Progressive Hypothesis Transformer for 3D Human Mesh Recovery

Huang-Ru Liao[†‡], Jen-Chun Lin[†], and Chun-Yi Lee[‡]
[†]Academia Sinica, Taipei, Taiwan
[‡]Elsa Lab, Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
{mimiliao2000, jenchunlin}@gmail.com, cylee@cs.nthu.edu.tw

Abstract

Recent advancements in Transformer-based human mesh reconstruction (HMR) are commendable. However, these models often lift 2D images directly to 3D vertices without explicit intermediate guidance. In addition, the global attention mechanism tends to spread attention across larger body areas and even unrelated background regions during human mesh estimation, rather than focusing on critical local regions such as human body joints. This tendency leads to inaccurate and unrealistic results for complex activities. To address these challenges, we introduce the Progressive Hypothesis Transformer, which employs 2D and 3D pose predictions to progressively guide our model. Moreover, we propose a mechanism that generates multiple plausible hypotheses for both 2D and 3D poses to mitigate potential inaccuracies arising from intermediate pose estimations. Our model also incorporates inter-intra attention to capture correlations between joints and hypotheses. Experimental results demonstrate that our method surpasses existing image-based approaches on Human3.6M [13] and 3DPW [36] with fewer parameters and relatively lower computational costs.

1. Introduction

Estimating a 3D human mesh from a 2D RGB image is a long-standing and pivotal challenge, primarily due to its widespread applications in virtual reality, augmented reality, motion capture, and interactive gaming. In recent years, Transformer-based architectures [6, 21, 22] have gained significant attention among researchers due to their prowess in modeling long-range dependencies. This shift is partly attributed to the limitations of CNN-based methods: while they have made considerable advancements, their designs primarily focus on local feature extraction, which constrains their ability to capture non-local correlations. Existing 3D human mesh estimation methodologies can be broadly categorized into parametric [3, 14–16, 19, 30] and non-parametric [6, 7, 17, 28] approaches. The former primarily gears towards estimating parameters from established human body models, such as the Skinned Multi-Person Linear (SMPL) model [25]. On the other hand, the latter aims to directly regress the 3D coordinates of human mesh vertices without relying on any predefined model. Even with the promise of Transformers, however, there are inherent challenges in their application to HMR, necessitating further exploration and refinement.

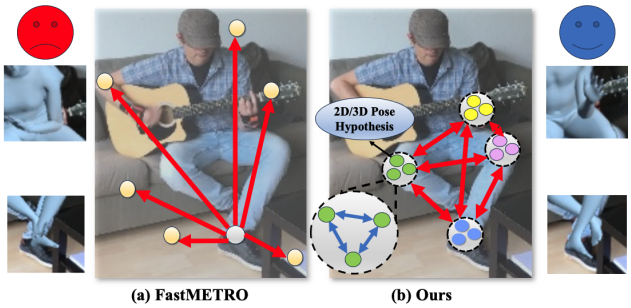

Figure 1. (a) Prior Transformer-based methods [6] that directly lift 2D images to 3D vertices by global attention might distribute attention across body areas and unrelated backgrounds, leading to incorrect poses. (b) Our method progressively generates 2D/3D pose hypotheses and captures the inter-intra relations of body joints.

Fig. 1 sheds light on the challenges encountered by previous Transformer-based human mesh reconstruction (HMR) endeavors. Two primary issues stand out. First, even with the glocal attention inherent in Transformer architectures, there is a tendency to prioritize holistic relationships. While they inherently possess a broad field of view, in the absence of proper guidance, they might overlook crucial localized features. This can cause them to spread attention across larger body areas and even unrelated background regions during human mesh estimation, rather than focusing on local regions such as specific body joints. The second issue pertains to the intrinsic complexities of the task: achieving a direct transformation from an image to a 3D human mesh is far from straightforward. For instance, addressing depth ambiguity and complex human pose rotations is challenging due to the significant representational differences between 2D images and 3D meshes. Because of this, some existing methods may experience difficulties when estimating com-

plex human poses or activities. Moreover, without utilizing geometric insights from models like SMPL, these methods could even generate unrealistic results. While attempts have been made to include additional SMPL regressor heads to enhance pose predictions, such modifications can occasionally compromise performance [6, 17]. In light of these findings, the primary objective of this work is to progressively guide the Transformer-based human pose and shape estimation and model the correlations of body parts to ensure a more gradual and concentrated focus on essential local specifics.

To address the aforementioned challenges, this paper introduces a new framework distinguished by three primary innovations. First, we present a *progressive pose-guided learning* approach. Rather than regressing directly to the final 3D human mesh, this technique utilizes 2D and 3D poses as intermediate representations, and predicts them in a stepwise manner before deriving the final SMPL parameters. During this process, the intermediate representations are used as guidance to extract joint-related features, and the predictions evolve and progressively mature from their initial 2D form to a refined 3D understanding. This strategy offers two salient benefits: (1) it enables progressive refinement of features supervised by 2D/3D pose ground truths, and (2) it emphasizes the extraction of localized, joint-centric features that are largely invariant to background distractions. Our second innovation further facilitates the learning of these intermediate representations, which revolves around the *2D/3D pose hypothesis generation*. Recognizing potential inaccuracies from intermediate pose estimations, our framework generates multiple plausible hypotheses for both 2D and 3D poses, instead of a single pose. By predicting and then sampling from a distribution of potential poses, our framework harnesses information from multiple possible poses, thereby enhancing the robustness of its final prediction. The final innovation is our *Inter-Intra Joint-Hypothesis Transformer*. By leveraging the joint features generated by pose-guided sampling, this Transformer integrates both inter and intra joint attentions to capture body part correlations and aggregate our pose hypotheses. Experimental evidence reveals that our proposed methodology is able to outperform existing image-based approaches on the Human3.6M [13] and 3DPW [36] datasets with much fewer parameters and computational costs. This not only substantiates the effectiveness of our methodology but also demonstrates how the synergy of these innovations significantly elevates the overall performance of the framework. Our primary contributions are summarized as follows:

