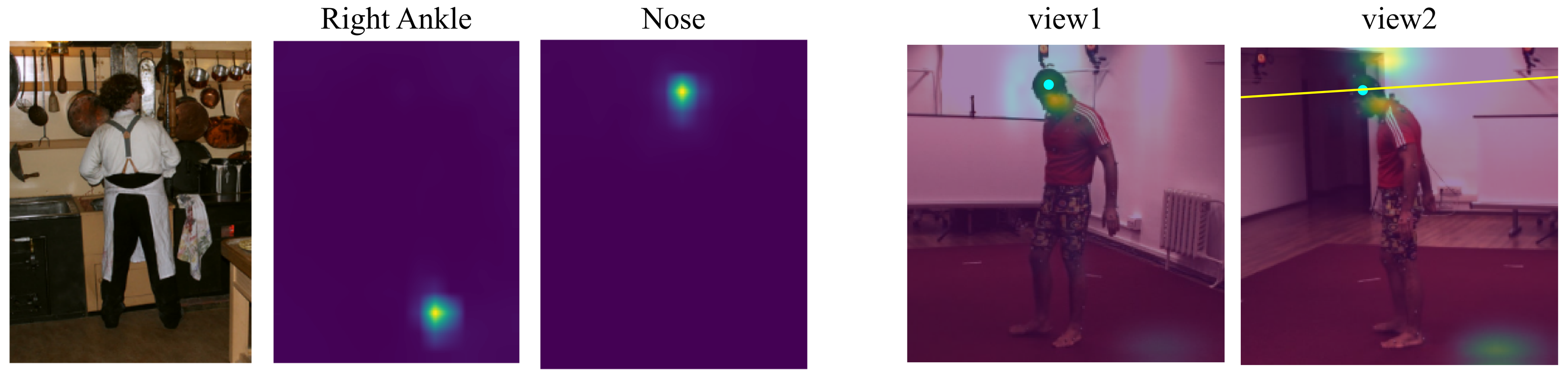
PPT: token-Pruned Pose Transformer for monocular and multi-view human pose estimation

Haoyu Ma[1], Zhe Wang[1], Yifei Chen[2], Deying Kong[1], Liangjian Chen[3], Xingwei Liu[1], Xiangyi Yan[1], Hao Tang[3], and Xiaohui Xie[1]
[1]University of California, Irvine, [2]Tencent, [3]Meta Reality Lab, Meta AI

Code is available!

➢ Motivation

- Dense global attention is computationally extensive
- Hard to scale up to high-resolution features and many views
- Attention map is very sparse in pose transformers [1][2]



