Global-to-Local Modeling for Video-based 3D Human Pose and Shape Estimation

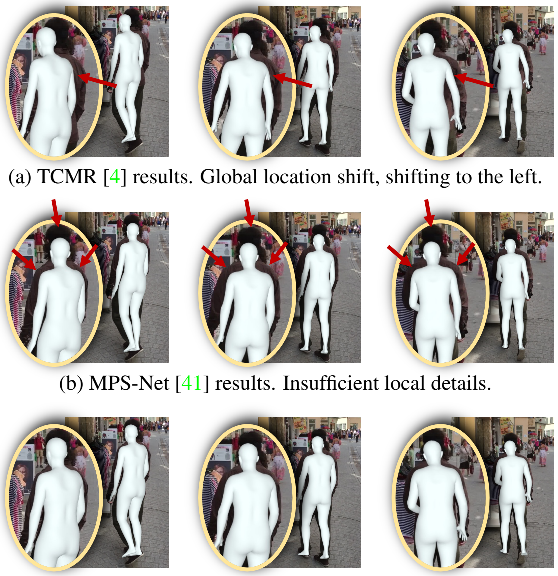
Xiaolong Shen[1,2*], Zongxin Yang[1], Xiaohan Wang[1], Jianxin Ma[2], Chang Zhou[2], Yi Yang[1]
[1] ReLER, CCAI, Zhejiang University [2] DAMO Academy, Alibaba Group

Abstract

Video-based 3D human pose and shape estimations are evaluated by intra-frame accuracy and inter-frame smoothness. Although these two metrics are responsible for different ranges of temporal consistency, existing state-of-the-art methods treat them as a unified problem and use monotonous modeling structures (e.g., RNN or attention-based block) to design their networks. However, using a single kind of modeling structure is difficult to balance the learning of short-term and long-term temporal correlations, and may bias the network to one of them, leading to undesirable predictions like global location shift, temporal inconsistency, and insufficient local details. To solve these problems, we propose to structurally decouple the modeling of long-term and short-term correlations in an end-to-end framework, Global-to-Local Transformer (GLoT). First, a global transformer is introduced with a Masked Pose and Shape Estimation strategy for long-term modeling. The strategy stimulates the global transformer to learn more inter-frame correlations by randomly masking the features of several frames. Second, a local transformer is responsible for exploiting local details on the human mesh and interacting with the global transformer by leveraging cross-attention. Moreover, a Hierarchical Spatial Correlation Regressor is further introduced to refine intra-frame estimations by decoupled global-local representation and implicit kinematic constraints. Our GLoT surpasses previous state-of-the-art methods with the lowest model parameters on popular benchmarks, i.e., 3DPW, MPI-INF-3DHP, and Human3.6M. Codes are available at https://github.com/sxl142/GLoT.

1. Introduction

Automatically recovering a sequence of human meshes from a monocular video plays a pivotal role in various applications, *e.g.*, AR/VR, robotics, and computer graphics. This technology can potentially reduce the need for motion capture devices and manual 3D annotations, providing human motion templates for downstream tasks, *e.g.*, the animation of 3D avatars. By utilizing parametric human models (*i.e.*, SMPL [26]) with well-defined artificial joint and shape structures, the popular procedure for video-based human mesh recovery involves indirectly regressing the SMPL parameters. However, effectively integrating deep neural networks with parametric artificial models to leverage multi-knowledge representations [42] for better estimation accuracy still remains an open problem.


(a) TCMR [4] results. Global location shift, shifting to the left.

(b) MPS-Net [41] results. Insufficient local details.

(c) Our results.

Figure 1. Our motivation. With the help of global-local cooperative modeling, our results avoid the global location shift and complement local details on intra-frame human meshes.

In video-based human mesh recovery, temporal understanding poses a crucial challenge that necessitates maintaining both intra-frame accuracy and inter-frame smoothness. Previous methods [4, 16, 41] mainly design deep networks to model long-term and short-term correlations simultaneously. For instance, VIBE [16] utilizes Recurrent Neural Network (RNN) [3] to model correlations. TCMR [4] and MPS-Net [41] consist of a temporal en-

[*] This work was done during an internship at Alibaba.

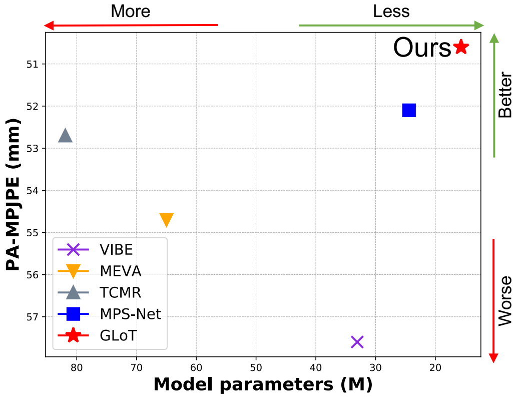


Figure 2. Model parameters vs. Performance.

coder and a temporal integration. The temporal encoder includes two types, RNN-based or attention-based methods, while the temporal integration employs attentive methods, *i.e.*, one-step or multi-step integration, to aggregate the representation extracted by the temporal encoder.

