TORE: Token Reduction for Efficient Human Mesh Recovery with Transformer

Zhiyang Dou[1†] Qingxuan Wu[2†] Cheng Lin[3‡] Zeyu Cao[4‡] Qiangqiang Wu[5]
Weilin Wan[1] Taku Komura[1] Wenping Wang[6]

[1]The University of Hong Kong [2]University of Oxford [3]Tencent Games
[4]University of Cambridge [5]City University of Hong Kong [6]Texas A&M University

Abstract

In this paper, we introduce a set of simple yet effective TOken REduction (TORE) strategies for Transformer-based Human Mesh Recovery from monocular images. Current SOTA performance is achieved by Transformer-based structures. However, they suffer from high model complexity and computation cost caused by redundant tokens. We propose token reduction strategies based on two important aspects, i.e., the 3D geometry structure and 2D image feature, where we hierarchically recover the mesh geometry with priors from body structure and conduct token clustering to pass fewer but more discriminative image feature tokens to the Transformer. Our method massively reduces the number of tokens involved in high-complexity interactions in the Transformer. This leads to a significantly reduced computational cost while still achieving competitive or even higher accuracy in shape recovery. Extensive experiments across a wide range of benchmarks validate the superior effectiveness of the proposed method. We further demonstrate the generalizability of our method on hand mesh recovery. Visit our project page at `https://frank-zy-dou.github.io/projects/Tore/index.html`.

1. Introduction

Human Mesh Recovery (HMR) has been extensively researched in recent years, given its wide real-world applications [15, 81, 11, 62, 69, 17, 42, 84]. This task becomes more challenging when the input is a monocular 2D image, due to the large pose and shape variation, large appearance variation, partial observation, and self-occlusion.

There has been steady progress in 3D human mesh recovery [30, 50, 36, 24, 60, 39, 40, 8, 5]. Recently, Transformer [67] has shown state-of-the-art (SOTA) results on a wide variety of tasks due to its strong capability of capturing long-range dependency for more accurate predictions [3, 67, 76, 74]. Using tokens constructed from local features extracted by a convolutional neural network (CNN) [20, 70] to query the joint and mesh vertex positions, Transformer-based methods [39, 40, 8] achieved SOTA performance.

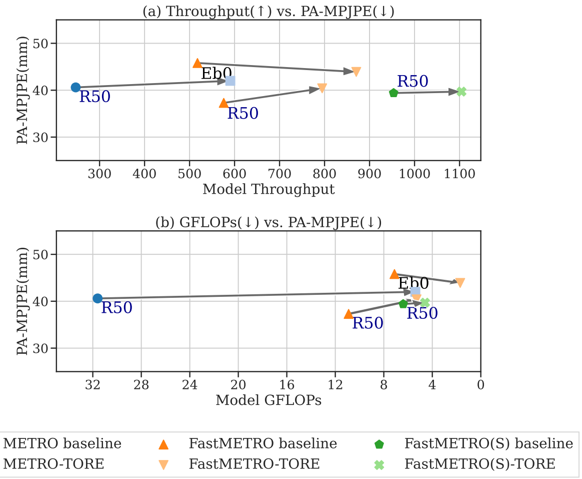


Figure 1. Throughput v.s. Accuracy and GFLOPs v.s. Accuracy on Human3.6M [23]. Our method dramatically saves GFLOPs and improves throughput while maintaining highly competitive accuracy. The x-axis of the bottom GFLOPs figure is reversed for demonstration. Eb0 and R50 represent EfficientNet-b0 [65] and ResNet-50 [20] backbones, respectively.

However, improved performance comes with costs: the increased expressivity of Transformers comes with quadratically increasing computational costs since all pairwise interactions are taken into account [14]. The space and time complexity of a QKV-attention operation is known to be $O(M^2)$, where M is the number of tokens. The token number thus plays a vital role in time efficiency: a large number of tokens inevitably leads to a heavy computation burden. Unfortunately, almost all the existing Transformer-based SOTA methods for HMR [39, 40, 8] are suffering from redundant tokens. This incurs a high model complexity and computational costs, which prevents the current Transformer-based HMR methods from achieving their full potential in real-world applications.

In this paper, we make key observations from two important aspects: the 3D geometry structure and 2D image

†, ‡ denote equal contributions.

feature, to reveal the problem of token redundancy. First, to recover the 3D geometry, all existing methods use both mesh vertices and skeleton joints as tokens for the feature interaction between input and body shape. Whereas a body mesh contains numerous vertices, they can be abstracted by a small number of joints on the skeleton. For instance, when animating a SMPL [44] avatar, the skeleton joints, together with a blend shape binding the joints and corresponding mesh vertices of the local body part, are able to describe various body meshes. Therefore, the joints can already be viewed as an underlying structure of a body shape, which intrinsically encodes the human mesh geometry. Second, for image-based input, most existing methods indiscriminately use all the feature patches to capture pose, shape and appearance variance. However, although the human body exhibits large variance, the important features for shape inference are dominantly clustered within the body area in an RGB image. Most features, e.g., image background, are not informative, thus bringing about redundancy.

Given the aforementioned insights, we argue that the Pareto-front of accuracy and efficiency for Transformer-based HMR could be further improved by reducing the number of tokens [59, 47, 56]. To this end, we introduce a set of simple yet effective token reduction strategies mainly from two aspects corresponding to our observations. First, for 3D mesh recovery, instead of querying both vertices and joints with input features simultaneously, we consider learning a small set of body tokens at the skeleton level for each body part. To recover corresponding mesh vertices, we use an efficient Neural Shape Regressor (NSR) to infer the mesh from the body features encoded by these tokens. This query process can also be interpreted as an attention matrix decomposition, by which we effectively leverage the geometric insights encoded at the skeleton level to infer the mesh structure hierarchically. Second, for the input image feature, we introduce a learnable token pruner to prune the tokens of patch-based features extracted by a CNN. We employ a clustering-based strategy to identify discriminative features, which results in two appealing properties: 1) the end-to-end learning of the pruner is unsupervised, avoiding the need for additional data labeling; 2) it learns semantically consistent features across various images, thus further benefiting the geometry reasoning and enhancing the capability of generalizability. These token reduction strategies substantially reduce the number of query tokens involved in the computation without sacrificing the important information. An overview is shown in Figure 2.

We conduct extensive experiments across wide benchmarks [23, 68, 85], including both the human body and hand mesh recovery, to validate the proposed method. Compared to SOTA methods, our framework faithfully recovers body meshes with fewer tokens, which considerably reduces memory and computation overhead while maintaining competitive geometric accuracy.

