

# Dual-Branch Graph Transformer Network for 3D Human Mesh Reconstruction from Video

Tao Tang<sup>1</sup>, Hong Liu<sup>1\*</sup>, Yingxuan You<sup>1</sup>, Ti Wang<sup>1</sup> and Wenhao Li<sup>1</sup>

**Abstract**—Human Mesh Reconstruction (HMR) from monocular video plays an important role in human-robot interaction and collaboration. However, existing video-based human mesh reconstruction methods face a trade-off between accurate reconstruction and smooth motion. These methods design networks based on either RNNs or attention mechanisms to extract local temporal correlations or global temporal dependencies, but the lack of complementary long-term information and local details limits their performance. To address this problem, we propose a Dual-branch Graph Transformer network for 3D human mesh reconstruction from video, named DGTR. DGTR employs a dual-branch network including a Global Motion Attention (GMA) branch and a Local Details Refine (LDR) branch to parallelly extract long-term dependencies and local crucial information, helping model global human motion and local human details (e.g., local motion, tiny movement). Specifically, GMA utilizes a global transformer to model long-term human motion. LDR combines modulated graph convolutional networks and the transformer framework to aggregate local information in adjacent frames and extract crucial information of human details. Experiments demonstrate that our DGTR outperforms state-of-the-art video-based methods in reconstruction accuracy and maintains competitive motion smoothness. Moreover, DGTR utilizes fewer parameters and FLOPs, which validate the effectiveness and efficiency of the proposed DGTR. Code is publicly available at <https://github.com/TangTao-PKU/DGTR>.

## I. INTRODUCTION

3D human mesh reconstruction is a crucial yet challenging task in computer vision and human-robot interaction [1], [2], [3], with a wide range of applications such as assisting household robots [4] and interactive mechanical arms [5]. 3D human mesh reconstruction is essential for higher-level intelligent human-robot interaction. These intelligent assistants should perceive the position of the human body, thereby effectively interacting with human and ensuring the safety of human-robot collaboration.

Many methods [6], [7], [8], [9], [10] have been proposed to recover the 3D human mesh from a single image, which doesn't require complex and expensive motion capture equipment. Simultaneously, compared with 3D human pose estimation [11], [12], [13], the human mesh can provide more information about the human body (e.g., body surface positions, body shape), which is crucial for many downstream applications (e.g., human-robot interaction, motion capture). Nevertheless, directly extracting detailed 3D human body

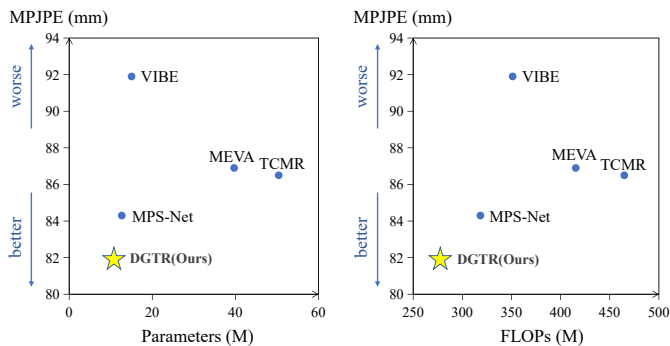


Fig. 1. Comparison between accuracy (MPJPE) and parameters (left), FLOPs (right) of video-based methods. All methods are evaluated on the 3DPW dataset.

mesh from images remains challenging due to the depth ambiguities, occlusions, and background interference.

While existing image-based methods [6], [7], [8], [9], [10] can generate remarkably accurate 3D human mesh from individual images, they often struggle to estimate smooth 3D human pose and shape from videos due to the absence of modeling temporal continuity in human motion. To tackle this challenge, some methods [14], [15], [16], [17], [18] have extended image-based methods to video scenarios. These methods primarily leverage a pre-trained Convolutional Neural Network (CNN) to obtain the static features from images, and then serially utilize RNN-based or attention-based networks to capture temporal information.

However, these methods have several limitations. The use of serial spatial-temporal networks in video-based methods results in a trade-off between accuracy and smoothness. RNN-based methods [15], [16], [17] employs Gate Recurrent Units (GRUs) [20] to capture local temporal information, resulting in over-smoothed human motion and inaccurate human mesh. In contrast, the attention-based method [18] effectively captures global temporal dependencies but lacks local human details. Moreover, the coupled spatial-temporal features make it challenging for the network to balance the accuracy and smoothness of human mesh.

