FoGMesh: 3D Human Mesh Recovery in Videos with Focal Transformer and GRU

Yihao He
6201910023@stu.jiangnan.edu.cn
Xiaoning Song*
x.song@jiangnan.edu.cn
Tianyang Xu
tianyang_xu@163.com
Yang Hua
7211905018@stu.jiangnan.edu.cn
Xiaojun Wu
wu_xiaojun@jiangnan.edu.cn

Jiangnan University
, China

Abstract

Dense 3D human shape recovery plays an essential role in many computer vision and human-computer interaction tasks. However, accurate and robust 3D human body reconstruction in the wild is very challenging due to non-rigid deformation, occlusion, high-speed motion, etc., restricting the practical applications of the existing 3D human body recovery methods. To alleviate these issues, we propose a novel method using FOcal and Gated Recurrent Unit (GRU) encoders for high-precision 3D human Mesh reconstruction (FoGMesh) in video sequences. Specifically, we first design a new human body prior encoder based on the focal attention mechanism to learn fine-grained local and coarse-grained global interactions. Then, we build a multi-scale feature fusion module to fuse the context information and adaptively adjust the attention weights of small-scale body parts, such as hands. Last, we use the GRU encoder to connect the relevance and implement the proposed FoGMesh method in an end-to-end trainable framework. The proposed method achieves excellent performance on several benchmarking datasets, demonstrating its merits and superiority over the state-of-the-art approaches.

1 Introduction

3D human mesh recovery, a.k.a human pose estimation or shape reconstruction, is one of the most popular research topics in computer vision. With the rapid development of the area, high-precision 3D human body reconstruction has become a key technology in many downstream applications. The existing 3D human body reconstruction methods can be divided into two main categories, model-based and learning-based methods. Most of the classical approaches are model-based, which fit the 3D parametric human body and joint models and other clues by minimizing a predefined cost function with an optimization method. Recently, with the success of deep neural networks, state-of-the-art approaches are learning-based and data-driven. These methods usually learn a deep network that regresses the parameters of a 3D morphable human body model directly from an input image, achieving promising 3D reconstruction performance.

In general, it is a challenging task to infer accurate 3D human body meshes from 2D images since the information is incomplete and the process involves several sub-problems. In addition to the common difficulties, such as inadequate inference from 2D to 3D, lack of sufficient 3D annotated data, occlusion, self-occlusion, etc, the challenges are also posed by high-speed motion blur and obscure characteristics of small-scale body parts such as the hands. In recent years, one of the most widely studied frameworks is to learn a deep network from a large number of training samples, so the network can accurately regress the pose and shape parameters of a parametric human body model. The pose parameters control the rotation angle and position of the root joint while the shape parameters adjust the deformation of the human body. The well-known Skinned Multi-Person Linear model (SMPL) [28] has been widely used by existing approaches [19, 20, 32, 33]. However, the performance of these methods highly relies on the representative capability of the 3D parametric human body model. In contrast to these methods, we use an alternative approach for high-performance 3D human body recovery by using a 3D parametric human body model only as the prior information.

In this paper, we propose a novel method, namely FOcal encoder and GRU for Mesh reconstruction (FoGMesh), which uses a GRU encoder and multi-layer Transformer encoder to reconstruct accurate 3D human mesh from multiple sequential image frames. Recent studies demonstrate that Transformer has achieved great success in many computer vision tasks due to the self-attention mechanism, which can effectively capture global relevance among any two elements in a feature map. Based on the Transformer encoder structure, the method in this paper incorporates the idea of focal attention to capture global relevance information and effectively uses the local information of the relation between vertices. In addition, we propose the Contextual Augmentation and spatial Attention (CAA) module that integrates multi-scale features and highlights important elements via spatial attention. CAA effectively fuses the context pose information and automatically adjusts the attention weights of small-scale body parts for robust and discriminative feature extraction. Furthermore, GRU is a kind of Recurrent Neural Network (RNN), which has been proven to be able to obtain promising performance when processing sequence data. Therefore, we use the GRU encoder to learn historical pose information, so that our method can maintain robustness in high-speed motion scenes. Last, the above modules are integrated in an end-to-end trainable manner for ease of implementation and use.

