2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-4509-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/CVPR46437.2021.00199

End-to-End Human Pose and Mesh Reconstruction with Transformers

Kevin Lin Lijuan Wang Zicheng Liu
Microsoft
{keli, lijuanw, zliu}@microsoft.com

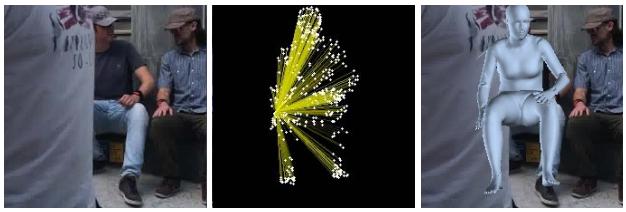
Abstract

We present a new method, called MEsh TRansfOrmer (METRO), to reconstruct 3D human pose and mesh vertices from a single image. Our method uses a transformer encoder to jointly model vertex-vertex and vertex-joint interactions, and outputs 3D joint coordinates and mesh vertices simultaneously. Compared to existing techniques that regress pose and shape parameters, METRO does not rely on any parametric mesh models like SMPL, thus it can be easily extended to other objects such as hands. We further relax the mesh topology and allow the transformer self-attention mechanism to freely attend between any two vertices, making it possible to learn non-local relationships among mesh vertices and joints. With the proposed masked vertex modeling, our method is more robust and effective in handling challenging situations like partial occlusions. METRO generates new state-of-the-art results for human mesh reconstruction on the public Human3.6M and 3DPW datasets. Moreover, we demonstrate the generalizability of METRO to 3D hand reconstruction in the wild, outperforming existing state-of-the-art methods on FreiHAND dataset.

1. Introduction

3D human pose and mesh reconstruction from a single image has attracted a lot of attention because it has many applications including virtual reality, sports motion analysis, neurodegenerative condition diagnosis, etc. It is a challenging problem due to complex articulated motion and occlusions.

Recent work in this area can be roughly divided into two categories. Methods in the first category use a parametric model like SMPL [24] and learn to predict shape and pose coefficients [12, 21, 34, 17, 19, 29, 39, 18]. Great success has been achieved with this approach. The strong prior encoded in the parametric model increases its robustness to environment variations. A drawback of this approach is that the pose and shape spaces are constrained by the limited exemplars that are used to construct the parametric model. To overcome this limitation, methods in the second category do not use any parametric models [20, 7, 27]. These methods either use a graph convolutional neural network to model neighborhood vertex-vertex interactions [20, 7], or use 1D heatmap to regress vertex coordinates [27]. One limitation with these approaches is that they are not efficient in modeling non-local vertex-vertex interactions.



(a) (b) (c)

Figure 1: METRO learns non-local interactions among body joints and mesh vertices for human mesh reconstruction. Given an input image in (a), METRO predicts human mesh by taking non-local interactions into consideration. (b) illustrates the attentions between the occluded wrist joint and the mesh vertices where brighter color indicates stronger attention. (c) is the reconstructed mesh.

Researchers have shown that there are strong correlations between non-local vertices which may belong to different parts of the body (e.g. hand and foot) [50]. In computer graphics and robotics, inverse kinematics techniques [2] have been developed to estimate the internal joint positions of an articulated figure given the position of an end effector such as a hand tip. We believe that learning the correlations among body joints and mesh vertices including both short range and long range ones is valuable for handling challenging poses and occlusions in body shape reconstruction. In this paper, we propose a simple yet effective framework to model global vertex-vertex interactions. The main ingredient of our framework is a transformer.

Recent studies show that transformer [48] significantly improves the performance on various tasks in natural language processing [3, 8, 35, 36]. The success is mainly attributed to the self-attention mechanism of a transformer,

978-1-6654-4509-2/21/$31.00 ©2021 IEEE
DOI 10.1109/CVPR46437.2021.00199

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

which is particularly effective in modeling the dependencies (or interactions) without regard to their distance in both inputs and outputs. Given the dependencies, transformer is able to *soft-search* the relevant tokens and performs prediction based on the important features [3, 48].

In this work, we propose METRO, a multi-layer Transformer encoder with progressive dimensionality reduction, to reconstruct 3D body joints and mesh vertices from a given input image, simultaneously. We design the Masked Vertex Modeling objective with a transformer encoder architecture to enhance the interactions among joints and vertices. As shown in Figure 1, METRO learns to discover both short- and long-range interactions among body joints and mesh vertices, which helps to better reconstruct the 3D human body shape with large pose variations and occlusions.

Experimental results on multiple public datasets demonstrate that METRO is effective in learning vertex-vertex and vertex-joint interactions, and consequently outperforms the prior works on human mesh reconstruction by a large margin. To the best of our knowledge, METRO is the first approach that leverages a transformer encoder architecture to jointly learn 3D human pose and mesh reconstruction from a single input image. Moreover, METRO is a general framework which can be easily applied to predict a different 3D mesh, for example, to reconstruct a 3D hand from an input image.

In summary, we make the following contributions.

