathbf{PC}_r$, and $\mathbf{TC}_r$.

After adding coordinate information, three linear projections, f_Q, f_K, f_V, are applied to transfer $\mathbf{C}_m$ and $\mathbf{F}^l_{input}$ into three matrices of equal size, namely the query $\mathbf{Q}$, the key $\mathbf{K}$, and the value $\mathbf{V}$, respectively. The CAA operation is then calculated by:

$$\begin{cases} \mathbf{Q} = f_Q(\mathbf{F}_{input}), \\ \mathbf{K} = f_K(\mathbf{C}_m), \\ \mathbf{V} = f_V(\mathbf{F}_{input}), \end{cases} \tag{5}$$

$$CAA(\mathbf{C}_m, \mathbf{F}_{input}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}})V. \tag{6}$$

As shown in Fig. 3, in our proposed CEL, H heads of CAA are applied to $\mathbf{C}_m$ and $\mathbf{F}^l_{input}$. Therefore, the output of the lth CEL, $\mathbf{F}^l_{output} \in \mathbb{R}^{T \times H \times W}$, can be compute as

$$\mathbf{F}' = LN(\mathbf{F}^l_{input} + CAA(\mathbf{F}^l_{input})), \tag{7}$$
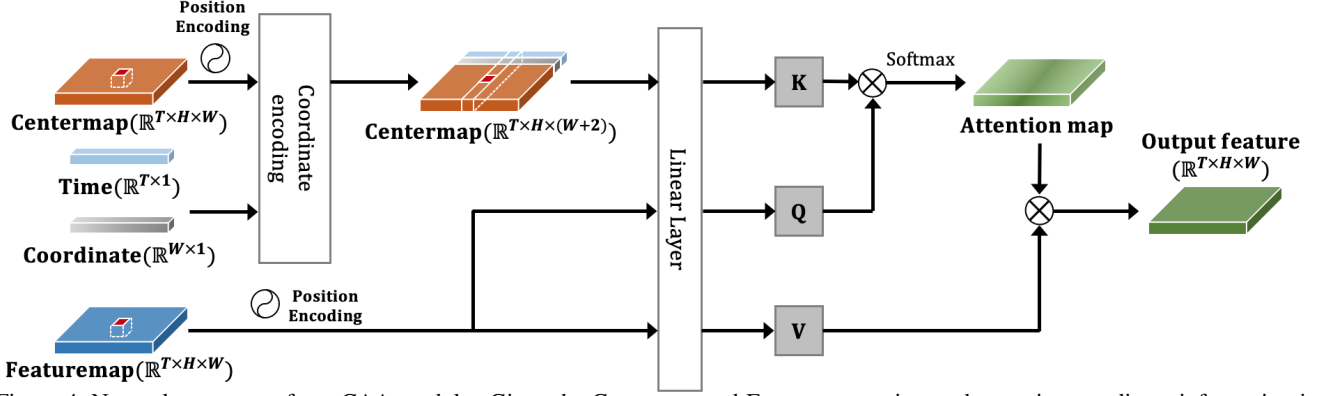
Figure 4. Network structure of our CAA module. Given the Centermap and Featuremap as input, the precise coordinate information is encoded in the Centermap by coordinate-encoding and the rough position information is encoded in the Featuremap by Position-Encoding. Then, K, Q and V are computed for scaled dot-product attention. With powerful position information as key, CAA can capture high-quality spatial-temporal correspondence among multiple persons.

$$\mathbf{F}^l_{output} = LN(\mathbf{F}' + FFN(\mathbf{F}')), \qquad (8)$$

where LN indicates the Layer Normalization [4] and FFN indicates a feed-forward network. The output of layer l, $\mathbf{F}^l_{output}$, is then provided as input to the next layer, i.e. becomes $\mathbf{F}^{(l+1)}_{input}$. Through multiple CELs, our Spatial-Temporal Transformers receives sufficient global location information for implicit feature matching.

