POTTER: Pooling Attention Transformer for Efficient Human Mesh Recovery — Supplementary Material

Ce Zheng[1]*, Xianpeng Liu[2], Guo-Jun Qi[3,4], Chen Chen[1]
[1]Center for Research in Computer Vision, University of Central Florida
[2] North Carolina State University
[3] OPPO Seattle Research Center, USA [4] Westlake University
cezheng@knights.ucf.edu; xliu59@ncsu.edu; guojunq@gmail.com; chen.chen@crcv.ucf.edu

In this supplementary material, we provide the following sections:

- Section A: Broader Impact and Limitations.
- Section B: Human Mesh Visualization on in-the-wild data.
- Section C: Memory and Computational Costs of One PAT Block.
- Section D: More Experiments (image classification and HMR) and Implementation Details.
- Section E: Generalization to 3D Hand Reconstruction

A. Broader Impact and Limitations

We anticipate that our POTTER can be used for widespread applications such as motion capture in animation and movies, virtual AI assistants, and VR/AR content. Currently, motion capture devices are mandatory for these applications, which are usually expensive, time-consuming, and complicated to set up. In contrast, one of the biggest advantages of our method is that POTTER can reconstruct 3D human mesh directly from 2D images/videos without extra devices. With the reliable reconstructing quality as depicted in Section B, POTTER shows a promising impact as a lightweight model for real-world applications.

There are also a few limitations of POTTER. Although POTTER can estimate reliable human mesh for in-the-wild scenarios, the performance would be downgraded when a severe occlusion exists. Another challenge is POTTER may fail for the rare and complicated pose scenarios due to limited training data. We will tackle these issues in future work.

*Work conducted during an internship at OPPO Seattle Research Center, USA.

B. Human Mesh Visualization on in-the-wild data

POTTER achieves superior performance on Human3.6M [3] and 3DPW [11] datasets as described in the main paper. However, it is critical to evaluate the actual performance of our POTTER on in-the-wild data. Reconstructing accurate human mesh on in-the-wild data is an extremely challenging task due to the different human shapes, scales, pose variations, and backgrounds from the training data.

In Fig. 1, we show the qualitative comparison with SOTA transformer-based method METRO [6] in this challenging scenarios (images are taken from in-the-wild dataset COCO [7]). Following METRO, we use the SMPL gender-neutral model [9] for all visualization. Our POTTER clearly outperforms METRO in many challenging cases, where the red circles highlight the area where POTTER is more accurate than METRO.

As an image-based method, POTTER can also reconstruct human mesh sequences given the input videos. In Fig. 2, we select several frames of the reconstructed human mesh to illustrate the performance of POTTER. We also provide the **video demo** of the entire reconstructed sequence in the supplementary material, which demonstrates the effectiveness of POTTER given the **in-the-wild** videos.

Since POTTER is a data-driven approach, the performance can not be guaranteed if the image is very different from the training data (i.e. data distribution shift), such as complicated pose and heavy occlusion (see Fig. 3 for an example). How to tackle these issues would be our future work. One potential solution is to use the domain adaptation method to make the trained model adapt to the target domain for better mesh recovery.