To address the issues above, we propose a **Dual-branch Graph Transformer** network for 3D human mesh **Reconstruction from video (DGTR)**, which parallelly handle the global temporal and local crucial information. Our method mainly consists of two branches: the Global Motion Attention (GMA) branch and the Local Details Refine (LDR) branch. In GMA, the transformer is employed to capture

This paper is supported by the National Natural Science Foundation of China (No.62373009)

\*Corresponding authors: hongliu@pku.edu.cn (Hong Liu).

<sup>1</sup>Tao Tang, Hong Liu, Yingxuan You, Ti Wang and Wenhao Li are with State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, China

global temporal information (e.g., long-term human motion). In LDR, we introduce a local information aggregation module and a crucial feature extraction module to capture local human details. Compared with state-of-the-art video-based method MPS-Net [18], DGTR reduces the MPJPE by 2.3mm, 2.2mm, 2.2mm on 3DPW, MPI-INF-3DHP, and Human3.6M datasets, respectively. Additionally, as shown in Fig. 1, our DGTR network has fewer parameters and FLOPs among video-based methods, which is more efficient for human-robot interaction. Our main contributions are as follows:

- We propose a Dual-branch Graph Transformer network for 3D human mesh Reconstruction from video (DGTR) that parallelly captures global human motion and local human details with a dual-branch network.
- We introduce the Global Motion Attention (GMA) branch to extract long-term human motion. Besides, we propose local information aggregation and crucial feature extraction in the Local Details Refine (LDR) branch, which aggregates local human details presented in video frames and uses Modulated GCN to capture the crucial information of local human motion.
- We conduct extensive experiments on 3DPW, MPI-INF-3DHP and Human3.6M datasets. The results demonstrate that our DGTR surpasses previous video-based methods while using fewer parameters and FLOPs, which is efficient for practical applications.

## II. RELATED WORKS

### A. Human Mesh Reconstruction

**Image-based human mesh Reconstruction.** Most human mesh reconstruction methods use a single image as input and regress the pose and shape parameters of the human model. For instance, HMR [6] proposed an end-to-end framework that reconstructs the 3D human mesh from a single RGB image without relying on intermediate 2D key points. SPIN [7] proposed a self-improving network that consists of an SMPL parameter regressor and an iterative fitting framework. HKMR [8] explicitly leveraged the hierarchical structure and joint inter-dependencies of the parametric model. PyMAF [10] leveraged a feature pyramid and an explicit parameter rectification loop to improve the reliability of spatial features.

**Video-based human mesh Reconstruction.** In contrast to image-based methods, video-based methods need to simultaneously model accurate reconstruction and smooth human motion. HMMR [14] proposed a representation of 3D human dynamics from video sequences, enabling smooth 3D mesh motion prediction. VIBE [15] employed a bidirectional GRU-based motion generator and an adversarial motion discriminator with the actual human motion dataset AMASS [21] to capture human motion. MEVA [16] employed a two-step encoding process involving a GRU-based motion generator and a residual refinement, effectively capturing both general human motion and person-specific details. TCMR [17] proposed a three-branch temporal encoder that effectively utilizes temporal information from past and future

frames with GRUs to constrain the target frame. MPS-Net [18] replaced GRU with an attention module to capture non-local temporal features, which can learn global and long-term human motion. However, these RNN-based methods [15], [16], [17] still suffer from inadequate temporal information extracted by GRUs, resulting in motion jitter and inaccurate mesh estimation. Although attention-based MPS-Net effectively extracts long-term temporal information, the lack of local human details results in insufficient human details. Therefore, we find that the lack of complementary global motion information or local human details limits their performance, which leads to the trade-off between the accuracy and smoothness of human mesh reconstruction.

### B. Graph Convolutional Networks

Graph Convolutional Networks (GCNs) [22] generalize the capabilities of CNNs by performing convolution operations on graph-structured data. Recently, GCNs have been widely applied to 3D human pose and shape reconstruction [23], [24], [25], [26], primarily for extracting and integrating information from various human joints or meshes. GraphCMR [23] utilized a Graph-CNN to encode the template mesh structure and process image-based features on the mesh, which enables it to regress the 3D locations of mesh vertices directly rather than predicting the model parameters (SMPL [27]). Pose2Mesh [24] employed GCNs to directly estimate the 3D coordinates of human mesh vertices from the 2D human pose, addressing appearance domain gap issues and challenging pose estimation. Modulated GCN [28] disentangled feature transformations for different body joints through weight modulation and extends the graph structure beyond the human skeleton using affinity modulation while keeping relatively small parameters.