3.3. Spatial-Temporal Decoder

Building on the coordinate-awareness induced by CAA, we leverage the Spatial-Temporal Transformers to learn the spatial and temporal constraints, respectively. As shown in Fig. 3, the Spatial-Temporal Decoder forms a residual structure and first establishes spatial feature relationships, before modelling the temporal connections.

Spatial Transformer Module. Since the representation patterns of the Body Center Heatmap bring spatial information to the feature due to its similarity around the body center, we first use the Spatial Transformer to extract the corresponding spatial information. Given the input $\mathbf{F}_{input}$, the Spatial Transformer performs a CAA operation on each frame, where Q, K, V are $\in \mathbb{R}^{k_t \times E_{k_t}}$, and where k and E_{k_t} indicate the number of tokens and the length of the token embedding, respectively.

Temporal Transformer Module. After building the spatial relationships, the Temporal Transformer is used to ensure consistency in the temporal relationships. Given the input $\mathbf{F}_{input}$, the Temporal Transformer performs a CAA operation on all frames, where Q, K, V are $\in \mathbb{R}^{(T \cdot k_t) \times E_{k_t}}$.

Coordinate Information Fusion. Since the Coordinate encoding only adds spatial coordinates for one dimension at a time in the 2D image, we observe improvements by transposing $\mathbf{C}_m$ at alternating layers, thereby infusing coordinate information along both spatial dimensions.

More specifically, each Transformer has $2L$ CELs. At every L^{th}_{2N} layer where $N = [1, 2, 3..., L]$, $\mathbf{C}_m$ will be transposed to $\mathbf{C}_{mt} \in \mathbb{R}^{T \times W \times H}$ to add precise coordinate infor-

mation, resulting in

$$\begin{cases} \mathbf{F}^{l+1}_{output} = CEL(\mathbf{C}_{mt}, \mathbf{F}^l_{input}), & l = 2N \\ \mathbf{F}^{l+1}_{output} = CEL(\mathbf{C}_m, \mathbf{F}^l_{input}), & otherwise. \end{cases} \qquad (9)$$

Through multiple layers of CEL, the Transformer learns the correspondence along both dimensions.

3.4. Loss Functions

The loss function of CoordFormer consists of a set of temporal and spatial loss functions that ensure temporal consistency and spatial accuracy, respectively.

Temporal loss L_{tem}. We add L_{tem} to maintain the similarity of adjacent frames via

$$L_{tem} = w_{accel}L_{accel} + w_{aj3d}L_{aj3d} + w_{sm}L_{sm}, \qquad (10)$$

where L_{accel} and L_{aj3d} are the Accel error [20] and the L_2 loss of the 3D joints offsets, respectively, and L_{sm} is a regular L_1 loss between consecutive frames, preventing mutation of $\mathbf{C}_m$ and $\mathbf{F}_m$. For each loss item, $w_{(\cdot)}$ indicates the corresponding weight.

Spatial losses L_{spa}. For spatial accuracy, we follow the previous methods [17, 32] to add loss functions on SMPL parameters, 3D body joints, 2D body joints and Center Body heatmap. Specifically, L_{cm} is the focal loss [32] of the Center Body heatmap. L_θ and L_β are L_2 loss of SMPL pose $\overrightarrow{\theta}$ and shape $\overrightarrow{\beta}$ parameters respectively. L_{prior} is the Mixture Gaussian prior loss [5, 27] of the SMPL parameters for supervision of prior knowledge. To supervise the accuracy of the joint prediction, L_{j3d} and L_{pj2d} are added. L_{j3d} consist of L_{mpj} and L_{pmpj}, where L_{mpj} is the L_2 loss of predicted 3D joints $\overrightarrow{J}$ and L_{pmpj} is the L_2 loss of the predicted 3D joints after Procrustes alignment with the ground truth [32, 34]. L_{pj2d} is the L_2 loss of the 2D projection of the 3D joints $\overrightarrow{J}$. For each loss item, $w_{(\cdot)}$ indicates the

corresponding weight and L_{spa} can be computed as,

$$L_{spa} = w_{cm}L_{cm} + w_{pose}L_{pose} + w_{shape}L_{shape} \\ + w_{prior}L_{prior} + w_{j3d}L_{j3d} + w_{pj2d}L_{pj2d}. \quad (11)$$