In summary, our contribution is three-fold:

- We reveal the issues of token redundancy in the existing Transformer-based methods for HMR.
- We propose effective strategies for token reduction by incorporating the insights from the 3D geometry structure and 2D image feature into the Transformer design.
- Our method achieves SOTA performance on various benchmarks with less computation cost. For instance, for the Transformer Encoder structure [39] and the Transformer Encoder-Decoder structure [8] with ResNet-50 [20] backbone, our method maintains competitive accuracy while saving 82.9%, 50.5% GFLOPs and improving 139.1%, 39.8% throughput, respectively; see Figure 1 for an overview.

2. Related Work

Human Mesh Recovery (HMR) from monocular images has achieved great progress in the past years [4, 30, 50, 36, 10, 40, 24, 60, 5, 39, 8, 82, 71, 58]. Given a monocular RGB image, the goal of HMR is to recover the 3D body shape, typically using a human body model, e.g., SMPL [44]. We refer readers to [66] for a comprehensive review. Existing methods fall into two categories: Non-Transformer-based and Transformer-based approaches.

Non-Transformer-based Human Mesh Recovery To recover the human body shape, one could predict a few parameters, i.e., joint rotation and body shape, to drive the parametric model (parametric approach) or directly regress the vertex positions of the body (non-parametric approach). *Parametric approaches* SMPLify [4] estimates human pose and shape by fitting SMPL to the detected 2D key points [55] by optimization. Kanazawa et al. [24] adopt adversarial prior knowledge of the 3D body shape into the neural network for HMR. SPIN [30] then combines regression-based and optimization-based methods during the training loop. Intermediate features such as key points [54], pixel-to-surface correspondences [77] and texture consistency [52], have been exploited. HybrIK [36] proposes a hybrid inverse kinematics via twist-and-swing decomposition [2]. Regressing the body parameters from the image is a highly non-linear mapping [50], which thus limits the performance of parametric-based approaches.

Non-parametric approaches Non-parametric approaches [50, 32, 9] recover spatial coordinates of a body shape directly from the image features. I2L-MeshNet [50] predicts the per-pixel likelihood on 1D heatmaps for vertex coordinates for a trade-off between accuracy and computational cost. GraphCMR [32] and Pose2Mesh [9] employ graph convolutional neural network (GCNN) [27]

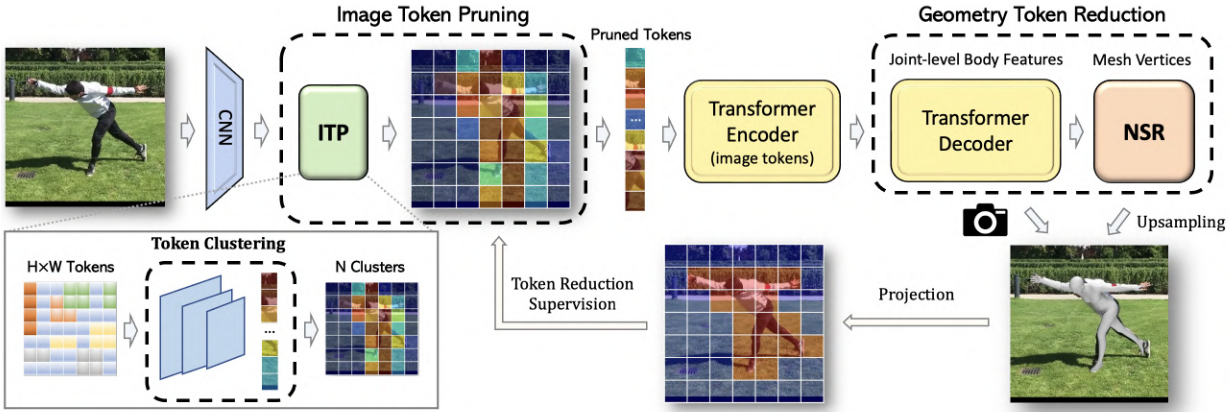


Figure 2. Overview of the proposed framework. Our goal is to reduce tokens for Transformer Encoder and Decoder which are critical modules in the whole pipeline. Image Token Pruner (ITP) and Neural Shape Regressor (NSR) are two lightweight components.

to regress mesh positions by modeling local interactions among vertices from the image [32] and 2D pose [9], respectively. Although steady progress has been made, SOTA accuracy is produced by Transformer-based methods.

Transformer-based Human Mesh Recovery Recently, Transformer structures have been successfully adopted in various vision tasks [12, 43, 6, 51, 66, 80]. In HMR, the SOTA performance is achieved by Transformer-based (attention mechanism) approaches [39, 29, 40, 8]. Specifically, METRO [39] recovers body mesh using a Transformer Encoder [67] to model vertex-vertex and vertex-joint interactions conducting dimension reduction from the image features to 3D shape. MeshGraphormer [40] injects Graph Convolutions [27] into the Transformer Encoder [39] blocks to improve local interactions among vertices. The recent work FastMETRO [8] employs a Transformer Encoder-Decoder architecture disentangling the image encoding and mesh estimation for model reduction and acceleration. However, all Transformer-based approaches [39, 40, 8] suffer from redundant tokens, causing heavy interaction between body geometry and image features, which is still cumbersome and computationally intensive. In this paper, we design strategies driven by the 3D body geometry and 2D image feature for token reduction for both Transformer Encoder and Transformer Encoder-Decoder structures.

Token Reduction for Transformers Token reduction for Transformer structures has been causing people's attention nowadays [56, 47, 59, 78, 16, 38, 46, 7]. In particular, DynamicViT [56] reduces less informative tokens hierarchically with supervision from a teacher network to save computational costs during inference. TokenLearner [59] generates a smaller number of tokens with spatial attention adaptively. Patch-to-Cluster [16] proposes to cluster over image patches with a token selector. Our method follows this way and adopts a token selection layer for clustering over input tokens, reducing tokens and improving generalization capability in challenging scenarios. Token pruning techniques have also been explored in single-view 2D pose estimation and multi-view 3D pose estimation [47] by selecting tokens based on the attention value in the Transformer. Differently, in HMR, besides the key points, body geometry, i.e., mesh vertices, also introduces a burden to the Transformer and needs to be carefully considered.

3. Methodology

We employ the popular body parametric model SMPL [44] to represent a human body. Formally, given a monocular image $\mathcal{I} \in \mathbb{R}^{224\times224}$, our goal is to recover the joint positions $\mathcal{J} \in \mathbb{R}^{J\times3}$ and mesh vertices $\mathcal{V} \in \mathbb{R}^{V\times3}$ where $J = 14, V = 6890$ are the numbers of joints and the mesh vertices, respectively. In addition to the human body, we also demonstrate our method on hand mesh recovery with MANO model [57] to show its generalizability.