- We introduce a progressive pose-guided learning approach that leverages 2D/3D poses as intermediate representations. This approach ensures that pose learning evolves progressively, maturing through continuous intermediate supervision with ground truths.
- We propose a mechanism that generates multiple plausible hypotheses for both 2D and 3D poses. This is achieved by estimating pose distributions and subsequently sampling potential poses from the distributions.
- We present the Inter-Intra Joint-Hypothesis Transformer, which integrates both inter and intra joint attentions. This is designed to capture correlations between body parts and aggregate pose hypotheses, which enables the framework to enhance its overall performance notably.

2. Related Work

2.1. Single Image 3D Human Mesh Reconstruction

The field of HMR from single images can be broadly categorized into two main approaches: parametric and non-parametric. Parametric methods focus on estimating shape and pose parameters of models like SMPL [25]. SMPLify [3] fitted the SMPL model to 2D body joints with constraints from body priors. The authors in [14] introduced an end-to-end HMR framework using a CNN and iterative regression, incorporating re-projection and adversarial losses. SPIN [16] augmented [14] by combining end-to-end regression models with optimization loops. The authors in [38] used pyramidal mesh alignment feedback to refine spatial learning. PARE [15] utilized a segmentation map to learn part attention masks enabling feature aggregation. Non-parametric methods directly regress images to non-parametric body shapes like voxels [34] or 3D vertices [17, 22, 28]. GraphCMR [17] captured the adjacent relations of vertices using Graph-CNN. I2L-MeshNet [28] used lixel-based heatmaps for vertex estimation. Transformer-based methods [6, 21, 22] captured non-local image-vertex relationships for vertex localization. Unfortunately, Transformer-based non-parametric approaches often led to unrealistic results and could become computationally intensive due to the large quantity of vertices.

2.2. Transformer-based 3D Mesh Reconstruction

Originally designed for natural language processing, Transformers have excelled in computer vision tasks [2, 4, 5, 8, 20, 24, 35, 39]. They leveraged self-attention to grasp long-range connections, which proved valuable for HMR tasks. METRO [21] pioneered Transformer usage for modeling vertex interactions and 3D mesh generation. The work in [22] introduced a graph-convolution-reinforced Transformer, blending global and local vertices relationships. FastMETRO [6] resolved computational inefficiency by designing an encoder-based architecture. Nevertheless, these methods often lacked detailed local focus since they directly map 2D images to 3D vertices with broad global attention on image features.

2.3. Intermediate Representations

The task of HMR from a single image is challenging. Several methods introduce intermediate estimations into their networks to alleviate the difficulty. The authors in [33] employed the detected 2D keypoint coordinates for body-skeleton disentanglement via a bilinear transformation technique. HoloPose [10] trained a multi-task network comprising 2D, 3D, and dense pose estimation with a part-based regression. DaNet [37] leveraged UVI maps to bridge the 2D-3D mapping. Pose2Mesh [7] estimated 3D mesh vertices from 2D poses via GraphCNN. HybrIK [19] converted 3D joints to body-part rotations for HMR via the twist-and-swing decomposition. While these methods employ such representations to bridge the gap between 2D images and 3D meshes, inaccuracies in the representations may accumulate.

3. Methodology

In this section, we introduce our methodology. We first provide the problem definition, followed by an overview of the proposed framework. Then, we detail the individual components within the framework and the training objectives.

3.1. Problem Definition

To set the stage for subsequent discussions, we first introduce the notations specific to our framework and delve into the SMPL model. SMPL is a parametric model capable of producing a 3D mesh. Specifically, SMPL defines a mapping, $\mathcal{M}(\theta, \beta)$ that takes an input pose θ and shape β to yield the human body mesh M. In this representation, M belongs to $\mathbb{R}^{N\times 3}$, $\mathbb{R}$ represents the real number space, and $N = 6,890$ indicates the count of vertices in the standard SMPL model. For each mesh output M, the body joints, expressed as J_{3D}, are derived as a linear combination of its vertices. This can be encapsulated by the equation: $J_{3D} = \mathcal{W}_J M$, where $\mathcal{W}_J$ is a pre-trained linear regressor. When presented with a 2D image I of dimensions $H \times W \times 3$, our framework, denoted as $F(I)$, aims to predict a tuple of SMPL parameters (θ, β).

3.2. Overview of the Framework

Fig. 2 provides an overview of our proposed framework. Designed to generate a 3D mesh from a given 2D RGB input image I, our framework leverages the technique of pose-guided sampling and operates in a progressive manner. Instead of directly transforming a 2D image I into a 3D mesh M, our method unfolds over multiple stages. This multi-stage process involves generating a 2D pose, progressing to its 3D counterpart, and eventually leading to the formation of the 3D mesh. We refer to these progressive outputs as *intermediate representations*. At each stage, the framework evaluates potential distributions for both 2D and 3D poses. These distributions are utilized to generate plausible hypotheses, which are then employed for performing pose-guided feature sampling. The sampled features guided by 2D/3D pose hypotheses are fed into our Inter-Intra Joint Hypothesis Transformers for deriving the interrelations of body parts.

