JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

POTTER: Pooling Attention Transformer for Efficient Human Mesh Recovery

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

Introduction

Human Mesh Recovery (HMR) which can estimate 3D human pose and shape of the entire human body has drawn increasing attention.

Recently, the attention mechanism in transformer demonstrates a strong ability to model global dependencies in comparison to the CNN.

SOTA HMR methods all utilize transformer to exploit non-local relations among different human body parts for achieving impressive performance.

Introduction

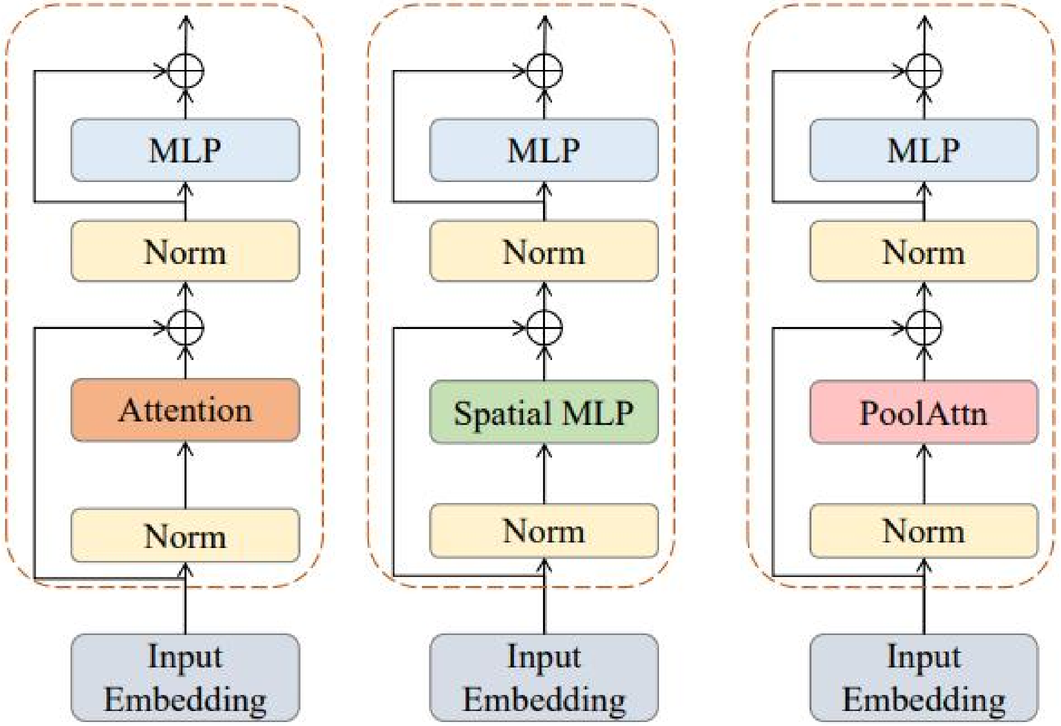
However, one significant limitation of these SOTA HMR methods is model efficiency.

- The large CNN backbones are needed for to extract features first.
- Computational and memory expensive transformer architectures are applied to process the extracted features for the HMR task.

Mainly pursuing higher accuracy is not an optimal solution

Introduction

Lightweight attention design: Pooling Attention Transformer (PAT)



(a) Attention-based (e.g. ViT, Swin) **(b) MLP-based (e.g. MLP-Mixer)** **(c) PAT (ours)**

Number of Patches: N
Embedding Dim: D

Module	Params	MACs
Attention	$4D^2+4D$	$4DN^2+2D^2N$
Spatial MLP	$4N^2+4N$	$4D^2N$
PoolAttn	30D	27DN