Existing Transformer-based HMR methods fall into two categories: a single Transformer Encoder structure [39, 40] and a Transformer Encoder-Decoder structure [8]. Accordingly, we propose two strategies to reduce tokens from 3D and 2D levels, namely, Geometry Token Reduction (GTR) and Image Token Pruning (ITP), which can accommodate different Transformer structures to improve their efficiency.

3.1. Geometry Token Reduction (GTR)

As an essential module, all the existing Transformer-based methods for HMR [39, 40, 8] have to model the interaction between body geometry and image features with an attention mechanism. Therefore, an effective token reduction method to improve the efficiency for inferring mesh geometry can benefit the design of Transformer based methods in a wide scope.

We observe existing methods invariably use both skeleton joints and mesh vertices as tokens for regressing their spatial coordinates. The numerous tokens inevitably lead to significant computational costs for a Transformer. In fact, as an underlying structure of the human body, the skeleton joints already provide strong priors for perceiving the body geometry. A human body avatar, like SMPL [44], can be driven by a small number of joints with the blend shape. This inspires us to regress the human body mesh in a hierarchical manner.

Our key idea is to decompose a heavy Transformer into smaller Transformer modules, each of which involves fewer tokens, avoiding expensive computation for image feature interaction. Specifically, our approach only queries the body tokens, whose count is equivalent to the number of joints. Since the number of joints is much less than the number of mesh vertices, we can significantly reduce the token number in complex 2D-3D feature interactions and this leads to a substantial enhancement in efficiency.

To faithfully recover body shape from the body tokens, we introduce a *Neural Shape Regressor (NSR)* module, which is a light Transformer structure, to query the vertex tokens from their interaction with body features. Let $F_J = \{f^j_1, f^j_2, ..., f^j_J\}$ denotes a set of learned body features and $T_\mathcal{V} = \{t^v_1, t^v_2, ..., t^v_V\}$ where t^v_i represents a set of vertex tokens which are the query tokens. We conduct cross-attention between query vertex tokens and body features to model the interaction between vertices and learned body features. Note that non-adjacent vertices are masked to improve the efficiency [8].

The feature size within NSR is even smaller compared with the main Transformer. Thanks to the informative geometric feature encoded in the body tokens, despite not incorporating image tokens, NSR still faithfully recovers the surface vertices. Another interesting discovery is that the learned attention scores reflect the correlation between joints and vertices at a body part level, similar to body blending weights (see Sec. 4.7.3). In this way, we can effectively reduce the cost of redundant interaction by decomposing the entire body into encoded constituent parts and efficiently recovering the whole body shape.

3.2. Image Token Pruning (ITP)

In addition to the geometric insights, for a Transformer Encoder-Decoder structure, image tokens also affect the computation overhead. However, the existing methods adopting this Transformer structure fail to avoid the token redundancy issue [8] because all available image feature patches are indiscriminately involved in human mesh recovery. Actually, some features, e.g., image background, are not informative, thus introducing redundancy and increasing the computational cost of the Transformer.

To tackle this issue, we introduce an effective token pruning strategy, namely, Image Token Pruning (ITP). Our key insight is that the informative features in an image for inferring 3D geometry are overwhelmingly clustered within the region of the human body. Inspired by recent advances in token pruning [16, 59, 56, 47], we propose to aggregate features into a small number of meaningful clusters.

Let $0 < \rho < 1$ denote the predefined token pruning ratio. Our goal is to learn a projection that maps feature map $F_\mathcal{I} \in R^{HW \times c}$ extracted from the given image with HW tokens to $Z_\mathcal{I} \in R^{T \times c}$ with T tokens, where $T = \lfloor \rho HW \rfloor$ and c is the feature dimension. The small number of clusters $Z_\mathcal{I}$ are expected to capture the fewer but more discriminative features in $F_\mathcal{I}$. Inspired by [16, 59], we implement the projection as a learnable CNN module. We first apply Conv2D to the input feature, where the kernel size, stride and zero padding of Conv2D are 3×3, 1 and 1, reducing the feature dimension from c to $c' = c/4$ with GELU [21] activation function applied. We further map the dimension c to N by MLP and Softmax, producing $M \in R^{HW \times N}$. Finally, the clustered token set is given by $Z_\mathcal{I} = M^T \cdot F_\mathcal{I}$. The mapping matrix M^T, in essence, produces a clustering over origin tokens. Each element $m_{ij} \in M^T$ depicts the contribution of j token to i-th cluster. Note that LayerNorm is employed after the clustering.

Token Reduction Supervision In order to encourage ITP to pay more attention to those discriminative feature regions, i.e., body parts. during token reduction, we introduce Token Reduction Supervision. Given the weak-perspective camera estimated the neural network, we compute the 2D projection of the ground truth 3D vertices $\hat{\mathcal{V}}$ as $s\Pi(\hat{\mathcal{V}}) + t$ where Π represents an orthographic projection, s and t are scale and translation estimated by the network. The projected results are downsampled to a discrete $H \times W$ grid corresponding to each token followed by a binary indicator function

$$F_d(x) = \begin{cases} 1, \text{if the cell contains a projected point.} \\ 0, \text{if the cell does not contain a projected point.} \end{cases} \tag{1}$$

Then the supervision for token pruning is given by

$$L_P = -\frac{1}{NHW} \sum_i^N (F_d(s\Pi(\hat{\mathcal{V}}) + t) \cdot M_{[:,i]}). \tag{2}$$

With this supervision, the weights of background tokens in $M[:, i]$ for each i-th cluster will be penalized during training, which thus encourages the ITP to learn the discriminative features.

The proposed ITP has several noticeable advantages. First, in contrast to other methods relying on explicit supervision, e.g., using a pretrained teacher network to facilitate the pruning [56, 78, 33], our token pruner is trained in an unsupervised manner; Note that no ground truth masking is required for Token Reduction Supervision. Second, unlike previous clustering-based pruning methods [16] that use pre-defined class labels, ITP aims to adaptively identify discriminative features within the body region. Therefore, ITP is not limited to fixed semantic labels but is able to induce higher-level semantics according to the learning target.

Finally, we find the clustered informative features learned by ITP improve the model generalizability, especially for challenging datasets, e.g., in-the-wild dataset 3DPW [68] as shown in Table 3. We elaborate on these properties in Sec. 4.3.2 and Sec. 4.7.1 with extensive experiments.