- We introduce a new transformer-based method, named METRO, for 3D human pose and mesh reconstruction from a single image.
- We design the Masked Vertex Modeling objective with a multi-layer transformer encoder to model both vertex-vertex and vertex-joint interactions for better reconstruction.
- METRO achieves new state-of-the-art performance on the large-scale benchmark Human3.6M and the challenging 3DPW dataset.
- METRO is a versatile framework that can be easily realized to predict a different type of 3D mesh, such as 3D hand as demonstrated in the experiments. METRO achieves the first place on FreiHAND leaderboard at the time of paper submission.

2. Related Works

Human Mesh Reconstruction (HMR): HMR is a task of reconstructing 3D human body shape, which is an active research topic in recent years. While pioneer works have demonstrated impressive reconstruction using various sensors, such as depth sensors [28, 43] or inertial measurement units [15, 49], researchers are exploring to use a monocular camera setting that is more efficient and convenient. However, HMR from a single image is difficult due to complex pose variations, occlusions, and limited 3D training data.

Prior studies propose to adopt the pre-trained parametric human models, *i.e.*, SMPL [24], STAR [30], MANO [38], and estimate the pose and shape coefficients of the parametric model for HMR. Since it is challenging to regress the pose and shape coefficients directly from an input image, recent works further propose to leverage various human body priors such as human skeletons [21, 34] or segmentation maps [29], and explore different optimization strategies [19, 17, 46, 12] and temporal information [18] to improve reconstruction.

On the other hand, instead of adopting a parametric human model, researchers have also proposed approaches to directly regress 3D human body shape from an input image. For example, researchers have explored to represent human body using a 3D mesh [20, 7], a volumetric space [47], or an occupancy field [41, 42]. Each of the prior works addresses a specific output representation for their target application. Among the literature, the relevant study is GraphCMR [20], which aims to regress 3D mesh vertices using graph convolutional neural networks (GCNNs). Moreover, recent proposed Pose2Mesh [7] is a cascaded model using GCNNs. Pose2Mesh reconstructs human mesh based on the given human pose representations.

While GCNN-based methods [7, 20] are designed to model neighborhood vertex-vertex interactions based on a pre-specified mesh topology, it is less efficient in modeling longer range interactions. In contrast, METRO models global interactions among joints and mesh vertices without being limited by any mesh topology. In addition, our method learns with self-attention mechanism, which is different from prior studies [7, 20].

Attentions and Transformers: Recent studies [31, 23, 48] have shown that attention mechanisms improve the performance on various language tasks. Their key insight is to learn the attentions to *soft-search* relevant inputs that are important for predicting an output [3]. Vaswani *et al*. [48] further propose a transformer architecture based solely on attention mechanisms. Transformer is highly parallelized using multi-head self-attention for efficient training and inference, and leads to superior performance in language modeling at scale, as explored in BERT [8] and GPT [35, 36, 4].

Inspired by the recent success in neural language field, there is a growing interest in exploring the use of transformer architecture for various vision tasks, such as learning the pixel distributions for image generation [6, 32] and classification [6, 9], or to simplify object detection as a set prediction problem [5]. However, 3D human reconstruction has not been explored along this direction.

In this study, we present a multi-layer transformer archi-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