If N = 196 and D = 512

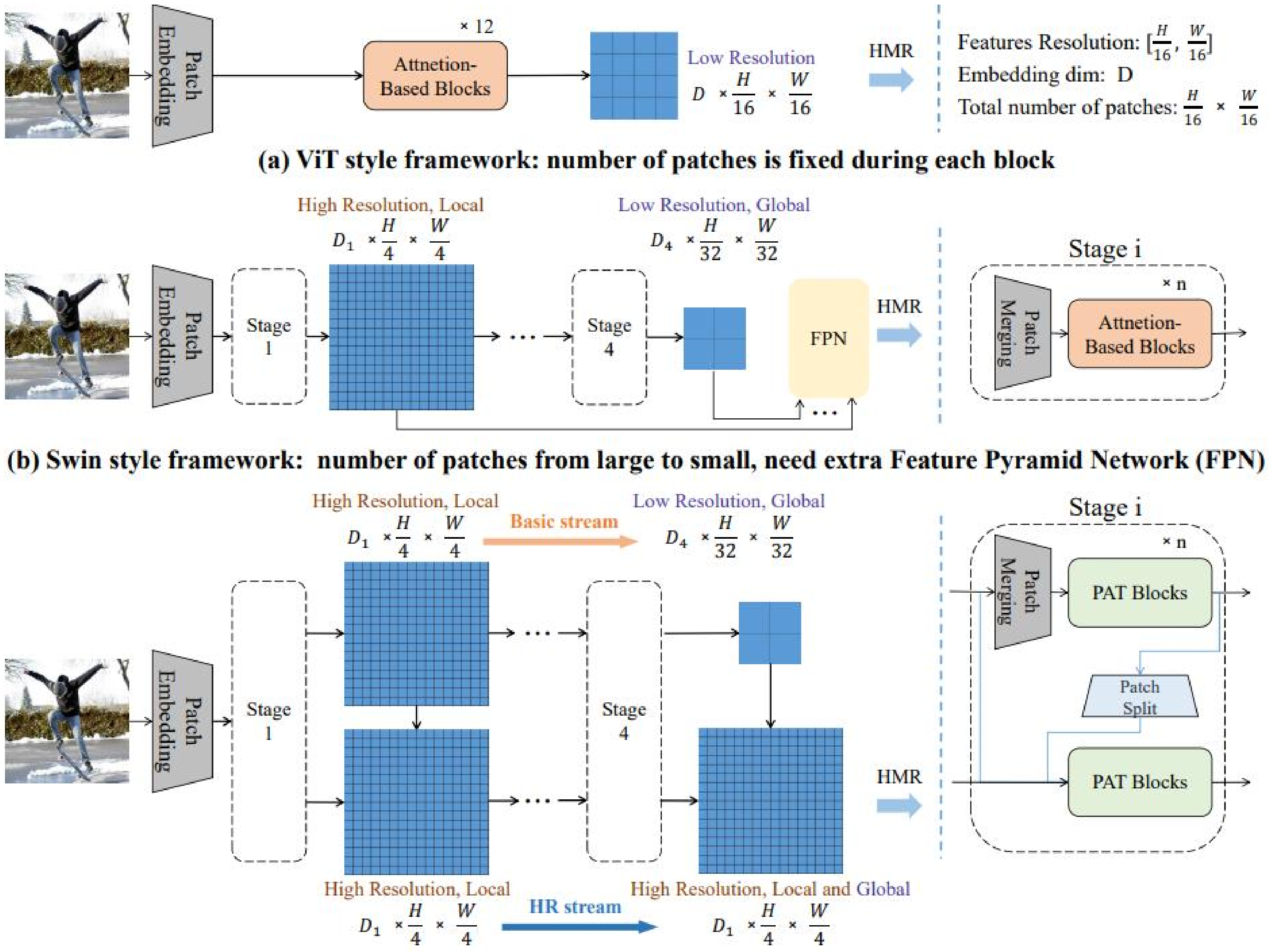
Module	Params	MACs
Attention	1.05 M	180 M
Spatial MLP	0.15 M	200 M
PoolAttn	0.02 M	4 M

(d) Parameters and MACs comparison of each block

Signidicantly reduce the Params and MACs significantly while maintaining high performance.