### C. Vision Transformer

Transformer [30] is first proposed for Natural Language Processing (NLP) tasks and rapidly outperforms CNN-based and RNN-based methods in many other tasks. Vision Transformer (ViT) [31] first replaced convolutional architecture with the transformer in the image classification task. Moreover, many researchers [32], [33] have employed transformers in the human mesh reconstruction task. For instance, METRO [32] utilized an image-based transformer encoder to jointly model vertex-vertex and vertex-joint interactions for the first time. GTRS [33] employed a pose-based graph transformer to exploit structured and implicit joint correlations. Metaformer [34] introduced a general architecture abstracted from the transformer without specifying the token mixer and could achieve competitive performance. Motivated by previous works, we leverage a transformer to extract global temporal features among video frames and employ GCNs based on the transformer framework to capture local crucial features of human motion in the video-based 3D human mesh reconstruction task.

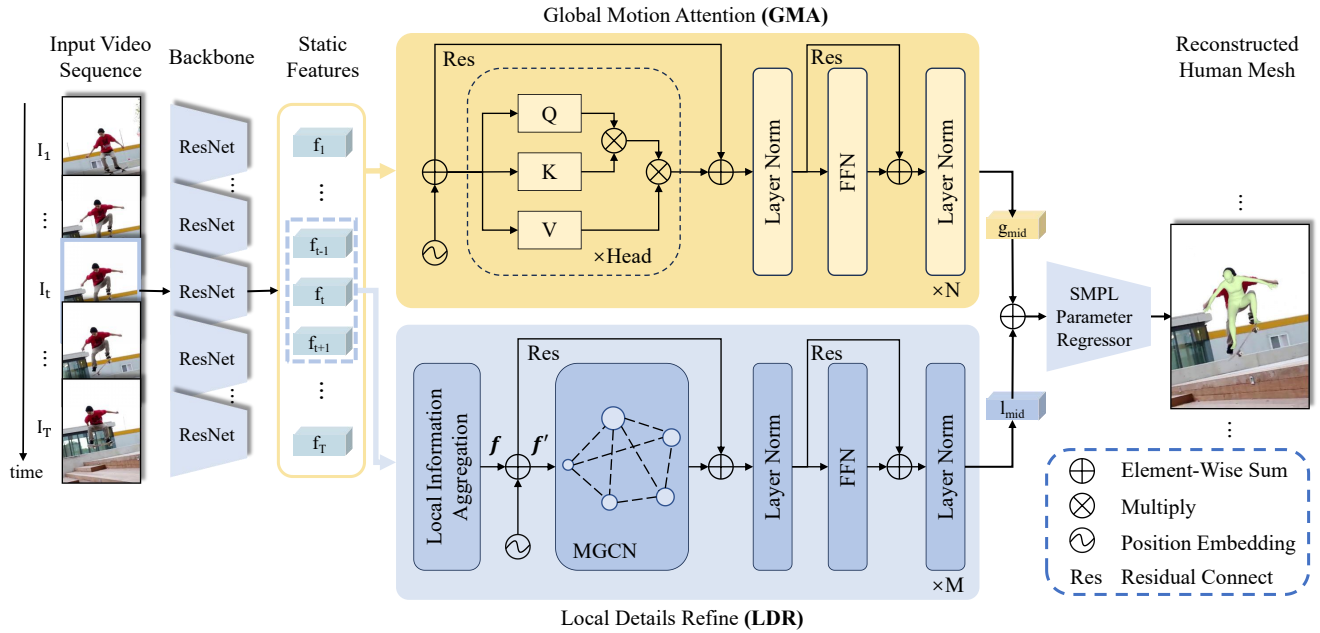


Fig. 2. The overview of dual-branch graph transformer network for 3D human mesh reconstruction from video (DGTR). Given a video sequence, ResNet [29] is utilized to extract the static features. The static features of all frames and adjacent 3 frames are separately fed into the GMA and LDR network. Then, GMA extracts global human motion, and local details of human motion are obtained by LDR. Finally, DGTR adds the output of the GMA and LDR branch and feeds it to the SMPL parameter regressor to generate the specific human mesh.