To summarize, the main contributions of this paper include:

- We present FoGMesh, taking the advantage of the focal attention mechanism to model both local and global interactions for 3D human mesh recovery.
- We propose the CAA module that merges multi-scale features and effectively amplifies the attention weights of small-scale body parts.
- FoGMesh outperforms the existing state-of-the-art methods on 3DPW and achieves better stability in high-speed motion scenes.

2 Related work

2.1 Model-based methods

A model-based method performs 3D human body reconstruction by fitting 3D parametric human models, silhouettes, or junctions to an input image with a pre-defined optimization cost. For shape estimation, early studies usually solve the problem by fixing the pose of the estimated 3D human body. The reason is that early-stage 3D parametric human body models lack the representation capability of recovering complicated pose deformations [11]. To address this issue, the Shape Completion and Animation for PEople (SCAPE) model [2] was proposed to achieve high-quality 3D mesh reconstruction of human body images and videos. Balan et al. [3] proposed to use SCAPE as the human body template to track moving persons in 3D, by fitting SCAPE to silhouettes. The method can handle rich pose variations and has achieved more realistic 3D reconstruction results. In addition, some methods reconstruct 3D shapes and track human motion using a pre-scanned human model from laser sensors [7, 9]. For example, by fitting an articulated template using segmented images [8], one can estimate the 3D human body skeleton to track the movements of humans. For pre-scanned models, some new non-rigid optimization algorithms [23, 24] were also proposed for better reconstruction results.

In general, the performance of a model-based method highly relies on a 3D parametric human body model. A 3D human mesh is reconstructed by fitting limited information such as joint points and silhouettes extracted from an input image. The model-based methods are classical and useful for human body tracking and reconstruction. Although they have achieved some success, this type of method is heavily dependent on the accuracy of the prior information. Also, due to the complexity of the optimization process and poor performance in fast-motion scenarios, their practical applications are limited.

2.2 Learning-based methods

Recently, deep learning has become the mainstream method in 2D and 3D human pose estimation [29, 30, 37, 41, 43]. The aim is to learn a powerful network that can regress the human pose via training the network with a large number of examples. In rent years, a variety of deep networks have been developed and investigated by the community. As a result, great progress has been made on publicly available human pose estimation datasets. In this paper, we focus on 3D human mesh recovery so we introduce 3D methods only.

3D human mesh recovery from a single image. Bogo et al. [4] proposed SMPLify, one of the earliest end-to-end learning-based methods, to predict the parameters of the SMPL model from 2D joints with convolutional neural networks. Since then, several studies have been proposed to directly regress the SMPL model parameters using deep neural networks [13, 17, 32, 33]. Moreover, due to the lack of 3D annotations, many existing approaches use a weakly supervised method by re-projecting the losses obtained by 2D key points [17], such as using a self-supervised method by enforcing consistency of the feature representations across different resolutions [44], or recording body and part segmentation as an intermediate representation [32, 33]. More recently, with the success of Transformer in computer vision, a variety of Transformer-based methods have been proposed for the task of 3D human mesh recovery [25, 26, 40, 46]. For instance, Lin et al. [25] introduced the prior knowledge of a parametric model by injecting the base template of SMPL into image features. After that, Lin et al. [26] introduced graph convolution on the basis of METRO [25], which makes local and global interaction modeling more robust. However, these methods could not well tackle temporal data, especially in high-speed motion scenes.

3D human mesh recovery from video sequences. Hogg [14] matched a simplified human body model with image features of walking persons in many early studies to estimate their body poses in video sequences. Classical methods also explored the use of Principal Component Analysis (PCA) to learn the prior knowledge of motion from data, but they were limited to simple motion postures. Many recently, deep learning methods [6, 15, 35] mainly focus on capturing joint positions of the human body in video sequences. Several recent studies have proposed to perform the task in an end-to-end manner [19, 31]. For example, Mehta et al. [31] adopted an end-to-end trainable network to directly regress 3D joint positions. Although this method performs well on indoor datasets, such as Human3.6M [16], its performance decreases significantly on in-the-wild datasets such as 3DPW [39]. To further improve the performance of 3D human mesh recovery in the wild, we advocate a novel method in this paper using focal and GRU encoders.

3 The proposed FoGMesh method

We first present the overall framework of FoGMesh in Section 3.1. Then, we introduce the proposed Contextual Augmentation and spatial Attention (CAA) module in Section 3.2. Last, we propose the Focal encoder to explore inner vertex relations in Section 3.3.