4. Experiments

4.1. Implementation Details

Network Architecture. To facilitate a fair comparison, we follow prior approaches [32] and leverage the HRNet-32 [7] as the backbone, similar to [32, 40].

Datasets. To ensure a fair comparison with previous methods, the training is conducted on well-known datasets. The image dataset that is used to train the spatial branch consists of two 3D pose datasets (MPI-INF-3DHP[29] and MuCo-3DHP [29]) and two in-the-wild 2D pose datasets (MPII [2] and LSP [14, 15]), while the video dataset consists of the 3DPW [37] and Human3.6M [12] datasets.

Evaluation. Evaluation is performed on the 3DPW [37] dataset as the Human3.6M [12] and MPI-INF-3DHP[29] datasets only contain one person per frame and can thus not be used to assess the performance in multi-person scenarios. Therefore, 3DPW [37] is employed as the main benchmark for evaluating the 3D mesh/joint error. Moreover, we follow [32] and divide the 3DPW dataset into three subsets, namely 3DPW-PC, 3DPW-OC and 3DPW-NC. These subsets represent subsets containing person-person occlusion, object occlusion and non-occluded/truncated cases, respectively, and are used to evaluate the performance under different occlusion scenarios.

Following prior approaches [20, 38], the quantitative performance is evaluated by computing the mean per joint position error (MPJPE), the Procrustes-aligned mean per joint position error (PAMPJPE), and the mean Per Vertex Error (PVE) for each frame.

Baselines. We compare CoordFormer to both single-image-based and video-based baseline methods. For single image-based methods, we include HMR [17], SPIN [22], CRMH [13], EFT [16], BMP [42], ROMP [32] and BEV [33]. For video-based methods, we include HMMR [18], Doersch *et al.* [10], DSD-SATN [34], VIBE [20], TCMR [8], MEVA [44], MPS-Net [38] and MotionBERT [46]. Note that MotionBERT [46] requires additional 2D skeletons motion information as input.

4.2. Comparisons to the State-of-the-Art

In-the-wild multi-person scenarios. To reveal the effectiveness of CoordFormer, we evaluate CoordFormer under the different in-the-wild scenarios of 3DPW. For a fair and comprehensive comparison, we follow [32] to adopt three evaluation protocols and then compare CoordFormer with state-of-the-art methods. As ROMP was originally trained on a considerably larger dataset, which included OH [43], the pseduo 3D labels from [16], and Pose-Track [1], we retrain ROMP on our dataset to ensure fair

comparisons. While we attempted the same with BEV, we observed that BEV did not converge due to the missing relative depth and age supervision that is used to learn BEV's centermap. For completeness, we still report the original results reported in [32] for both ROMP and BEV as reference.

To comprehensively verify the in-the-wild performance, we follow *Protocol 1* to evaluate models on the entire 3DPW dataset. Without any ground truth as input, single-person methods [20, 22] are equipped with a 2D human detector [6, 31]. As shown in Tab. 1, CoordFormer significantly outperforms all the baselines in MPJPE and PAM-PJPE, which reveals that CoordFormer can successfully learn the pixel-level feature representation and better model spatial-temporal relations through ST-Trans.