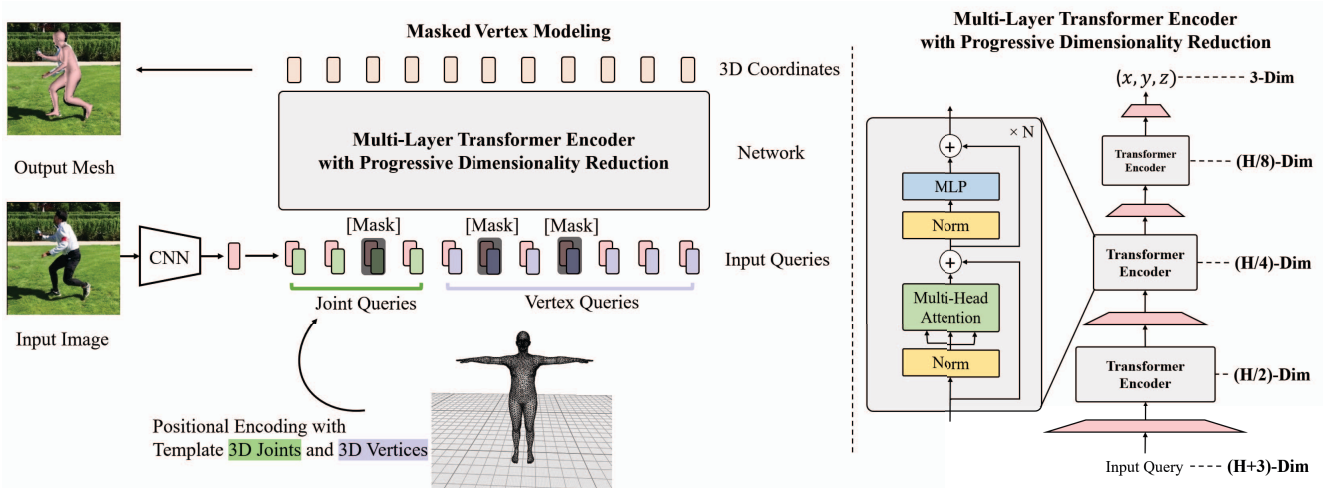


Figure 2: **Overview of the proposed framework.** Given an input image, we extract an image feature vector using a convolutional neural network (CNN). We perform position encoding by adding a template human mesh to the image feature vector by concatenating the image feature with the 3D coordinates (x_i, y_i, z_i) of every body joint i, and 3D coordinates (x_j, y_j, z_j) of every vertex j. Given a set of joint queries and vertex queries, we perform self-attentions through multiple layers of a transformer encoder, and regress the 3D coordinates of body joints and mesh vertices in parallel. We use a progressive dimensionality reduction architecture (right) to gradually reduce the hidden embedding dimensions from layer to layer. Each token in the final layer outputs 3D coordinates of a joint or mesh vertex. Each encoder block has 4 layers and 4 attention heads. H denotes the dimension of an image feature vector.

tecture with progressive dimensionality reduction to regress the 3D coordinates of the joints and vertices.

3. Method

Figure 2 is an overview of our proposed framework. It takes an image of size 224×224 as input, and predicts a set of body joints J and mesh vertices V. The proposed framework consists of two modules: *Convolutional Neural Network*, and *Multi-Layer Transformer Encoder*. First, we use a CNN to extract an image feature vector from an input image. Next, Multi-Layer Transformer Encoder takes as input the feature vector and outputs the 3D coordinates of the body joint and mesh vertex in parallel. We describe each module in details as below.

3.1. Convolutional Neural Network

In the first module of our framework, we employ a Convolutional Neural Network (CNN) for feature extraction. The CNN is pre-trained on ImageNet classification task [40]. Specifically, we extract a feature vector X from the last hidden layer. The extracted feature vector X is typically of dimension 2048. We input the feature vector X to the transformer for the regression task.

With this generic design, it allows an end-to-end training for human pose and mesh reconstruction. Moreover, transformer can easily benefit from large-scale pre-trained CNNs, such as HRNets [51]. In our experiments, we conduct analysis on the input features, and discover that high-resolution image features are beneficial for transformer to regress 3D coordinates of body joints and mesh vertices.

3.2. Multi-Layer Transformer Encoder with Progressive Dimensionality Reduction

Since we need to output 3D coordinates, we cannot directly apply the existing transformer encoder architecture [9, 5] because they use a constant dimensionality of the hidden embeddings for all the transformer layers. Inspired by [14] which performs dimentionality reduction gradually with multiple blocks, we design a new architecture with a progressive dimensionality reduction scheme. As shown in Figure 2 right, we use linear projections to reduce the dimensionality of the hidden embedding after each encoder layer. By adding multiple encoder layers, the model is viewed as performing self-attentions and dimensionality reduction in an alternating manner. The final output vectors of our transformer encoder are the 3D coordinates of the joints and mesh vertices.

As illustrated in Figure 2 left, the input to the transformer encoder are the body joint and mesh vertex queries. In the same spirit as positional encoding [48, 20, 11], we use a template human mesh to preserve the positional information of each query in the input sequence. To be specific, we concatenate the image feature vector $X \in \mathbb{R}^{2048\times 1}$ with

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

the 3D coordinates (x_i, y_i, z_i) of every body joint i. This forms a set of joint queries $Q_J = \{q_1^J, q_2^J, \ldots, q_n^J\}$, where $q_i^J \in \mathbb{R}^{2051\times 1}$. Similarly, we conduct the same positional encoding for every mesh vertex j, and form a set of vertex queries $Q_V = \{q_1^V, q_2^V, \ldots, q_m^V\}$, where $q_j^V \in \mathbb{R}^{2051\times 1}$.

3.3. Masked Vertex Modeling

Prior works [8, 44] use the Masked Language Modeling (MLM) to learn the linguistic properties of a training corpus. However, MLM aims to recover the inputs, which cannot be directly applied to our regression task.

To fully activate the bi-directional attentions in our transformer encoder, we design a Masked Vertex Modeling (MVM) for our regression task. We mask some percentages of the input queries at random. Different from recovering the masked inputs like MLM [8], we instead ask the transformer to regress all the joints and vertices.

In order to predict an output corresponding to a missing query, the model will have to resort to other relevant queries. This is in spirit similar to simulating occlusions where partial body parts are invisible. As a result, MVM enforces transformer to regress 3D coordinates by taking other relevant vertices and joints into consideration, without regard to their distances and mesh topology. This facilitates both short- and long-range interactions among joints and vertices for better human body modeling.

