Multi-view Fusion Transformer For 3D Whole-Body Mesh Recovery

Junman Zhou, Ziwei Zhou

***Abstract*—Current mainstream methods for human mesh recovery can now be designed as a single-stage framework based on transformers and work through a multi-stage pipeline based on neural networks, thanks to the most recent research on ViT. This effectively avoids bottlenecks caused by multi-stage pipelines and the strict requirements for late fusion. Nevertheless, the majority of current transformer-based single-stage frameworks lack post-correction capabilities and do not properly utilize the extra information offered by multi-view images. This work suggests a comprehensive single-stage representation model for full-body human mesh recovery that incorporates multi-view features and optimizes with posterior probabilities to overcome these shortcomings. First, we extract high-quality multi-scale features by capturing global correlations using a global encoder. Second, we use RoIAlign to create a cross-view encoder that integrates multi-view image characteristics and produces higher-resolution images. Lastly, the SMPL human joint distribution is used as a physical prior for posterior probability optimization, and an encoder with deformable attention is used to improve the regression of hands and face. The model can now more thoroughly understand the relative locations of human joints based on the body poses supplied by multi-view photos, thanks to this enhancement. The model's one-step procedure is straightforward but efficient, and it does not require manual post-processing or datasets that are exclusively focused on hand and facial features for training, which naturally prevents irrational human parameter predictions. Our enhancements show that our upgraded model has greater generalization performance and faster learning speed than the baseline model, as evidenced by at least 4.69%, 6.70%, and 3.41% better performance on MPVPE for the body, hands, and face respectively.**

***Index Terms*—3D human mesh recovery, Cross-view Attention, Multi-view Fusion, Differentiable RoIAlign.**

I. Introduction

UNDERSTANDING human geometry and successfully extracting relevant information from a variety of data sources, including RGB-D cameras and multi-view photos, are essential components of 3D human mesh recovery, a critical step in modeling human behavior [1]. More detailed descriptions of human bodies, including the locations and shapes of muscles and bones as well as the textures of skin, can be obtained by consistently refining algorithms[2]. Furthermore, accurate recovery of human facial expressions and gestures opens the door to further study of human behavior by revealing human intentions and emotions.

Historically, deep learning-based multi-stage pipelines would identify faces and hands independently, crop and resize each area, increase their resolution, feed them into different networks to predict parameters, and then combine the outcomes[3]. Although the fine features of faces and hands can be captured by this copy-paste process [4], the method becomes computationally difficult when various estimators are used. Furthermore, because it is impossible to rejoin various body parts precisely, the impeded communication between components invariably leads to incompatible configurations during fusion, resulting in awkward wrist positions and 3D rotations that are outside of permitted bounds. Choi [1] et al. developed an elbow joint twist compensating fusion as a post-fusion technique to solve these problems. To improve resilience against complicated backgrounds, zhu [5] et al. incorporated model-fitting steps during training. Pixel-aligned implicit functions were presented by Saito [6] et al.To reconstruct high-resolution clothed human models from multi-view photos. This method captures precise geometric aspects of clothing while ensuring consistency across various positions. The system's complexity is greatly increased by the model's limited comprehension of global structures and its inability to use post-compensation techniques to rectify irrational predictions due to the limited local receptive fields.

Recent developments in Vision Transformers (ViTs) have made it possible for single-stage pipelines to capture intricate patterns and structures in images globally. Traditional multi-stage pipelines can be replaced with self-attention mechanisms, which allow all pixels in a picture to be taken into consideration at once, creating a comprehensive view of the human body [7]. Using a single-stage pipeline approach, Lin [8] et al. presented an enriched representation of full-body human mesh recovery that effectively reconstructs human models while preserving and improving fine-grained information such as gestures and facial expressions. Motivated by this, we created an MFT transformer that consists of two decoders tailored to specific local components and a global encoder. Decoders with cross-attention handle dependencies between multiple views and fuse features, providing more accurate estimates. Another decoder incorporates a differentiable RoIAlign cropping scheme to extract part-specific high-resolution features and uses keypoint-guided deformable attention to accurately locate and estimate hand and face parameters. The encoder that uses human tokens as inputs captures global correlations, predicts human parameters, and provides high-quality feature maps for the decoders. We suggest three crucial improvement modules to improve mesh estimation accuracy and model performance to address the drawbacks of earlier single-stage pipelines that did not fully optimize long-term sequence long-range dependencies and did not fully utilize the extra information offered by multi-view images [9].

1) Multi-View Fusion Scheme: Collaboratively estimate 2D poses from many viewpoints, assign weights based on

Manuscript received January 20, 2025; revised May 10, 2025.

Junman Zhou is a Postgraduate of the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: zhouzhou2001man@163.com).

Ziwei Zhou is a Professor of the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: 381431970@qq.com).

global correlations[4] to each view's significance for the present job. After that, features from various perspectives are combined to make sure that more important information is given more emphasis and less important information is downplayed. Fusion with Differentiable ROIAlign: Upsample feature tokens that convey particular semantic information into numerous higher-resolution feature maps via deconvolution layers. Divide the ROIs into multiple sub-regions and perform ROI mapping. Then, use bilinear interpolation and average the results.

2) Fusion with Differentiable RoIAlign: Upsampled feature tokens that convey particular semantic information into several higher-resolution feature maps via deconvolution layers. Divide the ROIs into multiple sub-regions and perform ROI mapping. Then, use bilinear interpolation and average the results.

3) Deformable Attention Decoder for Low-Resolution RoIs: This method handles low-resolution ROIs in images by dynamically selecting randomly distributed sample sites from input feature maps, as opposed to typical attention techniques that are limited by fixed sampling positions. To learn a collection of offsets around each key point and ascertain the precise sample sites, it selects the center point of the ROI as the guiding key point for deformable attention.

The goal of this research is to increase the quality of human mesh recovery in a single-stage Transformer-based framework by making these enhancements, especially in areas like hand details, body shape, full-body stance, and resilience against partial occlusions or blurring. Our approach lowers the MPVPE metric to 70.34mm on the EHF dataset, which is a 4.69% improvement over the baseline model. These findings show that in human mesh recovery tasks, the suggested improvements produce a model with improved robustness and detection accuracy.