Moreover, to evaluate the ability in modeling temporal motion constrains, we follow *Protocol 2* on the 3DPW test set without fine-tuning on the 3DPW training set. In Tab. 1, CoordFormer takes the whole image as input and the temporal branch is only trained on the Human3.6 M [12] dataset, while multi-stage baseline methods can use the cropped single-person image as input and train on more video datasets, i.e. Human3.6 M [12], MPI-INF-3DHP [29], AMASS [28]. CoordFormer still outperforms all baselines. Finally, we follow *Protocol 3* to evaluate the models on the 3DPW test set with 3DPW fine-tuning. As shown in Tab. 2, CoordFormer outperforms all the methods in MPJPE and PAMPJPE, while being only slightly worse than MotionBERT in PVE. Note that MotionBERT requires additional 2D skeleton motion as input, while CoordFormer can directly be applied on in-the-wild images.

A qualitative comparisons to state-of-the-art methods is provided in Fig. 5, demonstrating the effectiveness of CoordFormer to precisely recover the mesh. Additional qualitative results are included in the supplementary material.

3DPW upper-bound performance. To show the upper-bound performance of the video-based methods on the in-the-wild multi-person video dataset, i.e. 3DPW, we compare CoordFormer with previous state-of-the-art video-based methods regardless of their training dataset and training setting. As shown in Tab. 3, CoordFormer achieves the best results, which demonstrates the effectiveness of Coord-Former for multi-person mesh recovery from videos.

Occlusion scenarios. As shown in Tab. 4, Coord-Former achieves superior performance on the 3DPW-NC and 3DPW-OC subset under non-occlusion and object occlusion cases according to PAMPJPE. Further comparisons show that CoordFormer outperforms ROMP [32] in MPJPE on all 3DPW subsets, demonstrating that precise coordinate information improves the performance under occlusion.

Runtime comparisons. In Tab. 5, all comparisons are performed on a desktop with a GTX 3090Ti GPU and a Intel(R) Xeon(R) Platinum 8163 CPU. All video-based models are tested on 8-frames video clips. CoordFormer is

Table 1. Comparisons to the state-of-the-art methods on 3DPW following *Protocol 1 and 2* (evaluate on the entire 3DPW dataset and on the test set only). * means that additional datasets are used for training [32] .

