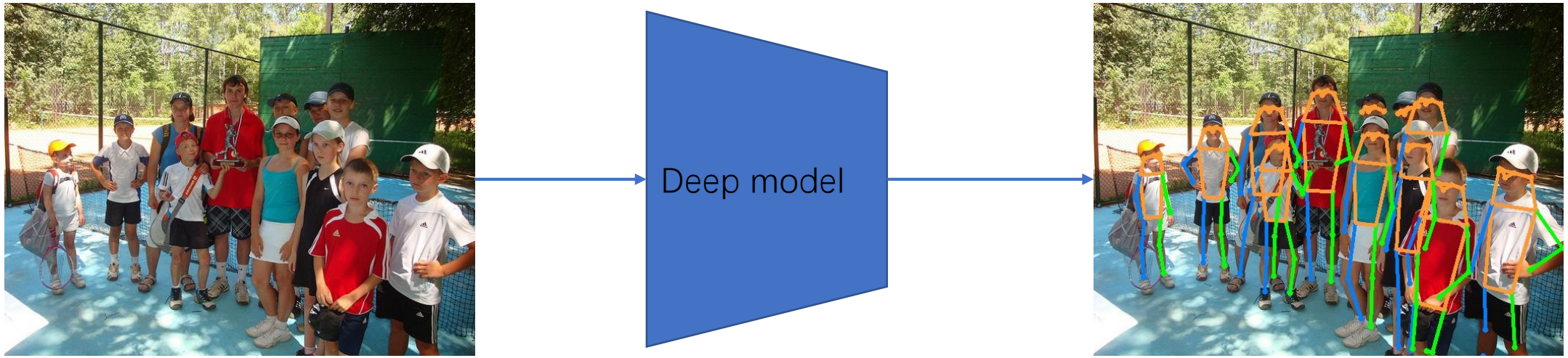
哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