II. Related Works

As of right now, the three primary methodologies for 3D human mesh recovery are hybrid, regression, and optimization-based approaches. By minimizing an objective function,optimization-based techniques usually take an iterative approach to estimating the parameters of the human model. Although this approach has the advantage of being physically sensible, it is computationally costly and prone to becoming trapped in local optima. As deep learning has advanced, regression-based techniques have progressively gained popularity. Two distinct regression-based frameworks are presented below, along with the issues they must resolve:

A. Multi-stage Pipeline 3D Human Mesh Recovery Algorithm Based on CNN

As deep learning has advanced, regression-based techniques have progressively gained popularity. These techniques, which have end-to-end learning capabilities, use neural networks to directly predict human model parameters from photos. By integrating 2D keypoints with image features, Kanazawa [2]et al.proposed HMR (Human Mesh Recovery), which employs a regression network to directly predict 3D posture and shape parameters. Based on the SMPL model, Kolotouros [10]et al. presented SPIN (SMPLify-X with a Neural Network), which uses a multi-stage optimization technique to gradually improve the 3D human mesh by optimizing position, shape, and deformation at various levels. A 3D human mesh recovery technique called Neural Body was proposed by Xu [11]et al. It uses deep implicit functions to gradually refine the 3D human mesh by combining time-series data and multi-view inputs.

Although these techniques can quickly provide high-quality 3D reconstruction results on huge datasets, the combination of several estimators in multi-stage pipelines results in a system with a high computational cost. Furthermore, because it is impossible to rejoin various body parts precisely, restricted communication between components invariably results in incompatible configurations during fusion, leading to abnormal wrist positions and 3D rotations that are beyond of permitted bounds [12]. By creating more intricate ensemble schemes or elbow joint twist compensating fusion modules between individual body parts, the aforementioned techniques aim to overcome these problems. In practice, nevertheless, this method greatly raises the complexity of the model system and has little potential to improve and rectify irrational forecasts.

B. Single-stage 3D Human Mesh Recovery Algorithm Based on Transformer

Transformer-based models have shown impressive results in a variety of domains, including computer vision and natural language processing. In recent years, their distinct benefit is the global self-attention mechanism, which circumvents the drawbacks of conventional convolutional neural networks [13], which are limited to local modeling, by capturing long-range global dependencies inside input images or point clouds. Parallel computing is another area in which these transformers shine. To improve the accuracy of pose estimation, Zhang [14] et al. presented the TransPose model, which combines 2D keypoints with picture features via cross-attention processes and employs a multi-head self-attention mechanism to capture global geometric information in images. Dosovitskiy [15] et al. employed ViT for global feature modeling of images and presented a multi-scale feature fusion mechanism that combines feature information from many levels to increase the accuracy of posture prediction. For the model to properly exploit multimodal data, Huang [16]et al. integrated cross-attention methods to connect 2D keypoints with picture features, facilitating information transmission between distinct feature spaces. These models are sensitive to the quality of input photos, which particularly affects 2D keypoint identification, even though they use lightweight network topologies. When dealing with a variety of human shapes and complicated settings, they may not be able to generalize well enough due to the influence of various body types, dress styles, or action positions [17]. They also lack the additional information about human stances that multi-view photos provide.

III. Methods

Four important elements of the MFT framework are thoroughly discussed in this section: First, we extract high-quality multi-scale features by using a global encoder to capture global correlations. This enables us to get a thorough

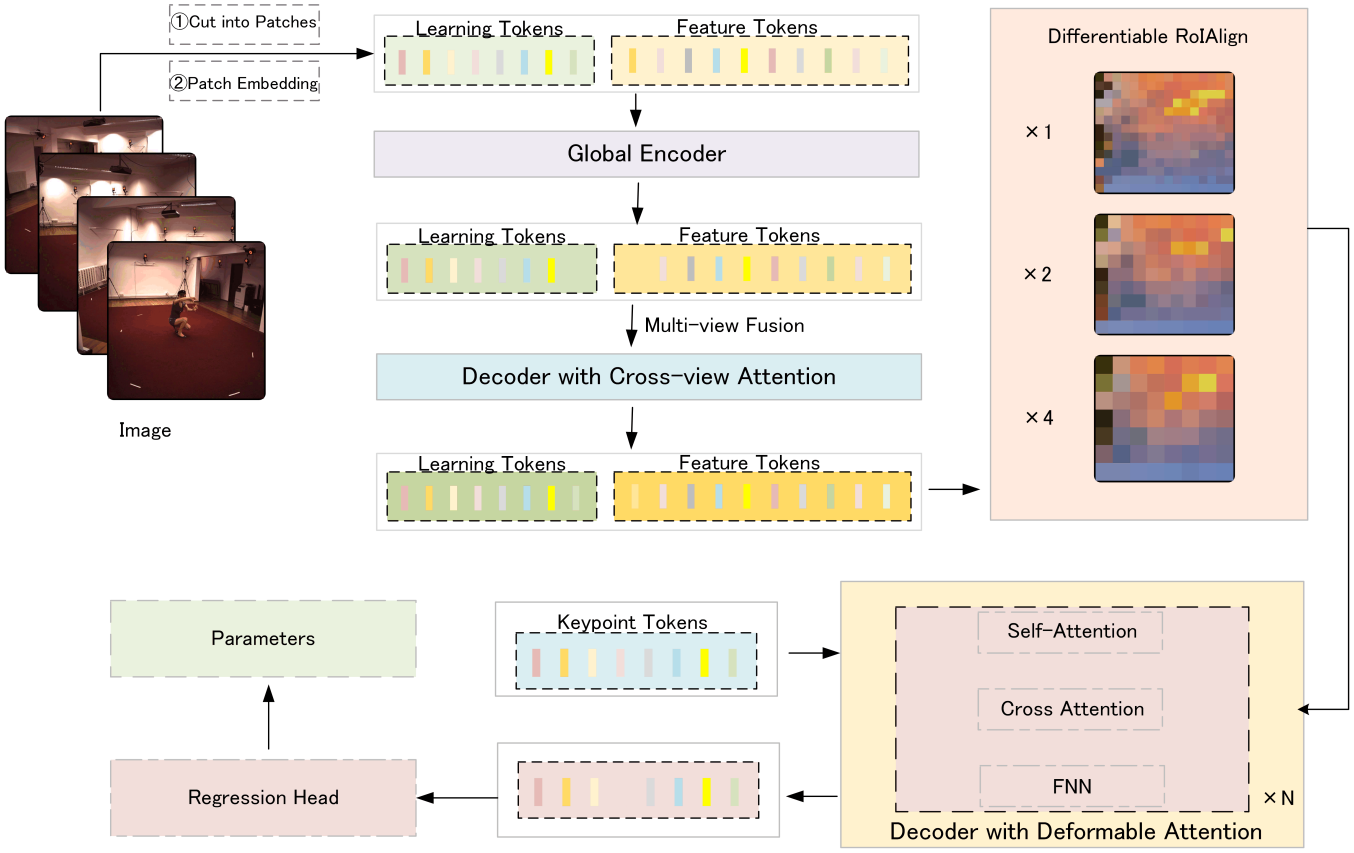

Fig. 1. Structure of the Multi-view Fusion Transformer (MFT)