3.3. Loss Functions

We supervise the network using L1 distance between predicted mesh vertices at three sampling levels: $L_{\mathcal{V}3D} = ||\mathcal{V}^l_{3D} - \hat{\mathcal{V}}^l||_1 + ||\mathcal{V}^m_{3D} - \hat{\mathcal{V}}^m||_1 + ||\mathcal{V}^h_{3D} - \hat{\mathcal{V}}_h||_1$, where $*^l, *^m, *^h$ denote low, middle and high body mesh resolution with vertex numbers to be 431, 1723 and 6890 for an SMPL body. The ground-truth 3D joints are used for supervising predicted 3D joints and the ones regressed from the vertices $\mathcal{V}^h_{3D}$ with a SMPL regression matrix $\mathcal{M}$: $L_{\mathcal{J}3D} = ||\mathcal{J}_{3D} - \hat{\mathcal{J}}_{3D}||_1, L^R_{\mathcal{J}3D} = ||\mathcal{M}(\mathcal{V}) - \hat{\mathcal{J}_{3D}}||_1$. A 2D projection loss L_{J2D} is used during the network training, where we employ a weak perspective camera model to project 3D joints to 2D for supervision: $L^R_{\mathcal{J}2D} = ||(s\Pi(\mathcal{M}(\mathcal{V})) + t) - \hat{\mathcal{J}_{2D}}||_1,$. s, t are scale and translation estimated by the network. Overall, together with the Token Pruning Supervision L_P, the total loss is:

$$L = \alpha \left[\lambda_{\mathcal{J}3D}(L^R_{\mathcal{J}3D} + L_{\mathcal{J}3D}) + \lambda_{\mathcal{V}3D}(L_{\mathcal{V}3D}) + \lambda_P L_P\right] \\ + \beta\lambda_{\mathcal{J}2D}L^R_{\mathcal{J}2D},$$

where α and β indicate the availability of the supervision. We set $\lambda_P, \lambda_{J2D}, \lambda_{V3D}, \lambda_{J3D}$ to be $1, 100, 100, 1000$.

4. Experimental Results

4.1. Datasets and Metrics

We evaluate our model in two scenarios: human body mesh recovery and hand mesh recovery. We adopt commonly-used metrics [25, 31, 39, 40, 8]: Mean Per-Joint Position Error (MPJPE), MPJPE after further alignment, i.e., Procrustes Analysis (PAMPJPE) and Mean Per-Vertex Error (MPVE). For human body mesh recovery, our network is trained with Human3.6M [23], MuCo-3DHP[48], UP-3D [34], COCO [41], MPII [1]. Following previous works [39, 40, 8], we use the pseudo mesh data in Human3.6M [23] for training, splitting subjects S1, S5, S6, S7, S8 for training and S9, S11 for testing. We also report our performance on 3DPW [68], a more challenging in-the-wild dataset. To further evaluate the generalization ability of the proposed method, we test our method on hand mesh recovery on FreiHAND [85] dataset. For analyzing the efficiency of our method, we report GFLOPs and throughput (image per second), strictly following [56, 49, 59, 47, 38].

Table 1. Comparison with the Transformer Encoder structure METRO (M) [39] on Human3.6M [23]. We test with ResNet-50 (R50) [20] and HRNet-W64 (H64) [70] as backbones. GFLOPs$^{\mathrm{T}}$ is GFLOPs of the transformer.

Method	GFLOPs ↓	GFLOPs$^{\mathrm{T}}$ ↓	Throughput (im/s) ↑	PAMPJPE ↓
M-H64 [39]	56.5	27.5	141.0	**36.7**
M-H64+GTR	**30.2 (-46.5%)**	**0.8 (-97.1%)**	**210.1 (+49.0%)**	37.1 (+1.1%)
M-R50 [39]	31.6	27.5	247.0	**40.6**
M-R50+GTR	**5.4 (-82.9%)**	**0.8 (-97.1%)**	**590.6 (+139.1%)**	42.0 (+3.4%)

4.2. Implementation Details

Network Training We implement the network using PyTorch. For the Transformer Encoder-Decoder structure, we set the learning rate to be 1×10^{-4}. We use the AdamW optimizer [45] and train for 60 epochs, with a batch size of 16 per GPU on 4 Nvidia A100 GPUs. When comparing with Transformer Encoder structure METRO [39] (see Table 1), we follow the setting of METRO where we train the models with a batch size of 30 per GPU on 8 Nvidia A100 GPUs in total. We adopt Adam optimizer [26] and train the models for 200 epochs. See more details in Appendix A.

Performance Evaluation To analyze the performance on a consumer-level GPU device, we measure the throughput on an NVIDIA RTX 3090 GPU with 24G VRAM. When comparing with encoder-decoder Transformer structure FastMETRO [8], we set the batch size to 32 following PPT [47]. The comparison with Transformer Encoder structure METRO [39] uses 16 as the batch size. Note that the batch size is limited by the size of the original METRO [39] model. To factor out the influence of the batch size, we provide a more detailed performance report in Appendix B.

4.3. Performance on Human Mesh Recovery

Currently, there only exist two types of representative Transformer structures for HMR: Encoder-only structure (METRO [39]) and Encoder-Decoder structure (FastMETRO [8]). To demonstrate the effectiveness of our method to both Transformer structures, we first conduct experiments by adding GTR to two structures: METRO and FastMETRO. Since ITP is designed for the encoder-decoder structure, we further report the results on FastMETRO with both ITP and GTR in Sec. 4.3.2. The overall comparison is shown in Table 5.

4.3.1 Transformer Encoder Structure

As shown in Table 1, GTR effectively reduces the computation costs while still producing competitive accuracy results. For the encoder-based Transformer model METRO [39] with HRNet-W64 [70] as a backbone, applying GTR saves the GFLOPs of the whole model for 46.5% and the Transformer part for 97.1%, with throughput im-

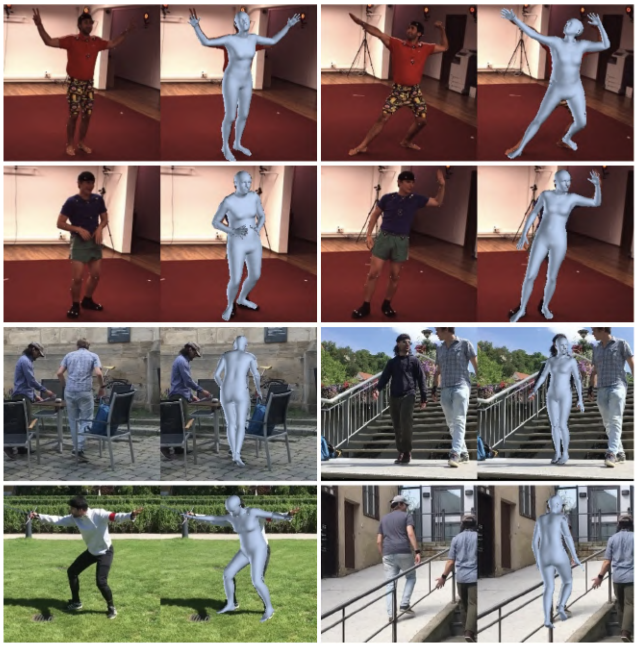


Figure 3. Qualitative results of GTR equipped Encoder-Decoder structure [39] (H64) on Human3.6M [23] and 3DPW [68].