图像人体2D姿态估计的研究

报告人：李浩然
导　师：姚鸿勋
2023年5月19日

人体2D姿态估计



人体2D姿态估计

Top-down

Person detector → Single Person pose estimation

Bottom-up

Joints detector → Joints association

Vilab
Visual Intelligence Lab

人体2D姿态估计

Top-down

Person detector → Single Person pose estimation

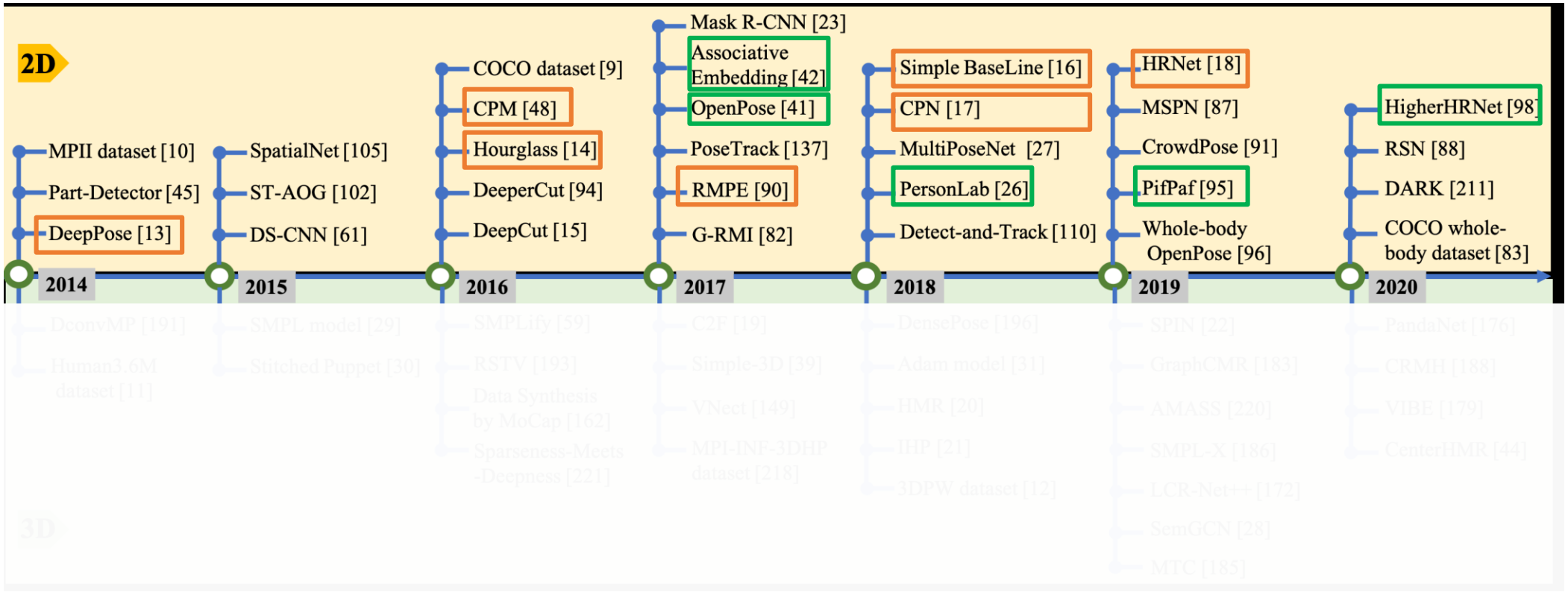
Bottom-up

Joints detector → Joints association

Single stage

Whole pose regression

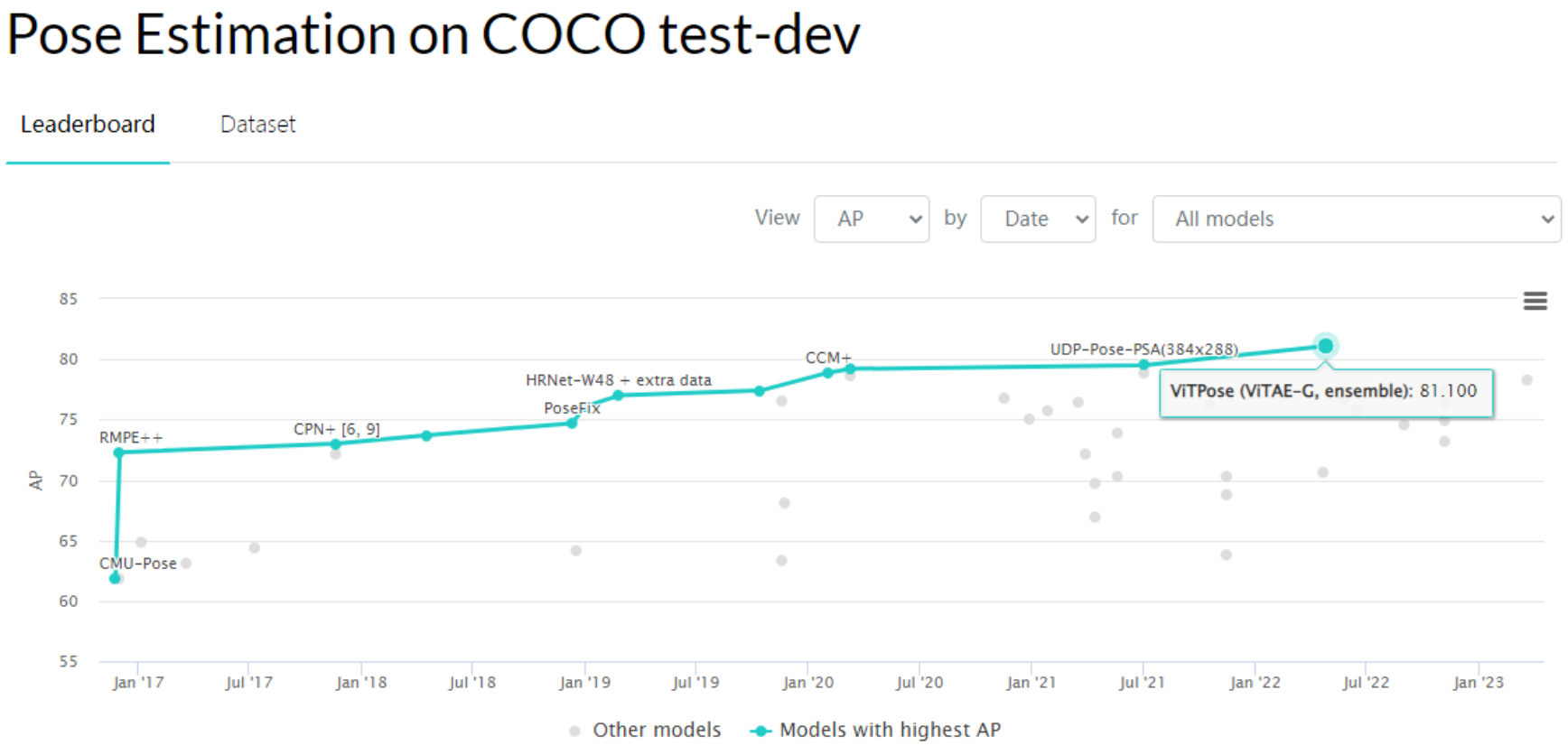
人体2D姿态估计



- Liu, Wu, et al.“ Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective. " *ACM Computing Surveys* 55.4 (2022): 1-41.