comprehension of the scene at different resolutions. Second, we create a cross-view encoder that uses RoIAlign to acquire higher-resolution pictures and integrate multi-view image characteristics. By utilizing data from many viewpoints, this integration stage is essential for improving the accuracy of feature representation [18]. Thirdly, we use fusion with differentiable RoIAlign, which uses deconvolution layers to upsample lower-resolution feature maps into higher-resolution pictures. The collected characteristics are sufficiently detailed for accurate localization and classification tasks thanks to this procedure. Last but not least, the SMPL body joint distribution is used as a physical prior for posterior probability optimization in an encoder with deformable attention to improve the regression of hand and facial landmarks [19]. A more accurate reconstruction of the human mesh results from this optimization, which helps the model comprehend the relative locations of human joints based on the human postures supplied by multi-view inputs. Figure III illustrates our MFT architecture.

A. MFT Global Encoder

Given the target prediction image $\mathbf{I} \in \mathcal{I}^{H\times W\times 3}$, before being fed into the encoder to obtain global relevance, preprocessing of the target image is required. Based on the physical information of the image, which can represent the perspective, the target image is divided into appropriately sized patches $\mathbf{P} \in \mathcal{I}^{\frac{HW}{M^2}\times(M^2\times 3)}$, M is the patch size. These patches $\mathbf{P}$ are then linearly projected through convolution operations, transforming each patch into a one-dimensional vector $\mathbf{v} = \mathbf{Reshape}(\mathbf{P})W^X \in \mathcal{V}^{l\times d}$, This representation is suitable for subsequent framework processing.

Next, the positional and viewpoint information of the patches in the original image are added to the one-dimensional vectors, ultimately obtaining feature tokens $\mathbf{T}^{'}$ that contain internal dependencies within the target image data. Prior knowledge about the joints of different parts of the human body is used as body tokens $\mathbf{T_b}$, which are predefined learnable parameters in the model. To enable the model to better understand and process information related to human structure and achieve a more precise description of human shape, we concatenate the feature tokens $\mathbf{T}^{'}$ with the body tokens $\mathbf{T_b}$.

The concatenated tokens $\mathbf{T}$ are fed into our global encoder, which consists of multiple transformer layers stacked together. Each encoder layer includes a multi-head self-attention mechanism that attends to all other positions in the input sequence, two linear transformations, and a feedforward network composed of an activation function, residual connections, and layer normalization.The multi-head self-attention mechanism projects the input sequence $\mathbf{T}$ into attention triplets $(\mathbf{Q_i}, \mathbf{K_i}, \mathbf{V_i})$:

$$(Q_i, K_i, V_i) = (XW_i^Q, XW_i^K, XW_i^V) \tag{1}$$

These triplets are split into $\mathbf{h}$ subspaces, such as query Q being a three-dimensional matrix obtained by reshaping the input sequence and multiplying it with weight matrices W, $\{\mathbf{Q_1}, \mathbf{Q_2}, ...\mathbf{Q_h}\} \in \mathcal{V}^{\mathbf{h}\times l\times \frac{d}{h}}$ used to confirm the relevance between different positions in the sequence. Dot-product attention is performed in each subspace and the results are concatenated into context-aware token sequences:

$$Z = Concat(Y_1, Y_2, \ldots, Y_h)W^Z \in \mathbf{R}^{l\times d} \tag{2}$$

including $Y_i = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d/h}}\right) V_i \in \mathbf{R}^{l \times \frac{d}{h}}$. The final sequence is input into a layer normalization sub-layer. This forms a higher-level, stronger representation form with enhanced expressive power. The feature tokens are updated to $\mathbf{T}^{'}$. This process involves using feature tokens derived from various views to improve the accuracy of body pose estimation. Afterward, specific tokens related to the face and hands are extracted and input into a specialized decoder. This decoder, guided by keypoints, employs deformable attention mechanisms to generate more detailed feature maps, which significantly improves the precision of hand and face regressions.

B. Decoder With Cross-view Attention

To fully exploit multi-view picture information and obtain precise depth information on human postures, we propose a cross-view attention strategy to dynamically select and combine significant information from several perspectives. Using the global encoder, we extract picture properties and tokens from each perspective. Assuming we have N viewpoints (typically N=4), each viewpoint contains an image feature map $\mathbf{F_i} \in \mathcal{I}^{B \times C \times H \times W}$ and tokens $\mathbf{T}^{'} = \{\mathbf{T_b}, \mathbf{T_h}, \mathbf{T_f}\} \in \mathcal{I}^{B \times Q \times D}$, where: B is the batch size, C is the number of channels, H and W are the height and width of the feature map, Q is the number of queries, D is the model dimension.

We use the current viewpoint as target tokens to obtain better-initialized component tokens than random initialization. For each viewpoint, we flatten the feature map and adjust the dimension order, then concatenate them together as the feature map for the global view:

$$\mathbf{A}, \mathbf{A_{weights}} = \mathbf{MultiheadAttention}(\mathbf{T_q}, \mathbf{T_k}, \mathbf{T_v}) \quad (3)$$

where $\mathbf{A} \in \mathcal{I}^{\mathbf{Q} \times \mathbf{B} \times \mathbf{D}}$ is the attention output with shape and $\mathbf{A_{weights}}$ is the attention weight matrix. To enhance the interaction between viewpoints, all viewpoints' feature maps and tokens can directly interact, capturing more complex cross-view relationships. We set the target viewpoint as query $\mathbf{T_q}$, other viewpoints' feature maps as keys $\mathbf{T_k}$ and values $\mathbf{T_v}$ and compute the similarity between queries and keys using the dot-product:

$$\mathbf{Score}(\mathbf{T_q}, \mathbf{T_k}) = \frac{\mathbf{T_q} \mathbf{T_k^T}}{\sqrt{\mathbf{d_k}}} \quad (4)$$

where Q is the query matrix, K is the key matrix, and $\mathbf{d_k}$ is the dimension of the key vector. Finally, the weighted sum of value vectors is combined to generate the final attention output:

$$\mathbf{Output} = \sum_{\mathbf{i}} \mathbf{softmax}(\mathbf{Score_i}) \mathbf{V_i} \quad (5)$$

This method seeks to improve the model's capacity to analyze and learn from multi-view image features to optimize the relative positions of human joints, enhance the model's perception of human meshes and poses, and improve the accuracy of 3D human mesh recovery under typical background conditions.