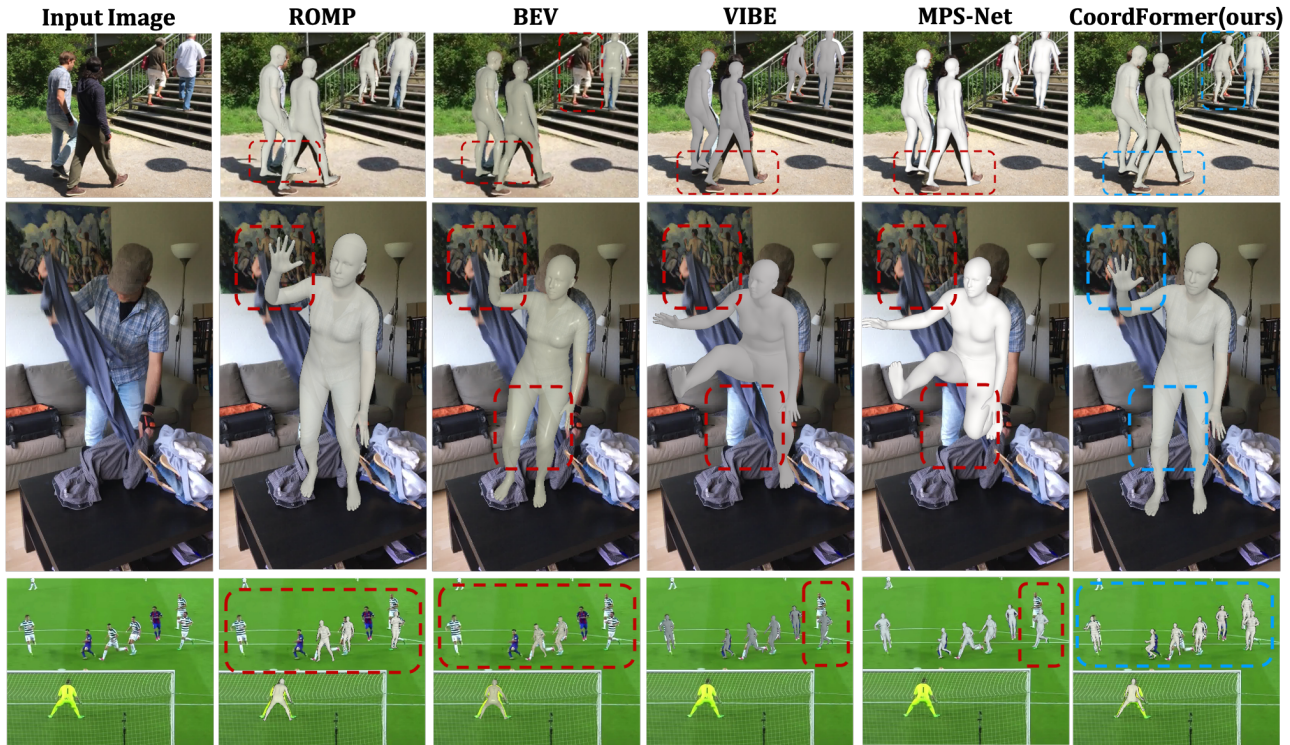| *Protocol 1*			*Protocol 2*			
Methods	MPJPE ↓	PAMPJPE ↓	Methods	MPJPE ↓	PAMPJPE ↓	PVE ↓
ROMP(ResNet-50)* [32]	87.0	62.0	ROMP(ResNet50)* [32]	91.3	54.9	108.3
Openpose + SPIN [22]	95.8	66.4	HMR [17]	130.0	76.7	-
			HMMR [18]	116.5	72.6	139.3
CRMH [13]	105.9	71.8	Arnab *et al.* [3]	-	72.2	-
YOLO + VIBE[20]*	94.7	66.1	GCMR [23]	-	70.2	-
			DSD-SATN [34]	-	69.5	-
BMP [42]*	104.1	63.8	SPIN [22]	96.5	59.2	116.4
ROMP [32]	90.87	61.34	ROMP [32]	96.96	57.48	**110.13**
CoordFormer(Ours)	**88.95**	**59.86**	CoordFormer(Ours)	**95.27**	**54.58**	110.35

Figure 5. Qualitative results of ROMP [32], BEV [33], VIBE [20], MPS-Net [38] and CordFormer on 3DPW and the internet videos.

slightly slower than image-based methods [32, 33] due to the overhead in spatial-temporal modeling, however, Co-ordFormer is significantly faster than the video-based methods [20, 38].

4.3. Ablation Study

To validate the effectiveness of the BCA and CAA modules in CoordFormer, we train CoordFormer under different settings and conduct ablation studies following *Protocol 3* to evaluate on 3DPW. Specifically, we evaluate the BCA module by replacing $\mathbf{F}_{focus}$ with $\mathbf{F}_{coord}$ without extra attention mechanism and evaluate the CAA module by skipping the Coordinate encoding in Fig. 4. As shown in

Tab. 6, CoordFormer with BCA and CCA achieves the best result in the in-the-wild scenarios, which fully demonstrates the effectiveness of BCA and CCA. Specifically, the results confirm that BCA can effectively enhance the perception of potential people in the multi-person scenario. Second, the ablation experiments strongly reflect the importance of precise coordinate information in videos. In summary, the results from Tab. 6 reveal the importance of capturing position information in the multi-person scenario and the effectiveness of the BCA and CCA modules. We further perform an additional ablation study on the Spatial-Temporal Transformer of CoordFormer. Results in Tab. 7 illustrate the benefit of exploiting temporal and spatial information jointly.

Table 2. Comparisons to the state-of-the-art methods on 3DPW following *Protocol 3* (fine-tuned on the training set). * means that additional datasets are used for training [32].