3.4. Training

To train the transformer encoder, we apply loss functions on top of the transformer outputs, and minimize the errors between predictions and ground truths. Given a dataset $D = \{I^i, \bar{V}_{3D}^i, \bar{J}_{3D}^i, \bar{J}_{2D}^i\}_{i=1}^T$, where T is the total number of training images. $I \in \mathbb{R}^{w\times h\times 3}$ denotes an RGB image. $\bar{V}_{3D} \in \mathbb{R}^{M\times 3}$ denotes the ground truth 3D coordinates of the mesh vertices and M is the number of vertices. $\bar{J}_{3D} \in \mathbb{R}^{K\times 3}$ denotes the ground truth 3D coordinates of the body joints and K is the number of joints of a person. Similarly, $\bar{J}_{2D} \in \mathbb{R}^{K\times 2}$ denotes the ground truth 2D coordinates of the body joints.

Let V_{3D} denote the output vertex locations, and J_{3D} is the output joint locations, we use L_1 loss to minimize the errors between predictions and ground truths:

$$\mathcal{L}_V = \frac{1}{M}\sum_{i=1}^{M} \left|\left|V_{3D} - \bar{V}_{3D}\right|\right|_1, \tag{1}$$

$$\mathcal{L}_J = \frac{1}{K}\sum_{i=1}^{K} \left|\left|J_{3D} - \bar{J}_{3D}\right|\right|_1. \tag{2}$$

It is worth noting that, the 3D joints can also be calculated from the predicted mesh. Following the common practice in literature [7, 17, 20, 19], we use a pre-defined regression matrix $G \in \mathbb{R}^{K\times M}$, and obtain the regressed 3D joints by $J_{3D}^{reg} = GV_{3D}$. Similar to prior works, we use L_1 loss to optimize J_{3D}^{reg}:

$$\mathcal{L}_J^{reg} = \frac{1}{K}\sum_{i=1}^{K} \left|\left|J_{3D}^{reg} - \bar{J}_{3D}\right|\right|_1. \tag{3}$$

2D re-projection has been commonly used to enhance the image-mesh alignment [17, 20, 19]. Also, it helps visualize the reconstruction in an image. Inspired by the prior works, we project the 3D joints to 2D space using the estimated camera parameters, and minimize the errors between the 2D projections and 2D ground truths:

$$\mathcal{L}_J^{proj} = \frac{1}{K}\sum_{i=1}^{K} \left|\left|J_{2D} - \bar{J}_{2D}\right|\right|_1, \tag{4}$$

where the camera parameters are learned by using a linear layer on top of the outputs of the transformer encoder.

To perform large-scale training, it is highly desirable to leverage both 2D and 3D training datasets for better generalization. As explored in literature [29, 17, 20, 19, 18, 7, 27], we use a mix-training strategy that leverages different training datasets, with or without the paired image-mesh annotations. Our overall objective is written as:

$$\mathcal{L} = \alpha \times (\mathcal{L}_V + \mathcal{L}_J + \mathcal{L}_J^{reg}) + \beta \times \mathcal{L}_J^{proj}, \tag{5}$$

where α and β are binary flags for each training sample, indicating the availability of 3D and 2D ground truths, respectively.

3.5. Implementation Details

Our method is able to process arbitrary sizes of mesh. However, due to memory constraints of current hardware, our transformer processes a coarse mesh: (1) We use a coarse template mesh (431 vertices) for positional encoding, and transformer outputs a coarse mesh; (2) We use learnable Multi-Layer Perceptrons (MLPs) to upsample the coarse mesh to the original mesh (6890 vertices for SMPL human mesh topology); (3) The transformer and MLPs are trained end-to-end; Please note that the coarse mesh is obtained by sub-sampling twice to 431 vertices with a sampling algorithm [37]. As discussed in the literature [20], the implementation of learning a coarse mesh followed by upsampling is helpful to reduce computation. It also helps avoid redundancy in original mesh (due to spatial locality of vertices), which makes training more efficient.

4. Experimental Results

We first show that our method outperforms the previous state-of-the-art human mesh reconstruction methods on Human3.6M and 3DPW datasets. Then, we provide ablation study and insights for the non-local interactions and model design. Finally, we demonstrate the generalizability of our model on hand reconstruction.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

Method	3DPW			Human3.6M	
	MPVE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
HMR [17]	–	–	81.3	88.0	56.8
GraphCMR [20]	–	–	70.2	–	50.1
SPIN [19]	116.4	–	59.2	–	41.1
Pose2Mesh [7]	–	89.2	58.9	64.9	47.0
I2LMeshNet [27]	–	93.2	57.7	55.7	41.1
VIBE [18]	99.1	82.0	51.9	65.6	41.4
METRO (Ours)	**88.2**	**77.1**	**47.9**	**54.0**	**36.7**

Table 1: Performance comparison with the state-of-the-art methods on 3DPW and Human3.6M datasets.

4.1. Datasets

Following the literature [29, 17, 20, 19, 18, 7, 27], we conduct mix-training using 3D and 2D training data. We describe each dataset below.