Even though these methods improve intra-frame accuracy or inter-frame smoothness to some extent, designing a coupled model to handle different ranges of temporal consistency, *i.e.*, long-term and short-term [43, 44], is ineffective and inefficient for this task, as shown in Figure 2. We empirically find that the single type of modeling structures leads to undesirable predictions as shown in Figure 1. For instance, RNN-based TCMR [4] suffers from a global location shift. Attention-based MPS-Net [41] captures the proper location of human meshes in the video sequence, but the mesh shape does not fit with the human in the images. We consider that the coupled modeling of long-term and short-term dependencies may not balance these two sides. It leads the RNN-based method to fall into the local dependency and does not capture the global location of the human mesh in the video. In contrast, the attention-based method tends to capture the long-term dependency and does not capture sufficient local details. Moreover, regressing the SMPL parameters from a coupled representation also confuses the model, *e.g.*, fine-grained intra-frame human mesh structure requires more short-term local details.

To solve the above problems, we propose to use information from both deep networks [37] and human prior structures [26, 39] in a joint manner. The proposed method, namely global-to-Local Transformer (GLoT), decouples the short-term and long-term temporal modeling. The framework is composed of Global Motion Modeling and Local Parameter Correction. In global modeling, we introduce a global transformer with a Masked Pose and Shape Estimation (MPSE) strategy for capturing the long-range global dependency (*e.g.*, proper global location, motion continuity). Specifically, the strategy randomly masks the features of several frames, and then the model predicts the SMPL parameters for the masked frames, which further helps the global transformer mine the coherent consistency of human motion and guides it to seize the inter-frame correlation from a global view. Under the local view, we introduce a local transformer and a Hierarchical Spatial Correlation Regressor (HSCR) for exploiting the short-term inter-frame detail and learning the intra-frame human mesh structure. To achieve this, we introduce nearby frames of the mid-frame and process them through the local encoder, which utilizes the mid-frame as a query to match the global transformer encoder's memory, generating a disentangled global-local representation of the mid-frame. Finally, HSCR employs human kinematic structures to constrain the refinement of decoupled global-local representation and improve global estimation.

With the help of global-local cooperative modeling, our model obtains the best intra-frame accuracy and inter-frame smoothness. For example, compared with the previous state-of-the-art method [41], our model significantly reduces the PA-MPJPE, MPJPE, and MPVPE by 1.5 mm, 3.6 mm, and 3.4 mm, respectively, on the widely used dataset 3DPW [38], while preserving the lowest Accel metric representing the inter-frame smoothness. Moreover, our model remarkably decreases the model parameters, as shown in Figure 2. Our contributions can be summarized as follows:

- To our best knowledge, we make the first attempt to decouple the modeling of long-term and short-term correlations in video-based 3D human pose and shape estimation. The proposed Global-to-Local Transformer (GLoT) merges the knowledge from deep networks and human prior structures, improving our method's accuracy and efficiency.
- In GLoT, we carefully design two components, *i.e.*, Global Motion Modeling and Local Parameter Correction, for learning inter-frame global-local contexts and intra-frame human mesh structure, respectively.
- We conduct extensive experiments on three widely-used datasets. Our results show that GLoT outperforms the previous state-of-the-art method [41], while achieving the lowest model parameters.

2. Related work

2.1. Image-based human pose and shape estimation

There are two lines of methods in the image-based human pose and shape estimation. The first one is the SMPL-based regression method. Kanazawa *et al.* [12] design an end-to-end framework named Human Mesh Recovery (HMR) for predicting the shape and 3D joint angles of the SMPL model while introducing adversarial training. Some methods [8, 17, 34, 36, 47] integrate prior knowledge, *i.e.*,