Vilab
Visual Intelligence Lab

Sotas



https://paperswithcode.com/sota/pose-estimation-on-coco-test-dev

Sotas top-down

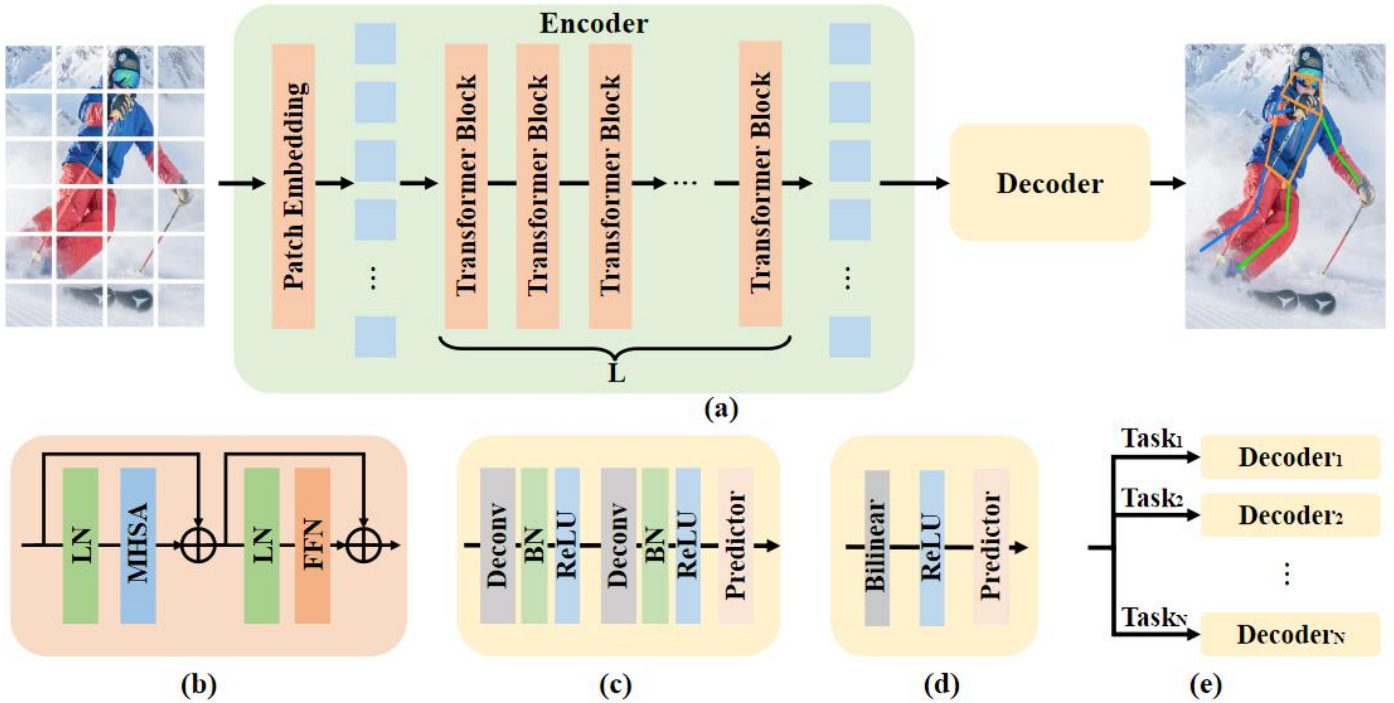
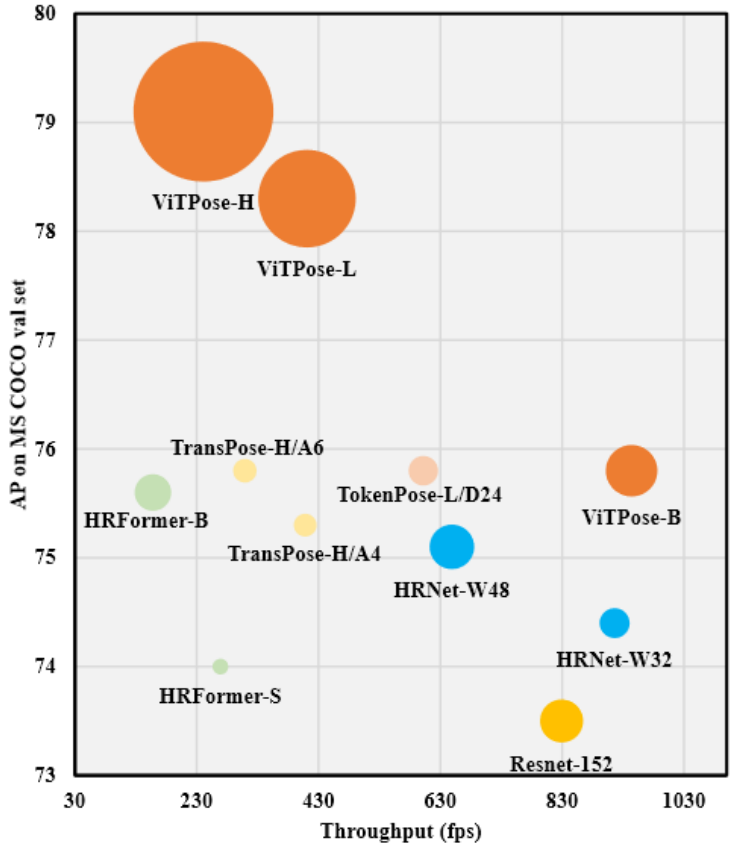