### III. METHOD

#### A. Overview of DGTR

The framework of DGTR is shown in Fig. 2. Given a video sequence  $V = \{I_t\}_{t=1}^T$  with  $T$  frames, we firstly use ResNet [29] pretrained by SPIN [7] to extract static feature sequence  $F = \{f_t\}_{t=1}^T$  of all frames, where  $f_t \in \mathbb{R}^{2048}$ . We regard  $(\lfloor T/2 \rfloor + 1)_{th}$  frame as the target frame,  $1, \dots, (\lfloor T/2 \rfloor)_{th}$  frames as past frames and  $(\lfloor T/2 \rfloor + 2)_{th}, \dots, T_{th}$  frames as future frames. Then, we feed the static feature of all frames and adjacent 3 frames separately into the Global Motion Attention (GMA) branch and the Local Details Refine (LDR) branch in parallel. These branches extract global temporal features and local fine-grained details, respectively. We add the global motion feature  $g_{mid}$  and local refined feature  $l_{mid}$  at the mid-frame, where  $g_{mid} \in \mathbb{R}^{2048}$ ,  $l_{mid} \in \mathbb{R}^{2048}$ . Finally, the combined information is fed into the pre-trained SMPL Parameter Regressor [7] to obtain the vertices of the human mesh. We introduce each branch in DGTR as follows.

#### B. Global Motion Attention

RNN-based networks like GRU often struggle to capture long-term dependencies adequately and may become trapped in local temporal modeling. On the contrary, the transformer architecture has gained a significant reputation for effectively capturing long-term temporal dependencies in various tasks owing to its multi-head attention mechanism and robust global information retention capabilities. Therefore, the transformer is an ideal choice for capturing global human motion across video frames.

As shown in Fig. 2, we employ the transformer encoder to extract global information from all adjacent 16 frames,

facilitating the generation of smooth human motion. Specifically, we employ only  $N = 2$  layers of the transformer encoder block encoding temporal feature, which makes the network remain lightweight. The  $\mathbb{R}^{2048}$  static features of each frame are used as input tokens and the learnable position encoding is added. We set the heads of multi-head attention mechanisms to  $Head = 8$  and the hidden dimension  $d$  to 512, which can be formulated as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{d})V. \quad (1)$$

Finally, we utilize the output features  $g_{mid} \in \mathbb{R}^{2048}$  corresponding to the target frame as the output of this branch.

#### C. Local Details Refine

**Local Information Aggregation.** The input of LDR is a static feature sequence  $F' = \{f_t\}_{t=1}^T \in \mathbb{R}^{T \times 2048}$  of nearby 3 frames. Due to the use of pooling operations in ResNet50 [29] for feature extraction from images, which loses a significant portion of the information relevant to the human body in the images. This harms the representation capability of static features and makes it extremely challenging to recover complex human body mesh from the static features. To address this problem, we introduce the local information aggregation module. This module leverages 1D convolution to aggregate the local human details among adjacent 3 frames in the static feature sequence. It effectively utilizes local redundancy in video frames to model local human details (e.g., local motion, tiny movement). This is achieved by sliding a small kernel along the input sequence and the weights within the kernel are learned during training. This

TABLE I

EVALUATION OF STATE-OF-THE-ART METHODS ON 3DPW, MPI-INF-3DHP, AND HUMAN3.6M DATASETS. ALL METHODS USE 3DPW FOR TRAINING, BUT DO NOT USE THE SMPL PARAMETERS OF HUMAN3.6M FROM MOSH [39]. THE TOP TWO BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED, RESPECTIVELY.

Method	3DPW				MPI-INF-3DHP			Human3.6M			input frames
	PA-MPJPE ↓	MPJPE ↓	MPVPE ↓	ACC-ERR ↓	PA-MPJPE ↓	MPJPE ↓	ACC-ERR ↓	PA-MPJPE ↓	MPJPE ↓	ACC-ERR ↓	
HMMR (CVPR'19) [14]	72.6	116.5	139.3	15.2	-	-	-	56.9	-	-	<b>16</b>
VIBE (CVPR'20) [15]	57.6	91.9	-	25.4	68.9	103.9	27.3	53.3	78.0	27.3	<b>16</b>
MEVA (ACCV'20) [16]	54.7	86.9	-	11.6	65.4	<u>96.4</u>	11.1	53.2	76.0	15.3	90
TCMR (CVPR'21) [17]	52.7	86.5	102.9	<b>6.8</b>	63.5	97.3	<b>8.5</b>	52.0	73.6	3.9	<b>16</b>
MPS-Net (CVPR'22) [18]	52.1	84.3	99.7	7.4	62.8	96.7	9.6	<u>47.4</u>	<u>69.4</u>	<b>3.6</b>	<b>16</b>
Zhang et. al (CVPR'23) [19]	<u>51.7</u>	<u>83.4</u>	<u>98.9</u>	<u>7.2</u>	<u>62.5</u>	98.2	<u>8.6</u>	51.0	73.2	<b>3.6</b>	<b>16</b>
DGTR (Ours)	<b>51.3</b>	<b>82.0</b>	<b>97.3</b>	7.6	<b>61.3</b>	<b>94.5</b>	<b>8.5</b>	<b>46.1</b>	<b>67.2</b>	<u>3.8</u>	<b>16</b>

process enhances the capacity of the network to focus on the crucial human details in images.