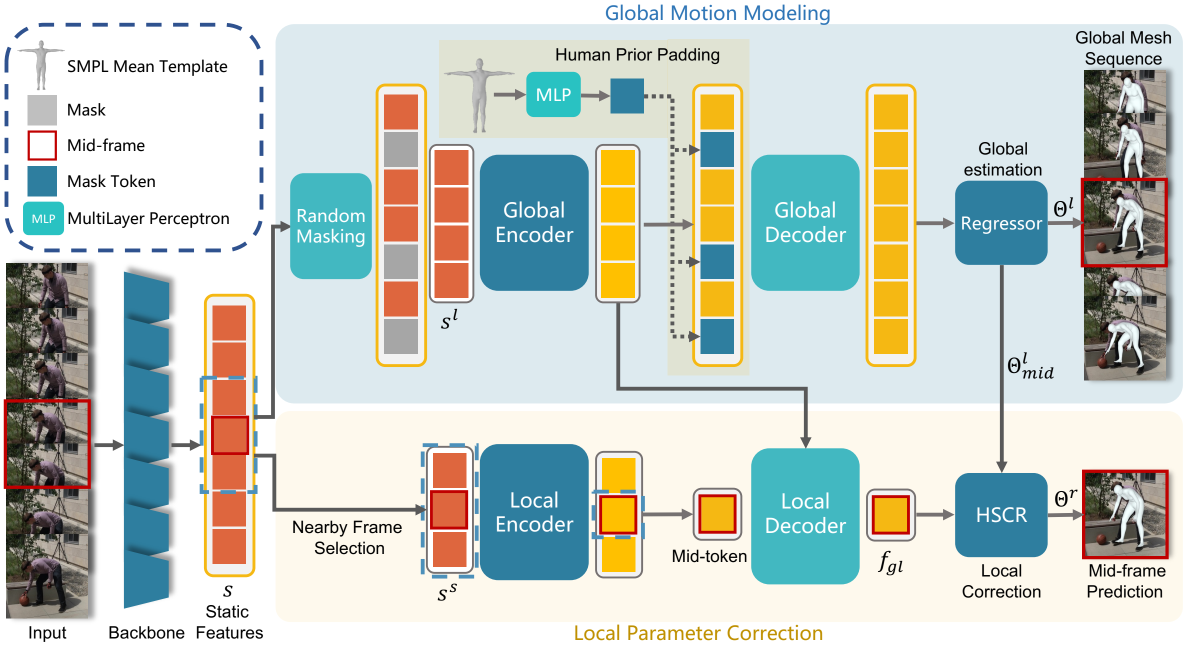


Figure 3. An overview of our GLoT. GLoT includes two branches, *i.e*., Global Motion Modeling (GMM) and Local Parameter Correction (LPC). We first extract static features from pretrained ResNet-50 [10], following [4, 16, 41]. Then the static features S are separately processed by random masking and nearby frame selection for feeding them (S^l, S^s) into global and local transformers. Last, Hierarchical Spatial Correlation Regressor (HSCR) corrects the global results Θ^l_{mid} obtained by GMM with the decoupled global-local representation f_{gl} and the inside kinematic structure. Note that our method utilizes T frames to predict the mid-frame, following [4, 41].

2D joint heatmaps and silhouette, semantic body part segmentation, multi-scale contexts, and kinematic prior, with a CNN model for estimating SMPL parameters. Moreover, 3D pose estimation [30, 35] is highly related to this task. For example, Li *et al*. [20] incorporate a 3D pose estimation branch to use inverse kinematics. Kolotouros *et al*. [18] unify two paradigms, *i.e*., optimization-based and regression-based methods, into a general framework. The other line [19, 23, 33] is to directly regress human mesh without a parametric model (SMPL). Even though these image-based methods achieve good results, they suffer from unstable human motion when applied to video sequences.

2.2. Video-based human pose and shape estimation

Video-based human pose and shape estimation mainly consists of SMPL-based regression methods. Kanazawa *et al*. [13] propose HMMR that learns a dynamics representation of humans from videos and allows the model to predict future frames. Kocabas *et al*. [16] introduce a motion prior provided by a large-scale Mocap dataset [29] to guide the model in learning a kinematically reasonable human motion structure via adversarial training. Although the motion prior helps the model to capture the human kinematic structure, it still suffers from temporally inconsistent. Hence, a series of methods emerged to solve this problem. MEVA [28] utilizes VAE [15] to encode the motion sequence and generate coarse human mesh sequences. The corresponding human meshes are then further refined via a residual connection. TCMR [4] and MPS-Net [41] can be unified into a general framework, *i.e*. temporal encoder, and temporal integration. TCMR utilizes GRU [3] to encode video representation of three different input lengths and then integrates three frames, *i.e*., mid-frame, front and rear frames of mid-frame, with an attentive module. MPS-Net replaces the GRU with a Non-local based [40] motion continuity attention (MoCA) module, which helps to learn non-local context relations. Moreover, instead of the one-step attentive aggregation of three frames in TCMR, MPS-Net utilizes a hierarchical attentive feature integration (HAFI) module for multi-step adjacent frames integration. Although these methods improve the per-frame accuracy or temporal consistency in some way, a coupled temporal modeling structure leads to some problems, *e.g*., global location shift, motion inconsistency, and intra-frame inaccuracy.

2.3. Vision Transformers

The Transformer [37] is first introduced in natural language processing (NLP). Transformer-based structures are effective for capturing long-range dependency and are suitable for learning global context due to the natural property of the self-attention mechanism. Inspired by the success of Transformer in NLP, Alexey *et al*. [6] first propose ViT and successfully obtain good results in image classification compared to convolutional architectures. Subsequently, many works [7, 24, 46, 49] have emerged to improve computing efficiency and enhance representation ability. Moreover, recent studies [21, 22, 43, 45, 50] have notably employed transformers to accomplish dense video tasks [31, 32]. Except for the design of the structure, some strategies for pretraining transformers, *e.g.*, masked Language modeling (MLM) [5] in NLP, masked image modeling (MIM) [2, 9] in CV, enhance the representation of transformers.

3. Method

3.1. Overview

Figure 3 shows an overview of our Global-to-Local Transformer (GLoT), which includes two branches, namely Global Motion Modeling (GMM) and Local Parameter Correction (LPC). Given an RGB video $\mathbf{I}$ with T frames, $\mathbf{I} = \{I_t\}_{t=1}^{T}$, we first utilize a pretrained ResNet-50 [10, 18] to extract static features $\mathbf{S} = \{s_t\}_{t=1}^{T}, s_t \in \mathbb{R}^{2048}$. Note the static features are saved on disk as we do not train the ResNet-50. We then feed these static features into GMM and LPC to obtain the final human mesh. Next, we elaborate on each branch of this framework as follows.

3.2. Global Motion Modeling

The Global Motion Modeling involves three components, *i.e.* a global transformer, a Masked Pose and Shape Estimation strategy, and an iterative regressor proposed by HMR [12]. For convenience in describing the overall process, we refer to the static features as static tokens.

Global Transformer. Recently, transformers have shown a powerful ability to model long-range global dependency owing to the self-attention mechanism. It is suitable for this task to capture long-term dependency and learn temporal consistency in human motion.

Masked Pose and Shape Estimation. Moreover, inspired by MIM [2, 9], we propose a simple yet effective strategy named Masked Pose and Shape Estimation, which helps the global transformer further mine the inter-frame correlation on human motion. Specifically, we randomly mask several static tokens and predict only the SMPL parameters of the masked locations by leveraging the learned correlation with other unmasked tokens during training.