C. Differentiable RoIAlign to Obtain High-Resolution Images

Traditionally, when processing low-resolution feature maps of hands and faces in human images to achieve high-resolution results, methods such as linear interpolation or transposed convolution are often used [20]. However, these methods can sometimes lead to issues like checkerboard artifacts and may not be the optimal choice, especially when precise control over the upsampling process is required. Traditional RoIPooling methods apply maximum pooling or average pooling for each grid to obtain fixed-size outputs. This pooling process is non-differentiable because it involves rounding operations, which can cause gradient propagation problems and potentially lose precise positional information.

To address the aforementioned issues, we propose a differentiable RoIAlign method to obtain high-resolution feature maps. Specifically, we reshape the output feature tokens $\mathbf{T}^{'}$ from the full-body encoder and upsample them through transposed convolution layers to multiple higher-resolution feature maps $\mathbf{T_h r}$. We then use FFNs (Feed-Forward Networks) to regress the feature tokens back to hand and face bounding boxes. For an input feature map $\mathbf{F} \in \mathcal{I}^{\mathbf{H} \times \mathbf{W} \times \mathbf{C}}$, first convert the RoI bounding box coordinates under the original image coordinate system to coordinates on the feature map. Assume the original image size is $((\mathbf{H_{img}}, \mathbf{W_{img}})$, and the feature map size is $(\mathbf{H}, \mathbf{W})$. The RoI coordinates on the feature map can be calculated using the following formula:

$$(\mathbf{x_{min}}, \mathbf{y_{min}}, \mathbf{x_{max}}, \mathbf{y_{max}}) = \left(\frac{\mathbf{x_{min}^{img}}}{\mathbf{s}}, \frac{\mathbf{y_{min}^{img}}}{\mathbf{s}}, \frac{\mathbf{x_{max}^{img}}}{\mathbf{s}}, \frac{\mathbf{y_{max}^{img}}}{\mathbf{s}}\right) \quad (6)$$

Here, $\mathbf{s}$ is the downsampling ratio of the feature map relative to the original image, and the sampling rate. Divide the RoI into $H \times W$ sub-regions. Each sub-region corresponds to one position on the output feature map. For each RoI sub-region, select four adjacent points for bilinear interpolation. Let the top-left corner coordinates of the sub-region be $(\mathbf{x_p}, \mathbf{y_p})$, then interpolate the four points of the bounding box:

$$\mathbf{I}(\mathbf{x_p}, \mathbf{y_p}) = \sum_{\mathbf{i=1}}^{\mathbf{2}} \sum_{\mathbf{j=1}}^{\mathbf{2}} \mathbf{a_{ij}} \mathbf{I}(\mathbf{x_i}, \mathbf{y_j}) \quad (7)$$

where $\mathbf{I}(\mathbf{x_i}, \mathbf{y_j})$ represents the value at point $(\mathbf{x_i}, \mathbf{y_j})$ on the feature map and

$$\mathbf{a_{11}} = (\mathbf{x_2} - \mathbf{x_p})(\mathbf{y_2} - \mathbf{y_p})$$
$$\mathbf{a_{12}} = (\mathbf{x_2} - \mathbf{x_p})(\mathbf{y_p} - \mathbf{y_1})$$
$$\mathbf{a_{21}} = (\mathbf{x_p} - \mathbf{x_1})(\mathbf{y_2} - \mathbf{y_p})$$
$$\mathbf{a_{22}} = (\mathbf{x_p} - \mathbf{x_1})(\mathbf{y_p} - \mathbf{y_1})$$

Finally, average the interpolated values for each sub-region to generate the final multi-scale hand feature map $\mathbf{F_h}$ and face feature map $\mathbf{F_f}$.

D. Deformable Attention Decoder for Hand and Face Features

While deformable attention mechanisms enable the network to learn from input feature maps to dynamically select a collection of unevenly distributed sampling sites instead of adhering to a predefined grid pattern, traditional attention

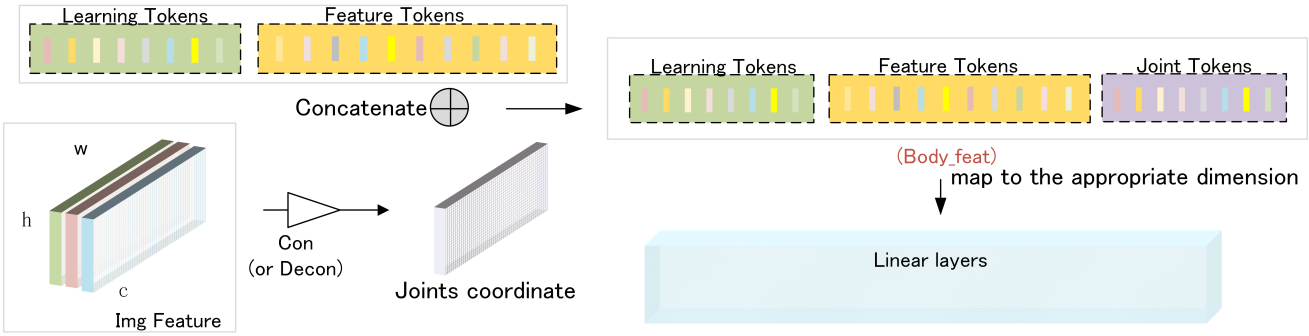

Fig. 2. Deformable Attention Decoder for Hand and Face Features

mechanisms are frequently constrained by fixed sampling places.