**Crucial Feature Extraction.** Following local information aggregation, we obtain a local crucial feature sequence. Subsequently, we leverage a crucial feature extraction module to extract crucial human details (e.g., local motion, tiny movement) from the nearby frames related to the target frame. Inspired by MetaFormer [34] and the good crucial feature extraction capability of the transformer framework, we introduce a novel module by integrating Modulated GCN [28] into the transformer framework. Modulated GCN [28] has demonstrated its effectiveness in capturing local crucial features of the human body with two components: weight modulation and adjacency modulation. Weight modulation utilizes modulation vectors to adjust or modulate the shared weight matrix for each node individually while maintaining small model parameters. On the other hand, adjacency modulation overcomes the limitation of relying on a pre-defined adjacency matrix to propagate and fuse information among all nodes. Modulated GCN appends a learnable adjacency matrix to capture strong correlations between long-term video frames. Modulated GCN can be expressed as:

$$Y = \text{sigmoid}(D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}} X(W \odot V)), \quad (2)$$

where  $X$  represents the input of all nodes in graph,  $\tilde{A}$  donates a learnable adjacency matrix,  $D$  is a degree matrix,  $W$  is a fusion matrix that needs to be learned,  $V$  represents modulation vectors, and  $Y$  is the output of Modulated GCN.

As shown in Fig. 2, we replace the multi-head attention mechanism with Modulated GCN in the transformer encoder, which enables us to effectively enforce local constraints while significantly reducing network parameters and FLOPs. Specifically, each video frame is treated as a graph node, and the adjacency matrix of the graph is initialized as all ones and can be learned through extensive training on mixed 3D and 2D human datasets. The specifics of the crucial feature extraction module can be calculated as follows:

$$f' = f + PE, \quad (3)$$

$$m = \text{Norm}(\text{MGCN}(f')), \quad (4)$$

$$l_{mid} = \text{Norm}(\text{FFN}(m) + m), \quad (5)$$

where  $PE$  represents the position embedding of each frame, Norm denotes the layer normalization, MGCN denotes the Modulated GCN, FFN denotes feedforward networks. Besides, We only employ  $M = 1$  layer of the LDA module.

#### D. Loss Function

Following previous methods [15], [17], [18], we apply SMPL shape loss, SMPL pose loss, 3D joint location loss, and 2D joint reprojection loss to help the network learn the accurate human body from images. To further constrain the network to generate smooth human motion, we employ two additional losses to utilize the velocity of the predicted 3D/2D joint locations. More details about the loss function are available on our project page.

## IV. EXPERIMENTS

### A. Implementation Details

Consistent with the previous methods [15], [17], [18], we set the input sequence length  $T$  to 16. We utilize the pre-trained ResNet50 from SPIN [7] to extract static features relevant to the human body of each frame. The SMPL parameter regressor consists of a single fully connected layer with 1024 neurons, followed by an output layer that predicts SMPL pose, shape, and camera parameters. We initialize the learning rate to  $1 \times 10^{-4}$  and use the batch size of 64. Additionally, we employ a Cosine Annealing scheduler with the linear warm-up for the Adam optimizer [35]. We train the entire network for 50 epochs, all experiments are conducted on a single NVIDIA GTX 3090 GPU. PyTorch is utilized for code implementation.

### B. Datasets and Evaluation metrics

**Datasets.** Following previous methods [15], [17], [18], we train our model on 3DPW [36], Human3.6M [37], MPI-INF-3DHP [38], and InstaVariety [14]. Subsequently, we evaluate the performance of our model on the 3DPW, Human3.6M, and MPI-INF-3DHP datasets that include 3D joint annotations.

TABLE II  
COMPARISON OF THE PARAMETERS AND FLOPS.

Method	Parameters (M)	FLOPs (M)
VIBE [15]	15.01	351.19
MEVA [16]	39.70	415.43
TCMR [17]	50.43	464.80
MPS-Net [18]	12.65	318.39
DGTR (Ours)	<b>10.89</b>	<b>277.56</b>

TABLE III  
ABLATION RESULTS FOR DIFFERENT BRANCHES OF DGTR ON 3DPW DATASET.