Human Prior Padding. To reduce computation costs, the global transformer does not encode the masked tokens. Therefore, we need to pad tokens onto the masked location during the decoding phase. We propose to use the SMPL mean template as padding for the masked tokens. The SMPL template represents the mean of the SMPL parameter distribution and thus has the smallest average difference to a random pose sampled from the distribution. Considering Transformers are built on a stack of residual connections, the learning will be easier from an input initialization which has a smaller residual difference from the output prediction. To handle the SMPL mean template, we utilize an MLP for dimension transformation and refer to it as an SMPL token.

Iterative Regressor. The iterative regressor, first proposed by HMR [12], has been shown to provide good estimations in some recent works [4, 16, 41]. It iteratively regresses the residual parameters of the previous step, starting from the mean SMPL parameters. Although the regressor may overlook some local details, *e.g.*, human kinematic structure, it is well-suited to provide an initial global estimation under the global-to-local framework.

The complete process of GMM can be described as follows, we first randomly mask some static tokens along the temporal dimension, $\mathbf{S}^l \in \mathbb{R}^{(1-\alpha)T\times 2048}$, α is a mask ratio. Then we apply the global encoder to these unmasked tokens. During the global decoder phase, following [9], we pad the SMPL tokens onto the masked location and send the whole sequence into the global decoder to obtain a long-term representation. Finally, we apply the iterative regressor to the long-term representation to achieve global initial mesh sequences. We formulate the SMPL parameters obtained by GMM as follows, $\Theta^l = \{\theta^l, \beta^l, \phi^l\}, \theta^l \in \mathbb{R}^{T\times(24\times 6)}, \beta^l \in \mathbb{R}^{T\times 10}, \phi^l \in \mathbb{R}^{T\times 3}$. θ and β are the pose and shape parameters that control the joint rotation and mesh shapes by the parametric model SMPL [26]. Due to insufficient 3d data annotation, ϕ represents a set of pseudo camera parameters predicted by the model, which project the 3d coordinates onto the 2d space for weakly supervising the model by abundant 2d annotated data.

3.3. Local Parameter Correction

Local Parameter Correction consists of two components, *i.e.* a local transformer and a Hierarchical Spatial Correlation Regressor.

Local Transformer. As it is well known, the motion of a human body in a mid-frame is significantly influenced by its nearby frames. To capture the short-term local details in these frames more effectively, we introduce a local transformer. Specifically, we select the nearby frames for short-term modeling, $\mathbf{S}^s = \{s_t\}_{t=\frac{T}{2}-w}^{\frac{T}{2}+w}, s_t \in \mathbb{R}^{2048}$, w is the length of nearby frames. We utilize the local encoder on these selected tokens. The local decoder is different from the global transformer, which decodes the features representing not only global-wise human motion consistency but also local-wise fine-grained human mesh structure through a cross-attention mechanism. The cross-attention function can be defined as follow,

$$CrossAtten(Q_s^{mid}, K_l, V_l) = Softmax(\frac{Q_s^{mid} K_l^T}{\sqrt{C}})V_l, \tag{1}$$

where Q_s^{mid} is a query of the mid-token, K_l and V_l are key

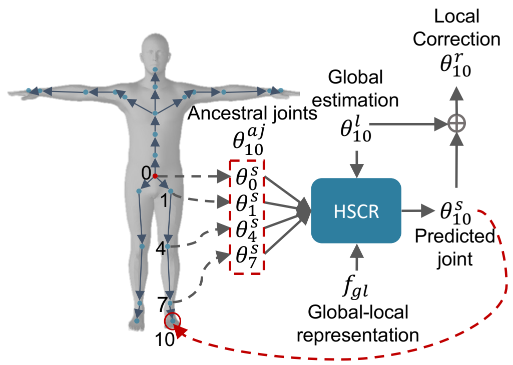


Figure 4. Human kinematic structure and an example of regressing and correcting one joint location.

and value of the global encoder.

Hierarchical Spatial Correlation Regressor. Previous methods [4, 16, 41] utilize the regressor from the pioneering work HMR [12] to estimate the SMPL parameters. However, this approach overlooks the inherent human joint correlation, *i.e.*, kinematic structure. Although the iterative regressor can produce a good estimation of SMPL parameters, it still requires local details to create a kinematically reasonable and rendered accurate human mesh.

Hence, inspired by [39], we propose a Hierarchical Spatial Correlation Regressor. Considering that directly regressing θ according to the kinematic structure may cause the model to fall into sub-optimal solutions due to a local view. Therefore, we add initial global prediction and decoupled global-local representation to this regressor, which allows the model to focus on adjusting the initial global estimation from a global-to-local perspective. Specifically, modeling the local intra-frame human mesh structure need to learn the joint correlation inside the kinematic structure. As shown in Figure 4, the kinematic structure can be described as the current joint rotation matrix being constrained by its ancestral joints. For example, when estimating the child joint (10), we need to compute the parent joints (0, 1, 4, 7) step-by-step. The total process of regressing parameters can be formulated as follows,

$$\begin{aligned}
\theta_i^{aj} &= Concat(\{\theta^s_{i,j}\}_{j=1}^{len(aj)}) \\
\theta_i^s &= M_{\theta_i^s}(f_{gl}, \theta_i^{aj}, \theta_i^l) \\
\theta^r &= \theta^l + \theta^s, \\
\beta^r &= \beta^l + M_\beta(f_{gl}, \beta^l), \\
\phi^r &= \phi^l + M_\phi(f_{gl}, \phi^l),
\end{aligned} \tag{2}$$

where $M_{\theta_i^s}$, M_β and M_ϕ are Multilayer Perceptron (MLP), $len(aj)$ is the number of the ancestral joints of the current joint i, f_{gl} is a decoupled global-local representation. Note θ is composed of 6D rotation representations of 24 joints. $\theta = \{\theta_i\}_{i=1}^{24}, \theta_i \in \mathbb{R}^6$.