For example, to handle facial features $\mathbf{F_f}$, we first use the center point of the facial bounding box as the key point to guide the deformable attention mechanism. Around each key point, the network learns a set of offsets to determine the actual sampling positions. The query $\mathbf{Q}$ is initialized as component tokens guided by key points, which are obtained through reference key point features, pose position embedding, and learnable embedding. Next, by combining multi-scale feature maps, information can be extracted from different abstraction levels. Multi-scale Feature Map $\{\mathbf{F_i}\}_{\mathbf{i=1}}^{\mathbf{L}}$ are feature maps from different layers, with each layer's feature map represented as $\mathbf{F_i} \in \mathcal{I}^{\mathbf{H_l} \times \mathbf{W_l} \times \mathbf{C}}$.

For each query and each scale, the network predicts a set of sampling points. These sampling points are around the reference key point position $\mathbf{P_k}$ and have learnable offsets $\mathbf{\Delta P_{lk}} \in \mathcal{I}^{\mathbf{2}}$.In multi-head self-attention layers, the query undergoes multi-head self-attention computation to capture relationships between queries. In multi-scale deformable cross-attention layers, multi-scale feature map information is extracted. Given the query $\mathbf{Q}$ and multi-scale feature map $\{\mathbf{F_i}\}_{\mathbf{i=1}}^{\mathbf{L}}$, the deformable attention decoder updates the representation of the query by computing the relationship between the query and the corresponding features of each sampling point:

$$\mathbf{CA}(\mathbf{Q}, \mathbf{F_i}, \mathbf{P_k}) = \sum_{\mathbf{l=1}}^{\mathbf{L}} \sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{W_{lk}} \mathbf{M} \mathbf{F_i}(\phi_{\mathbf{i}}(\mathbf{p_k}) + \mathbf{\Delta P_{lk}}) \tag{8}$$

Here, $\mathbf{W_{lk}}$ is the attention weight, $\mathbf{M}$ is the linear transformation matrix, $\phi_{\mathbf{i}}(\mathbf{P_k})$ represents the offset of the key point position in the l-th layer feature map, $\mathbf{\Delta P_{lk}}$ represents the learnable offset. After processing with multi-scale deformable cross-attention, the results are further processed through a feed-forward neural network (FFN), which consists of two fully connected layers with a ReLU activation function in between. Joint coordinates are derived from image features through convolution operations. The features processed by the deformable attention decoder are concatenated with these joint coordinates to form body features as Figure III-C. These body features are then flattened and passed through a linear layer for dimensionality mapping, ultimately serving as parameters for the human model.

E. Incorporating SMPL as a Physical Prior for Posterior Probability Constraints

Since SMPL is trained on a lot of scan data, many unlikely or severely distorted human configurations are naturally excluded from its parameter space. Thus, the accuracy and resilience of body shape and pose estimation can be further improved by reducing outliers or inaccurate estimations in forecasts by imposing posterior probability restrictions using SMPL as prior knowledge. We define the prior distribution of shape parameters β and pose parameters θ based on the SMPL model as the prior probability, while the error between the 2D keypoints predicted from input images and those generated by SMPL serves as the likelihood function to construct the posterior probability:

$$P(\beta, \theta | I) = \frac{P(I|\beta, \theta) P(\beta, \theta)}{P(I)} \tag{9}$$

Here, $P(I)$ is the evidence term, which is constant for optimization problems, so we directly optimize the unnormalized posterior probability:

$$\hat{\beta}, \hat{\theta} = \arg\max_{\beta, \theta} P(I|\beta, \theta) P(\beta, \theta) \tag{10}$$

Gradient descent optimization algorithms are used to determine the SMPL parameters that optimize the posterior probability. To guarantee the physical plausibility of the final recovered human mesh, extra regularization terms are introduced during optimization, including the regression loss of 3D full-body keypoints, 2D full-body keypoints, and 2D bounding boxes of the left/right hands and face. These constraints are implemented by adding corresponding penalty terms to the loss function:

$$\mathcal{L} = -\log(P(I|\beta, \theta)) - \log(P(\beta, \theta)) + \lambda_1 \mathcal{L}_{phy}$$

Here, $\mathcal{L}_{phy} = \mathcal{L}_{smplx} + \mathcal{L}_{kpt3D} + \mathcal{L}_{kpt2D} + \mathcal{L}_{bbox2D}$ and λ_1 is a tuning coefficient. All terms are calculated as L1 distances between ground truth and predictions. Specifically, $\mathcal{L}_{smplx}$ provides explicit supervision for SMPL-X parameters, whereas $\mathcal{L}_{kpt3D}$, $\mathcal{L}_{kpt2D}$, and $\mathcal{L}_{bbox2D}$ represent the regression losses for 3D full-body keypoints, projected 2D full-body keypoints, and 2D bounding boxes of the left/right hands and face respectively.

IV. Experiment

A. Environment Setup

The experimental equipment setup are displayed in Table I. All photos were resized to 512×384 pixels to comply with

TABLE I
EXPERIMENTAL EQUIPMENT SETUP

Hardware/Frameworks	Parameters and Versions
CPU	Intel Xeon(R) Gold 6430
GPU	NVIDIA RTX 4090
RAM	200GB
Operating System	Ubuntu20.04
Program Language	Python 3.8
ML Library	PyTorch 1.11.0
CUDA	11.3

the model's input specifications. The computer hardware-corresponding batch size was set to 8.

B. Experimental Dataset

We have chosen to use EHF as our testing dataset and Human3.6M and MSCOCO as our training datasets. Our choice of training datasets excludes extra datasets designed to improve hand and face mesh recovery, in contrast to earlier multistage pipeline training methods.

1) Human3.6M: With more than 3.6 million frames of data, the "Human 3.6 Million" dataset [21] is one of the biggest publicly accessible datasets for human pose estimation and is frequently used to train and assess human pose estimation models. Eleven professional actors—seven men and four women—perform a range of movements in this dataset, which includes over 70 action sequences of various everyday activities like sitting, walking, hugging, shaking hands, and more. A high-precision Vicon motion capture technology was used to record each action from four distinct angles, yielding precise three-dimensional data on the position of human joints.