Methods	MPJPE↓	PAMPJPE↓	PVE↓
ROMP(ResNet-50)* [32]	84.2	51.9	100.4
ROMP(HRNet-32)* [32]	78.8	48.3	94.3
BEV* [33]	78.5	46.9	92.3
EFT [16]	-	51.6	-
VIBE [20]	82.9	51.9	99.1
MPS-Net [38]	84.3	52.1	99.7
MotionBERT [46]	80.9	49.1	**94.2**
ROMP [32]	81.06	49.07	96.74
CoordFormer(Ours)	**79.41**	**46.58**	94.44

Table 3. Comparisons of best result to the state-of-the-art video-based methods for in-the-wild scenarios on 3DPW.

Methods	MPJPE↓	PAMPJPE↓	PVE↓
HMMR [18]	116.5	72.6	139.3
Doersch *et al.* [10]	-	74.7	-
Arnab *et al.* [3]	-	72.2	-
DSD-SATN [34]	-	69.5	-
VIBE [20]	82.9	51.9	99.1
MEVA [44]	86.9	54.7	-
TCMR [8]	86.5	52.7	103.2
GLAMR [40] + SPEC [21]	-	54.9	-
GLAMR [40] + KAMA [19]	-	51.1	-
MPS-Net [38]	84.3	52.1	99.7
CoordFormer(Ours)	**79.41**	**46.58**	**94.44**

Table 4. Comparisons to state-of-the-art methods on the person-occluded (3DPW-PC), object-occluded (3DPW-OC) and non-occluded/truncated (3DPW-NC) subsets of 3DPW. * means that additional datasets are used for training.

Metric	Method	3DPW-PC↓	3DPW-NC↓	3DPW-OC↓
	ROMP*	75.8	57.1	67.1
	CRMH [13]	103.55	65.7	78.9
PAMPJPE	VIBE [20]	103.9	57.3	65.9
	ROMP	**77.64**	56.67	66.6
	CoordFormer	79.30	**54.13**	**64.47**
MPJPE	ROMP	103.70	95.53	100.79
	CoordFormer	**101.51**	**93.17**	**97.25**

Table 5. Run-time comparison on a 3090 GPU.

Methods	Time per frame(s)↓	FPS↑	Backbbone	Using Temporal information
ROMP [32]	**0.01329**	**75.26**	HRNet-32	✗
BEV [33]	0.01448	69.04	HRNet-32	✗
VIBE [20]	0.07881	12.68	HRNet-32	✓
MPS-Net [38]	0.08013	12.47	HRNet-32	✓
CoordFormer	**0.01867**	**53.55**	HRNet-32	✓

Table 6. Ablation study under 3DPW *Protocol 3*.

Methods	MPJPE↓	PAMPJPE↓	PVE↓
CoordFormer w/o CAA	83.19	50.62	99.21
CoordFormer w/o BCA	82.20	48.84	98.23
CoordFormer	**79.41**	**46.58**	**94.44**

Table 7. Ablation study of spatial and temporal Transformer on 3DPW. S means only training the spatial branch, ST means fine-tuning the temporal branch on Human3.6 M, ST-fine means fine-tuning on the 3DPW training set.

Evaluation	Methods	MPJPE↓	PAMPJPE↓	PVE↓
On entire 3DPW	S	95.05	63.22	115.90
	ST	88.95	59.86	103.88
On test set only	S	103.95	58.03	120.67
	ST	95.27	54.58	110.35
	ST-fine	**79.41**	**46.58**	**94.44**

The reason for the decline in performance when only leveraging the spatial branch can be attributed to two factors: the inability to utilize temporal information and the fact that CAA lacks temporal coordinate information.

5. Conclusion

We proposed CoordFormer to achieve single-stage multi-person mesh recovery from videos. CoordFormer incorporates implicit multi-person detection, tracking, and spatial-temporal modeling. Two critical novelties are the Coordinate-Aware Attention mechanism for pixel-level feature learning and the Body Center Attention for person-focused feature selection. CoordFormer paves the way for various downstream applications related to perceiving group behavior, including but not limited to virtual reality and physical therapy.