Table 2. Comparison with the Transformer Encoder-Decoder structure FastMETRO (FM) [8] on Human3.6M [23]. We test with EfficientNet-b0 (Eb0) [65], ResNet-50 (R50) [20] and HRNet-W64 (H64) [70] as backbones. GFLOPs^T stand for GFLOPs for the transformer.

Method	GFLOPs ↓	GFLOPs^T ↓	Throughput (im/s) ↑	PAMPJPE ↓
FM-H64 [8]	35.7	6.6	221.5	**33.7**
FM-H64+GTR	**30.2 (-15.4%)**	**0.7 (-89.4%)**	**249.2 (+12.5%)**	34.8 (+3.2%)
FM-R50 [8]	10.9	6.6	576.0	**37.3**
FM-R50+GTR	**5.4 (-50.5%)**	**0.7 (-89.4%)**	**805.3 (+39.8%)**	38.6 (+3.4%)
FM(S)-R50 [8]	6.4	2.2	953.5	39.4
FM(S)-R50+GTR	**4.6 (-28.1%)**	**0.3 (-86.4%)**	**1128.9 (+18.4%)**	**38.6 (-2.0%)**
FM-Eb0	7.1	6.6	517.6	45.8
FM-Eb0+GTR	**1.7 (-76.1%)**	**0.7 (-89.4%)**	**870.5 (+68.2%)**	**44.2 (-3.5%)**

proved by 49%. For the ResNet-50 [20] backbone, GTR helps save the GFLOPs of the whole model for 82.9% and the Transformer part for 97.1%, with throughput improved by 139.1%. These results validate the effectiveness of the proposed GTR on the Transformer Encoder structure. Qualitative results of METRO-H64+GTR can be found in Figure 3, in which the model produces high-quality results in human mesh recovery over various input monocular images.

4.3.2 Transformer Encoder-Decoder Structure

We experiment on the Transformer Encoder-Decoder structure FastMETRO [8] with its two variants: models with 1 and 3 Encoder-Decoder layers. For 1-layer FastMETRO, we denote it as FastMETRO(S). In Table 2, GTR equipped FastMETRO [39] produces competitive accuracy while reducing 15.4% in GFLOPs for the whole model, 89.4% in GFLOPs for the Transformer part and improving 12.5% in throughput with an HRNet-W64 [70] CNN backbone. When testing with ResNet-50 [20], GTR helps save 50.5% in GFLOPs for the whole model, 89.4% in GFLOPs for the Transformer part and improves 39.8% in throughput. Notably, our FastMETRO(S)-R50+GTR and FastMETRO(S)-Eb0+GTR yield 38.6, 44.2 mm PAMPJPE respectively, surpassing the corresponding baselines while greatly saving computational costs.

We then present the performance of the Transformer Encoder-Decoder structure with GTR and ITP using ResNet-50 [20] and EfficientNet-b0 [65] CNN backbones.

Table 3. Influence of ITP for monocular 3D human mesh recovery on 3DPW [68].

Method	MPVE	MPJPE	PAMPJPE
FastMETRO-H64+GTR	91.3	75.4	46.7
FastMETRO-H64+GTR+ITP@20%	**88.2**	**72.3**	**44.4**

As shown in Table 4, TORE (GTR+ITP) is effective for the Encoder-Decoder Transformer structure. Specifically, when pruning with 20%, 50% in ITP, the models save the GLOPs of the Transformer structure by 14.3%, 14.3% and 28.5%, 42.9% while receiving competitive accuracy or even higher accuracy i.e., FM-Eb0+GTR+ITP@20% improves PAMPJPE compared with the baseline.

Table 5. Comparison with the SOTA methods for monocular 3D human mesh recovery on 3DPW [68] and Human3.6M [23].

Method	3DPW			Human3.6M	
	MPVE	MPJPE	PAMPJPE	MPJPE	PAMPJPE
HMR [25]	–	130.0	76.7	88.0	56.8
GraphCMR [32]	–	–	70.2	–	50.1
SPIN [30]	116.4	96.9	59.2	62.5	41.1
I2LMeshNet [50]	–	93.2	57.7	55.7	41.1
PyMAF [83]	110.1	92.8	58.9	57.7	40.5
ROMP–R50 [64]	105.6	89.3	53.5	–	–
PARE–R50 [29]	99.7	82.9	52.3	–	–
DSR–R50 [13]	99.5	85.7	51.7	60.9	40.3
METRO–R50 [39]	–	–	–	56.5	40.6
METRO–H64 [39]	88.2	77.1	47.9	54.0	36.7
METRO–H64+GTR	87.9	75.5	46.6	57.6	37.1
FastMETRO-R50 [8]	90.6	77.9	48.3	53.9	37.3
FastMETRO-R50+GTR+ITP@20%	99.2	82.4	52.3	59.8	40.5
FastMETRO-H64 [8]	84.1	73.5	44.6	52.2	33.7
FastMETRO-H64+GTR+ITP@20%	88.2	72.3	44.4	59.6	36.4
FastMETRO-Eb0 [8]	112.5	93.8	60.2	69.2	45.8
FastMETRO-Eb0+GTR+ITP@20%	112.5	93.7	60.1	63.2	43.9

In addition, we find that ITP helps improve the accuracy on the challenging in-the-wild 3DPW [68] dataset, shown in Table 3. Specifically, when further equipped with ITP, the model performance in MPVE, MPJPE and PAMPJPE are improved by 3.1mm, 3.1mm and 2.3 mm, respectively.

Table 4. Statistics of all proposed components (GTR + ITP) for Encoder-Decoder structure FastMETRO (FM) [8] on Human3.6M [23]. The backbones are EfficientNet-b0 (Eb0) [65] and ResNet-50 (R50) [20]. GFLOPs^T stands for GFLOPs of the Transformer.