Diving deeper into the workflow of the framework, it first processes an image I through both a CNN backbone and a Transformer encoder. This results in extracted features, represented as $f_{enc} \in \mathbb{R}^{H'W'\times C}$, where $H' \times W'$ represents the dimensions of f_{enc} and C stands for the number of channels. The extracted f_{enc} is further refined by a convolutional layer to produce an embedding $f_{2D} \in \mathbb{R}^{H'W'\times C}$. This f_{2D} is then directed into a ***Hypothesis Generation Module (HGM)*** to generate 2D pose hypotheses $H_{2D\_heat} \in \mathbb{R}^{K\times N_J\times H'W'}$, where K denotes the number of hypotheses and N_J indicates the number of joints. Note that H_{2D_heat} is a heatmap representation of 2D pose. By leveraging the 2D pose distribution insights revealed in H_{2D_heat}, the framework conducts feature sampling on f_{enc}, and results in a sampled feature embedding denoted as $f_{2D\_sampled} \in \mathbb{R}^{K\times N_J\times C}$. Benefiting from the insights encapsulated in H_{2D_heat}, the features in $f_{2D\_sampled}$ exhibit enhanced localization and contain pertinent joint information. Subsequently, an ***Inter-Intra Joint-Hypothesis Transformer*** ($\mathcal{T}_{inter-intra}$) captures the inter- and intra-correlations of $f_{2D\_sampled}$, and aggregates these features to generate the embedding $f_{3D} \in \mathbb{R}^{N_J\times C}$.

In the second half of Fig. 2, the framework F transitions to the 3D pose generation stage. The generated 3D pose acts as guidance to further enhance the learning process. This stage starts by feeding f_{3D} into HGM, which in turn produces 3D pose hypotheses $H_{3D} \in \mathbb{R}^{K\times N_J\times 3}$. It is important to note that for the subsequent 2D feature sampling, H_{3D} is projected back into 2D coordinates. Gaussian blobs are then applied to generate 2D heatmaps, denoted as $H'_{2D\_heat} \in \mathbb{R}^{K\times N_J\times H'W'}$. By leveraging the distribution information of body joints revealed from H'_{2D_heat}, the framework is able to effectively sample features from f_{enc}. This yields the feature embedding $f_{3D\_sampled} \in \mathbb{R}^{K\times N_J\times C}$ enriched with 3D pose insights. This $f_{3D\_sampled}$ then similarly traverses through an Inter-Intra Joint-Hypothesis Transformer, to derive our final feature embedding $f_{HMR} \in \mathbb{R}^{N_J\times C}$. To enrich f_{HMR} with 3D information, it undergoes an element-wise addition with the preceding 3D feature f_{3D}. The final stage involves flattening the embedding, after which the framework utilizes a series of linear layers to finalize the SMPL pose and shape parameters. These parameters are then processed through an SMPL [3] model to produce the final mesh M. Furthermore, a pre-trained linear regressor is employed to generate 3D body joints, denoted as J_{3D}.

3.3. 2D/3D Pose Hypothesis Generation Module

Fig. 3 illustrates the flow of HGM, which is primarily designed to generate a diverse set of pose hypotheses as our intermediate representations. Instead of directly transforming

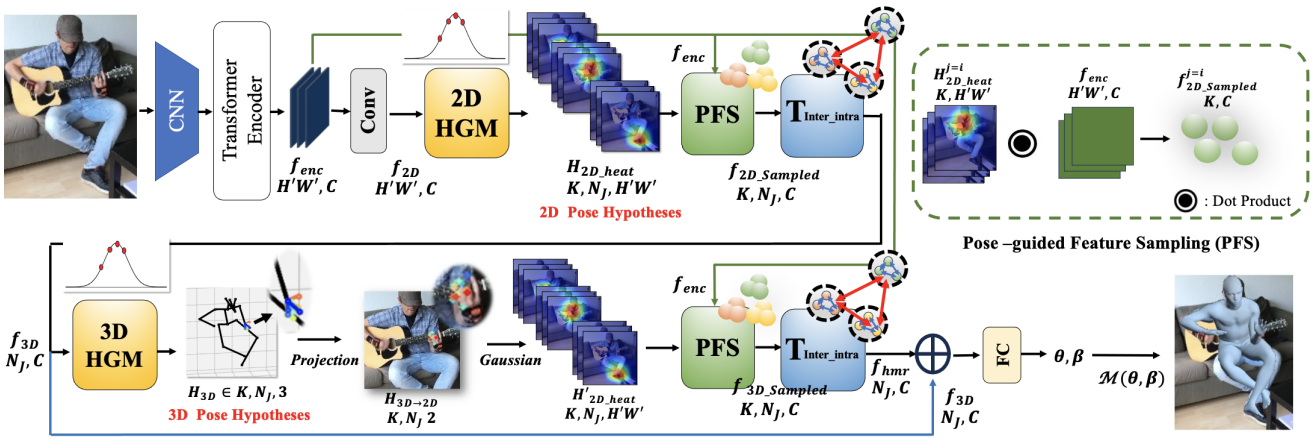

Figure 2. Overview of the proposed framework. Our framework estimates multiple 2D/3D pose hypotheses and leverages them for feature sampling in a progressive manner. Subsequently, our Inter-Intra Joint-Hypothesis Transformer ($\mathcal{T}_{inter-intra}$) captures the correlations among body joints. The final enhanced feature is then processed through several FC layers to generate the 3D human body pose and shape.