Figure 2: (a) The framework of ViTPose. (b) The transformer block. (c) The classic decoder. (d) The simple decoder. (e) The decoders for multiple datasets.



- Xu, Yufei, et al. "Vitpose: Simple vision transformer baselines for human pose estimation." *arXiv preprint arXiv:2204.12484* (2022).

Vilab
Visual Intelligence Lab

Sotas top-down

Table 9: Comparison of ViTPose and SOTA methods on MS COCO val set. * denotes the models are trained under the multi-dataset setting.

Model	Backbone	Params (M)	Speed (fps)	Input Resolution	Feature Resolution	COCO val AP	COCO val AR
SimpleBaseline [42]	ResNet-152	60	829	256x192	1/32	73.5	79.0
HRNet [36]	HRNet-W32	29	916	256x192	1/4	74.4	78.9
HRNet [36]	HRNet-W32	29	428	384x288	1/4	75.8	81.0
HRNet [36]	HRNet-W48	64	649	256x192	1/4	75.1	80.4
HRNet [36]	HRNet-W48	64	309	384x288	1/4	76.3	81.2
UDP [18]	HRNet-W48	64	309	384x288	1/4	77.2	82.0
TokenPose-L/D24 [27]	HRNet-W48	28	602	256x192	1/4	75.8	80.9
TransPose-H/A6 [44]	HRNet-W48	18	309	256x192	1/4	75.8	80.8
HRFormer-B [48]	HRFormer-B	43	158	256x192	1/4	75.6	80.8
HRFormer-B [48]	HRFormer-B	43	78	384x288	1/4	77.2	82.0
ViTPose-B	ViT-B	86	944	256x192	1/16	75.8	81.1
ViTPose-B*	ViT-B	86	944	256x192	1/16	77.1	82.2
ViTPose-L	ViT-L	307	411	256x192	1/16	78.3	83.5
ViTPose-L*	ViT-L	307	411	256x192	1/16	78.7	83.8
ViTPose-H	ViT-H	632	241	256x192	1/16	79.1	84.1
ViTPose-H*	ViT-H	632	241	256x192	1/16	79.5	84.5

- The speed of all methods is recorded on a single A100 GPU with a batch size of 64.

- Xu, Yufei, et al. "Vitpose: Simple vision transformer baselines for human pose estimation." *arXiv preprint arXiv:2204.12484* (2022).