3.4. Loss function

In GMM, we apply $\mathcal{L}_2$ loss to the SMPL parameters Θ^l and 3D/2D joints location, following the previous method [4, 16, 41]. Note that we only compute the loss of masked location. Moreover, we empirically discover that constraints on the velocity of the predicted 3D/2D joint location can help the model learn motion consistency and capture the long-range dependency better when applying the Masked Pose and Shape Estimation strategy to the global transformer. The velocity loss can be defined as follows,

$$\begin{aligned}
\mathcal{L}_{vel\_2d} &= \sum_{t=1}^{T-1} m_t||(jt_{2d}^{t+1} - jt_{2d}^t) - (gt_{2d}^{t+1} - gt_{2d}^t)||_2 \\
\mathcal{L}_{vel\_3d} &= \sum_{t=1}^{T-1} m_t||(jt_{3d}^{t+1} - jt_{3d}^t) - (gt_{3d}^{t+1} - gt_{3d}^t)||_2
\end{aligned} \tag{3}$$

where jt is a predicted 2d/3d joint location, gt is the ground truth of the 2d/3d joint location. $\mathbf{M} = \{m_i\}_{i=1}^{T-1}, \mathbf{M} \in \mathbb{R}^{T-1}$ is a mask vector. m_i is 1 when masking the i location, otherwise, m_i is 0. In LPC, we apply the same loss as the GMM, except that we only constrain the mid-frame.

4. Implementation details

Following previous methods [4, 16, 41], we set the input length T to 16. and use the same data processing procedure as TCMR [4]. We set the mini-batch N to 64 and initialize the learning rate to 1e-4. Moreover, we apply the Cosine Annealing [27] scheduler and linear warm-up to the Adam [14] optimizer. Our model is trained on one Nvidia Tesla V100 GPU. Since our model predicts the mid-frame of the input sequence, the predicted frame may fall outside the clip boundary. To handle this issue, we use nearest-padding, which duplicates the nearest boundary frame to pad the missing part. For example, when estimating frame 0, we duplicate frame 0 for 7 times. More details are provided in the supplementary material.

5. Experiments

We begin by introducing the evaluation metrics and datasets. Next, we report the evaluation results of our model and compare them with previous state-of-the-art methods. Finally, we provide ablation studies and qualitative results to further support our findings.

Evaluation Metrics. Following previous methods [4, 13, 16, 41], we report several intra-frame metrics, including Mean Per Joint Position Error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE), and Mean Per Vertex Position Error (MPVPE). Additionally, we provide a result for acceleration error (Accel) to verify inter-frame smoothness.

Datasets. The training datasets include 3DPW [38], Human3.6M [12], MPI-INF-3DHP [30], InstaVariety [13],

Method	3DPW				MPI-INF-3DHP			Human3.6M			number of
	PA-MPJPE ↓	MPJPE ↓	MPVPE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	Accel ↓	input frames
VIBE [16]	57.6	91.9	-	25.4	68.9	103.9	27.3	53.3	78.0	27.3	**16**
MEVA [28]	54.7	86.9	-	11.6	65.4	96.4	11.1	53.2	76.0	15.3	90
TCMR [4]	52.7	86.5	102.9	7.1	63.5	97.3	8.5	52.0	73.6	3.9	**16**
MPS-Net [41]	52.1	84.3	99.7	7.4	62.8	96.7	9.6	47.4	69.4	**3.6**	**16**
GLoT(Ours)	**50.6**	**80.7**	**96.3**	**6.6**	**61.5**	**93.9**	**7.9**	**46.3**	**67.0**	**3.6**	**16**

Table 1. Evaluation of state-of-the-art video-based methods on 3DPW [38], MPI-INF-3DHP [41], and Human3.6M [11]. All methods use 3DPW training set in training phase, but do not utilize Human3.6M SMPL parameters from Mosh [25]. The number of input frames follows the protocol of each paper.

	Method	3DPW			
		PA-MPJPE ↓	MPJPE ↓	MPVPE ↓	Accel ↓
single image	HMR [12]	76.7	130.0	-	37.4
	GraphCMR [19]	70.2	-	-	-
	SPIN [18]	59.2	96.9	116.4	29.8
	I2L-MeshNet [33]	57.7	93.2	110.1	30.9
	PyMAF [47]	58.9	92.8	110.1	-
video	HMMR [13]	72.6	116.5	139.3	15.2
	VIBE [16]	56.5	93.5	113.4	27.1
	TCMR [4]	55.8	95.0	111.5	7.0
	MPS-Net [41]	54.0	91.6	109.6	7.5
	GLoT(Ours)	**53.5**	**89.9**	**107.8**	**6.7**

Table 2. Evaluation of state-of-the-art methods on 3DPW [38]. All methods do not use 3DPW [38] on training.