Human3.6M [16] is a large-scale dataset with 2D and 3D annotations. Each image has a subject performing a different action. Due to the license issue, the groundtruth 3D meshes are not available. Thus, we use the pseudo 3D meshes provided in [7, 27] for training. The pseudo labels are created by model fitting with SMPLify-X [33]. For evaluation, we use the groundtruth 3D pose labels provided in Human3.6M for fair comparison. Following the common setting [45, 20, 17], we train our models using subjects S1, S5, S6, S7 and S8. We test the models using subjects S9 and S11.

3DPW [49] is an outdoor-image dataset with 2D and 3D annotations. The training set consists of $22K$ images, and the test set has $35K$ images. Following the previous state-of-the-arts [18], we use 3DPW training data when conducting experiments on 3DPW.

UP-3D [21] is an outdoor-image dataset. Their 3D annotations are created by model fitting. The training set has $7K$ images.

MuCo-3DHP [26] is a synthesized dataset based on MPI-INF-3DHP dataset [25]. It composites the training data with a variety of real-world background images. It has $200K$ training images.

COCO [22] is a large-scale dataset with 2D annotations. We also use the pseudo 3D mesh labels provided in [19], which are fitted with SMPLify-X [33].

MPII [1] is an outdoor-image dataset with 2D pose labels. The training set consists of $14K$ images.

FreiHAND [53] is a 3D hand dataset. The training set consists of $130K$ images, and the test set has $4K$ images. We demonstrate the generalizability of our model on this dataset. We use the provided set for training, and conduct evaluation on their online server.

4.2. Evaluation Metrics

We report results using three standard metrics as below. The unit for the three metrics is millimetter (mm).

MPJPE: Mean-Per-Joint-Position-Error (MPJPE) [16] is a metric for evaluating human 3D pose [17, 19, 7]. MPJPE measures the Euclidean distances between the ground truth joints and the predicted joints.

PA-MPJPE: PA-MPJPE, or Reconstruction Error [52], is another metric for this task. It first performs a 3D alignment using Procrustes analysis (PA) [10], and then computes MPJPE. PA-MPJPE is commonly used for evaluating 3D reconstruction [52] as it measures the errors of the reconstructed structure without regard to the scale and rigid pose (*i.e.*, translations and rotations).

MPVE: Mean-Per-Vertex-Error (MPVE) [34] measures the Euclidean distances between the ground truth vertices and the predicted vertices.

4.3. Main Results

We compare METRO with the previous state-of-the-art methods on 3DPW and Human3.6M datasets. Following the literature [18, 19, 17, 20], we conduct mix-training using 3D and 2D training data. The results are shown in Table 1. Our method outperforms prior works on both datasets.

First of all, we are interested in how transformer works for in-the-wild reconstruction of 3DPW. As shown in the left three columns of Table 1, our method outperforms VIBE [18], which was the state-of-the-art method on this dataset. It is worth noting that, VIBE is a video-based approach, whereas our method is an image-based approach.

In addition, we evaluate the performance on the in-door scenario of Human3.6M. We follow the setting in the prior arts [19, 27], and train our model without using 3DPW data. The results are shown in the right two columns of Table 1. Our method achieves better reconstruction performance, especially on PA-MPJPE metric.