Method + Pruning Rate	#Tokens	GFLOPs ↓	GFLOPs^T ↓	Throughput (im/s) ↑	PAMPJPE ↓
FM-R50+GTR	49	5.4	0.7	805.3	**38.6**
FM-R50+GTR+ITP@20%	39	**5.3 (-1.8%)**	**0.6 (-14.3%)**	794.7 (−1.3%)	40.5 (+4.9%)
FM-R50+GTR+ITP@50%	24	**5.1 (-5.5%)**	**0.5 (-28.5%)**	806.1 (+0.1%)	40.7 (+5.4%)
FM-Eb0+GTR	49	1.7	0.7	870.5	44.2
FM-Eb0+GTR+ITP@20%	39	**1.6 (-5.9%)**	**0.6 (-14.3%)**	876.3 (+0.7%)	**43.9 (-0.%)**
FM-Eb0+GTR+ITP@50%	24	**1.4 (-17.6%)**	**0.4 (-42.9%)**	870.4 (−0.1%)	**44.0 (-0.5%)**

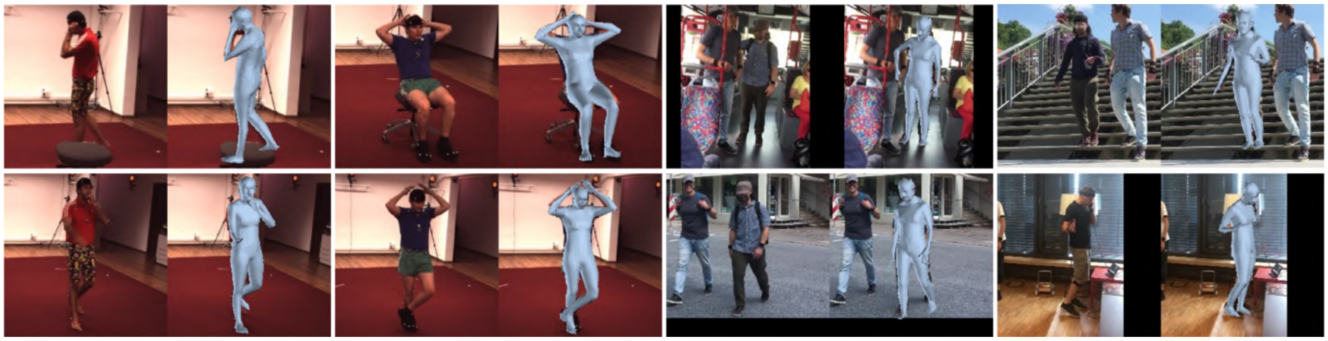
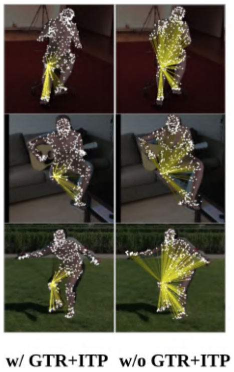


Figure 4. Qualitative results of FastMETRO+GTP+ITP@20% on Human3.6M [23] and 3DPW [68].

These results indicate that the ITP module learns more discriminative features and thus enhances the capability of generalization, allowing methods with ITP to achieve better performance in more challenging in-the-wild scenarios. The strong representation capability of ITP can also be seen in the clustering process, which promotes competitive accuracy with fewer tokens while producing semantically meaningful clustering results, as demonstrated in Sec. 4.7.1.

We conduct comparisons with existing SOTA approaches (both Transformer and Non-Transformer based) on monocular human mesh recovery. The results are summarized in Table 5. It is notable that, with effective token reduction techniques, our method produces competitive or higher accuracy on Human3.6M [23] and 3DPW [68] datasets. For instance, on 3DPW, our METRO-H64+GTR achieves 87.9mm MPVE, 75.5mm MPJPE and 46.6mm PAMPJPE surpassing METRO–H64. In general, token reduction results in information loss, which leads to a slight drop in accuracy. For MPVE, using an HRNet-W64 backbone, GTR typically causes a larger error increase of 2.6 mm (from 84.1 to 86.7), while ITP is 1.5 mm (from 86.7 to 88.2). This suggests GTR sacrifices more accuracy for efficiency, while ITP has less impact. Qualitative results of FastMETRO+GTR+ITP@20% are visualized in Figure 4, where our method produces accurate and robust human mesh recovery from monocular images. In the inset, we also provide a qualitative comparison of FastMETRO+GTR+ITP@20% (w/ GTR+ITP) and the baseline model FastMETRO (w/o GTR+ITP), where the joint-vertex attention is similar to the blending weights in SMPL, which properly captures the shape structure. However, the model w/o GTR+ITP redundantly correlates local joints with distant vertices, leading to additional interaction costs. In summary, extensive experiments validate the effectiveness of proposed strategies for token reduction across different Transformer structures (Encoder-based METRO [39] and Encoder-Decoder-based FastMETRO [8]) and different Transformer model sizes (FastMETRO and FastMETRO(S)).



4.4. ITP v.s. TokenLearner [59]

Table 6. Comparison with TokenLearning [59] in ITP for token reduction on 3DPW [68] and Human3.6M [23].

Method	GFLOPs	3DPW			Human3.6M	
		MPVE	MPJPE	PAMPJPE	MPJPE	PAMPJPE
TokenLearner [59]	5.7	99.3	82.4	52.6	61.2	45.4
Image Token Pruning	**5.3**	**99.2**	**82.4**	**52.3**	**59.8**	**40.5**

We compare ITP with another popular pruning strategy Tokenlearner [59] for HMR on Human3.6M and 3DPW in Table 6. We use Encoder-Decoder structure [8] with ResNet-50 [20] as a backbone. The pruning rate is 20%. In

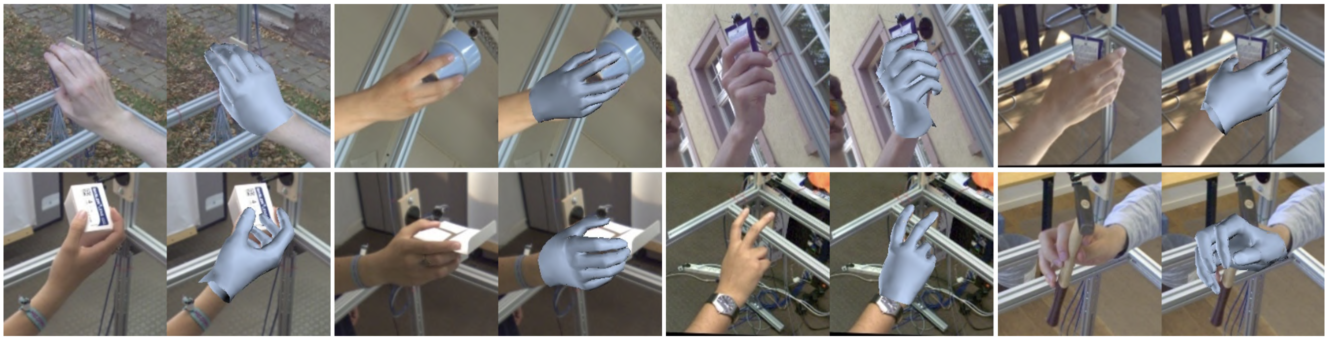


Figure 5. Qualitative results on FreiHAND [85] by FastMETRO+H64+GTR+ITP@20% model.