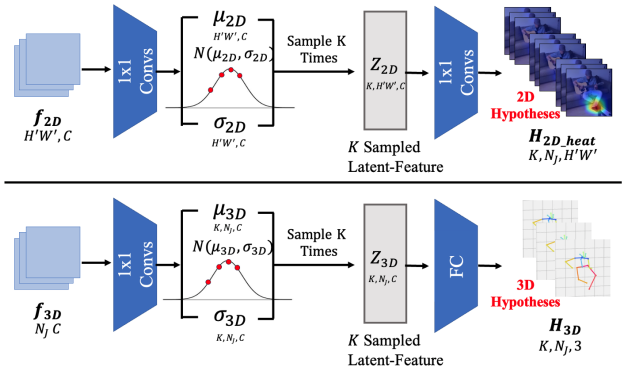

Figure 3. The proposed Hypothesis Generation Module (HGM).

2D images into 3D meshes, our methodology progressively estimates the 3D human mesh by leveraging these 2D and 3D pose representations. Nevertheless, such intermediate representations can be prone to inaccuracies, which might potentially lead to error propagation that could adversely influence subsequent results. Recognizing this challenge, our framework deviates from the traditional method of relying solely on a single pose prediction. HGM is tailored to learn a distribution of feasible 2D/3D poses, drawing insights from the principles of uncertainty and ensemble learning. This approach ensures that the framework systematically addresses uncertainties at each stage, and accommodates the inherent variability and discrepancies of intermediate representations.

For the generation of 2D pose hypotheses, f_{2D} first undergoes several 1×1 convolutions. This results in a mean μ_{2D} and a standard deviation σ_{2D}, both with dimensions $H'W' \times C$ based on a Gaussian distribution. Once this distribution is constructed, HGM samples K latent features for each joint, implying that K hypotheses are generated. These sampled latent features are as a whole denoted as $z_{2D} \in \mathbb{R}^{K \times H'W' \times C}$. The generation process can be represented as the following:

$$\begin{aligned} \mu_{2D}, \sigma_{2D} &= \text{Convs}(f_{2D}), \\ \{z_{2D}^i\}_{i=1}^{K} &\sim \mathcal{N}(\mu_{2D}, \sigma_{2D}^2), \end{aligned} \tag{1}$$

where $\{z_{2D}^i\}$, $i = 1, \ldots, K$, represent the set of sampled latent features. The distribution $\mathcal{N}(\mu_{2D}, \sigma_{2D}^2)$ is a Gaussian characterized by mean μ_{2D} and variance σ_{2D}^2. The latent features z_{2D} are passed through a series of 1×1 convolutions to regress the 2D pose hypotheses $H_{2D\_heat} \in \mathbb{R}^{K \times N_J \times H'W'}$.

The procedure for 3D Pose Hypothesis Generation in HGM closely mirrors that of the 2D Pose Hypothesis Generation. The feature embedding f_{3D} is first processed through a series of 1×1 convolutions to generate a mean μ_{3D} and a standard deviation σ_{3D}, both with dimensions $N_J \times C$. In a manner consistent with the 2D process, K hypotheses are sampled for each joint. This leads to the formulation of latent features, denoted by $z_{3D} \in \mathbb{R}^{K \times N_J \times C}$. Following this, several fully connected (FC) layers regress the 3D pose hypotheses $H_{3D} \in \mathbb{R}^{K \times N_J \times 3}$ using the latent features z_{3D}.

The Progressive Pose-guided Feature Sampling (PFS) method is designed to extract joint-related localized features by leveraging the 2D/3D pose hypotheses, which serve as our intermediate representations. The concept of pose-guided feature sampling is anchored in our framework's capability of harnessing guidance from the 2D and 3D heatmaps during feature sampling and aggregation. Specifically, our approach emphasizes a progressive prediction when transitions from 2D to 3D poses. As depicted in the framework overview, the method starts with 2D pose-guided feature sampling and subsequently advances to 3D pose-guided feature sampling. These 2D and 3D intermediate representations strategically guide the framework to sequentially extract and aggregate pertinent joint features, and thus ensure a smoother

and more effective 2D-to-3D transition. With respect to 2D pose-guided feature sampling, our objective revolves around extracting joint-centric features from f_{enc}, under the guidance of the 2D pose hypotheses H_{2D_heat} predicted by HGM. A significant step involves a dot product operation between $H_{2D\_heat} \in \mathbb{R}^{K\times N_J\times H'W'}$ and $f_{enc} \in \mathbb{R}^{H'W'\times C}$, resulting in $f_{2D\_Sampled} \in \mathbb{R}^{K\times N_J\times C}$. Please note that a softmax operation is applied to H_{2D_heat} prior to the dot product operation. The procedure is expressed as follows:

$$f_{2D\_Sampled} = Softmax(H_{2D\_heat}) \cdot f_{enc}. \quad (2)$$

In the case of 3D pose-guided feature sampling, the framework employs 3D pose hypotheses to further refine the features. Given that our feature sampling inherently takes place in a 2D space, a preliminary projection from 3D to 2D is indispensable. As a result, the framework converts the 3D pose coordinates from 3D hypotheses $H_{3D} \in \mathbb{R}^{K\times N_J\times 3}$ into 2D pose coordinates $H_{3D->2D} \in \mathbb{R}^{K\times N_J\times 2}$ using the predicted camera parameters cam, which are regressed from f_{enc} via several FC layers. To construct 2D heatmaps from the deduced 2D coordinates $H_{3D->2D}$, the system adopts Gaussian blobs centered at these coordinates to generate H'_{2D_heat}. Drawing from the methodology of 2D pose-guided feature sampling outlined previously, a dot product between the softmax-treated H'_{2D_heat} and the extracted f_{enc} is performed, resulting in $f_{3D\_Sampled} \in \mathbb{R}^{K\times N_J\times C}$.