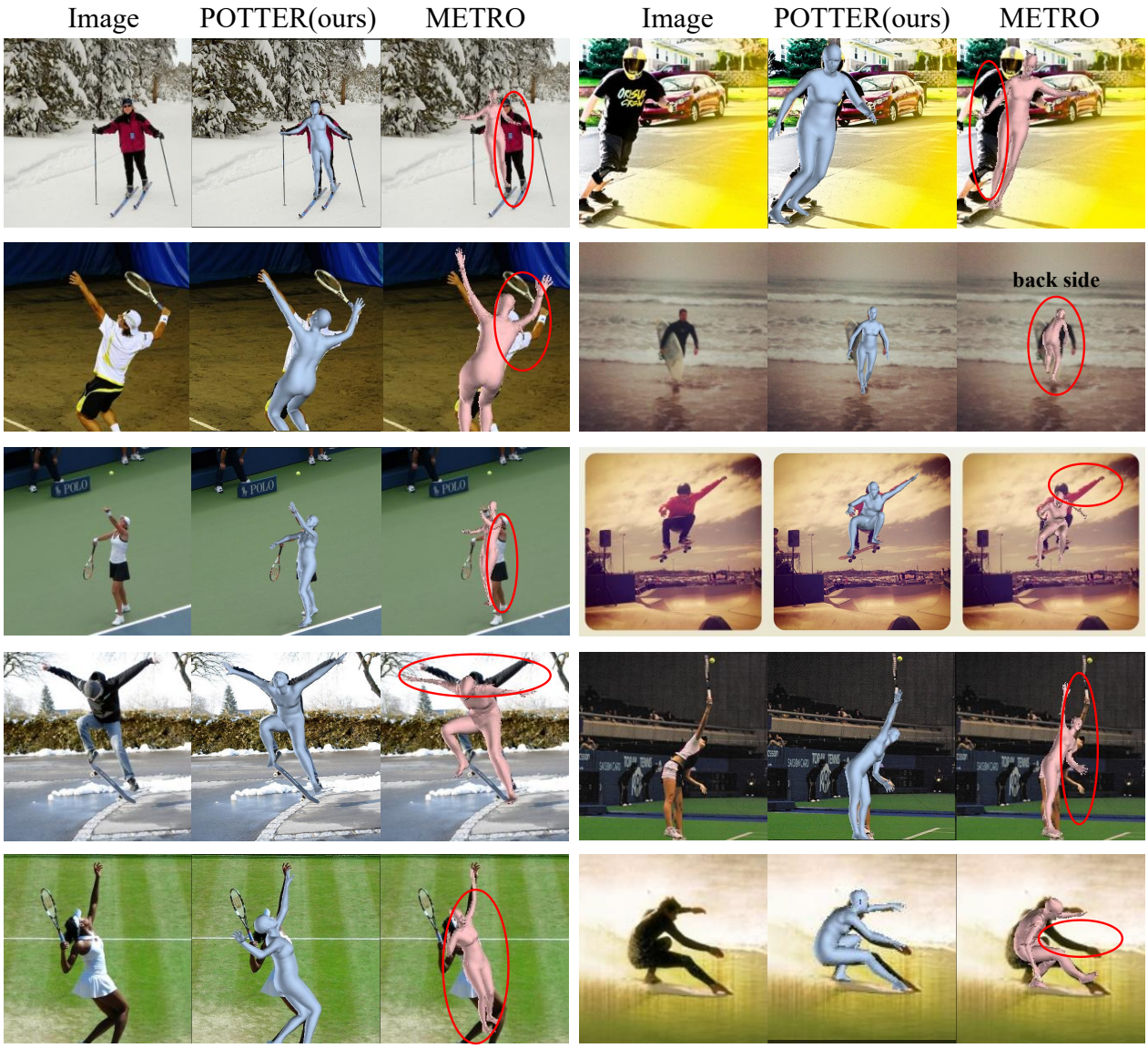

Figure 1. Qualitative comparison with SOTA transformer-based method METRO [6]. The red circles highlight regions where our POTTER generates more accurate mesh recoveries than METRO. Images are taken from the in-the-wild COCO [7] dataset.

Table 1. Total parameters and MACs of one PAT block.

Layer				Params	MACs
PAT	PoolAttn	Patch-wise	Pooling1 Pooling2 MatMul Proj1	$10D$	$9DN$
		Embed-wise	Pooling3 Pooling4 MatMul Proj2	$10D$	$9DN$
		Projection	Proj3	$10D$	$9DN$
	FFN		MLP1	$4D^2$	$4D^2N$
			MLP2	$4D^2$	$4D^2N$

C. Memory and Computational Costs of One PAT Block

To achieve model efficiency, one PAT block in the proposed method consists of one PoolAttn module with a Feed-forward Network (FFN). For the layers such as pooling, layer normalization, and matrix multiplication operations for the squeezed features, the required memory and computational costs can be ignored when compared with the projection or FFN layer. Thus, The total parameters and MACs of one PAT block given the input $[D, h, w]$ can be estimated as in Table 1, where the number of patches $N = h \times w$. To save the memory and computational costs, we utilize depth-wise convolution [1] served as the "Proj1", "Proj2", and "Proj3". The PoolAttn only requires $10D$ params and $9DN$