Introduction

Architecture design for HMR task



(c) POTTER: maintains high-resolution while capturing both local and global correlations for HMR

Overview of POTTER

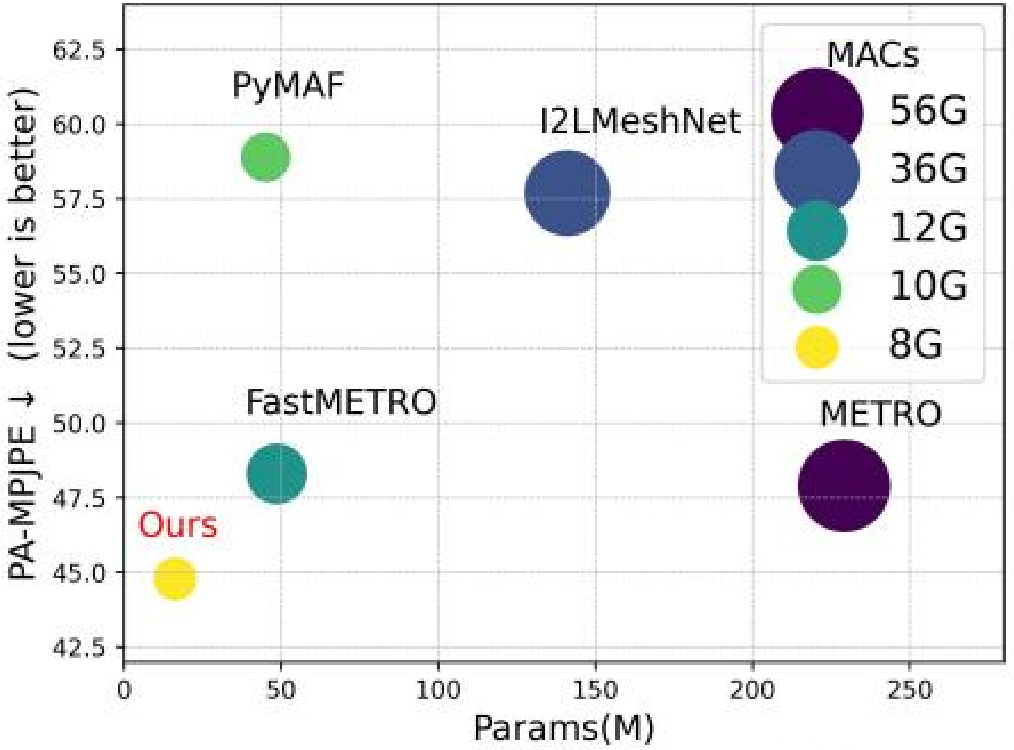


Figure 1. HMR performance comparison with Params and MACs on 3DPW dataset. We outperform SOTA methods METRO [18] and FastMETRO [3] with much fewer Params and MACs. PA-MPJPE is the Procrustes Alignment Mean Per Joint Position Error.