3.4. Inter-Intra Joint-Hypothesis Transformer

With the 2D sampled features captured, the next challenge lies in deriving their 3D counterparts. This translation is far from trivial. As the human body joints are interrelated, capturing and accurately estimating both inter- and intra-joint relations is essential. Towards this end, we incorporate the *Transformer in Transformer* architecture [11] into our model, and name it the Inter-Intra Joint-Hypothesis Transformer, abbreviated as $\mathcal{T}_{inter-intra}$. To demonstrate the operation of $\mathcal{T}_{inter-intra}$, we use the 2D sampled features $f_{2D\_Sampled}$ in the derivation of this section, although the procedure is equally applicable to $f_{3D\_Sampled}$. The features are divided into N_J patches: $f_{2D\_Sampled} = [f^1_{2D\_Sampled}, \ldots, f^{N_J}_{2D\_Sampled}] \in \mathbb{R}^{N_J\times K\times C}$, where each $f^i_{2D\_Sampled}$ denotes the set of hypotheses for the i^{th} joint.

Our Inter-Intra Joint-Hypothesis Transformer block encompasses two attention mechanisms: intra-joint and inter-joint attention, as illustrated in Fig. 4. The intra-joint attention focuses on capturing correlations within a single joint for effective hypothesis aggregation. During this process, each $f^i_{2D\_Sampled}$ undergoes a linear transformation to produce queries $\mathcal{Q} \in \mathbb{R}^{K\times d_q}$, keys $\mathcal{K} \in \mathbb{R}^{K\times d_k}$, and values $\mathcal{V} \in \mathbb{R}^{K\times d_k}$, where d_q and d_k denote the channel dimensions of $\mathcal{Q}$ and $\mathcal{K}$, respectively. The scaled dot-product attention can then be described by $Softmax(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}})V$. Utilizing the multi-head self-attention mechanism [35], our

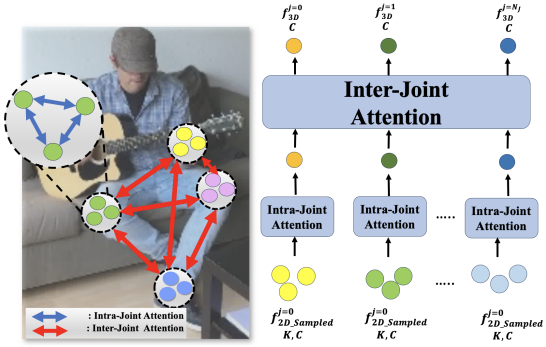

Figure 4. The Inter-Intra Joint-Hypothesis Transformer structure. Note that here we employ notation from the first $\mathcal{T}_{inter-intra}$.

approach divides the channel dimensions of $\mathcal{Q}$, $\mathcal{K}$, and $\mathcal{V}$ into h distinct segments, and performs self-attention on each separately. Subsequent to this, the output features from each head are concatenated and flattened. To further refine the representation, several FC layers are employed to condense the dimensions from $K \times C$ down to C. This transformation relects that the K hypotheses are aggregated through the attention mechanism into a singular feature for each joint.

For inter-joint attention, our objective is to capture the correlations among features pertaining to different joints. Analogous to the intra-joint attention mechanism, the features from disparate joints undergo transformations to form queries, keys, and values, followed by the application of multi-head self-attention. The resultant features embody the 3D characteristics of different joints, and encapsulate both inter- and intra-joint relationships. Please note that our framework incorporates two instances of $\mathcal{T}_{inter-intra}$. The first one derives f_{3D} from $f_{2D\_Sampled}$, and the second one produces f_{hmr} from $f_{3D\_Sampled}$, as illustrated in Fig. 2.

3.5. Training Objectives

Our training objective comprises four distinct loss terms: L_{hmr}, L_{2D}, L_{3D}, and L_{reg}. The total loss L_{total} is defined as:

$$L_{\text{total}} = \lambda_{\text{hmr}} \cdot L_{\text{hmr}} + \lambda_{2D} \cdot L_{2D} + \lambda_{3D} \cdot L_{3D} + \lambda_{\text{reg}} \cdot L_{\text{reg}}. \quad (3)$$

The term L_{hmr} encompasses the losses for human mesh reconstruction adopted in the prior work [7, 14, 17], given by:

$$L_{\text{hmr}} = \lambda_{\text{pose}} \cdot L_{\text{pose}} + \lambda_{\text{smpl}} \cdot L_{\text{smpl}}, \quad (4)$$

where L_{pose} measures the L1 distance between J_{3D} and its ground truth 3D pose, as well as the projected 2D pose and its ground truth 2D pose. Meanwhile, L_{smpl} denotes the L2 loss associated with the SMPL parameters. The last three terms are used for the intermediate representation, where L_{2D} computes the L2 losses for the 2D heatmap, and L_{3D} computes the L1 losses of the 3D coordinates. The final L_{reg} is a regularization term, which is used to regularize the learning distribution to generate hypotheses. This is

achieved by computing the KL divergence between (μ, σ) and a standard normal distribution $N(0, 1)$, expressed as:

$$L_{\text{reg}} = \text{KL}\left((\mu_{\text{dim}}, \sigma_{\text{dim}})||N(0,1)\right), \text{dim} \in \{2D, 3D\}. \quad (5)$$

4. Experimental Results

In this section, we first detail the experimental setups. This is followed by an examination of qualitative results and ablation studies. Finally, we present qualitative comparisons.