3.1 The Overall Framework

As shown in Figure 1, given a video sequence $V = \{I_t\}_{t=1}^{T}$ of length T, our network consists of three parts.

Firstly, we follow the standard formulation to extract the grid and global features from continuous video frames I_t using CNN. In particular, we establish our feature extraction module on the basis of an existing large-scale network, HRNet (High-Resolution Net) [36], which has illustrated consistent advantages in dense prediction tasks in computer vision [25, 26]. However, we believe that the feature layers of different scales are not well integrated, which suppress the small scale components. Therefore, we propose the CAA module to enrich the context information and adaptively adjust the attention weights of small-scale body parts through the spatial attention module. The details of the CAA module will be presented in Section 3.2. Similar to [26], our HRNet-CAA outputs 1024-Dim 7×7 grid features and 2048-Dim global feature vectors. After that, we tokenize the obtained grid features to 49 tokens and concatenate the global feature vectors across the temporal dimension.

Secondly, to reflect temporal variations, we fuse the obtained feature vector with previous frames by the GRU encoder to output the hidden vectors of the key frame. The GRU encoder facilitates the contextual posture information fusion across the temporal dimension. Hence, we can extract the key frame hidden vectors, performing positional encoding by the 3D coordinates of each vertex and body joint in a human template mesh. Besides, to adjust the input features, we first force all input tokens to 2051 dimensions by MLP (Multilayer Perceptron). Then we sample the input features into a new eigenspace $\in R^{494 \times 1024}$ via a linear layer.

Thirdly, we use the Multi-layer Focal Encoder (MFE) to regress the 3D human body joints and grid vertices. As shown in Figure 1, the MFE input consists of grid feature queries, joint queries, and vertex queries. For the joint queries, we use 14 key points to train the

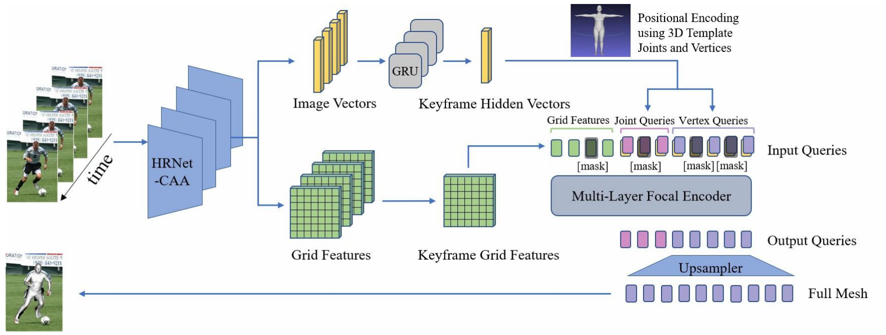


Figure 1: FoGMesh for 3D Human Mesh Reconstruction. Our framework takes a temporal sequence as input to obtain grid features and global feature vectors using a CNN. The global vectors are fed into the GRU encoder to obtain keyframe hidden vectors. The keyframe grid features and keyframe global vectors are tokenized and fed into the MFE for 3D human mesh.

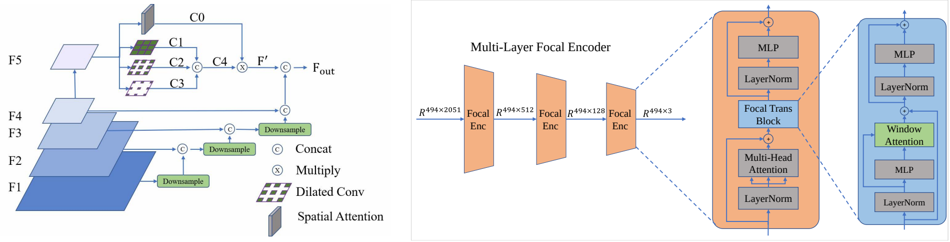
model, including right ankle, right knee, right hip, left hip, left knee, left ankle, right wrist, right elbow, right shoulder, left shoulder, left elbow, left wrist, neck, and head in order. For the vertex queries, as suggested by [25] and [26], we use a coarse human mesh of 431 vertices to accelerate the training stage. In particular, we construct the MFE with three progressively decreasing range transformer encoder blocks, which will be described in Section 3.3. The MFE output includes 3D coordinates of 431 vertices and 14 key points. At the final stage of FoGMesh, the predicted coarse mesh model (431 vertices) is restored to the SMPL model (6890 vertices) by upsampling via MLP. In addition, Mask Vertex Modeling (MVM) is used to address the occlusion issue. We randomly mask some percentages of the input queries to simulate occlusion, such that the learning the recovered input can improve the robustness of the model. Nevertheless, unlike recovering masked input like Masked Language Modeling (MLM) [18], MFE is required to regress all joints and vertices.