Method	3DPW			
	PA-MPJPE ↓	MPJPE ↓	MPVPE ↓	ACC-ERR ↓
DGTR wo. GMA	52.3	84.4	99.7	7.9
DGTR wo. LDR	52.0	82.6	98.8	<b>7.6</b>
DGTR	<b>51.3</b>	<b>82.0</b>	<b>97.3</b>	<b>7.6</b>

**Evaluation metrics.** We utilize metrics for both accuracy and smoothness. To evaluate reconstruction accuracy, we employ the mean per joint position error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE), and mean per vertex position error (MPVPE). These metrics measure the difference between predicted position and ground truth in millimeters ( $mm$ ). To evaluate reconstruction smoothness, we utilize the acceleration error (ACC-ERR) proposed in HMMR [14]. This metric calculates the average difference in acceleration of body joints, which is measured in ( $mm/s^2$ ).

### C. Comparison with state-of-the-art methods

**Comparison with video-based methods.** As shown in Table I, we conduct extensive experiments to compare our DGTR with previous video-based methods [14], [15], [16], [17], [18], [19] that report acceleration error. The results show that our DGTR outperforms state-of-the-art methods in almost all metrics, achieving the best reconstruction accuracy and competitive motion smoothness. This demonstrates that our dual-branch graph transformer network is proficient in concurrently modeling global dependencies (e.g., long-term human motion) and capturing local details (e.g., local motion, tiny movement). Since Zhang et al.[19] didn't release their code, making it impossible to compare qualitative experimental results and the parameters of their model. Therefore, we primarily compare our method with MPS-Net[18]. In terms of reconstruction accuracy, DGTR achieves the most accurate human mesh reconstruction compared with the previous methods. Compared to MPS-Net [18], our DGTR improves MPJPE by  $2.3mm$  (from  $84.3mm$  to  $82.0mm$ ),  $2.2mm$  (from  $96.7mm$  to  $94.5mm$ ) and  $2.2mm$  (from  $69.4mm$  to  $67.2mm$ ) on the 3DPW, MPI-INF-3DHP, and Human3.6M datasets, respectively. The dual-branch network significantly improves the accuracy of reconstruction. The

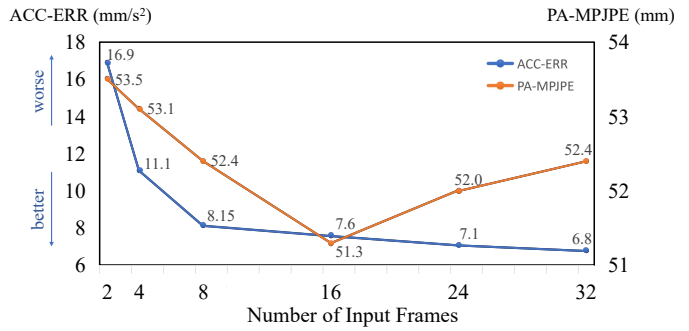


Fig. 3. Different number of input frames of DGTR under acceleration error and PA-MPJPE.

local details refine branch efficiently aggregates local details among adjacent frames, emphasizing the local information of human details. Regarding the smoothness of human motion, our DGTR achieves the best acceleration error on the MPI-INF-3DHP dataset and competitive results on the other two datasets. Although attention-based MPS-Net achieves similar motion smoothness to our DGTR, the reliability of global attention modules and the absence of local human details result in inaccurate reconstruction. On the contrary, our dual-branch DGTR alleviates the trade-off between smoothness and accuracy, achieving the best performance among video-based methods.

**Comparison in parameters and FLOPs.** As shown in Table II, taking the example of reconstructing one frame with an input of  $T = 16$  frames. Compared to previous video-based methods, our DGTR achieves fewer parameters and FLOPs while outperforming previous methods in almost all metrics.

### D. Ablation Analysis

**Effectiveness of GMA and LDR branch.** As shown in Table III, we conduct experiments to validate the effectiveness of each branch of DGTR. We evaluate the model on the outdoor 3DPW dataset. Experimental results demonstrate that the performance of only using the LDR branch is limited due to the utilization of only one layer of GCN, resulting in insufficient capacity to fit the high-dimensional mapping from images to the human mesh. When only the GMA branch is employed, the model achieves great motion smoothness. However, similar to MPS-Net [18], the lack of local human details leads to inaccurate reconstruction. The DGTR achieves the best performance by incorporating the LDR branch and the GMA branch.