4.1. Experimental Setups

Implementation Details. We employ ResNet-50 [12] as our backbone for feature extraction, which is further enhanced by a 3-layer transformer encoder. Within our Inter-Intra Joint-Hypothesis Transformer, two layers of inter-intra joint attention are incorporated. We set the number of joints N_J to 24, which aligns with the superset of joints across our training datasets. Moreover, we establish the number of hypotheses K to 81. A detailed discussion regarding the choice of K can be found in Section 4.3. We follow the configuration from [6] and utilize the AdamW optimizer [26] with a learning rate of 10^{-4}, a weight decay of 10^{-4}, β_1 set to 0.9, and β_2 set to 0.999. We assign the loss function coefficients as $\lambda_{hmr} = 60$, $\lambda_{pose} = 5$, $\lambda_{smpl} = 1$, $\lambda_{3D} = 300$, $\lambda_{2D} = 200$, and $\lambda_{reg} = 10$. In the preprocessing stage, input images are cropped and resized to dimensions of 224×224 pixels, using the data augmentation techniques outlined in [6, 16, 21, 22]. The training spans 60 epochs with a batch size of 64. All computations are performed on a single NVIDIA RTX 3090 GPU. Note that we have developed the entire framework using PyTorch [29].

Datasets. To ensure a fair comparison, our experimental setup mirrors that of previous transformer-based approaches [6, 21, 22]. Initially, our model is pre-trained on a collection of datasets including Human3.6M [13], UP-3D [18], MuCo-3DHP [27], COCO [23], and MPII [1]. Following this, we evaluate its performance using the P2 protocol on Human3.6M. The model is further fine-tuned on the 3DPW [36] training set and then tested on its corresponding test set. It is important to note that Human3.6M serves as a significant indoor benchmark for 3D pose estimation, where training is conducted on five subjects (i.e., S1, S5, S6, S7, and S8) and testing on two (i.e., S9 and S11). On the other hand, 3DPW stands out as a challenging outdoor benchmark, which incorporates annotations for 3D body poses and meshes, and features a vast array of poses set against dynamic backgrounds and diverse scenarios.

Evaluation Metrics. We assess the performance of our model using three evaluation metrics: MPJPE [13], PA-MPJPE [40], and MPVPE [31]. MPJPE (Mean-Per-Joint-Position-Error) quantifies the Euclidean distance in millimeters between the predicted and the ground-truth joint coordinates. PA-MPJPE builds upon MPJPE by first aligning the estimated joint coordinates to the ground truth using Procrustes Analysis (PA) [9], then computing the error. On the other hand, MPVPE (Mean-Per-Vertex-Position-Error) measures the Euclidean distance between the estimated vertex coordinates and their corresponding ground-truth values.

Baselines. In this work, we evaluate and compare the experimental results of our method against Transformer-based methods [6, 21, 22], as well as CNN-based techniques that leverage intermediate representations, including [7, 10, 15, 19]. The comparison with [6] is particularly worth noting, as it represents the current state-of-the-art in Transformer-based techniques. Specifically, the objective is to highlight the efficacy of our progressive pose-guided learning and our Inter-Intra Joint-Hypothesis Transformer. Furthermore, we compare our method with [7, 19]. The former employs 2D pose as an intermediate representation, while the latter employs 3D pose. These comparisons aim to emphasize the importance of generating multiple hypotheses as intermediate representations, a strategy introduced in our framework that can effectively mitigate error propagation.

4.2. Comparison with Image-Based Methods

In Table 1, we compare our method with prior image-based human mesh recovery methods, spanning both CNN-based and Transformer-based approaches. Our evaluations are conducted on the 3DPW and Human3.6M datasets. On the Human3.6M dataset, our method achieves 49.61 MPJPE and 36.73 PA-MPJPE. This performance surpasses all preceding Transformer-based models and other representative methods when deploying on the same Resnet-50 backbone. It is worth noting that our method maintains a consistent performance trend on the 3DPW dataset, a more rigorous benchmark due to its outdoor setting. Our model achieves results comparable to the state-of-the-art FastMETRO-L with merely 70% of its parameters, and even surpasses it in several metrics. This trend persists even when adopting the larger HRNet-W64 backbone [32] across both datasets.

The outperformance of our method over Transformer-based approaches [6, 21, 22] suggests the effectiveness of the proposed progressive pose-guided learning and the Inter-Intra Joint-Hypothesis Transformer. Moreover, our method outperforms contemporary CNN-based approaches. Specifically, we surpass the method from [7] that uses a 2D pose as an intermediate representation, as well as the approach from [19] that utilizes a 3D pose as an intermediate representation. These comparisons highlight the significance of generating multiple progressive hypotheses for intermediate representations, a strategy adopted by our framework for mitigating error propagation. Our substantial gains in MPJPE can be attributed to our progressive usage of 2D and 3D poses as intermediate representations, which provide the framework with potent cues for accurate joint localization.

Table 1. Comparison with the state-of-the-art (SOTA) methods for 3D human mesh recovery on the Human3.6M [13] and 3DPW [36] datasets. Please note that the gray background indicates Transformer-based methods, and best results for each metric are highlighted in bold.