Module	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
GMM	52.6	84.1	101.0	6.9
GMM + LPC	**50.6**	**80.7**	**96.3**	**6.6**

Table 3. Ablation studies of Global Motion Modeling (GMM) and Local Parameter Correction (LPC).

Penn Action [48], and PoseTrack [1], following [4]. We use 3DPW, Human3.6M, and MPI-INF-3DHP for evaluation. More details are provided in the supplementary material.

5.1. Comparison with state-of-the-art methods

Video-based methods. As shown in Table 1, our GLoT surpasses existing methods on 3DPW, MPI-INF-3DHP, and Human3.6M, for both intra-frame metrics (PA-MPJPE, MPJPE, MPVPE) and inter-frame metric (Accel). This indicates that our proposed Global-to-Local Modeling method is effective for modeling the long-range dependency (*e.g.*, the proper global location, coherent motion continuity) and learning the local details (*e.g.*, intra-frame human mesh structure). For example, GLoT reduces PA-MPJPE by 1.5 mm , MPJPE by 3.6 mm, MPVPE by 3.4 mm, and Accel by 0.8 mm/s^2 compared with MPS-Net [41] on 3DPW. Next, we analyze the existing problems of the previous method. First, VIBE [16], TCMR [4], and MPS-Net [41] all design a single type of modeling structure to simultaneously model long-term and short-term, leading to undesirable results like global location shift and insufficient local details shown in Figure 1. Second, MEVA follows a coarse-to-fine schema to model human motion, which requires too many frames as input and is ineffective for short videos.

Single image-based and video-based methods. We compare our GLoT with the previous single image-based and video-based methods on the challenging in-the-wild 3DPW [38]. Note that all methods do not utilize the 3DPW in the training phase. As shown in Table 2, our GLoT outperforms existing single image-based and video-based methods on all metrics, which validates the effectiveness of our method. For example, our model surpasses the video-based MPS-Net on PA-MPJPE, MPJPE, MPVPE, and Accel by 0.5 mm, 1.7 mm, 1.8 mm and 0.8 mm/s^2, respectively. In summary, our model achieves a smooth mesh sequence and accurate human meshes compared with previous methods. The results confirm that our proposed GLoT is effective for modeling long-range dependencies and learning local details, leading to better performance on the challenging 3DPW dataset.

5.2. Ablation studies

To validate the effectiveness of our GLoT, we conduct a series of experiments on 3DPW [38] with the same setting as Table 1. Initially, we verify two branches proposed in GLoT, *i.e.*, the Global Motion Modeling and Local Parameter Correction. Subsequently, we delve into the Masked Pose and Shape Estimation strategy, in addition to analyzing the types of mask tokens employed in the Global Motion Modeling phase. Lastly, we validate the influence of varying lengths of nearby frames and the effectiveness of the Hierarchical Spatial Correlation Regressor.

Global Motion Modeling (GMM) and Local Parameter Correction (LPC). We first study the GMM branch as shown in Table 3. Our results indicate that the GMM branch alone achieves competitive results, reducing the MPJPE by 0.2 mm and Accel by 0.5 mm compared to MPS-Net [41]. Furthermore, our method outperforms TCMR [4] in all metrics. Overall, our model shows the powerful learning long-

Mask ratio (%)	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
0	51.7	82.3	98	7.3
12.5	51.3	82	97.4	7.1
25	51.0	82.1	97.5	7.0
37.5	51.0	81.4	97.2	6.8
50	**50.6**	**80.7**	**96.3**	6.6
62.5	50.8	81.7	97.7	6.6
75	51.4	82.9	99.1	6.8
87.5	52.8	86.6	104.1	**6.5**

Table 4. Ablation studies of different mask ratios.

Types of mask token	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
SMPL Token	**50.6**	**80.7**	**96.3**	**6.6**
Learnable Token	51.5	82.5	98.5	6.9

Table 5. Ablation studies of types of mask tokens.

Module	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
Residual	51.5	81.7	97.2	6.7
HSCR	**50.6**	**80.7**	**96.3**	**6.6**

Table 6. Ablation studies of Hierarchical Spatial Correlation Regressor (HSCR).

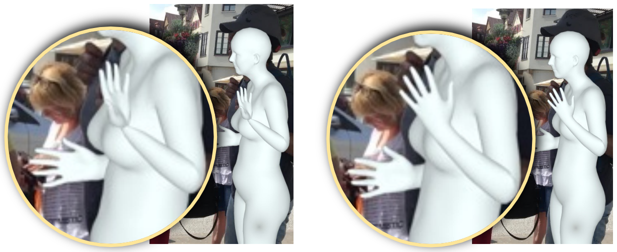
dependency ability of the global transformer and its potential to improve intra-frame accuracy. Next, we combine the LPC branch into this framework, which surpasses all methods with a larger margin in terms of both intra-frame accuracy and inter-frame smoothness. It shows that our global-to-local cooperative modeling helps the model to learn coherent motion consistency and inherent human structure.

Masked Pose and Shape Estimation Strategy. We randomly mask several static features during training. When testing, we do not mask any features. Table 4 shows the influence of different mask ratios. We find that (1) From the perspective of the Accel metric, applying the masked strategy helps the model to capture the long-range dependency. The Accel metric generally shows a reducing tendency when we gradually increase the mask ratio. The 87.5% mask ratio obtains the best Accel performance, which is in line with the intuitive understanding that the model tends to learn overly smooth meshes when applying a higher mask ratio. (2) In intra-frame metrics, the mined long-term dependency also provides global contexts for the next local modeling phase. When selecting a 50% mask ratio, our model achieves the best performance. A mask ratio between 37.5 % to 62.5 % is appropriate for improving the model performance.