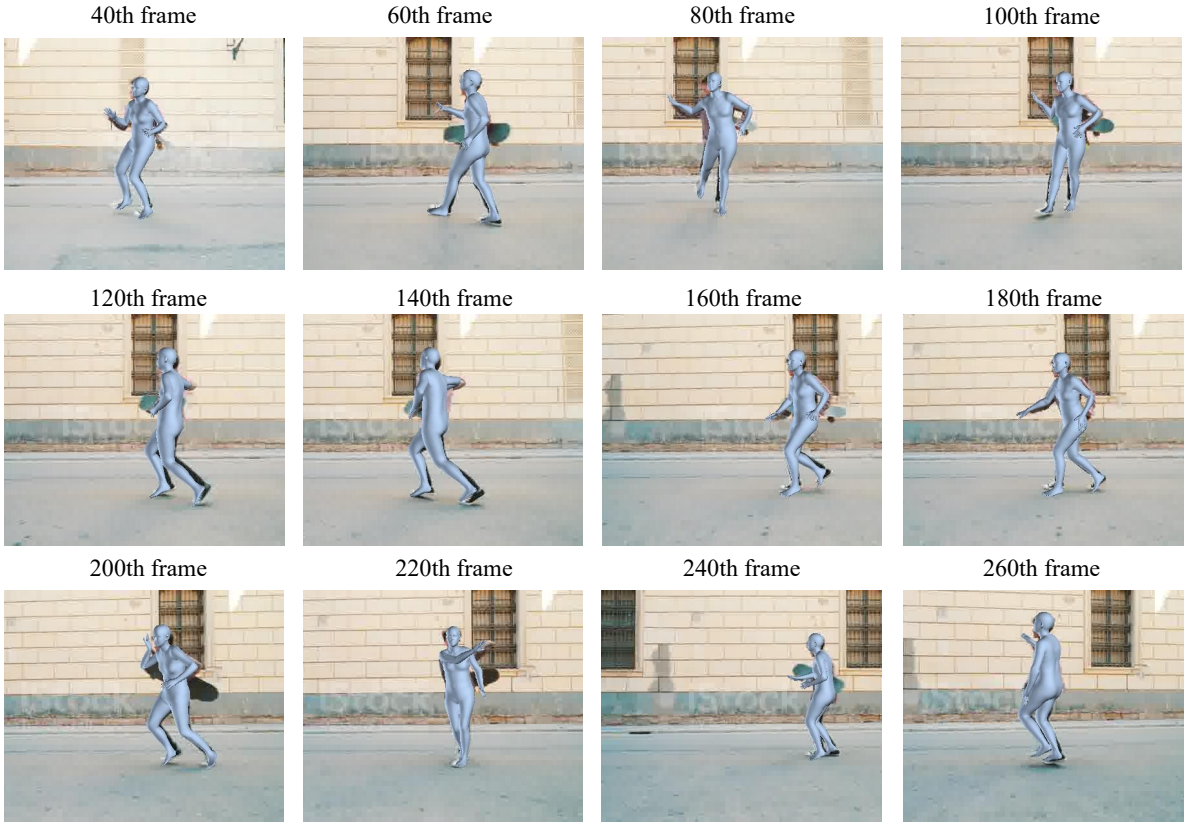

Figure 2. Qualitative results of using POTTER to reconstruct human mesh from an in-the-wild video. Although POTTER is an image-based method, the frame-by-frame reconstruction still works well. Please refer to our **video demo** for the reconstructed mesh sequences.

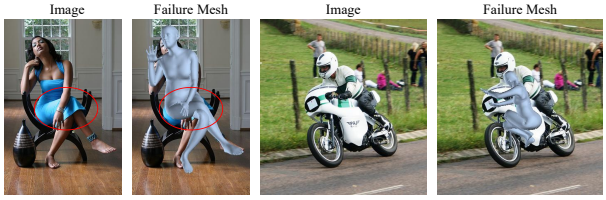

Figure 3. Failure cases. POTTER may not perform well due to severe occlusion.

MACs. Compared with the conventional attention module which requires $(4D^2 + 4D)$ params and $(4DN^2 + 2D^2N)$ MACs, our PoolAttn significantly reduce the complexity from $\mathcal{O}(D^2)$ to $\mathcal{O}(D)$.

D. More Experiments and Implementation Details

D.1. Image Classification

For the image classification task, we follow the same training scheme as PoolFormer [14]. Our model POTTER_cls is trained for 300 epochs with a cosine learning rate schedule (The number of warm-up epochs is 5). The AdamW optimizer [4] is used with weight decay 0.05 and peak learning rate $lr = 1e^{-3}$ and batch size 1024. The input image is with the size of [224, 224]. For POTTER_cls, the number of blocks for each stage is [2,2,6,2], which is the same as PoolFormer-S12. POTTER_cls outperforms PoolFormer-S12 by 1.8 % without increasing the memory and computational costs.

To further verify that our pooling attention design can significantly reduce the memory and computational cost without sacrificing performance, we increase the number of blocks for each stage as [4,4,12,4], named POTTER_cls_S24. The result is shown in Table 2. With the same hierarchical architecture, POTTER_cls_S24 (with PoolAttn) surpasses Swin-Tiny (with conventional attention) by requiring 72% of Params and 78% of MACs.