Sotas bottom-up

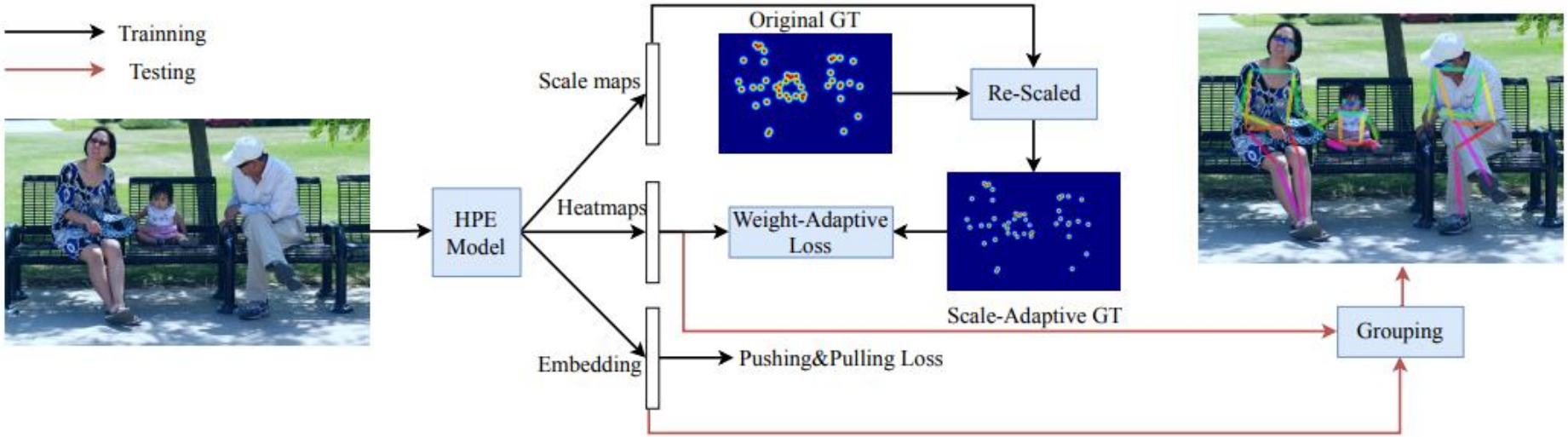


Figure 2. During training, the ground-truth heatmaps are firstly scaled according to predicted scale maps and then are used to supervise the whole model via weight-adaptive loss. During testing, the predicted heatmaps and associative embeddings are used for grouping of individual persons.

- Luo, Zhengxiong, et al. "Rethinking the heatmap regression for bottom-up human pose estimation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

Vilab
Visual Intelligence Lab

Sotas bottom-up

Methods	Backbone	Input Size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
w/o mutli-scale test									
OpenPose [4]	-	-	-	-	61.8	84.9	67.5	57.1	68.2
Hourglass [25]	Hourglass	512	277.8	206.9	56.6	81.8	61.8	49.8	67
PersonLab [27]	ResNet-152	1401	68.7	405.5	66.5	88.0	72.6	62.4	72.3
PifPaf [17]	-	-	-	-	66.7			62.4	72.9
HrHRNet [7]	HRNet-W32	512	28.5	47.9	66.4	87.5	72.8	61.2	74.2
HrHRNet [7] + SWAHR	HRNet-W32	512	28.6	48.0	67.9	88.9	74.5	62.4	75.5
HrHRNet [7]	HRNet-W48	640	63.8	154.3	68.4	88.2	75.1	64.4	74.2
HrHRNet [7] + SWAHR	HRNet-W48	640	63.8	154.6	**70.2**	**89.9**	**76.9**	**65.2**	**77.0**
w/ mutli-scale test									
Hourglass [25]	-	512	277.8	206.9	63.0	85.7	68.9	58.0	70.4
PersonLab [27]	-	1401	68.7	405.5	65.5	86.8	72.3	60.6	72.6
HrHRNet [7]	HRNet-W48	640	63.8	154.3	70.5	89.3	77.2	66.6	75.8
HrHRNet [7] + SWAHR	HRNet-W48	640	63.8	154.6	**72.0**	**90.7**	**78.8**	**67.8**	**77.7**

Table 1. Results on COCO test-dev2017. Top: without multi-scale test. Bottom: with multi-sale test (scale factors are 0.5, 1.0, and 1.5).

- Luo, Zhengxiong, et al. "Rethinking the heatmap regression for bottom-up human pose estimation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

Sotas single stage

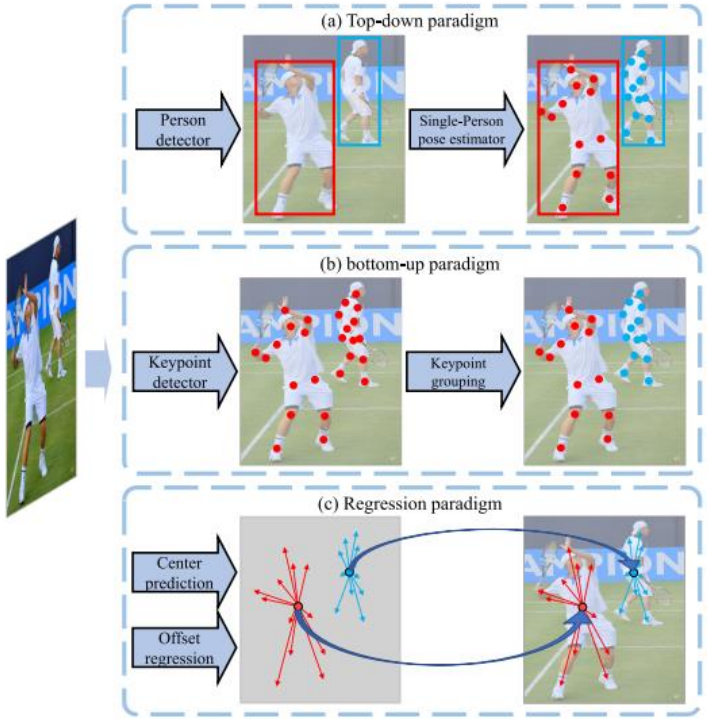


Fig. 1. Illustrations of different paradigms of multi-person pose estimation, (a) top-down paradigm, (b) bottom-up paradigm and (c) regression based paradigm.