Despite CoordFormer's robust performance to recover multi-person meshes, its current version lacks the ability to recover completely occluded meshes. We plan to explore this exciting area by leveraging the continuity along the temporal dimension of the body center heatmap.

---

[1]https://www.mindspore.cn/

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, pages 5167–5176. IEEE, 2018. 6

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693. IEEE, 2014. 6

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, pages 3395–3404. IEEE, 2019. 2, 7, 8

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv*, 2016. 5

[5] 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*, pages 561–578. Springer, 2016. 1, 5

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299. IEEE, 2017. 6

[7] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, pages 5386–5395. IEEE, 2020. 6

[8] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, pages 1964–1973. IEEE, 2021. 1, 2, 6, 8

[9] Vasileios Choutas, Lea Müller, Chun-Hao P Huang, Siyu Tang, Dimitrios Tzionas, and Michael J Black. Accurate 3d body shape regression using metric and semantic attributes. In *CVPR*, pages 2718–2728. IEEE, 2022. 2

[10] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *NeurIPS*, 2019. 6, 8

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 4

[12] 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*, pages 1325–1339, 2013. 6

[13] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588. IEEE, 2020. 2, 6, 7, 8

[14] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, page 5. Citeseer, 2010. 6

[15] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, pages 1465–1472. IEEE, 2011. 6

[16] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *ECCV*, pages 68–84. IEEE, 2020. 2, 6, 8

[17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131. IEEE, 2018. 1, 5, 6, 7

[18] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, pages 5614–5623. IEEE, 2019. 1, 2, 6, 7, 8

[19] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *3DV*, pages 918–927. IEEE, 2020. 8

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263. IEEE, 2020. 1, 2, 3, 5, 6, 7, 8

[21] 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*, pages 11035–11045. IEEE, 2021. 8

[22] 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*, pages 2252–2261. IEEE, 2019. 1, 2, 6, 7

[23] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510. IEEE, 2019. 7

[24] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, pages 13147–13156. IEEE, 2022. 2

[25] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *NeurIPS*, 2018. 3

[26] Shuying Liu, Wenbin Wu, Jiaxian Wu, and Yue Lin. Spatial-temporal parallel transformer for arm-hand dynamic estimation. In *CVPR*, pages 20523–20532. IEEE, 2022. 2, 4

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, pages 1–16, 2015. 2, 5

[28] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451. IEEE, 2019. 6

[29] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516. IEEE, 2017. 6

[30] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *CVPR*, pages 803–812. IEEE, 2019. 2

[31] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 6

[32] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188. IEEE, 2021. 2, 3, 5, 6, 7, 8

[33] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, pages 13243–13252. IEEE, 2022. 2, 6, 7, 8

[34] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, pages 5349–5358. IEEE, 2019. 2, 5, 6, 7, 8

[35] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117. IEEE, 2017. 1

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[37] 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*, pages 601–617. Springer, 2018. 6

[38] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *CVPR*, pages 13211–13220. IEEE, 2022. 1, 2, 3, 6, 7, 8

[39] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *ICCV*, pages 7760–7770. IEEE, 2019. 1

[40] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, pages 11038–11049. IEEE, 2022. 1, 2, 6, 8

[41] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for 10x efficient 2d and 3d pose estimation. *arXiv*, 2022. 1

[42] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, pages 546–556. IEEE, 2021. 2, 3, 6, 7

[43] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, pages 7376–7385. IEEE, 2020. 6

[44] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, pages 11656–11665. IEEE, 2021. 1, 2, 4, 6, 8

[45] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753. IEEE, 2019. 3

[46] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: Unified pretraining for human motion analysis. *arXiv*, 2022. 4, 6, 8