D.2. Human Mesh Recovery

For HMR task, the SMPL model [9] is utilized for reconstructing human mesh. Given the predicted pose parameters θ and the shape parameters β, the SMPL model can return the body mesh $M \in \mathbb{R}^{N\times 3}$ with $N = 6890$ vertices by the function $M = SMPL(\theta, \beta)$. After obtaining the body mesh M, the body joints J can be regressed by the predefined joint regression matrix W, which means

Figure 4. Qualitative results of our POTTER for reconstructing hand mesh.

Table 2. Performance of different types of models on ImageNet-1K classification task. All these models are only trained on the ImageNet1K training set. The top-1 accuracy on the validation set is reported in this table.

	Image Size	Params (M)	MACs (G)	Top-1 Acc ↑
RSB-ResNet-50 [13]	224	26	4.1	79.8
DeiT-S [13]	224	22	4.6	79.8
MLP-Mixer-B/16 [10]	224	59	12.7	76.4
PVT-Small [12]	224	25	3.8	79.8
ResMLP-S24 [14]	224	30	6.0	79.4
PoolFormer-S24 [14]	224	21	3.4	80.3
Swin-Mixer-T/D6 [8]	224	23	4.0	79.7
Swin-Tiny [8]	224	29	4.5	81.3
POTTER_cls_S24	224	21	3.5	**81.4**

$J \in \mathbb{R}^{k\times 3} = W \cdot M$, where k is the number of joints. The overall loss during the HMR task can be defined as:

$$\mathcal{L}_{HMR} = w_1\|\beta - \beta^*\| + w_2\|\theta - \theta^*\| + w_3\|J - J^*\| \quad (1)$$

where * denote the ground-truth value. In our experiments, we set $w_1 = 0.01$, $w_2 = 0.01$, and $w_3 = 1$.

Our POTTER is trained for 80 epochs with a step learning rate schedule with $lr = 5e-4$ and $lr_{decay} = 0.1$. The Adam [4] optimizer is utilized for training and the batch size is 32. The input image is resized to 256×256. We show more qualitative results for POTTER on images from Human3.6M and 3DPW datasets in Fig. 5.

Specifically, we compare our POTTER with THUNDR [15] in Table 3. Since the code of THUNDR is not released, we are unable to compute the MACs. POTTER achieves on-par results compared with THUNDR with 65 % of total parameters as shown in table 3. We also notice that THUNDR uses the more recent GHUM Model for the human mesh regression, while our POTTER and other methods such as SPIN [5], DSR [2], and TCFormer [16] use the SMPL Model for human mesh regression. This might be the reason that THUNDR achieves better performance.

Table 3. 3D Pose and Mesh performance comparison with SOTA methods on Human3.6M and 3DPW datasets.

			Human3.6M		3DPW		
	Params (M)	MACs (G)	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPVE
METRO	229.2	56.6	54.0	36.7	77.1	47.9	88.2
THUNDR	25	-	**48.0**	**34.9**	**74.8**	51.5	88.0
POTTER	16.3	7.8	56.5	35.1	75.0	**44.8**	**87.4**

E. Generalization to 3D Hand Reconstruction

POTTER can be also generalized for other mesh reconstruction tasks such as 3D hand reconstruction. To demonstrate this capability, we conduct the experiment on the hand mesh dataset FreiHand [17]. Without involving extra training data, POTTER can reconstruct reliable hand mesh. Unfortunately, due to the FreiHand online evaluation server being closed (The CodaLab website announced that the server is no longer accepting new challenges not new submissions

to old challenges), we are not able to report the test results. Here we provide the hand mesh visualization of POTTER in Fig. 4, which demonstrate that POTTER can generalize well for other tasks such as hand mesh reconstruction.

Figure 5. More qualitative results of our POTTER for HMR. Images are taken from Human3.6M and 3DPW datasets

References

[1] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2

[2] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J Black. Learning to regress bodies from images using differentiable semantic rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11250–11259, 2021. 4

[3] 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, jul 2014. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3, 4

[5] 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. 4

[6] 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. 1, 2

[7] 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. 1, 2

[8] 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. *International Conference on Computer Vision (ICCV)*, 2021. 4

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 1, 3

[10] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021. 4

[11] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 1

[12] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 4

[13] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 4

[14] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10819–10829, 2022. 3, 4

[15] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12971–12980, 2021. 4

[16] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 4

[17] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 4