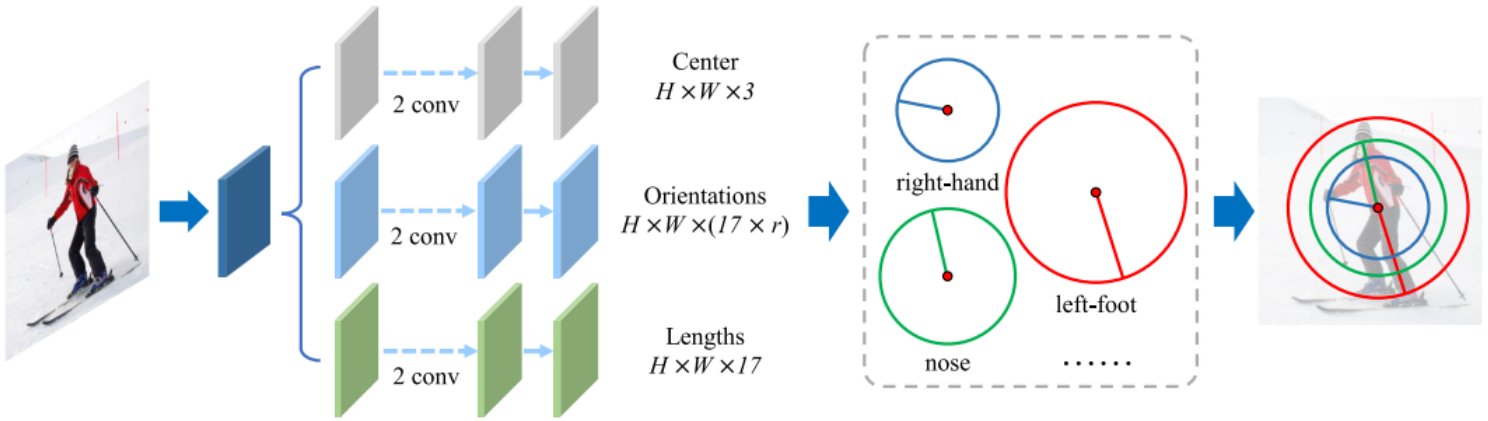


Fig. 3. Overview of the proposed PolarPose for multi-person pose estimation. PolarPose consists of a backbone feature extractor and three head branches for person center detection, orientation classification and length classification, respectively. H and W denote the height and width of feature maps, r is the number of quantized orientations. 17 equals to the number of keypoints annotated in the COCO dataset. Regression result of each keypoint is represented as an orientation and a length.

- Li, Jianing, Yaowei Wang, and Shiliang Zhang. "PolarPose: Single-Stage Multi-Person Pose Estimation in Polar Coordinates." *IEEE Transactions on Image Processing* 32 (2023): 1108-1119.