3.2 The CAA Module

As shown in Figure 2(a), we propose to obtain different receptive fields context information through the dilated convolution with different rates, which are 1,3 and 5, respectively. In specific, the dilated convolution uses different padding rates to keep the size of the output features consistent, and concatenates these output features sampled via fully connected layer. At first, the feature maps with different scales from the backbone network are taken as input, which are F1$(bs,c1,56,56)$, F2$(bs,c2,28,28)$, F3$(bs,c3,14,14)$, F4$(bs,c4,7,7)$, and F5$(bs,c5,7,7)$, respectively.

Then, F5 generates feature maps C1, C2, and C3 with the same size through parallel dilated convolution layers, and the three feature maps are concatenated to form feature map C4. Meanwhile, F5 obtains vector C0 through the spatial attention module [42], and C0 is multiplied with C4 to finally get the output feature map F$'\left(bs,f',7,7\right)$.

In parallel, F1 is downsampled and concatenated with F2, with the resulting feature map being downsampled and concatenated with F3, and so on. And finally concat with F$'$ to get F$_{\text{out}}\left(bs,f_{out},7,7\right)$. F$_{\text{out}}$ goes through an Average Pooling layer and then is flattened to a 1D vector which is the image vector.



(a)

(b)

Figure 2: The architecture of the proposed CAA Module and MFE module. (a) The feature map F5 is processed by the spatial attention module and the dilated convolution with rates of 1, 3, and 5 respectively. The dilated convolutions use different padding rates to keep the size of the output features consistent. Feature maps F1, F2, F3, and F4 are fused through downsampling and concatenating, and then concatenated with the processed feature map based on the feature map F5. (b) The MFE consists of three focal encoder blocks with the same number of input tokens.

3.3 The Multi-layer Focal Encoder

Inspired by the idea of [45], we combine fine-grained local and coarse-grained global interaction in our proposed MFE encoder module as shown in Figure 2(b). The overall structure of the focal encoder is similar to the traditional Transformer encoder, and we incorporate the window attention [45] to model fine-grained local interactions.

MFE consists of three encoder blocks with the same number of tokens, including 49 grid feature tokens, 14 joint queries, and 431 vertex queries. But their hidden dimensions are different. When the input tokens are given and the contextual features are generated by the Multi-Head Self-Attention (MHSA) proposed by Vaswani et al. [38], we can improve the local interactions with the help of window attention. In brief, firstly, the input features are divided into different levels, corresponding to the different granularity of attention. Secondly, different sizes of windows are set for different levels, and the self-attention operation is performed at the window level. Finally, the extracted fine-grained and coarse-grained tokens are concatenated to obtain local and global information. After that, our MFE sequentially reduces the dimensions to map the inputs to 3D joints and mesh vertices, simultaneously.

Although MHSA facilitates extracting long-range dependencies by using multiple self-attention functions in parallel to learn context representation, it is inefficient in capturing fine-grained local information in complex data structures. To explore the informative potentials of vertex-vertex relationships, we propose to design the Focal Transformer Block (FTB). As shown in Figure 2(b), FTB performs window attention where the self-attention is conducted at the window level. Specifically, FTB can effectively capture both short-term and long-term dependencies via performing fine-grained self-attention in the local regions and coarse-grained self-attention in the global regions.

Moreover, we use a template human mesh to reflect the positional information of vertex-vertex interactions, which is inspired by the widely discussed positional encoding [12, 21, 25, 26]. Similar to [26], the grid features are tokenized to 49 tokens, with each token being a 1024-Dim vector when the keyframe grid features and hidden vectors are given. The 2048-Dim keyframe hidden vectors are concatenated with the positional encoding vectors, using the 3D coordinates of each vertex and body joint in a human template mesh. Finally, we apply MLP to make the size of all the input tokens consistent.