**Impact of sequence lengths.** To investigate the impact of different numbers of input frames on the performance of DGTR, we conduct ablation experiments by setting the input length to 2, 4, 8, 16, 24, and 32. As shown in Fig. 3, the results demonstrate that the accuracy tends to decrease as the input sequence length surpasses 16 frames. The transformer captures excessive temporal information from distant frames, leading to a reduction in single-frame reconstruction accuracy. Regarding the smoothness of human



Fig. 4. Qualitative comparison of MPS-Net and DGTR on 3DPW dataset. Including fast motion, self-occlusion, and object occlusion scenarios.



Fig. 5. Qualitative results of DGTR on Internet video. Including complex backgrounds, motion blur, and multi-person scenarios.

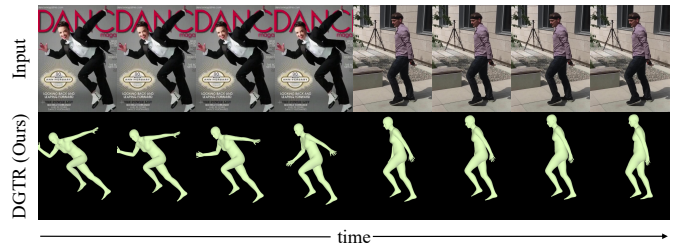


Fig. 6. Qualitative results of DGTR on a stitched video.

motion, the smoothness is improved as the input sequence length increases. This can be attributed to the fact that the transformer benefits from longer input sequences by acquiring more temporal information, which enhances temporal consistency in the motion. To ensure a fair comparison with previous video-based methods [14], [15], [17], [18], we set the input sequence length to 16.

### E. Qualitative Evaluation

**Comparison on experimental dataset.** As shown in Fig. 4, compared to MPS-Net [18], DGTR achieves more accurate human body reconstruction in fast basketball playing, slow climbing motion, and outdoor sports with multiple occlusions. As shown in the 7<sup>th</sup> column of Fig. 4, the parallel dual-branch DGTR can more effectively utilize global and local information under slight occlusion, which reconstructs better human details.

**Visualization on Internet videos.** To evaluate the generalization capability of DGTR, we employ our model on videos from the Internet with various motions and scenes. As shown in Fig. 5, the results demonstrate that our DGTR can generate accurate human mesh and smooth human motion in single-human, motion blur, and multi-human scenes.

**Analysis on a stitched video.** Qualitative evaluation above mainly focuses on per-frame reconstructed accuracy. To demonstrate the robust temporal feature extraction and motion jitter reduction capability of our DGTR, we conduct

experiments on a stitched video. We use two entirely different images and repeated them 30 times to create a stitched video. As shown in Fig. 6, frames 27<sup>th</sup> to 34<sup>th</sup> are displayed. Although the input human motion suddenly changes between 30<sup>th</sup> and 31<sup>st</sup> frames, we can note that the motion of the arms and legs changes gradually over time in the reconstructed human mesh.

## V. CONCLUSIONS

In this paper, we present DGTR, a Dual-branch Graph Transformer network for 3D human mesh Reconstruction from videos. We introduce a novel network with two branches: Global Motion Attention (GMA) and Local Details Refine (LDR), which can parallelly model global human motion and local human details. The GMA branch utilizes the transformer encoder to model long-term temporal information (e.g., long-term human motion). The LDR branch introduces a novel CNN-based information aggregation module and GCN-based transformer framework, which can effectively capture crucial information of human details (e.g., local motion, tiny movement). Compared with state-of-the-art methods, our DGTR achieves the best reconstruction accuracy and competitive motion smoothness while using fewer parameters and FLOPs. Moreover, experiments verify the efficiency of our DGTR for human mesh reconstruction, which shows its potential for practical applications.