Table 8, ITP saves more computational costs while achieving the highest accuracy in terms of both human mesh recovery and joint estimation.

4.5. NSR v.s. GCN [27] and MLP

We discuss the effectiveness of NSR in GTR. We compare NSR with other implementations, including Multi-Layer Perceptron and Graph Convolutional Network [27] following Pose2Mesh [9]. We condition the backbone on ResNet-50 [20] using Encoder-Decoder structure [8]. In Table 7, NSR achieves higher performance on Human3.6M and 3DPW. This indicates that the attention mechanism in NSR provides a stronger modeling capability of vertices given the learned body features, which thus improves the quality of recovered mesh vertices.

Table 7. Comparison of different network structures of NSR for GTR on 3DPW [68] and Human3.6M [23].

Model	3DPW			Human3.6M	
	MPVE	MPJPE	PAMPJPE	MPJPE	PAMPJPE
Multi-Layer Perceptron	99.0	80.4	49.6	57.7	38.9
Graph Convolutional Network	98.8	81.5	49.8	58.2	38.8
Neural Shape Regressor	**95.9**	**79.2**	**49.2**	**57.2**	**38.6**

4.6. Generalization on Hand Mesh Recovery

To investigate the generalizability of our framework, we conduct an experiment on monocular hand mesh recovery, which is summarized in Table 8. Following FastMETRO [8], we report PAMPJPE, F-score@15mm (F@15mm) on FreiHand [85] together with GFLOPs^{T} and throughput. The qualitative results are provided in Figure 5. The hand vertex-joint interactions are visualized in Appendix C.

4.7. Further Discussion

4.7.1 Analysis of Image Token Pruning

Image Token Pruning, in essence, achieves body-aware clustering results encoding discriminative body features.

Table 8. Comparison with the SOTA methods on hand mesh recovery on FreiHAND [85].

Method	GFLOPs^{T} ↓	Throughput ↑	PAMPJPE ↓	F@15mm ↑
METRO-H64 [39]	13.1	186.7	6.8	0.981
FastMETRO-H64 [8]	3.3	228.4	**6.5**	**0.982**
FastMETRO-H64+GTR+ITP@20%	**1.0**	**260.2**	6.7	0.980

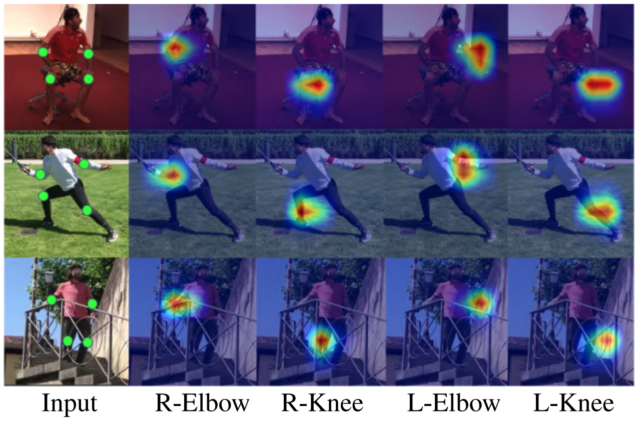


Figure 6. Visualization of learned semantics by ITP. Note that the clustering results are consistent across different identities, i.e, different clusters correspond to different body joints.

We visualize the heatmap of the predicted clustering scores in Figure 6, where the model is with ITP at 50% pruning rate (resulting in 24 image tokens) and GTR. The backbone is ResNet-50. As shown in Figure 6, clustering-based pruning maps original tokens to a fewer number of clusters with semantics corresponding to the body joints. Note that the semantics is consistent across different identities.

4.7.2 ITP with Token Reduction Supervision

In Figure 7, we visualize the scores predicted by ITP in HMR. The mask supervision is generated by projecting mesh vertices to image patches using estimated camera parameters as shown in Figure 7 (b)(c). ITP effectively learns to pay attention to a human body region in the given image. Quantitatively, the projection supervision improves PAM-

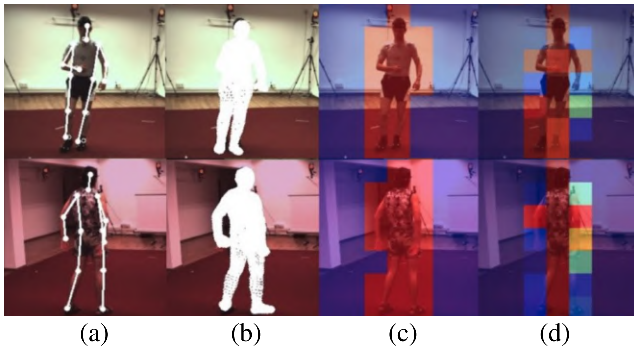


Figure 7. Visualization of learned semantics by Image Token Pruner. (a) projected joints. (b) projected mesh vertices. (c) mask supervision. (d) scores predicted by ITP.

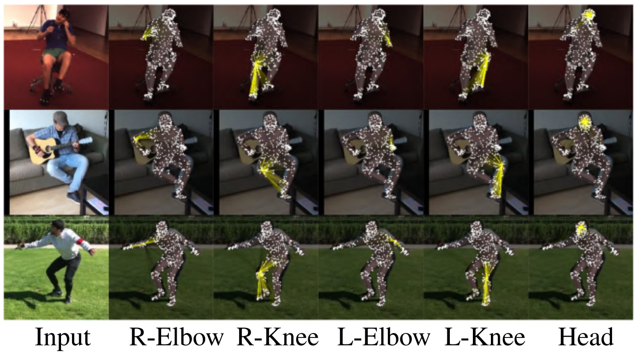


Figure 8. Visualization of cross-attention between joint and vertices. Samples are from Human3.6M and 3DPW.

PJPE from 41.6 to 40.6 on Human3.6M.

4.7.3 Vertex-Body Feature Interactions

We show the interactions between vertices and joints modeled by cross attention within Neural Shape Regressor; See Figure 8, where the heatmap is obtained by averaging attention scores across all heads of the multi-head cross-attention between query vertices and body features. As shown in Figure 8, the interactions of mesh vertices and joints are at the body part level, e.g., elbow, knee, head, which are similar to the way of the blending of a human body model, i.e., SMPL [44] and thus validates our claims.