The two datasets Human3.6M and 3DPW have different

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

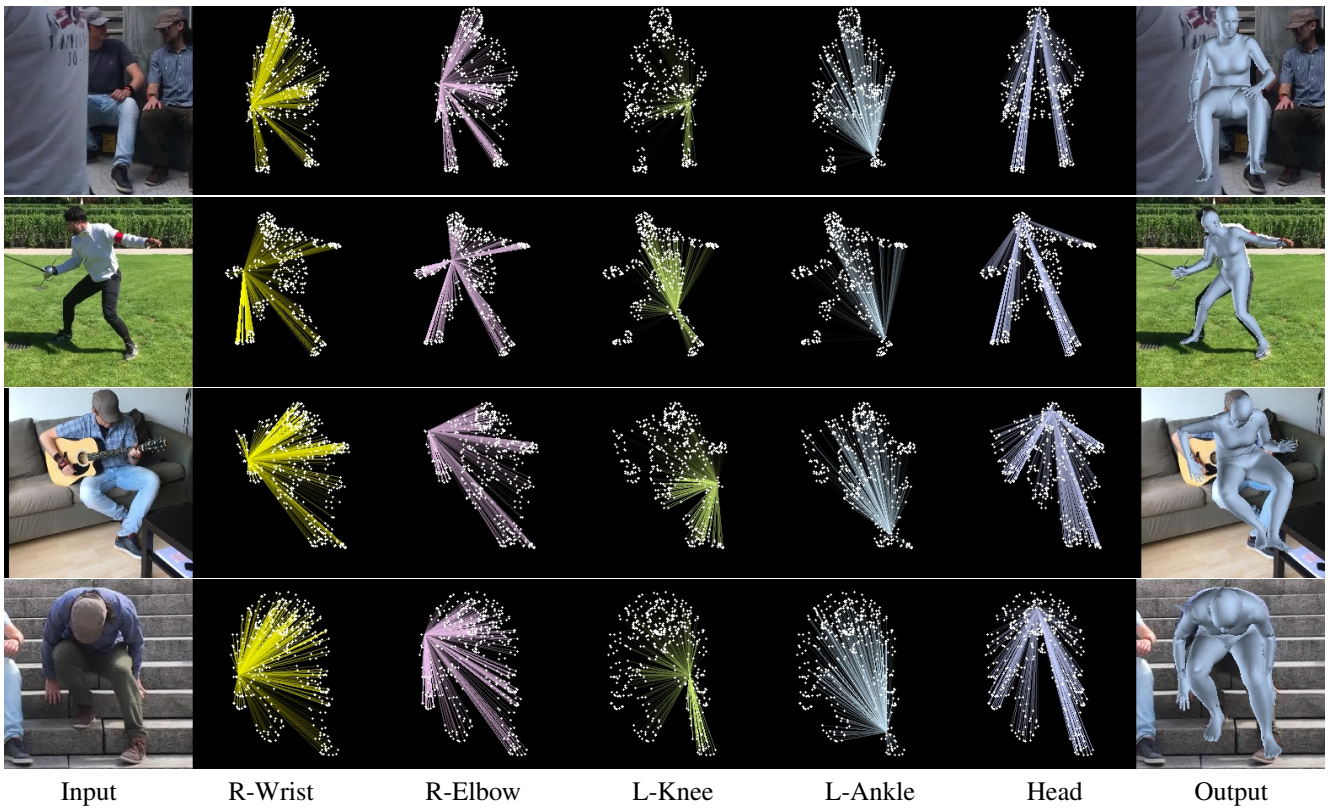


Figure 3: Qualitative results of our method. Given an input image (left), METRO takes non-local interactions among joints and vertices into consideration for human mesh reconstruction (right). We visualize the self-attentions between a specified joint and all other vertices, where brighter color indicates stronger attention. We observe that METRO discovers rich, input-dependent interactions among the joints and vertices.

challenges. The scenes in 3DPW have more severe occlusions. The scenes in Human3.6 are simpler and the challenge is more on how to accurately estimate body shape. The fact that METRO works well on both datasets demonstrates that it is both robust to occlusions and capable of accurate body shape regression.

4.4. Ablation Study

Effectiveness of Masked Vertex Modeling: Since we design a Masked Vertex Modeling objective for transformer, one interesting question is whether the objective is useful. Table 2 shows the ablation study on Human3.6M. We observe that Masked Vertex Modeling significantly improves the results. Moreover, we study how many percentage of query tokens should be masked. We vary the maximum masking percentage, and Table 3 shows the comparison. As we increase the number of masked queries for training, it improves the performance. However, the impact becomes less prominent if we mask more than 30% of input queries. This is because large numbers of missing queries would make the training more difficult.

	MPJPE ↓	PA-MPJPE ↓
w/o MVM	61.0	39.1
w/ MVM	**54.0**	**36.7**

Table 2: Ablation study of the Masked Vertex Modeling (MVM) objective, evaluated on Human3.6M.

Max Percentage	0%	10%	20%	30%	40%	50%
PA-MPJPE	39.1	37.6	37.5	**36.7**	38.2	37.3

Table 3: Ablation study of the Masked Vertex Modeling objective using different percentages of masked queries, evaluated on Human3.6M. The variable $n\%$ indicates we mask randomly from 0% to $n\%$ of input queries.

Non-local Interactions: To further understand the effect of METRO in learning interactions among joints and mesh vertices, we conduct analysis on the self-attentions in our transformer.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

Method	PA-MPVPE ↓	PA-MPJPE ↓	F@5 mm ↑	F@15 mm ↑
Hasson et al [17]	13.2	–	0.436	0.908
Boukhayma et al. [20]	13.0	–	0.435	0.898
FreiHAND [19]	10.7	–	0.529	0.935
Pose2Mesh [7]	7.8	7.7	0.674	0.969
I2LMeshNet [27]	7.6	7.4	0.681	0.973
METRO (Ours)	**6.3**	**6.5**	**0.731**	**0.984**

Table 4: Performance comparison with the state-of-the-art methods, evaluated on FreiHAND online server. METRO outperforms previous state-of-the-art approaches by a large margin.

Figure 3 shows the visualization of the self-attentions and mesh reconstruction. For each row in Figure 3, we show the input image, and the self-attentions between a specified joint and all the mesh vertices. The brighter color indicates stronger attention. At the first row, the subject is severely occluded and the right body parts are invisible. As we predict the location of right wrist, METRO attends to relevant non-local vertices, especially those on the head and left hand. At the bottom row, the subject is heavily bended. For the head position prediction, METRO attends to the feet and hands (6th column at the bottom row). It makes sense intuitively since the hand and foot positions provide strong cues to the body pose and subsequently the head position. Moreover, we observe the model performs self-attentions in condition to the input image. As shown in the second row of Figure 3, when predicting the location of right wrist, METRO focuses more on the right foot which is different from the attentions in the other three rows.