Sotas single stage

TABLE VI
COMPARISONS WITH HEATMAP BASED AND REGRESSION BASED METHODS ON THE COCO2017 TEST-DEV SET

Method	Backbone	Input size	FPS	single-scale test					multi-scale test				
				AP	AP_{50}	AP_{75}	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_M	AP_L
Heatmap based													
Mask-RCNN [2]	ResNet50+FPN	800	14.2	63.1	87.3	68.7	57.8	71.4	-	-	-	-	-
RMPE [3]	-	-	10.6	61.8	83.7	69.8	58.6	67.6	-	-	-	-	-
Openpose* [7]	-	-	10.1	61.8	84.9	67.5	57.1	68.2	-	-	-	-	-
HGG [26]	Hourglass	800	-	60.4	83.0	66.2	84.0	69.8	67.6	85.1	73.7	62.7	74.6
AE [4]	Hourglass	512	8.1	56.6	81.8	61.8	49.8	67.0	63.0	85.7	68.9	58.0	70.4
AE* [4]	Hourglass	512	8.1	62.8	84.6	69.2	57.5	70.6	65.5	86.8	72.3	60.6	72.6
PersonLab [24]	ResNet152	1401	5.0	66.5	88.0	72.6	62.4	72.3	68.7	89.0	75.4	64.1	75.5
Bottom-up HRNet [5]	HRNet-W32	512	6.5	64.1	86.3	70.4	57.4	73.9	-	-	-	-	-
PifPaf [25]	ResNet152	-	-	66.7	-	-	62.4	72.3	-	-	-	-	-
FCPose [41]	ResNet101	800	15.2	65.6	87.9	72.6	62.1	72.3	-	-	-	-	-
HigherHRNet [5]	HRNet-W32	512	6.0	66.4	87.5	72.8	61.2	74.2	-	-	-	-	-
HigherHRNet [5]	HRNet-W48	640	4.9	68.4	88.2	75.1	64.4	74.2	70.5	89.3	77.2	66.6	75.8
Regression based													
CenterNet [8]	Hourglass104	512	13.1	55.0	83.5	59.7	49.4	64.0	-	-	-	-	-
CenterNet-jd [8]	Hourglass104	512	12.2	63.0	86.8	69.6	58.9	70.4	-	-	-	-	-
LS-Net [11]	ResNeXt101-DCN	-	5.9	59.0	83.6	65.2	53.3	67.9	-	-	-	-	-
SPM [9]	Hourglass	512	17.2	-	-	-	-	-	66.9	88.5	72.9	62.6	73.1
Integral [18]	ResNet101	-	-	67.8	88.2	74.8	63.9	74.0	-	-	-	-	-
End2end PRTR [28]	ResNet101	-	-	63.4	86.2	69.4	59.3	72.0	-	-	-	-	-
End2end PRTR [28]	HRNet-W48	-	-	64.9	87.0	71.7	60.2	72.5	-	-	-	-	-
PSA [10]	HRNet-W48	800	4.8	66.3	87.7	73.4	64.9	70.0	68.7	89.9	76.3	64.8	75.3
PSA+RLE [42]	HRNet-W48	800	4.8	67.4	87.5	73.9	-	-	-	-	-	-	-
DEKR [14]	HRNet-W32	512	5.1	67.3	87.9	74.1	61.5	76.1	69.8	89.0	76.6	65.2	76.5
DEKR [14]	HRNet-W48	640	4.2	70.0	89.4	77.3	65.7	**76.9**	71.0	89.2	78.0	67.1	**76.9**
PolarPose	HRNet-W32	512	**24.5**	67.5	88.0	74.5	61.7	75.7	68.9	88.3	75.7	65.0	75.2
PolarPose	HRNet-W48	640	21.5	**70.2**	**89.5**	**77.5**	**66.1**	76.4	**71.3**	**89.5**	**78.4**	**67.3**	76.6

* indicates using extra single-person pose estimation model for refinement.

- Li, Jianing, Yaowei Wang, and Shiliang Zhang. "PolarPose: Single-Stage Multi-Person Pose Estimation in Polar Coordinates." *IEEE Transactions on Image Processing* 32 (2023): 1108-1119.