In addition, we employ a similar training strategy as used in [25]. On this basis, we use L_1 loss to train our model for these regressed 3D vertices, 3D joints, and 2D projected body joints. The relevant function names are $\mathcal{L}_V$, $\mathcal{L}_J$, and $\mathcal{L}_J^{proj}$, respectively. Notably, the 3D joints can be obtained from the predicted 3D vertices by using a pre-defined regression matrix [5, 17, 21]. Besides, we also use L_1 loss for supervising these regressed 3D joints, termed as $\mathcal{L}_J^{reg}$. The overall objective is defined as:

$$\mathcal{L} = \alpha \times \left(\mathcal{L}_V + \mathcal{L}_J + \mathcal{L}_J^{reg}\right) + \beta \times \mathcal{L}_J^{proj}, \tag{1}$$

where α and β indicate the availability of 3D and 2D ground truths, respectively.

4 Experimental Results

In this section, we first introduce the benchmarking datasets and metrics used for evaluation. Then, we report the results obtained by our method on all the benchmarks to compare with the state-of-the-art approaches. Finally, we analyze the impact of each innovative component of our FoGMesh method in the ablation study.

4.1 Datasets and Experimental Settings

We evaluate our model on two benchmarking datasets, **3DPW** [39] and **Human3.6M** [16].

3DPW is an outdoor-image dataset with 2D and 3D annotations. 3DPW contains 60 video sequences with 22K images for training and 35K images for testing. Following the previous methods [5, 17, 19, 20], we follow the released train-test splits when conducting experiments on 3DPW.

Human3.6M is an indoor and large-scale dataset containing accurate 3D human poses. Each image has a subject performing a different action under 4 different viewpoints, resulting in 3.6M images in total. However, the ground-truth 3D mesh is inaccessible due to the license issue. Therefore, we use the pseudo-labels generated by SMPLify-X [34]. Following the common setting, we use the subjects S1, S5, S6, S7, and S8 for training, and keep the subjects S9 and S11 for testing.

We compare our results with metrics as follows. The unit for the metrics is millimeters (mm).

- **MPVE.** Mean-Per-Vertex-Error (MPVE) [33] measures the Euclidean distances between the ground truth vertices and the predicted vertices.
- **MPJPE.** Mean-Per-Joint-Position-Error (MPJPE) [16] measures the Euclidean distances between the ground truth joints and the predicted joints.
- **PA-MPJPE** [47]. It computes MPJPE after processing 3D alignment using Procrustes Analysis (PA) [10] to ignore the scale and rigid pose.

Method	3DPW			Human3.6M	
	MPVE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
HMR [17]	-	-	81.3	88.0	56.8
SPIN [20]	116.4	-	59.2	-	41.1
RSC-Net [44]	-	96.4	59.0	-	-
Pose2Mesh [5]	-	89.2	58.9	64.9	47.0
VIBE [19]	99.1	82.0	51.9	65.6	41.4
METRO [25]	88.2	77.1	47.9	54.0	36.8
Graphormer [26]	87.7	74.7	45.6	**51.2**	**34.6**
Ours	**85.2**	**74.1**	**45.5**	54.6	35.6

Table 1: Comparisons with relevant methods on 3DPW and Human3.6M datasets.

4.2 Main Results

We compare our method with previous related approaches by reconstructing human mesh on 3DPW and Human3.6M datasets, as listed in Table 1. We reproduced the results by using the provided checkpoints. On 3DPW, our method achieves better performance compared to the two relevant methods, METRO [25] and Graphormer [26]. For instance, our method improves MPVE by 2.5 and 3 mm compared to Graphormer and METRO, respectively. Besides, FoGMesh obtains remarkable improvement on all metrics compared to other methods.

As for Human3.6M, most compared models were trained on the same datasets, except METRO and Graphormer, which are trained on a large-scale multiple 2D and 3D dataset, including Human3.6M, COCO [27], MUCO [31], UP3D [22], MPII [1], and evaluated on Human3.6M. Therefore, METRO and Graphormer exhibit advantages by involving more training datasets. In contrast, other models only use the Human3.6M training set. However, our approach can still achieve competitive results with Graphormer. More importantly, our approach also performs better than METRO and other methods in terms of PA-MPJPE.