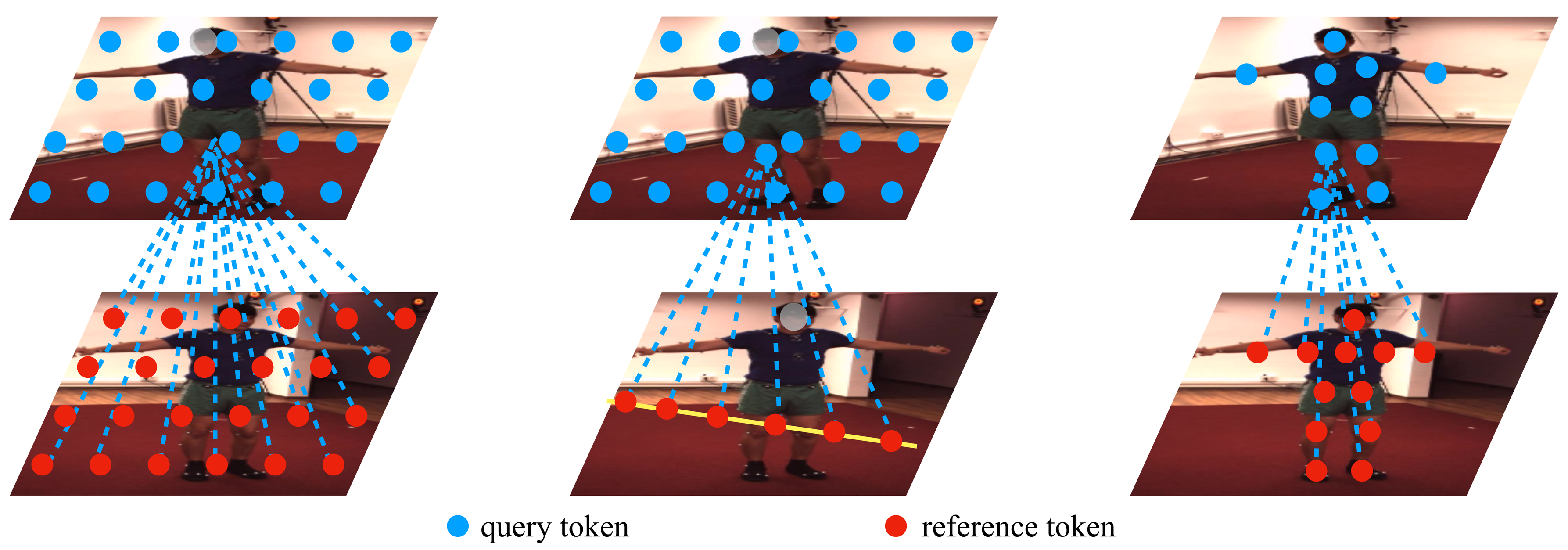
TokenPose TransFusion

➢ Methodology

- Human Token Identification
- Select informative visual tokens with the attention scores of keypoint tokens
- Attention value determines how much information of each visual token is fused into the output

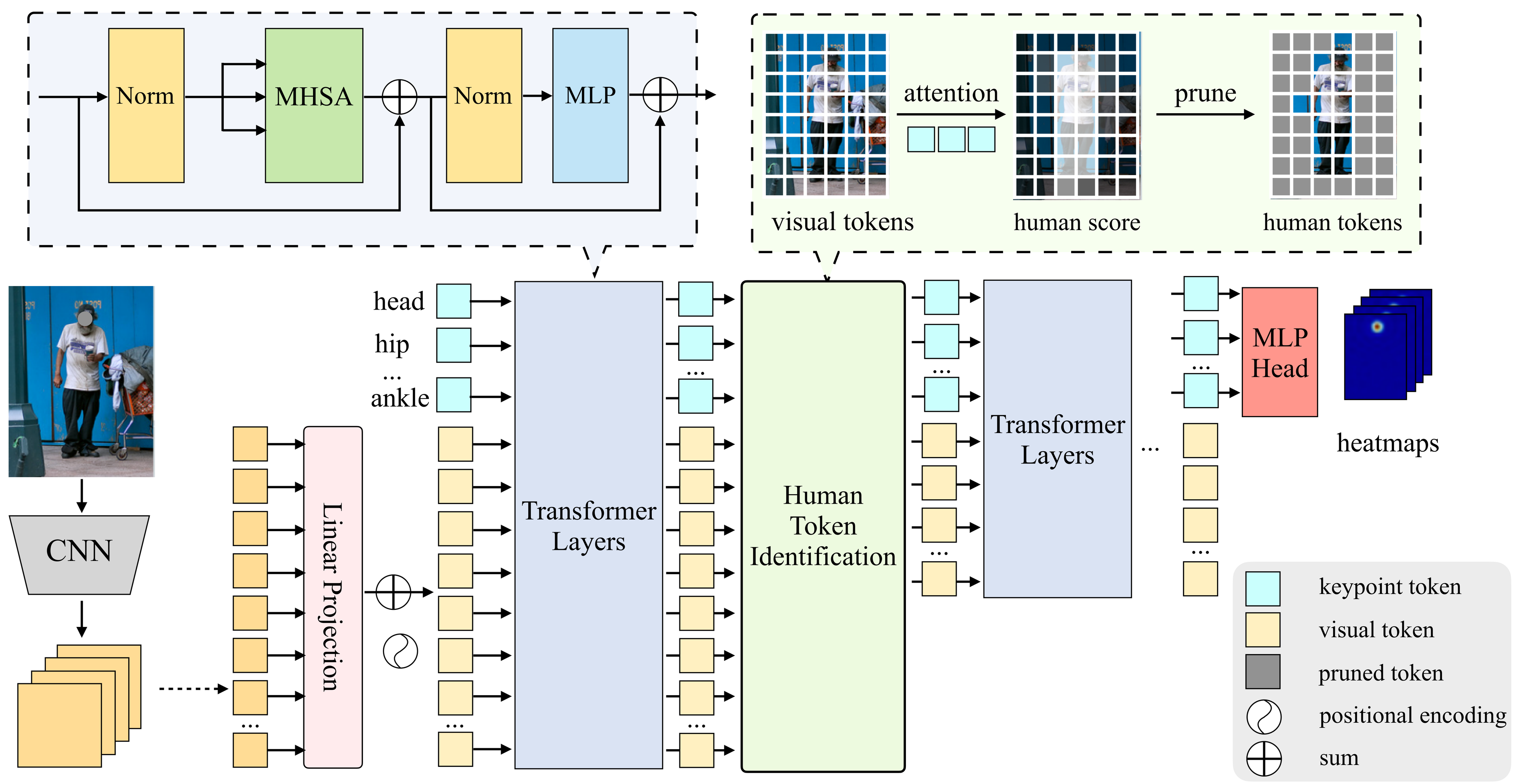
$$\mathrm{Softmax}\left(\frac{\mathbf{q}_k^j \mathbf{K}_v^T}{\sqrt{D}}\right)\mathbf{V}_v = \mathbf{a}^j \mathbf{V}_v$$