Different types of the mask token. As shown in Table 5, we draw several conclusions from the experiments with different mask tokens. (1) Simply applying a learnable token to masked locations obtains better performance than the previous state-of-the-art method, MPS-Net. (2) When employing the SMPL token processed from the SMPL mean template, our model obtains the best result. (3) We consider that applying the SMPL token complements a human mesh prior, helping the model to learn human inherent structure.

Length	PA-MPJPE↓	MPJPE↓	MPVPE↓	Accel↓
2	51.9	83	98.7	6.7
3	51.2	84.2	100.0	6.7
4	**50.6**	**80.7**	**96.3**	**6.6**
5	51.3	81.7	97.2	6.7
6	51.7	82.4	98.4	6.6

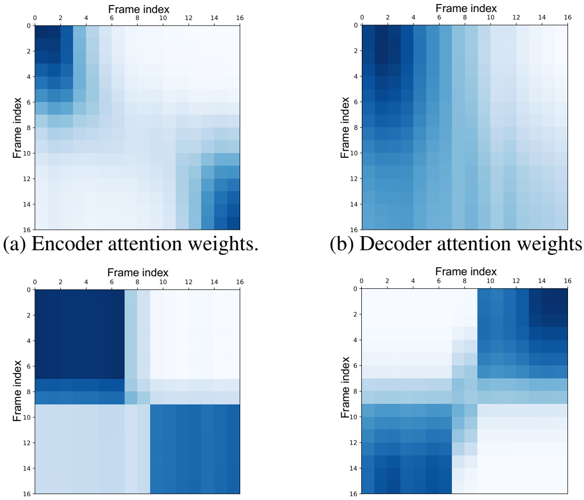
Table 7. Ablation studies of different lengths of nearby frames. The length of 2 means the total frames feeding to the local transformer will be $2 + 2 + 1 = 5$, including the mid-frame.



(a) Our results of GMM.

(b) Our results via LPC.

Figure 5. Qualitative results of Hierarchical Spatial Correlation Regressor (HSCR). (a) The wrist rotation is unreasonable. (b) HSCR corrects the wrist rotation. It shows that our model learns implicit kinematic constraints.



(a) Encoder attention weights.

(b) Decoder attention weights.

(c) Encoder attention weights.

(d) Decoder attention weights.

Figure 6. Attention visualizations. (a),(b). The results without using the masking strategy. (c), (d). The results of the masking strategy.

Hierarchical Spatial Correlation Regressor. Table 6 shows the evaluation results of our HSCR. Residual means that we do not utilize the implicit kinematic constraints in Sec. 3.3 and use the residual connection to replace them.

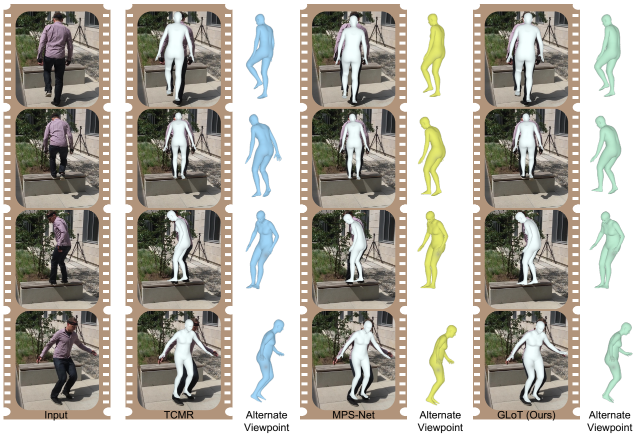


Figure 7. Qualitative comparison with previous state-of-the-art methods [4, 41].

We find that HSCR significantly reduces the PA-MPJPE, MPJPE, and MPVPE by 0.9 mm, 1.0 mm, and 0.9 mm, respectively, when compared with the residual connection. Moreover, our model achieves better results compared with other previous video-based methods when merely using the residual connection. This indicates that our Global-to-Local framework substantially improves the performance of video-based 3D human mesh recovery.

The different lengths of nearby frames. As shown in Table 7, we observe the following: (1) when setting the nearby length to four, our model achieves the best performance on all metrics. (2) the length of four means the input frames of the local transformer are nine, nearly half of the input frames of the global transformer. We consider that setting it to half of the input frames of the global transformer is a good solution. (3) Although other lengths result in worse performance than four frames, they are still competitive with other methods [4, 16, 41]. For example, when the length is set to two, our method surpasses the previous state-of-the-art MPS-Net with 0.2 mm PA-MPJPE, 1.3 mm MPJPE, 1.0 mm MPVPE, 0.7 mm/s^2 Accel.

5.3. Qualitative evaluation

Hierarchical Spatial Correlation Regressor (HSCR). In Figure 5, we study the proposed HSCR from a visualization perspective. It shows that the kinematic failure in global estimation is corrected by HSCR, which validates the effectiveness of our method.