2) MSCOCO: More than 330,000 photos from 91 item categories make up the Microsoft Common Objects in Context (MSCOCO) dataset [22], 80 of which are utilized to train models. An average of five object instances are annotated in each image, and each object instance has 17 human keypoints. Models may learn to identify things in complex situations thanks to the rich contextual information provided by the photos, which are taken from real-world scenes. MSCOCO provides more practical use cases by emphasizing how items are positioned inside particular contexts and how they relate to one another. We utilize a pre-trained model on the MSCOCO dataset to identify each observed individual's 17 body keypoints using the recently developed 2D pose estimator ViT-Body-only. We determine the maximum and minimum values of the identified left and right hand keypoints to generate approximate hand-bounding boxes. The hand areas are then cropped for additional processing after these bounding boxes are fine-tuned using an Intersection over Union (IoU) matching approach to produce accurate hand bounding boxes.

C. Evaluation Metrics

To objectively evaluate the quality of human mesh recovery and model performance, we use three widely adopted metrics: MPVPE (Mean Per Vertex Position Error), PA-MPVPE (Procrustes-aligned Mean Per Vertex Position Error), and PA-MPJPE (Procrustes-aligned Mean Per Joint Position Error).

1) MPVPE: MPVPE [23]measures the vertex position differences between the predicted human mesh and the ground truth human mesh.

$$\mathbf{MPVPE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{v}_i^{\mathbf{pred}} - \mathbf{v}_i^{bfgt}\|_2 \tag{11}$$

2) PA-MPVPE: PA-MPVPE [24]involves aligning the predicted model with the ground truth through similarity transformations (such as rotation and translation) before calculating the MPVPE.

$$\mathbf{PA-MPVPE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{R}(\mathbf{v}_i^{\mathbf{pred}} - \mathbf{t}) - (\mathbf{v}_i^{\mathbf{gt}} - \mathbf{t}_{\mathbf{gt}})\|_2 \tag{12}$$

3) PA-MPJPE: PA-MPJPE quantifies the differences between the predicted 3D human joint positions and the actual joint positions.

$$\mathbf{PA-MPJPE} = \frac{1}{M}\sum_{j=1}^{M} \|\mathbf{R}(\mathbf{j}_j^{\mathbf{pred}} - \mathbf{t}) - (\mathbf{j}_j^{\mathbf{gt}} - \mathbf{t}_{\mathbf{gt}})\|_2 \tag{13}$$

D. Experimental Results

Taking inspiration from Hand4Whole and MMT, we take OSX as the basis for this research and progressively add more enhancement modules. The findings of the ablation study below show how important and beneficial these modules are. While the incorporation of a differentiable RoAlign improves high-resolution image processing to improve the clarity of hand and face estimations, the ablation experiments demonstrate that the cross-view attention decoder added in this paper effectively improves the model's accuracy for human body pose estimation. Furthermore, the model's capabilities are improved by posterior probability restrictions, which efficiently speed up the regression of different losses.

The suggested improved model continuously performs better in terms of accuracy and stability than the baseline model. These outcomes confirm the suggested model's accuracy and efficacy in 3D human mesh recovery. The model in this study performs better in human stance, face, and hand mesh recovery when compared to the baseline model, confirming the efficacy of the modifications made.

Given that OSX, the baseline model, has a straightforward design and performs well across a variety of benchmarks, it is important to think about whether MFT can effectively regress parameters and lessen departure from ground truth despite adding complexity to the system. Thus, we recorded different losses during their training processes, such as 2D coordinate loss, joint coordinate loss, shape parameter loss, pose parameter loss, and so on, and compared the present MFT with the basic model without the addition of a cross-view attention module. Figure IV-D, which displays the loss variation of both approaches during the training process,

TABLE II
COMPARISON OF EVALUATION METRICS AMONG MODELS ON EHF

Method	MPVPE(mm)			PA MPVPE(mm)			MPJPE(mm)		PA MPJPE(mm)	
	ALL	Hands	Face	ALL	Hands	Face	Body	Hands	Body	Hands
SMPLify-X	88.6	55.4	44.9	59.7	14.5	7.3	-	-	87.6	13.5
ExPose	77.1	51.6	35.0	54.5	12.8	5.8	-	-	62.8	12.7
FrankMocap	107.6	42.8	-	57.5	12.6	-	-	-	62.3	12.6
SPIN	81.5	49.6	38.1	58	14.2	5.7	-	-	102.9	-
PIXIE	89.2	42.8	32.7	55.0	11.1	4.6	91.0	-	61.3	-
Hand4Whole	79.2	43.2	25.0	53.1	12.1	5.8	79.2	39.8	53.1	12.8
OSX	73.8	40.7	26.4	48.7	15.9	6.0	74.7	-	45.1	-
MFT(Ours)	70.34	39.1	25.3	45.5	13.2	6.7	70.44	40.2	49.1	18.06

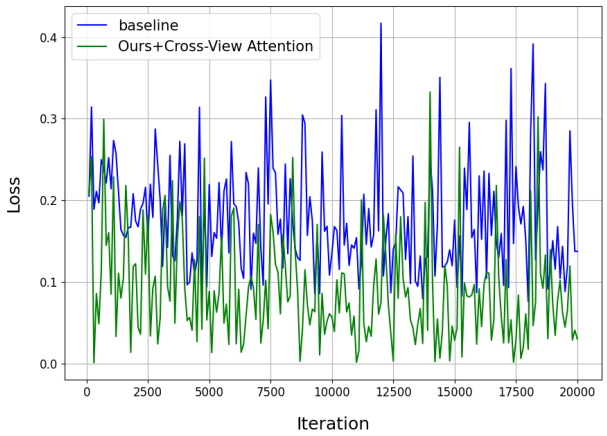

Fig. 3. Comparison of the Loss changes with the baseline(without the cross-view attention module)

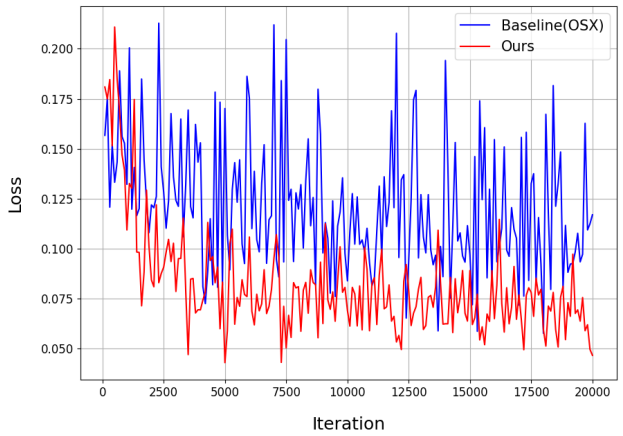

Fig. 4. Comparison of the Loss summary and changes with the baseline(OSX)

illustrates the results. During iterations, both curves displayed random fluctuation characteristics rather than any discernible upward or downward tendencies. It is clear that compared to MFT, the baseline approach without the cross-attention module showed greater volatility and overall losses. MFT with the cross-attention module was noticeably better, particularly in the early phases of iteration, when there were fewer than 5000 iterations. This research leads us to the conclusion that the cross-attention module must be included and that it can successfully improve model performance.