- Summary of cross-view Fusion Strategies
- Global Fusion: each pixel in each view calculates attention with respect to all n pixels of other m − 1 views, $O(m^2n^2)$
- Epipolar-based Fusion: each pixel in each view calculates attention with k pixels along the corresponded epipolar lines of other m − 1 views, $O(m^2nk)$
- **Human area fusion** (ours): dense global attention among k' human foreground pixels of m views, $O(m^2k'^2)$

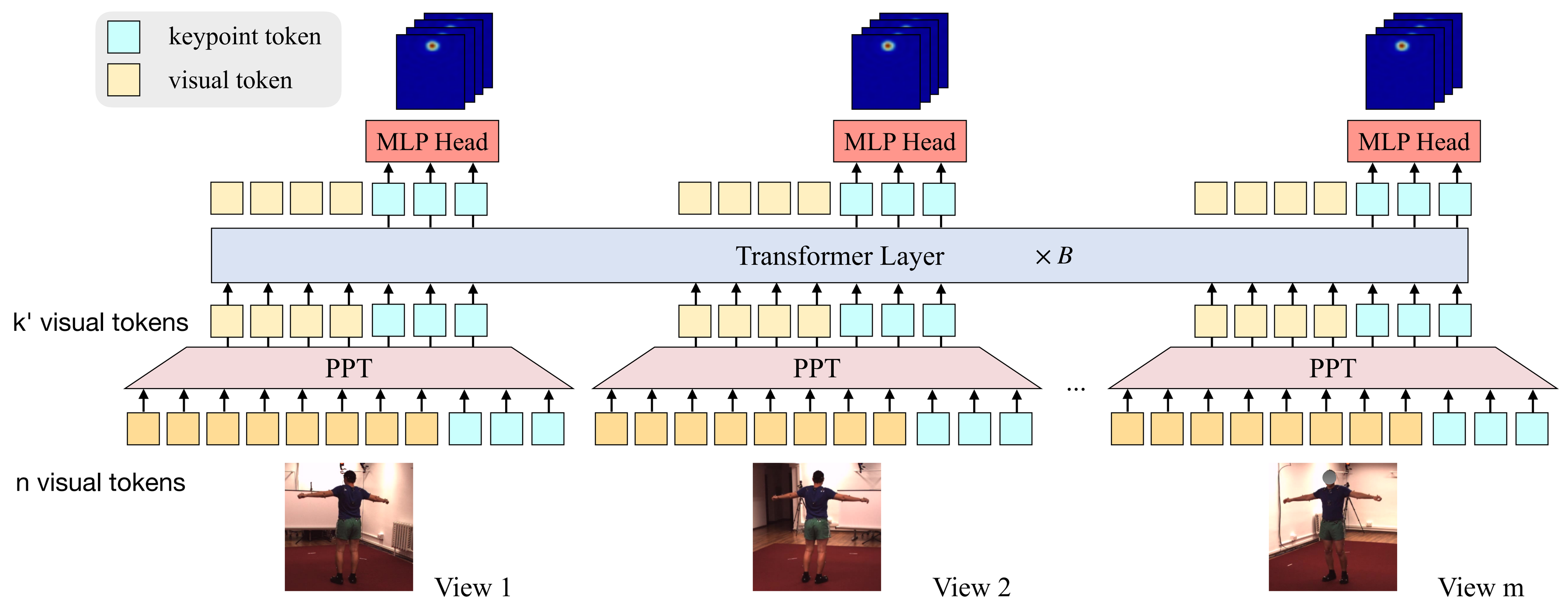


(a) Global fusion (b) Epipolar-based fusion (c) Human area fusion (ours)

