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

Yihao He, Xiaoning Song, Tianyang Xu, Yang Hua, Xiaojun Wu

Jiangnan University, China

Introduction

➢ **Motivation:**

- High-precision 3D human body recovery in the wild is a challenging task, especially in high-speed motion scenes.
- The attention weight of small-scale body parts is small in the reconstruction process, resulting in large deviation of the recovery results.

➢ **Contributions:**

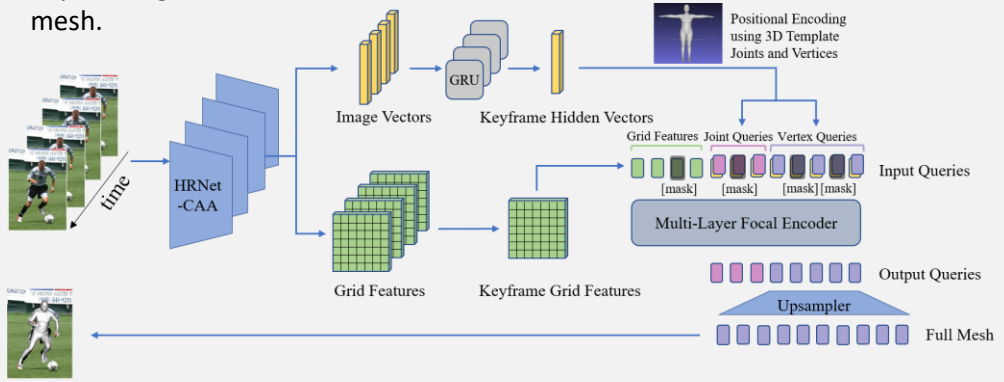
- A new method, FoGMesh, taking the advantage of the focal attention mechanism to model both local and global interactions for 3D human mesh recovery.
- We propose the CAA module that merges multi-scale features and effectively amplifies the attention weights of small-scale body parts.

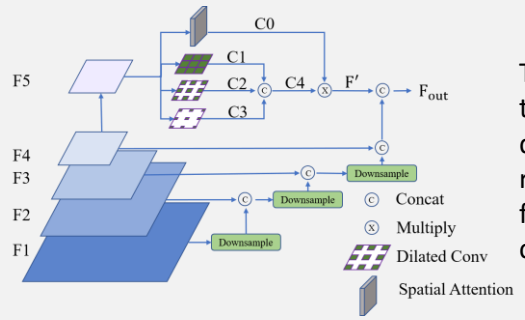
Our Results on 3DPW

Method	3DPW		
	MPVE↓	MPJPE↓	PA-MPJPE↓
HMR	-	-	81.3
SPIN	116.4	-	59.2
RSC-Net	-	96.4	59.0
Pose2Mesh	-	89.2	58.9
VIBE	99.1	82.0	51.9
METRO	88.2	77.1	47.9
Graphormer	87.7	74.7	45.6
Ours	**85.2**	**74.1**	**45.5**

The Proposed Method

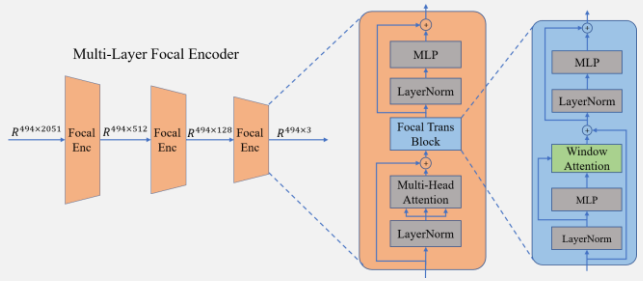
Our framework takes a temporal sequence as input to obtain grid features and global feature vectors using a CNN. The global vectors are fed into the GRU encoder to obtain keyframe hidden vectors. The keyframe grid features and keyframe global vectors are tokenized and fed into the MFE for 3D human mesh.



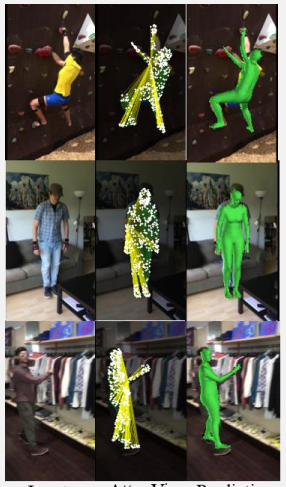


The feature map F5 is processed by the spatial attention module and the dilated convolution with different rates. Multi-scale feature maps are fused through downsampling and concatenating.

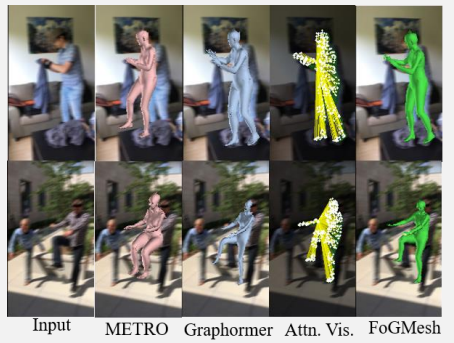
The MFE consists of three focal encoder blocks with the same number of input tokens. We combine fine-grained local and coarse-grained global interaction with the proposed MFE encoder module.



Reconstruction Results



↑ Prediction results on the clear original images



↑ Comparison of prediction results on high-speed motion blurred images

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] 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.

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