**Limitation.** Main limitation of our method comes from severe occlusion and truncation of human body.

## REFERENCES

- [1] P. Wang, J. Liu, F. Hou, D. Chen, Z. Xia, and S. Guo, "Organization and understanding of a tactile information dataset tacat for physical human-robot interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 7328–7333.
- [2] N. Feng, F. Hu, H. Wang, and Z. Zhao, "Hybrid graph convolutional networks for skeleton-based and eeg-based jumping action recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 4156–4161.
- [3] H. Xing and D. Burschka, "Understanding spatio-temporal relations in human-object interaction using pyramid graph convolutional network," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 5195–5201.
- [4] A. Roitberg, D. Schneider, A. Djamal, C. Seibold, S. Reiß, and R. Stiefelwagen, "Let's play for action: Recognizing activities of daily living by learning from life simulation video games," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 8563–8569.
- [5] J. Zhao, Y. Ma, J. Dong, and S. Hou, "Interactive mechanical arm control system based on kinect," in *35th Chinese Control Conference (CCC)*, 2016, pp. 5976–5981.
- [6] A. Kanazawa, M.J. Black, D.W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 7122–7131.
- [7] N. Kolotouros, G. Pavlakos, M.J. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2252–2261.
- [8] G. Georgakis, R. Li, S. Karanam, T. Chen, J. Košecák, and Z. Wu, "Hierarchical kinematic human mesh recovery," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 768–784.
- [9] M. Kocabas, C.H.P. Huang, O. Hilliges, and M.J. Black, "PARE: Part attention regressor for 3D human body estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 11127–11137.
- [10] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 11446–11456.
- [11] J. Choi, D. Shim, and H.J. Kim, "Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3773–3780.
- [12] S. Schaefer, D.F. Henning, and S. Leutenegger, "GloPro: Globally-Consistent Uncertainty-Aware 3D Human Pose Estimation & Tracking in the WildIn," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3803–3810.
- [13] I. Fürst-Walter, A. Nappi, T. Harbaum, et al, "Design Space Exploration on Efficient and Accurate Human Pose Estimation from Sparse IMU-Sensing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 10888–10893.
- [14] A. Kanazawa, J.Y. Zhang, P. Felsen, and J. Malik, "Learning 3D human dynamics from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5614–5623.
- [15] M. Kocabas, N. Athanasiou, and M.J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5253–5263.
- [16] Z. Luo, S.A. Golestaneh, and K.M. Kitani, "3d human motion estimation via motion compression and refinement," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.
- [17] H. Choi, G. Moon, J.Y. Chang, and K.M. Lee, "Beyond static features for temporally consistent 3d human pose and shape from a video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1964–1973.
- [18] W. Wei, J. Lin, T. Liu, and H. Liao, "Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13211–13220.
- [19] B. Zhang, K. Ma, S. Wu, et al, "Two-stage co-segmentation network based on discriminative representation for recovering human mesh from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5662–5670.
- [20] K. Cho, B. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, and H. Schwenk, "Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.
- [21] N. Mahmood, N. Ghorbani, N.F. Troje, G. Pons-Moll, and M.J. Black, "AMASS: Archive of motion capture as surface shapes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5442–5451.
- [22] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs. ArXiv Preprint ArXiv:1312.6203, 2013.
- [23] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4501–4510.
- [24] H. Choi, G. Moon, and K.M. Lee: Pose2mesh, "Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose," in *Proceedings of the European conference on computer vision (ECCV)*, 2020, pp. 769–787.
- [25] Y. You, H. Liu, X. Li, W. Li, T. Wang, and R. Ding, "Gator: Graph-Aware Transformer with Motion-Disentangled Regression for Human Mesh Recovery from a 2D Pose," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [26] T. Wang, H. Liu, R. Ding, W. Li, Y. You, and X. Li, "Interweaved Graph and Attention Network for 3D Human Pose Estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M.J. Black, "SMPL: A skinned multi-person linear model. ACM Transactions On Graphics (TOG), 34(6), 2015, pp. 1–16.
- [28] Z. Zou, W. Tang, "Modulated graph convolutional network for 3D human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 11477–11487.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al, "Attention is all you need", Conference on Neural Information Processing Systems (NIPS), 2017.
- [31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al, "An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv Preprint ArXiv:2010.11929, 2020.
- [32] K. Lin, L. Wang, Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 1954–1963.
- [33] C. Zheng, M. Mendieta, P. Wang, A. Lu, and C. Chen, "A lightweight graph transformer network for human mesh reconstruction from 2d human pose," in *Proceedings Of ACM International Conference On Multimedia (ACM MM)*, 2022, pp. 5496–5507.
- [34] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10819–10829.
- [35] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
- [36] T. Marcard, R. Henschel, M.J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 601–617.
- [37] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 36(7), 2013, pp. 1325–1339.
- [38] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, and W. Xu, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *international conference on 3D vision (3DV)*, 2017, pp. 506–516.
- [39] M. Loper, N. Mahmood, and M.J. Black, "MoSh: Motion and shape capture from sparse markers. ACM Transactions On Graphics (TOG), 33(6), 2014, pp. 220:1–220:13.