We further conduct quantitative analysis on the non-local interactions. We randomly sample 5000 images from 3DPW test set, and estimate an overall self-attention map. It is the average attention weight of all attention heads at the last transformer layer. We visualize the interactions among 14 body joints and 431 mesh vertices in Figure 4. Each pixel shows the intensity of self-attention, where darker color indicates stronger attention. Note that the first 14 columns are the body joints, and the rest of them represent the mesh vertices. We observe that METRO pays strong attentions to the vertices on the lower arms and the lower legs. This is consistent with the inverse kinematics literature [2] where the interior joints of a linked figure can be estimated from the position of an end effector.

Input Representations: We study the behaviour of our transformer architecture by using different CNN backbones. We use ResNet50 [13] and HRNet [51] variations for this experiment. All backbones are pre-trained on the 1000-class image classification task of ImageNet [40]. For each backbone, we extract a global image feature vector $X \in \mathbb{R}^{2048\times1}$, and feed it into the transformer. In Table 5, we observe our transformer achieves competitive performance

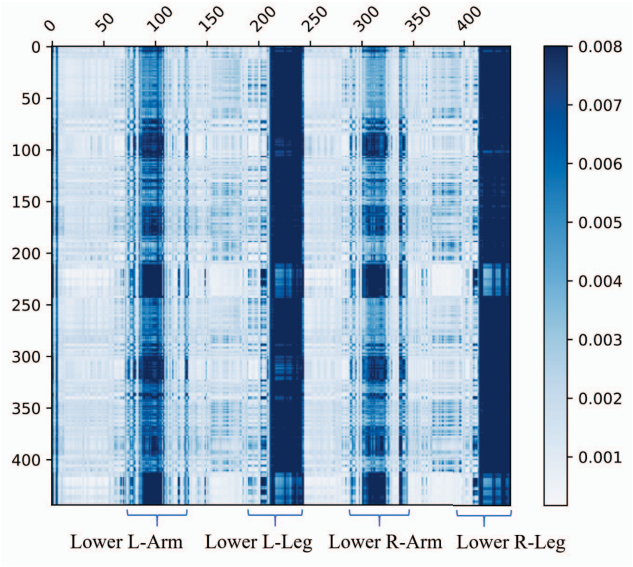


Figure 4: Visualization of self-attentions among body joints and mesh vertices. The x-axis and y-axis correspond to the queries and the predicted outputs, respectively. The first 14 columns from the left correspond to the body joints. The rest of columns correspond to the mesh vertices. Each row shows the attention weight $w_{i,j}$ of the j-th query for the i-th output. Darker color indicates stronger attention.

Backbone	MPJPE ↓	PA-MPJPE ↓
ResNet50	56.5	40.6
HRNet-W40	55.9	38.5
HRNet-W64	**54.0**	**36.7**

Table 5: Analysis on different backbones, evaluated on Human3.6M. All backbones are pre-trained on ImageNet. We observe that increasing the number of filters in the high resolution feature maps of HRNet is beneficial to mesh regression.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