The performance and different metrics of our MFT and the baseline model OSX are compared in the second section. First, we train both models on the same datasets (MSCOCO and Human3.6M), and then we evaluate each model's metrics (MPVPE, PA-MPVPE, MPJPE, and PA-MPJPE) using the EHF dataset. Second, we compare the two models' loss variations throughout training to examine how they change over time.

As shown in Figure IV-D, during the initial iteration phase, there is not much difference in performance and regression efficiency between our model MFT and OSX. However, our model demonstrates a faster rate of reducing losses, thereby enhancing performance, with less fluctuation in losses, indicating a more stable parameter learning process. Subsequently, we summarize the performances of several outstanding models in the domain of 3D human mesh recovery on the EHF dataset as shown in Table II. This comparison allows us to evaluate the effectiveness and superiority of our proposed MFT model against existing state-of-the-art models, providing insights into its capabilities and potential areas for improvement.

As shown in table II, Based on the experimental results, our proposed Transformer-based approach for 3D human mesh recovery achieved an MPVPE of 70.34 mm, with minor reductions in the hand and face areas, achieving 50.1 mm and 27.3 mm respectively. Our MFT also showed excellent performance in PA-MPVPE, with full-body, hand, and face PA-MPVPE scores of 45.5 mm, 13.2, and 6.7 respectively; similarly impressive results were observed in MPJPE, with body and hand MPJPE scores of 70.44 and 40.2 respectively. Compared to PIXIE, improvements in MPVPE across body, hand, and face areas were 14.95%, 14.80%, and 16.51% respectively; compared to Hand4Whole, improvements in MPVPE across these areas were 11.19%, 15.97%, and 9.2% respectively; compared to OSX, improvements in MPVPE across these areas were 4.69%, 6.70%, and 3.41%

Fig. 5. Comparison of the performance with other models

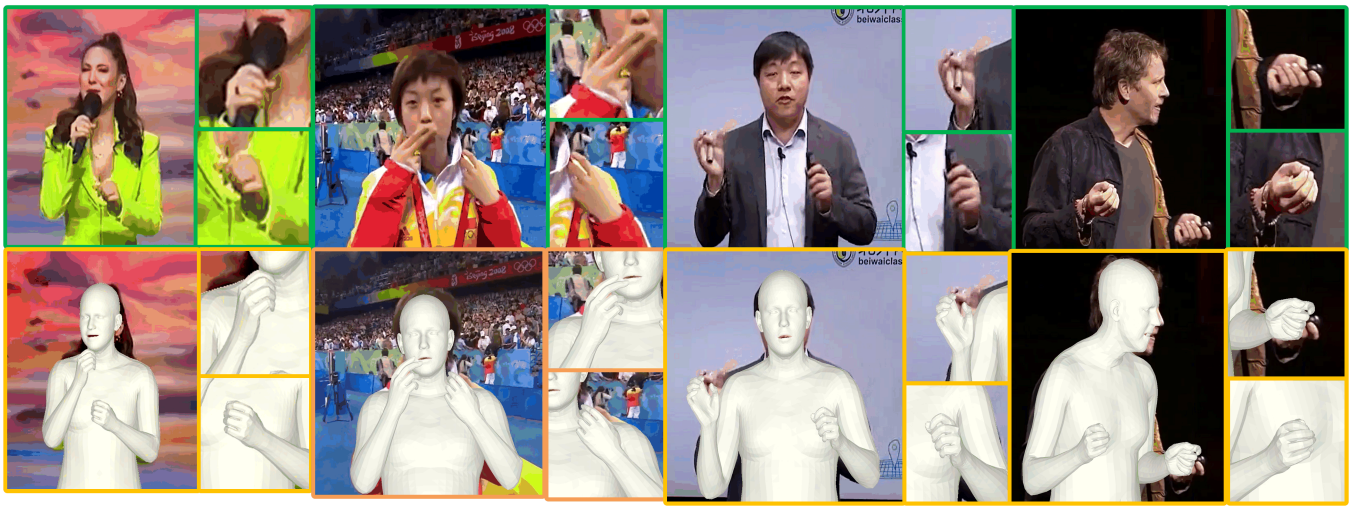

Fig. 6. Performance of MFT on Daily Life Images

TABLE III
COMPARISON OF HAND METRICS MPVPE OBTAINED BY MODELS THAT PREDICT 3D WRIST ROTATION FROM VARIOUS FEATURES ON EHF.

Input for 3D wrist prediction	MPVPE (Hands)
Body	50.6
Body + Hand GAP	45.2
Body + All hand joints	43.4
Body + feature joints (Ours)	**39.1**

TABLE IV
COMPARISON OF WHOLE-BODY METEICS PA MPVPE, OBTAINED BY MODELS THAT TAKE VARIOUS INPUT COMBINATIONS FOR THE 3D JOINT ROTATION PREDICTION ON EHF.

Input for 3D joint rotations	PA MPVPE (All)
GAP feat	59.2
Joint feat	52.3
2D joint coord	49.2
3D joint coord	47.1
3D joint coord.+joint feat.(Ours)	**45.5**

respectively, To provide the translation more intuitively as shown Figure IV-D. Improvements were also noted in other metrics such as PA-MPVPE and PA-MPJPE albeit to a lesser extent. These results indicate that the proposed improvement strategy effectively enhances the model's performance and improves its overall capabilities as shown in Figure IV-D.