Method	Backbone	Human3.6M		3DPW		
		MPJPE (↓)	PA-MPJPE (↓)	MPJPE (↓)	PA-MPJPE (↓)	MPVPE (↓)
HMR [14]	ResNet-50	88.0	56.8	130.0	81.3	-
SPIN [16]	ResNet-50	62.5	41.1	96.9	59.2	116.4
GraphCMR [17]	ResNet-50	-	50.1	-	70.2	-
HoloPose [10]	ResNet-50	60.3	46.5	-	-	-
DaNet [37]	ResNet-50	61.5	48.9	-	56.9	-
I2L-MeshNet [28]	ResNet-50	55.7	41.7	93.2	57.7	110.1
HybrIK [19]	ResNet-34	54.4	**34.5**	80.0	48.8	94.5
PyMAF [38]	ResNet-50	57.7	40.5	92.8	58.9	110.1
PARE [15]	ResNet-50	-	-	82.9	52.3	99.7
METRO [21]	ResNet-50	56.5	40.6	-	-	-
FastMETRO-S [6]	ResNet-50	55.7	39.4	79.6	49.3	91.9
FastMETRO-L [6]	ResNet-50	53.9	37.3	77.9	**48.3**	90.6
Ours	ResNet-50	**49.6**	36.7	**76.7**	48.8	**89.9**
Pose2Mesh [7]	HRNet-W48	64.9	47.0	89.5	56.3	105.3
METRO [21]	HRNet-W64	54.0	36.7	77.1	47.9	88.2
MeshGraphormer [22]	HRNet-W64	51.2	34.5	74.7	45.6	87.7
FastMETRO-L [6]	HRNet-W64	52.2	33.7	73.5	**44.6**	84.1
Ours	HRNet-W64	**47.9**	**33.4**	**71.6**	45.1	**83.9**

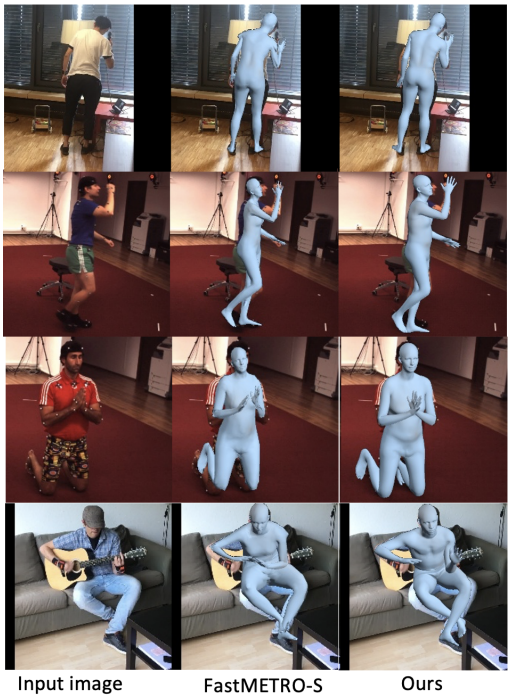

Figure 5. Qualitative comparisons of FastMETRO [6] and the proposed method on the Human3.6M and 3DPW datasets. Please note that our model size is comparable to that of FastMETRO-S.

4.3. Ablation Study

In this section, we ablatively analyze the effectiveness of our framework and validate our design choices. The framework is trained on the Human3.6M training set, and the evaluations are conducted on its corresponding test set.

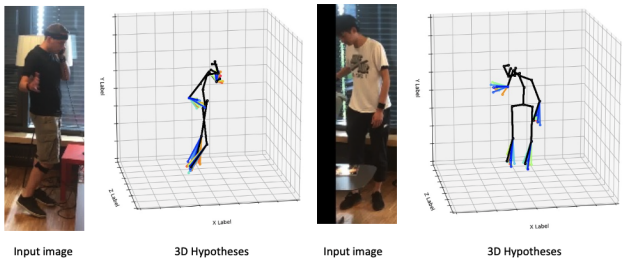

Figure 6. Visualization of the 3D Hypotheses. We visualize the left/right wrists and the left/right ankles. Various hypotheses are distinguished using distinct colors. In the case of occlusions, the model tends to generate a more diverse set of poses.

Table 2. Ablation study of the components introduced in our framework on Human3.6M. The baseline model is FatMETRO-S [6].

Progressive pose-guided learning	2D Hypothesis	3D Hypothesis	MPJPE (↓)	PA-MPJPE (↓)
✗	✗	✗	55.7	39.4
✓	✗	✗	51.69	37.5
✓	✓	✗	50.06	37.28
✓	✗	✓	49.99	37.24
✓	✓	✓	**49.61**	**36.73**

Table 3. Ablation study on different number of hypotheses.

Num. of Hypothesis	1	16	49	81	121
MPJPE(↓)	51.69	51.15	50.9	**49.61**	50.02

Effectiveness of the Components. To demonstrate the effectiveness of our progressive pose-guided learning and HGM, we compare the performance across five combina-

Table 4. Comparison of the computational complexity.

Method	FLOPs	Params	MPJPE (↓)
METRO [21]	56.8G	229.2M	54
FastMETRO-S [6]	8.9G	**32.7M**	55.7
FastMETRO-L [6]	11.8G	48.4M	53.9
Ours	**7.3G**	33.5M	**49.61**

Table 5. Ablation study on the design of inter-intra attention.

Method	MPJPE(↓)	PA-MPJPE(↓)
w/o intra-joint attention	50.54	37.11
w/o inter-joint attention	50.63	37.53
w/ both inter- and intra-joint attention	**49.61**	**36.73**

tions of different components as presented in Table 2. We select FastMETRO-S [6] as our baseline due to its comparable model size and its status as a state-of-the-art model. As seen in the first and second rows, our progressive pose-guided approach notably enhances the baseline performance by reducing the MPJPE from 55.7 to 51.69. This indicates that the predicted 2D/3D poses bridge the gap between 2D images and 3D meshes and our HGM generates diverse pose hypotheses to counter inaccuracies in intermediate representations. In the third row, it can be observed that incorporating 2D pose hypotheses alone brings a 1.6 MPJPE performance enhancement. Similarly, integrating 3D pose hypotheses results in a 1.7 MPJPE improvement. The most optimal performance is achieved when hypotheses are generated for both 2D and 3D poses concurrently. Such results highlight the significant advantages brought about by generating hypotheses for both 2D and 3D poses.