4.8. Limitations and Future Work

When an extremely high pruning rate is applied to ITP, the accuracy of the model drops dramatically, e.g, when the token number is pruned to be one, the accuracy of the model drops dramatically (12.1%) from 38.6 to 43.3 PAM-PJPE. For GTR, since we regress joints and vertices progressively, the quality of the recovered vertices by NSR depends on the learned body features. More failure cases are in Appendix D. In future work, one of the promising directions could be applying the shown enhanced efficiency of HMR from monocular images to methods exhibiting high complexity for improving the model efficiency, especially in tasks such as human-environment/object interaction that perceives environments [37, 22, 18, 75, 73, 19, 61], as well as HMR from videos that involve temporal information [28, 53, 35, 63, 79, 72].

5. Conclusion

In this paper, we investigate the issues of token redundancy in the existing Transformer-based methods for both body and hand mesh recovery tasks. To tackle the problem, we introduce two effective token reduction strategies for Transformers by incorporating insights from both 3D geometry structure and 2D image features. Specifically, we recover the body shape in a hierarchical manner and cluster for image features to feed fewer but more discriminative tokens to the Transformer. Our method dramatically reduces the high-complexity interactions in the Transformer, improving the Pareto-front of accuracy and efficiency. Extensive experiments validate the proposed strategies.

6. Acknowledgements

The authors would like to thank Gene Team for the fruitful discussion and the anonymous reviewers for their valuable comments and suggestions. This research is supported by Meta Reality Labs, Innovation and Technology Commission (Ref: ITS/319/21FP) and Research Grant Council (Ref: 17210222).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Paolo Baerlocher and Ronan Boulic. Parametrization and range of motion of the ball-and-socket joint. In *Deformable avatars*, pages 180–190. Springer, 2001.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] 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 *European conference on computer vision*, pages 561–578. Springer, 2016.

[5] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[7] Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2061–2070, 2023.

[8] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision (ECCV)*, 2022.

[9] 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 *European Conference on Computer Vision*, pages 769–787. Springer, 2020.

[10] MMHuman3D Contributors. Openmmlab 3d human parametric model toolbox and benchmark. `https://github.com/open-mmlab/mmhuman3d`, 2021.

[11] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9026–9035, 2019.

[12] 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.

[13] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J Black. Learning to regress bodies from images using differentiable semantic rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11250–11259, 2021.

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[15] Chuang Gan, Yi Yang, Linchao Zhu, Deli Zhao, and Yueting Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, 120(1):61–77, 2016.

[16] Ryan Grainger, Thomas Paniagua, Xi Song, and Tianfu Wu. Learning patch-to-cluster attention in vision transformer. *arXiv preprint arXiv:2203.11987*, 2022.

[17] Yong Guo, Zhiyang Dou, Nan Zhang, Xiyue Liu, Boni Su, Yuguo Li, and Yinping Zhang. Student close contact behavior and COVID-19 transmission in China's classrooms. *PNAS Nexus*, 2(5):pgad142, 05 2023.

[18] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[19] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[22] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022.

[23] 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, 2013.

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[25] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[27] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[28] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020.

[29] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021.

[30] 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.

[31] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019.

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

[33] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren,

Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pages 620–640. Springer, 2022.

[34] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017.

[35] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware human mesh recovery from video by learning part-based 3d dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12375–12384, 2021.

[36] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021.

[37] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022.

[38] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

[39] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021.

[40] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12939–12948, 2021.

[41] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[42] Xiyue Liu, Zhiyang Dou, Lei Wang, Boni Su, Tianyi Jin, Yong Guo, Jianjian Wei, and Nan Zhang. Close contact behavior-based covid-19 transmission and interventions in a subway system. *Journal of Hazardous Materials*, 436:129233, 2022.

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

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

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[46] Haoyu Ma, Liangjian Chen, Deying Kong, Zhe Wang, Xingwei Liu, Hao Tang, Xiangyi Yan, Yusheng Xie, Shih-Yao Lin, and Xiaohui Xie. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. *arXiv preprint arXiv:2110.09554*, 2021.

[47] Haoyu Ma, Zhe Wang, Yifei Chen, Deying Kong, Liangjian Chen, Xingwei Liu, Xiangyi Yan, Hao Tang, and Xiaohui Xie. Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation. In *ECCV*, 2022.

[48] 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 *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

[49] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022.

[50] 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 *European Conference on Computer Vision*, pages 752–768. Springer, 2020.

[51] Paschalis Panteleris and Antonis Argyros. Pe-former: Pose estimation transformer. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 3–14. Springer, 2022.

[52] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 803–812, 2019.

[53] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1485–1495, 2022.

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

[55] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[56] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[57] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017.

[58] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *2021 International Conference on 3D Vision (3DV)*, pages 432–441. IEEE, 2021.

[59] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. 2021.

[60] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11219–11229, 2021.

[61] Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning human mesh recovery in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17038–17047, 2023.

[62] Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, and Juwei Lu. Class semantics-based attention for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13739–13748, 2021.

[63] Zhaoqi Su, Weilin Wan, Tao Yu, Lingjie Liu, Lu Fang, Wenping Wang, and Yebin Liu. Mulaycap: Multi-layer human performance capture using a monocular video camera. *IEEE Transactions on Visualization and Computer Graphics*, 28(4):1862–1879, 2020.

[64] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021.

[65] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[66] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[68] 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 *European Conference on Computer Vision (ECCV)*, sep 2018.

[69] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022.

[70] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[71] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. *arXiv preprint arXiv:2303.13796*, 2023.

[72] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13211–13220, 2022.

[73] Qiangqiang Wu, Jia Wan, and Antoni B. Chan. Progressive unsupervised learning for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2993–3002, June 2021.

[74] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B. Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14561–14571, June 2023.

[75] Yan Xia, Qiangqiang Wu, Wei Li, Antoni B Chan, and Uwe Stilla. A lightweight and detector-free 3d single object tracker on point clouds. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[76] Yan Xia, Yusheng Xu, Shuang Li, Rui Wang, Juan Du, Daniel Cremers, and Uwe Stilla. Soe-net: A self-attention and orientation encoding network for point cloud based place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11348–11357, 2021.

[77] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7760–7770, 2019.

[78] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021.

[79] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022.

[80] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022.

[81] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *Euro-*

pean conference on computer vision, pages 37–54. Springer, 2020.

[82] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021.

[83] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[84] Nan Zhang, Li Liu, Zhiyang Dou, Xiyue Liu, Xueze Yang, Doudou Miao, Yong Guo, Silan Gu, Yuguo Li, Hua Qian, and Jianjian Wei. Close contact behaviors of university and school students in 10 indoor environments. *Journal of Hazardous Materials*, 458:132069, 2023.

[85] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019.