The table III shows that relying only on body information to forecast 3D wrist rotation has a poor effect. The model's prediction accuracy has increased by the addition of the hand GAP features. Accordingly, the prediction effect of the model can be further improved by include information about all hand joints. The accuracy of 3D wrist rotation prediction steadily improves with input feature optimization and enrichment. With the MPVPE value reaching 39.1, which is almost 22% lower than the baseline model that solely uses body information, Ours approach which was specifically suggested in this study—achieved the best prediction result by carefully choosing feature joints. This demonstrates that appropriate feature selection and use are crucial for enhancing model performance in the direction of 3D human mesh recovery.

The table IV shows that using GAP characteristics alone to predict 3D joint rotation has a poor result. The model's prediction accuracy has increased when joint characteristics were added, but much more may be done. Accordingly, using 2D joint coordinate data can improve the model's predictive power even more. Prediction accuracy can be further increased by using 3D joint coordinates as opposed to 2D joint coordinates. The accuracy of 3D joint rotation prediction can be greatly increased by our approach, which combines 3D joint coordinates with joint characteristics. The error is reduced by 24% when compared to the baseline model that uses GAP features and by 8% when compared to the baseline model that uses only 3D joint coordinates. This demonstrates that our approach may successfully increase 3D human mesh recovery accuracy and decrease error.

In addition to highlighting the benefits of using the cross-view attention module in our MFT framework, this evaluation also demonstrates how effective it is at enhancing model stability and performance efficiency when compared to the baseline method.

V. CONCLUSION

Without utilizing extra hand- or face-specific datasets [25], we obtain state-of-the-art performance across six benchmarks in our study, addressing the shortcomings of single-stage models in processing multi-view images [26]. Our contributions are put into practice clearly and effectively.

To improve the quality of 3D human mesh recovery, the changes suggested in this research are essential. By offering more detailed descriptions of body postures and shapes and encouraging the interpretation of human intentions and emotions through gestures and facial expressions[27], human mesh recovery contributes to a richer knowledge of human geometry. Models can be further investigated and optimized in future studies to fully utilize richer datasets for improved performance. Furthermore, incorporating additional cutting-edge deep learning methods may improve the models' efficacy and adaptability in subsequent applications.

REFERENCES

[1] H. Choi, G. Moon, and KM. Lee "Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose" in *ECCV 2020:16th European Conference*, pp.769-787.

[2] A. Kanazawa, M. Black, D. Jacobs, and J. Malik "End-to-End Recovery of Human Shape and Pose " in *CVPR: IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018*, pp 7122-7131.

[3] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis "Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the loop" in *ICCV: IEEE/CVF International Conference on Computer Vision 2019*, pp.2252-2261.

[4] Y. Deng, J. Yang, S. Xu, D.Chen, Y.Jia, and X. Tong "Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set" in *CVPRW: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops 2019*, pp.285-295.

[5] X. Zhu,W. Su,L. Lu,B. Li,X. Wang, and J. Dai "Deformable DETR:Deformable Transformers for End-to-End Object Detection" in *ICLR 2021*.

[6] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li "PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization" in *IEEE/CVF International Conference on Computer Vision 2019*, pp.2304-2314.

[7] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. Black "Monocular Expressive Body Regression Through Body-Driven Attention" in *Computer Vision ECCV - 2020*, pp.20-40.

[8] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li "One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer" in *CVPR: IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023*.

[9] Y. Rong, T. Shiratori, and H. Joo "FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration" in *IEEE/CVF International Conference on Computer Vision 2021*, pp.1749-1759.

[10] N. Kolotouros, G. Pavlakos, and K. Daniilidis "Convolutional Mesh Regression for Single-Image Human Shape Reconstruction" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019*, pp.4496-4505.

[11] S. Peng, C. Geng, Y. Zhang, Y. Xu "Implicit Neural Representations With Structured Latent Codes for Human Body Modeling" in *IEEE Transactions on Pattern Analysis and Machine Intelligence 2023*, pp.9895-9907.

[12] A. Kundu, Z. Huang,Y. Zhou,S. Zhu "Pose and Shape from Articulated Reconstruction Engine" in *European Conference on Computer Vision (ECCV) 2020*.
[13] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and J. Michael "Collaborative Regression of Expressive Bodies using Moderation" in *International Conference on 3D Vision 2021*.
[14] Y. Zhang, X. Zhou, and J. Zhu "TransPose: A Single-Stage Transformer for 3D Human Pose and Shape Estimation" in *International Conference on Computer Vision (ICCV) 2021*, pp.867-871.
[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and N. Houlsby "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" in *ICLR 2021*, pp.354-370.
[16] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges "A Spatio-temporal Transformer for 3D Human Motion Prediction" in *International Conference on 3D Vision 2021*.
[17] G. Moon, H. Choi, and K. Mu Lee "Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022*, pp.2308-2317.
[18] N. Carion, F. Massa, and S. Zagoruyko "End-to-End Object Detection with Transformers" in *Computer Vision–ECCV 2020*, pp.213-229.
[19] Y. Li, H. Mao, and K. He "Exploring plain vision transformer backbones for object detection" in *Computer Vision–ECCV 2022*, pp.280-496.
[20] H. Zhang, Y. Tian, and Y. Liu "PyMAF-X:Towards Well-Aligned Full-Body Model Regression From Monocular Images" in *IEEE Transactions on Pattern Analysis and Machine Intelligence 2023*, pp.12287-12303.
[21] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments" in *IEEE Transactions on Pattern Analysis and Machine Intelligence 2014*, pp.1325-1339.
[22] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo "Whole-Body Human Pose Estimation in the Wild" in *Computer Vision - ECCV 2020*, pp.196-214.
[23] N. Kolotouros, G Pavlakos, and K. Daniilidis "Learning to Reconstruct 3D Human Pose and Shape via Model-Fftting in the Loop" in *IEEE/CVF 2019*, pp.2252-2261.
[24] A. Zeng, X. Ju, and Q. Xu "DeciWatch: A Simple Baseline for 10xEfffcient 2D and 3D Pose Estimation" in *ECCV 2022*, pp.607-624.
[25] T. Karras, S. Laine, and T. Aila "A Style-Based Generator Architecture for Generative Adversarial Networks" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019*, pp.4401-4410.
[26] G. Moon, S. Yu, H. Wen, T. Shiratori, and K. Lee "InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image" in *Computer Vision - ECCV 2020*, pp.548-564.
[27] C. Zimmermann, and T. Brox "FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape From Single RGB Images" in *IEEE/CVF International Conference on Computer Vision 2019*, pp.813-822.