We also compare our method with other state-of-the-art methods including a video-based method, VIBE, as shown in Table 1. The results show that FoGMesh outperforms previous state-of-the-art methods with a large margin. Since our target is human mesh recovery in high-speed motion scenes, we also evaluated our method on the 3DPW test set that processed to simulate the high-speed motion blur as shown in Table 2. The last two rows indicate whether a GRU is used to learn the keyframe feature representations in our approach, the necessity of which will be discussed in Section 4.3. The results show that our method can maintain better generalization and stability in the high-speed motion scenes. Additionally, we use our method to make predictions for the outdoor data sampled from the processed 3DPW test set, and we simulate the high-speed motion blur by processing the data, as shown in Section 4.3.

4.3 Ablation Study

We conduct an ablation study on the 3DPW dataset to further evaluate the proposed method.

Effectiveness of CAA Module and Focal Encoder. As mentioned in Section 3.2, we propose to integrate the CAA module in HRNet, Table 3 reports the comparison between using and not using the CAA module in the backbone network. The results show that using the CAA module improves robustness in reconstructing 3D human mesh. To be specific, using HRNet-CAA as our backbone improves MPVE, MPJPE, and PA-MPJPE by 2.6, 2.7,

Method	Processed 3DPW		
	MPVE↓	MPJPE↓	PA- MPJPE↓
METRO	166.0	148.9	87.3
Graphormer	141.4	125.3	74.1
FoGMesh (- GRU)	150.6	132.9	80.2
FoGMesh (+ GRU)	**121.1**	**106.1**	**66.5**

Table 2: Comparisons with METRO and Graphormer on processed 3DPW. All models are not fine-tuned on processed 3DPW dataset. The bottom two rows indicate that our method does not use GRU and does use GRU, respectively.

Method	MPVE↓	MPJPE↓	PA-MPJPE↓
HRNet	88.2	77.1	47.9
HRNet+CAA	85.6	74.4	46.8
HRNet+CAA+Focal	**85.2**	**74.1**	**45.5**

Table 3: Ablation study with our proposed modules.

and 1.1 points, respectively, compared to HRNet. We further examine the focal encoder based on incorporating the CAA module into the backbone. As shown in Table 3, our method achieved our best results with the CAA module and focal encoder. Furthermore, we visualize the attention between specified joints and vertices, where brighter color indicates stronger attention, as shown in the fourth column in Figure 3. We observe that FoGMesh finds strong interactions between joints and adjacent vertices.

Analysis of GRU Encoder. To adapt our method to video data, we incorporate the GRU encoder so that the proposed method is no longer limited to single-frame image training. However, our experimental results show that using the GRU encoder can maintain better stability in video data, especially in a high-speed motion environment, as shown in Table 2. Our method outperforms METRO and Graphormer by a significant amount, and also proves that learning feature representations of keyframe by using GRU can maintain better robustness in high-speed motion scenes. Moreover, we used our method to make predictions for outdoor data that we sampled from the 3DPW dataset. When the background is not complex and the human features are obvious, all the methods will give slight deviation prediction results, as illustrated in the top two rows. However, FoGMesh also produces smoother results and more closely matches the target than other methods. In addition, FoGMesh outperforms even more than other methods when the scene is complicated and the targets are ambiguous, such as in the bottom three rows of Figure 3.

5 Conclusion

In this paper, we introduced FoGMesh, a new Transformer-based architecture that incorporates focal attention and the GRU encoder for reconstructing high-quality 3D human meshes from video sequences or a single image. The experimental results obtained on two well-known benchmarking datasets demonstrated a better generalization capability of the proposed method as compared with the existing state-of-the-art approaches. In addition, our experiments validated the effectiveness of cascading multi-scale features and modeling local interactions in Transformer in 3D human mesh recovery. Our FoGMesh model has the potential to facilitate 3D human reconstruction, but there is still room for improvement in