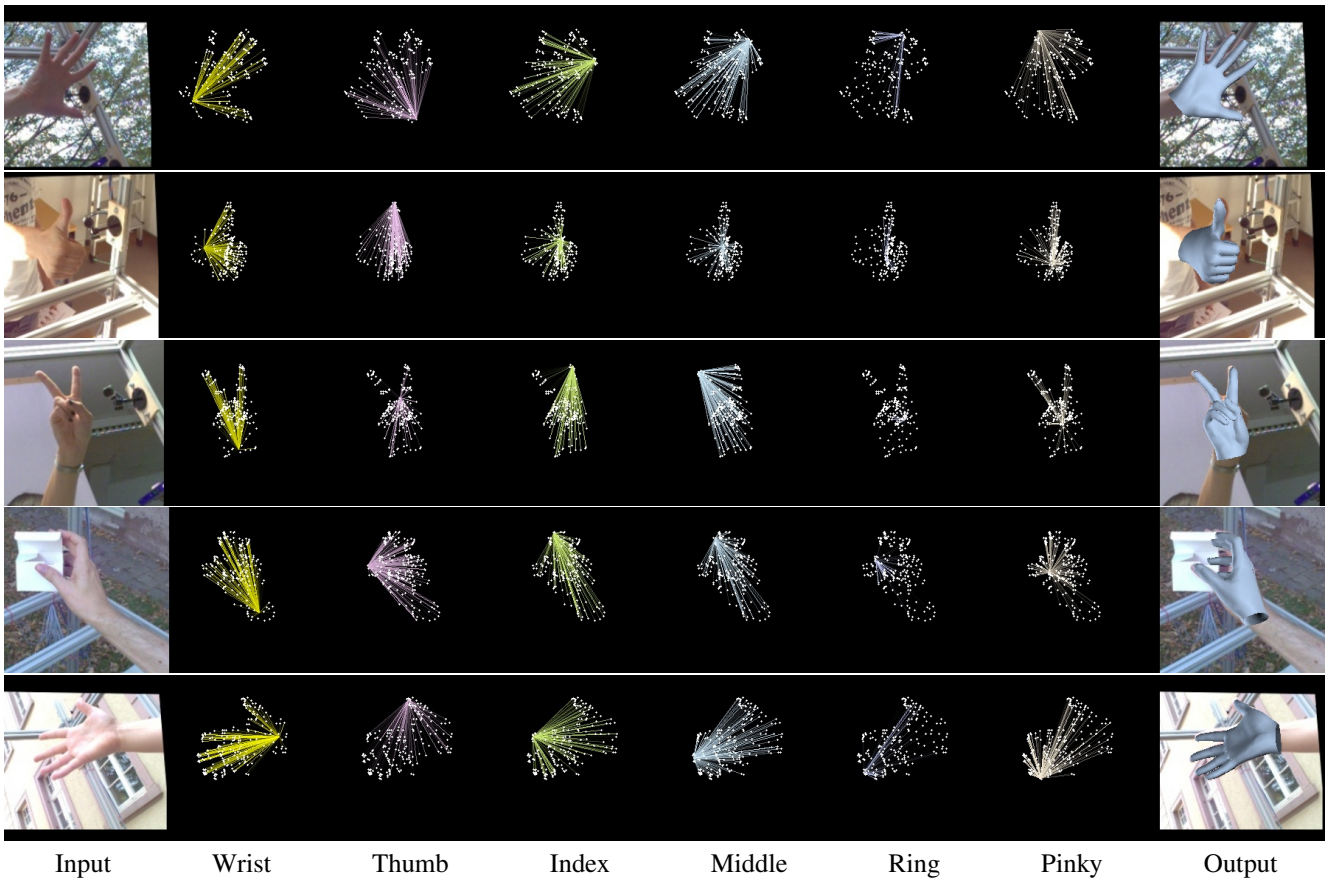


Figure 5: Qualitative results of our method on FreiHAND test set. We visualize the self-attentions between a specified joint and all the mesh vertices, where brighter color indicates stronger attention. METRO is a versatile framework that can be easily extended to 3D hand reconstruction.

when using a ResNet50 backbone. As we increase the channels of the high-resolution feature maps in HRNet, we observe further improvement.

Generalization to 3D Hand in-the-wild: METRO is capable of predicting arbitrary joints and vertices, without the dependencies on adjacency matrix and parametric coefficients. Thus, METRO is highly flexible and general for mesh reconstruction of other objects. To demonstrate this capability, we conduct experiment on FreiHAND [53]. We train our model on FreiHAND from scratch, and evaluate results on FreiHAND online server. Table 4 shows the comparison with the prior works. METRO outperforms previous state-of-the-art methods by a large margin. Without using any external training data, METRO achieved the first place on FreiHAND leaderboard at the time of paper submission[1].

Figure 5 shows our qualitative results with non-local interactions. In the supplementary material, we provide further analysis on the 3D hand joints, and show that the self-attentions learned in METRO are consistent with inverse kinematics [2].

[1]According to the official FreiHAND leaderboard in November 2020: https://competitions.codalab.org/competitions/21238

5. Conclusion

We present a simple yet effective mesh transformer framework to reconstruct human pose and mesh from a single input image. We propose the Masked Vertex Modeling objective to learn non-local interactions among body joints and mesh vertices. Experimental results show that, our method advances the state-of-the-art performance on 3DPW, Human3.6M, and FreiHAND datasets.

A detailed analysis reveals that the performance improvements are mainly attributed to the input-dependent non-local interactions learned in METRO, which enables predictions based on important joints and vertices, regardless of the mesh topology. We further demonstrate the generalization capability of the proposed approach to 3D hand reconstruction.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

References

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

[2] Andreas Aristidou, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir. Inverse kinematics techniques in computer graphics: A survey. In *Computer Graphics Forum*, 2018.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[6] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020.

[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 *ECCV*, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[9] 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.

[10] John C Gower. Generalized procrustes analysis. *Psychometrika*, 1975.

[11] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3d-coded: 3d correspondences by deep deformation. In *ECCV*, 2018.

[12] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*, 2009.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006.

[15] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser learning to reconstruct human pose from sparseinertial measurements in real time. *ACM Transactions on Graphics*, 37(6):185:1–185:15, Nov. 2018.

[16] 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 Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

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

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

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

[20] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

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

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

[23] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *ACL*, 2016.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248, 2015.

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

[26] 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 *3DV*, 2018.

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

[28] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

[29] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018.

[30] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A spare trained articulated human body regressor. In *ECCV*, 2020.

[31] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

[32] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.

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

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

[35] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Technical report*, 2018.

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *Technical report*, 2019.

[37] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *ECCV*, 2018.

[38] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017.

[39] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *ICCV*, 2019.

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[41] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019.

[42] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020.

[43] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *ICCV*, 2019.

[44] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 1953.

[45] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. In *CVPR*, 2016.

[46] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NeurIPS*, 2017.

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

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

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

[50] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[51] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[52] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[53] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, 2019.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:25:43 UTC from IEEE Xplore. Restrictions apply.