Masked Pose and Shape Estimation strategy. In Figure 6, we provide attention visualizations of the global transformer and make several observations. (1) Compared between (a) and (c), the encoder with masking strategy attends to more nearby frames and is more centralized, indicating that it is not limited by the several nearby frames like (a). (2) The masking strategy in the decoder pushes the model to focus on further frames, while the non-masked decoder is more distributed. (3) The encoder and decoder in (c) and (d) show a cooperative relationship. This achieves decoupled long-term and short-term modeling. The encoder captures the nearby frame correlations while the decoder attends to the further frame correlations.

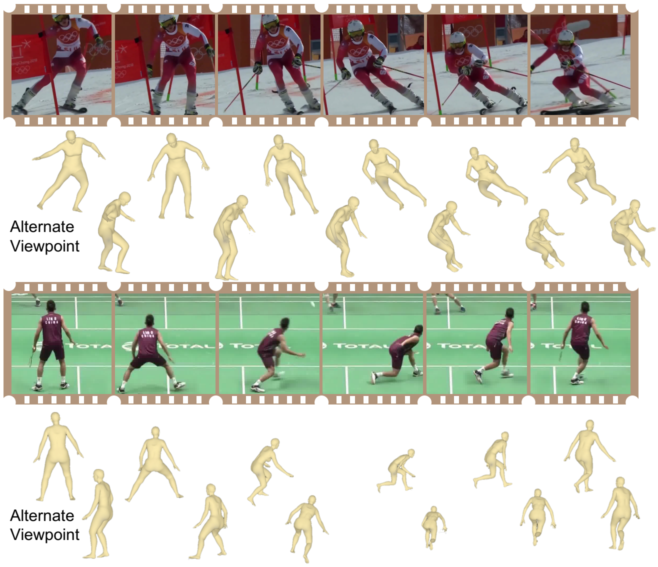


Figure 8. Qualitative results of GLoT on challenging internet videos.

Comparison with previous methods. In Figure 7, we compare our method to previous approaches [4, 41], and provide sequences of an alternate viewpoint. We observe that TCMR suffers global location shift and provides inaccurate meshes while MPS-Net captures the actual location, but the local details are insufficient.

Qualitative results on internet videos. In Figure 8, we validate our method on challenging internet videos and provide rendered meshes of an alternate viewpoint, demonstrating that our method successfully captures human motion from different perspectives. More qualitative results are provided in the supplementary material.

6. Conclusion

We propose a Global-to-Local Modeling (GloT) method for video-based 3d human pose and shape estimation, which captures the long-range global dependency and local details (*e.g.*, global location, motion consistency, and intra-frame human meshes) by combining deep networks and human prior structures. Through global-local cooperative modeling, GLoT achieves state-of-the-art performance on three widely used datasets. Furthermore, GLoT shows potential for handling in-the-wild internet videos, which could help the annotation of 3D meshes and provide various motion sequence templates for downstream tasks.

Acknowledgements. This work was supported by the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. *CVPR*, 2017. 6

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv*, 2021. 4

[3] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *empirical methods in natural language processing*, 2014. 1, 3

[4] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 8

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *north american chapter of the association for computational linguistics*, 2018. 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 3

[7] Matthijs Douze, Hugo Touvron, Matthieu Cord, Douze Matthijs, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *international conference on machine learning*, 2020. 4

[8] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. *ECCV*, 2020. 2

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2021. 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2015. 3, 4

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014. 6

[12] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 4, 5, 6

[13] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. *CVPR*, 2018. 3, 5, 6

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 5

[15] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013. 3

[16] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. *CVPR*, 2020. 1, 3, 4, 5, 6, 8

[17] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 2

[18] 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. 3, 4, 6

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 3, 6

[20] 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 *CVPR*, pages 3383–3393, 2021. 3

[21] Liulei Li, Tianfei Zhou, Wenguan Wang, Lu Yang, Jianwu Li, and Yi Yang. Locality-aware inter-and intra-video reconstruction for self-supervised correspondence learning. In *CVPR*, pages 8719–8730, 2022. 4

[22] Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware transformer for language-guided video segmentation. *PAMI*, 2023. 4

[23] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 3

[24] 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. *ICCV*, 2021. 4

[25] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6):220–1, 2014. 6

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 1, 2, 4

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv*, 2016. 5

[28] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. *ACCV*, 2020. 3, 6

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. *ICCV*, 2019. 3

[30] 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. *international conference on 3d vision*, 2016. 3, 5

[31] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, pages 21033–21043, 2022. 4

[32] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, pages 4133–4143, 2021. 4

[33] 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*, pages 752–768. Springer, 2020. 3, 6

[34] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. *international conference on 3d vision*, 2018. 2

[35] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *CVPR*, 2016. 3

[36] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. *CVPR*, 2018. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017. 2, 3

[38] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate {3D} human pose in the wild using {IMUs} and a moving camera. *ECCV*, 2018. 2, 5, 6

[39] Ziniu Wan, Zhengjia Li, Tian Maoqing, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. *ICCV*, 2021. 2, 5

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2017. 3

[41] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *CVPR*, pages 13211–13220, 2022. 1, 2, 3, 4, 5, 6, 8

[42] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021. 1

[43] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NIPS*, volume 34, pages 2491–2502, 2021. 2, 4

[44] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *PAMI*, 44(9):4701–4712, 2021. 2

[45] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *NIPS*, 2022. 4

[46] 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 *ICCV*, pages 558–567, 2021. 4

[47] 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 *ICCV*, pages 11446–11456, 2021. 2, 6

[48] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. *ICCV*, 2013. 6

[49] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 970–981, 2022. 4

[50] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *PAMI*, 2022. 4