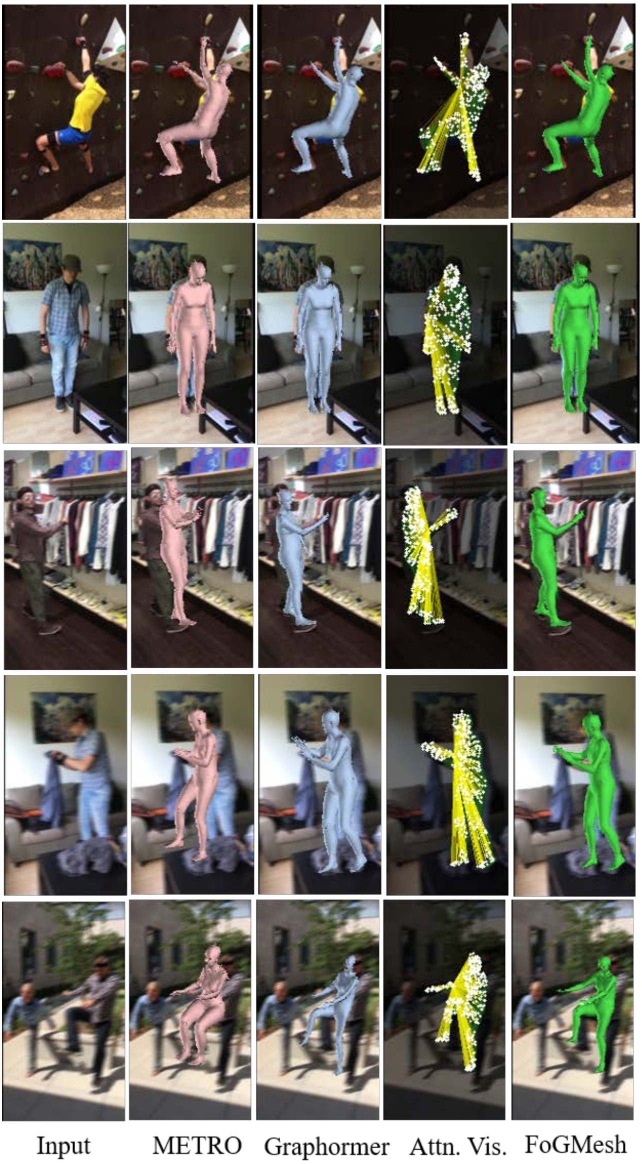


Figure 3: Qualitative results of FoGMesh. The top two rows of the first column are the original image, and the bottom three rows are the images that are processed to simulate high-speed motion. All models are not fine-tuned on processed 3DPW dataset. The last two columns are attention visualizations and predictions of FoGMesh.

our method. For example, we will further improve the generalization ability of our model in reconstructing human body parts and explore the method for better integrating Transformer and CNN in 3D human mesh recovery in our future work.

Acknowledgements

This material is based upon work supported by the Major Project of National Social Science Foundation of China (No. 21&ZD166), the National Natural Science Foundation of China (No. 61876072, 62106089) and the Natural Science Foundation of Jiangsu Province (No. BK20221535).

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] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.

[3] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[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] 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.

[6] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–683, 2018.

[7] Edilson de Aguiar, Christian Theobalt, Carsten Stoll, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *Untitled Event*. ACM, 2008.

[8] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753. Ieee, 2009.

[9] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753. Ieee, 2009.

[10] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[11] Daniel Grest, Dennis Herzog, and Reinhard Koch. Human model fitting from monocular posture images. In *Proc. of VMV*, pages 665–1344. Citeseer, 2005.

[12] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3d-coded: 3d correspondences by deep deformation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 230–246, 2018.

[13] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019.

[14] David Hogg. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1(1):5–20, 1983.

[15] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018.

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339, 2013.

[17] 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.

[18] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Hao Li, Robert W Sumner, and Mark Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Computer graphics forum*, volume 27, pages 1421–1430. Wiley Online Library, 2008.

[24] Hao Li, Bart Adams, Leonidas J Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics (ToG)*, 28(5):1–10, 2009.

[25] 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.

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

[27] 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.

[28] 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.

[29] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017.

[30] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)*, 36(4):1–14, 2017.

[31] 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.

[32] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

[33] 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.

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[35] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[37] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[39] 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[40] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021.

[41] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[43] Rongchang Xie, Chunyu Wang, and Yizhou Wang. Metafuse: A pre-trained fusion model for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13686–13695, 2020.

[44] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, László A Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image with self-supervised learning. In *European Conference on Computer Vision*, pages 284–300. Springer, 2020.

[45] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

[46] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for 10x efficient 2d and 3d pose estimation. *arXiv preprint arXiv:2203.08713*, 2022.

[47] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):901–914, 2018.