• Token-Pruned Pose Transformer (PPT)

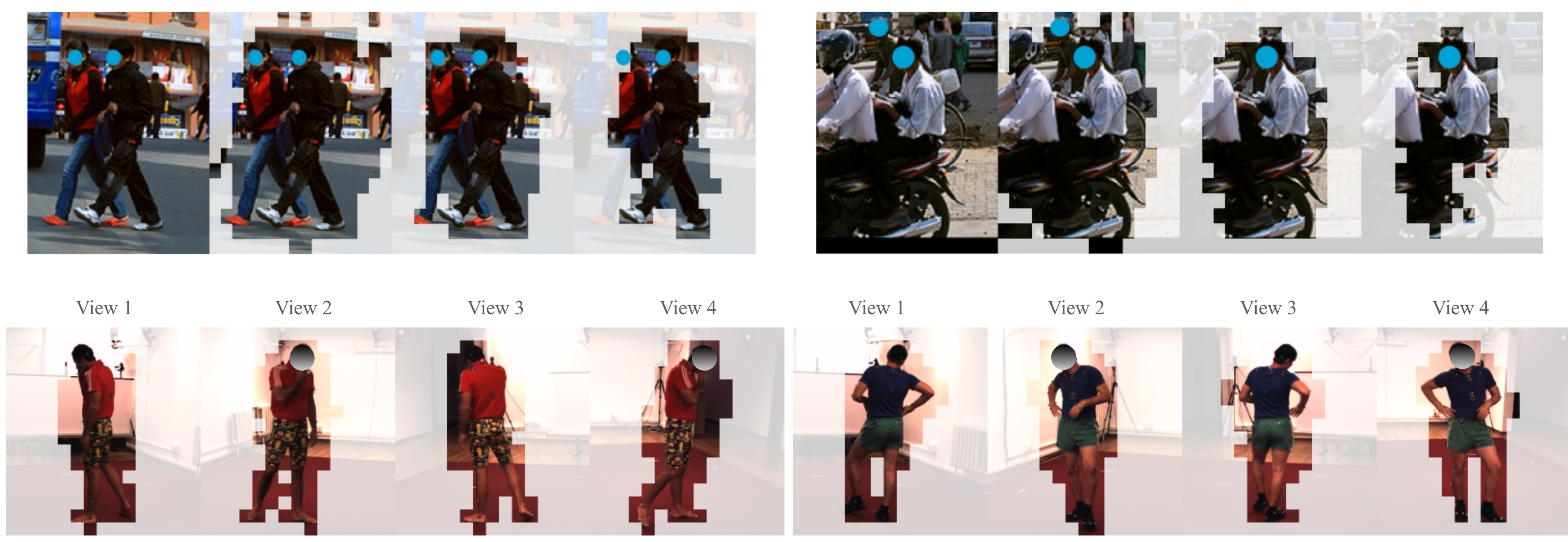


• Multi-view PPT



➢ Results

- selected tokens after each HTI module
- human areas are gradually refined as the network deepens



▪ Results on COCO

- PPT achieves significant acceleration while matching its accuracy on 2D pose estimation
- Pruning background tokens doesn't hurt the accuracy
- Attention among foreground tokens is sufficient