Effectiveness of the Attention Mechanism. We next assess the effectiveness of the two types of attention mechanisms within our $\mathcal{T}_{inter-intra}$. The results are presented in Table 5. It can be observed that removing the intra-joint attention and substituting it with several FC layers results in a decrease in performance by 0.93 MPJPE. A decline is also observed when we exclude the inter-joint attention. These highlight the significance of both inter- and intra-joint attentions, as they play a crucial role in capturing the relationships within each local joint and across different joints effectively.

Computational Cost and Model Parameters. Table 4 compares our approach's number of parameters and computational costs with several preceding Transformer-based models. These earlier models employ global attention to directly convert 2D images into 3D vertices. Such a conversion necessitates a significant number of parameters and results in high computational expenses for performance improvements. Our method requires fewer computational resources and a smaller model size to achieve performance comparable to or exceeding previous models. This reveals our capability to produce competitive outcomes while efficiently reducing both computational overhead and the model's overall size, thus outpacing the prior techniques.

Number Of Hypotheses. In Table 3, we evaluate the impact of employing different numbers of hypotheses. Since our approach is based on learning Gaussian distributions, we have the flexibility to select various quantities of hypotheses. It can be observed that there is a consistent improvement in performance as we increase the number of hypotheses, from 1 to 16, 49, and finally 81. Nevertheless, increasing the count further to 121 leads to a slight decline in performance. As a result, we select K=81 for our final experimental setting.

4.4. Qualitative Results

Fig. 5 presents the qualitative results of both FastMETRO [6] and our method on the Human3.6M and 3DPW datasets. As illustrated in the figure, FastMETRO exhibits artifacts and yields unrealistic outcomes for certain parts of the human body. Moreover, it occasionally produces inaccurate poses for complex activities, such as leg bending, overlapping, and occlusions. In contrast, our method consistently and accurately estimates 3D poses across different scenarios.

4.5. Visualization of Hypotheses

In Fig. 6, we visualize the 3D pose hypotheses for the left/right wrist and left/right ankles. As depicted in the figure, our HGM produces sets of plausible poses. It is worth noting that our model exhibits more diverse results in regions impacted by occlusion, indicative of areas with higher uncertainty. This enables our model to explore a broader spectrum of possibilities, leading to more precise decisions.

5. Conclusion

In this work, we introduced a multi-stage framework that progressively transforms 2D RGB images into 3D meshes. Through the generation of intermediate representations, our approach evaluates pose distributions and formulates plausible hypotheses. These are further refined by our Inter-Intra Joint Hypothesis Transformers. We benchmarked our approach against image-based human mesh recovery methods across the 3DPW and Human3.6M datasets. The results revealed that our method consistently outperforms Transformer-based and CNN-based predecessors even with fewer parameters. Our ablation study validated the robustness and efficiency of our framework's components.

6. Acknowledgments

The authors gratefully acknowledge the support from the National Science and Technology Council (NSTC) in Taiwan under grant numbers MOST 111-2223-E-007-004-MY3 and 110-2221-E-001-016-MY3. This work is also supported by Academia Sinica under grant number AS-TP-111-M02. The authors would also like to express their gratitude to the National Center for High-Performance Computing for providing the necessary computational and storage resources.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. 6

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViVit: A video vision transformer. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 6836–6846, 2021. 2

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 561–578, 2016. 1, 2, 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 213–229, 2020. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 9650–9660, 2021. 2

[6] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 342–359, 2022. 1, 2, 6, 7, 8

[7] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 769–787, 2020. 1, 3, 5, 6, 7

[8] 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 preprint arXiv:2010.11929*, 2020. 2

[9] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975. 6

[10] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10884–10894, 2019. 3, 6, 7

[11] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. 34:15908–15919, 2021. 5

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013. 1, 2, 6, 7

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 1, 2, 5, 7

[15] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3d human body estimation. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 11127–11137, 2021. 1, 2, 6, 7

[16] 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 *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2252–2261, 2019. 1, 2, 6, 7

[17] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4501–4510, 2019. 1, 2, 5, 7

[18] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6050–6059, 2017. 6

[19] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 1, 3, 6, 7

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1833–1844, 2021. 2

[21] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 1, 2, 6, 7, 8

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 12939–12948, 2021. 1, 2, 6, 7

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 740–755, 2014. 6

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Hierarchical vision transformer using shifted windows. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 10012–10022, 2021. 2

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph. (TOG)*, 34(6), oct 2015. 1, 2

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

[27] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation

from monocular rgb. In *Proc. Int. Conf. on 3D Vision (3DV)*, pages 120–130, 2018. 6

[28] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 752–768, 2020. 1, 2, 7

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[30] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1

[31] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 6

[32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 6

[33] 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 *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 5349–5358, 2019. 3

[34] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 20–36, 2018. 2

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 2, 5

[36] 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 *Proc. European Conf. on Computer Vision (ECCV)*, pages 601–617, 2018. 1, 2, 6, 7

[37] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. DaNet: Decompose-and-aggregate network for 3D human shape and pose estimation. In *Proc. ACM Int. Conf. Multimedia (ACMMM)*, pages 935–944, 2019. 3, 7

[38] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 11446–11456, 2021. 2, 7

[39] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021. 2

[40] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. MonoCap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)*, 41(4):901–914, 2018. 6