Overview of POTTER

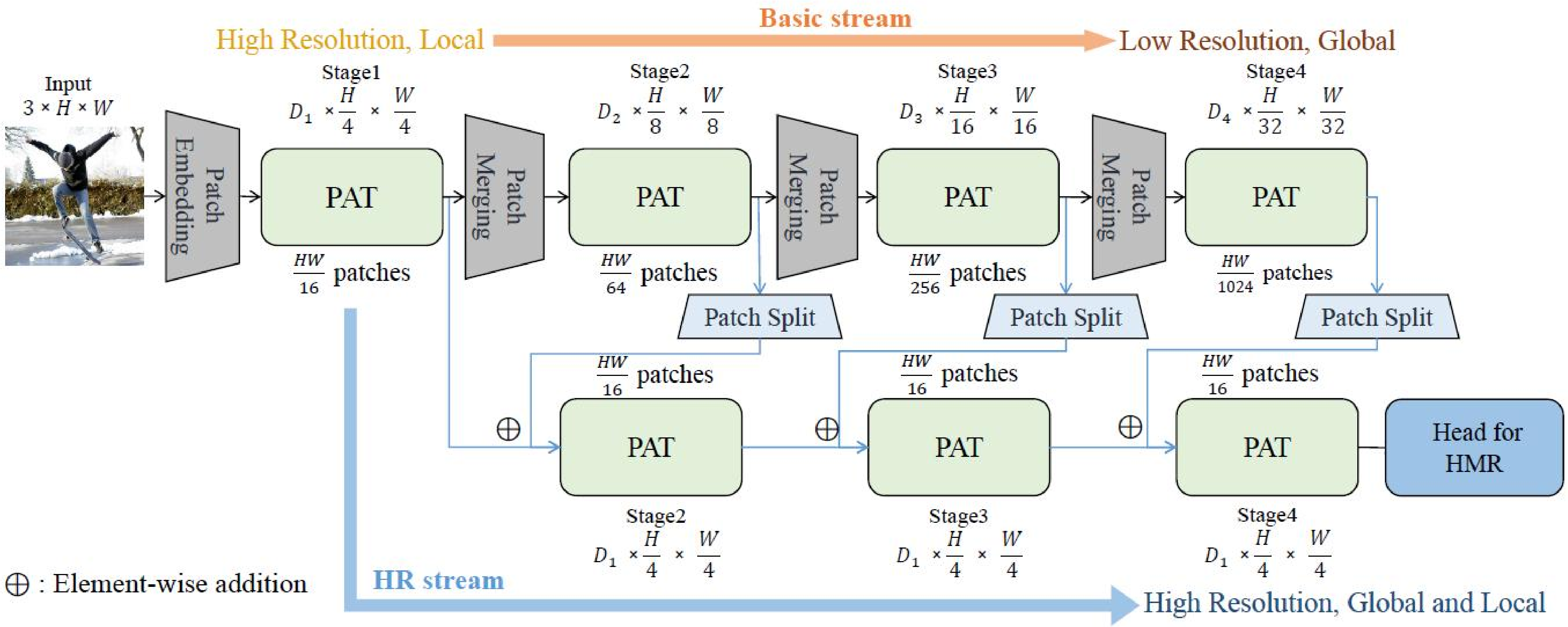


Figure 4. The overall architecture of our POTTER. PAT is our proposed Pooling Attention Transformer block. The basic stream of POTTER adopts hierarchical architecture with 4 stages [21], where the number of patches is gradually reduced for capturing more global information with low-resolution features ($\frac{H}{32} \times \frac{W}{32}$). Our proposed HR stream maintains the high-resolution ($\frac{H}{4} \times \frac{W}{4}$) feature representation at each stage. The global features from the basic stream are fused with the local features by patch split blocks in the HR stream. Thus, the high-resolution local and global features are utilized for the HMR task.

Experiment Results

Table 2. 3D Pose and Mesh performance comparison with SOTA methods on Human3.6M and 3DPW datasets. FastMETRO-S and FastMETRO-L are the FastMETRO using the small transformer encoder and large transformer encoder, respectively. * indicates that HybrIK uses ResNet34 as the backbone and with predicted camera parameters.

					Human3.6M		3DPW		
	Model	Year	Params(M)	Macs(G)	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPVE ↓
CNN-based	HMR [9]	CVPR 2018	-	-	88.0	56.8	130.0	76.7	-
	GraphCMR [14]	CVPR 2019	-	-	-	50.1	-	70.2	-
	SPIN [13]	ICCV 2019	-	-	62.5	41.1	96.9	59.2	116.4
	VIBE [12]	CVPR 2020	-	-	65.6	41.4	82.9	51.9	99.1
	I2LMeshNet [27]	ECCV 2020	140.5	36.6	55.7	41.1	93.2	57.7	-
	HybrIK* [16]	CVPR2021	27.6	12.7	57.3	36.2	75.3	45.2	87.9
	ProHMR [15]	ICCV 2021	-	-	-	41.2	-	59.8	-
	PyMAF [42]	ICCV 2021	45.2	10.6	57.7	40.5	92.8	58.9	110.1
	DSR [5]	ICCV 2021	-	-	60.9	40.3	85.7	51.7	99.5
	OCHMR [10]	CVPR 2022	-	-	-	-	89.7	58.3	107.1
Transformer -based	METRO [17]	CVPR 2021	229.2	56.6	54.0	36.7	77.1	47.9	88.2
	GTRS [44]	ACM MM 2022	71.5	3.8	64.3	45.4	88.5	58.9	106.2
	TCFormer [41]	CVPR 2022	-	-	62.9	42.8	80.6	49.3	-
	FastMETRO-S [3]	ECCV 2022	32.7	8.9	57.7	39.4	79.6	49.3	91.9
	FastMETRO-L [3]	ECCV 2022	48.5	11.8	**53.9**	37.3	77.9	48.3	90.6
	POTTER		16.3	7.8	56.5	**35.1**	**75.0**	**44.8**	**87.4**

Qualitative comparison with SOTA method METRO (in-the-wild images)

Generalization to 3D Hand Reconstruction

Figure 14. Qualitative results of our POTTER for reconstructing hand mesh.

Thanks for watching!