Method	#Params	GFLOPs	GFLOPs^T	AP	AP^{50}	AP^{75}	AP^M	AP^L	AR
SimpleBaseline-Res50 [63]	34M	8.9	-	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline-Res101 [63]	53M	12.4	-	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline-Res152 [63]	68.6M	15.7	-	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32 [50]	28.5M	7.1	-	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48 [50]	63.6M	14.6	-	75.1	90.6	82.2	71.5	81.8	80.4
Lite-HRNet-18 [69]	1.1M	0.20	-	64.8	86.7	73.0	62.1	70.5	71.2
Lite-HRNet-30 [69]	1.8M	0.31	-	67.2	88.0	75.0	64.3	73.1	73.3
EfficientPose-B [72]	3.3M	1.1	-	71.1	-	-	-	-	-
EfficientPose-C [72]	5.0M	1.6	-	71.3	-	-	-	-	-
TransPose-R-A4 [67]	6.0M	8.9	3.38	72.6	89.1	79.9	68.8	79.8	78.0
TransPose-H-S [67]	8.0M	10.2	4.88	74.2	89.6	80.8	70.6	81.0	79.5
TransPose-H-A6 [67]	17.5M	21.8	11.4	75.8	90.1	82.1	71.9	82.8	80.8
TokenPose-S [31]	6.6M	2.2	1.44	72.5	89.3	79.7	68.8	79.6	78.0
TokenPose-B [31]	13.5M	5.7	1.44	74.7	89.8	81.4	71.3	81.4	80.0
TokenPose-L/D6 [31]	20.8M	9.1	0.72	75.4	90.0	81.8	71.8	82.4	80.4
PPT-S (ours)	6.6M	1.6(**-27%**)	0.89(-38%)	72.2(**-0.3**)	89.0	79.7	68.6	79.3	77.8
PPT-B (ours)	13.5M	5.0(**-12%**)	0.89(-38%)	74.4(**-0.3**)	89.6	80.9	70.8	81.4	79.6
PPT-L/D6 (ours)	20.8M	8.7(**-4%**)	0.50(-31%)	75.2(**-0.2**)	89.8	81.7	71.7	82.1	80.4

Table 1. Results on COCO validation dataset. The input size is 256 × 192. GFLOPs^T means the GFLOPs for the transformers only following equations from [29], as our method only focus on accelerating the transformers.

▪ Results on Human3.6M

- Human area fusion is better than global attention in both accuracy and efficiency for multi-view pose

Method	#V	MACs	shlder	elb	wri	hip	knee	ankle	root	belly	neck	nose	head	Avg
ResNet50 [63]	1	51.7G	97.0	91.9	87.3	99.4	95.0	90.8	100.0	98.3	99.4	99.3	99.5	95.2
TransPose [67]	1	43.6G	96.0	92.9	88.4	99.0	95.0	91.8	100.0	97.5	99.0	99.4	99.6	95.3
TokenPose [31]	1	11.2G	96.0	91.3	85.8	99.4	95.2	91.5	100.0	98.1	99.1	99.4	99.1	94.9
Epipolar Transformer [19]	2	51.7G	97.0	93.1	91.8	99.1	96.5	91.9	100.0	99.3	99.8	99.8	99.3	96.3
TransFusion [36]	2	50.2G	97.2	96.6	93.7	99.0	96.8	91.7	100.0	96.5	98.9	99.3	99.5	96.7
Crossview Fusion [43]	4	55.1G	97.2	94.4	92.7	**99.8**	97.0	92.3	100.0	98.5	99.1	99.1	99.1	96.6
TokenPose+Transformers	4	11.5G	97.1	97.3	95.2	99.2	98.1	93.1	100.0	98.8	99.2	99.3	99.1	97.4
PPT	1	9.6G	96.0	91.8	86.5	99.2	95.6	92.2	100.0	98.4	99.3	99.5	99.4	95.3
Multi-view PPT	2	9.7G	97.1	95.5	91.9	99.4	96.4	92.1	100.0	99.0	99.2	99.3	99.0	96.6
Multi-view PPT	4	9.7G	97.6	**98.0**	96.4	99.7	98.4	93.8	100.0	99.0	**99.4**	**99.5**	**99.5**	**97.9**
Multi-view PPT + 3DPE	4	9.7G	**98.0**	**98.0**	96.4	99.7	**98.5**	**94.0**	**100.0**	**99.1**	99.2	99.4	99.3	**98.0**

Table 4. 2D pose estimation on Human3.6M. The metric is JDR on original image. All inputs are resized to 256 × 256. #V means the number of views used in cross-view fusion step. The FLOPs is the total computation for each view and cros-view fusion.

➢ Reference

- [1] Li, Yanjie, et al. "Tokenpose: Learning keypoint tokens for human pose estimation." ICCV 2021.
- [2] Ma, Haoyu, et al. "Transfusion: Cross-view fusion with transformer for 3d human pose estimation." BMVC 2021

Contact: haoyum3@uci.edu , Paper ID: 46