Sotas single stage

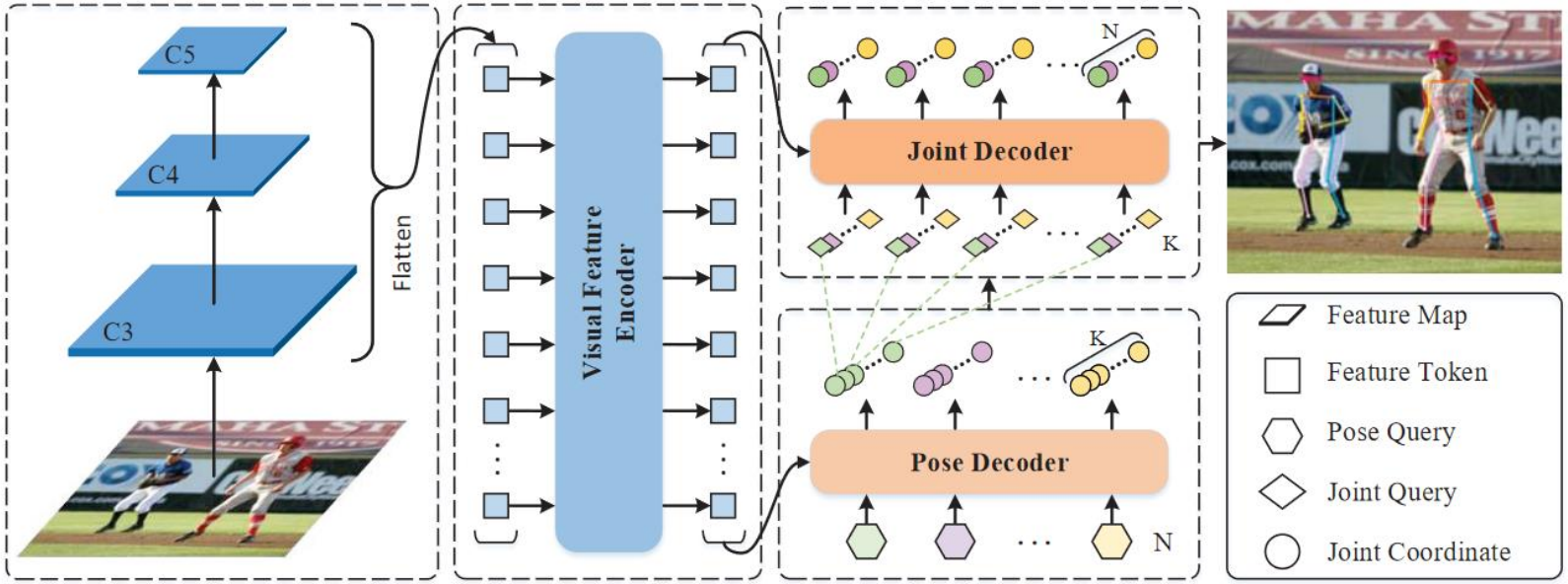


Figure 2. **The overall architecture of PETR**. C3 to C5 are multi-scale feature maps extracted from the backbone network (*e.g.*, ResNet-50). The visual feature encoder takes the flattened image features as inputs and refines them. Given N pose queries and the refined multi-scale feature tokens, pose decoder predicts N full-body poses in parallel. After that, an additional joint decoder takes each scattered pose (*i.e.*, kinematic joints of each pose) as its reference points and outputs the refined pose as final results. K is the number of keypoints for each instance (*e.g.*, $K = 17$ in COCO [21] dataset).

- Shi, Dahu, et al. "End-to-end multi-person pose estimation with transformers." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Sotas single stage

	Single-stage methods							
Non end-to-end	DirectPose [34]	ResNet-50	62.2	86.4	68.2	56.7	69.8	74
	FCPose [26]	ResNet-50	64.3	87.3	71.0	61.6	70.5	68
	InsPose [32]	ResNet-50	65.4	88.9	71.7	60.2	72.7	80
	DirectPose [34]	ResNet-101	63.3	86.7	69.4	57.8	71.2	-
	FCPose [26]	ResNet-101	65.6	87.9	72.6	62.1	72.3	93
	InsPose [32]	ResNet-101	66.3	89.2	73.0	61.2	73.9	100
	CenterNet [41]	Hourglass-104	63.0	86.8	69.6	58.9	70.4	160
	Point-Set Anchors†‡ [36]	HRNet-w48	68.7	89.9	76.3	64.8	75.3	-
Fully end-to-end	PETR (Ours)	ResNet-50	67.6	89.8	75.3	61.6	76.0	89
	PETR‡ (Ours)	ResNet-50	69.2	90.5	77.1	64.2	76.4	-
	PETR (Ours)	ResNet-101	68.5	90.3	76.5	62.5	77.0	95
	PETR‡ (Ours)	ResNet-101	70.0	90.9	78.2	65.3	77.1	-
	PETR (Ours)	Swin-L	70.5	91.5	78.7	65.2	78.0	133
	PETR‡ (Ours)	Swin-L	71.2	91.4	79.6	66.9	78.0	-

- Shi, Dahu, et al. "End-to-end multi-person pose estimation with transformers." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

轻量化模型

Top-down

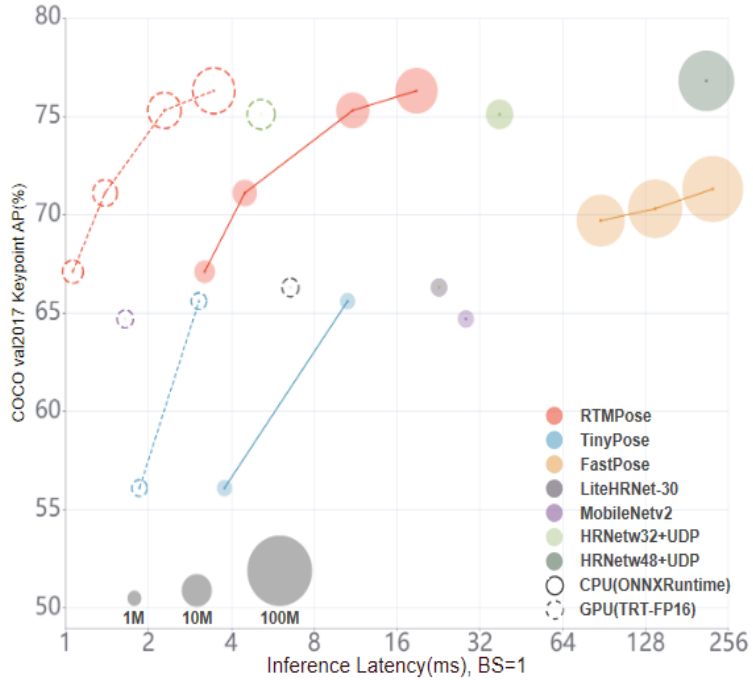


Figure 1. Comparison of RTMPose and open-source libraries on COCO val set regarding model size, latency, and precision. The circle size represents the relative size of model parameters.

- Jiang, Tao, et al. "RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose." arXiv preprint arXiv:2303.07399 (2023).

Bottom-up

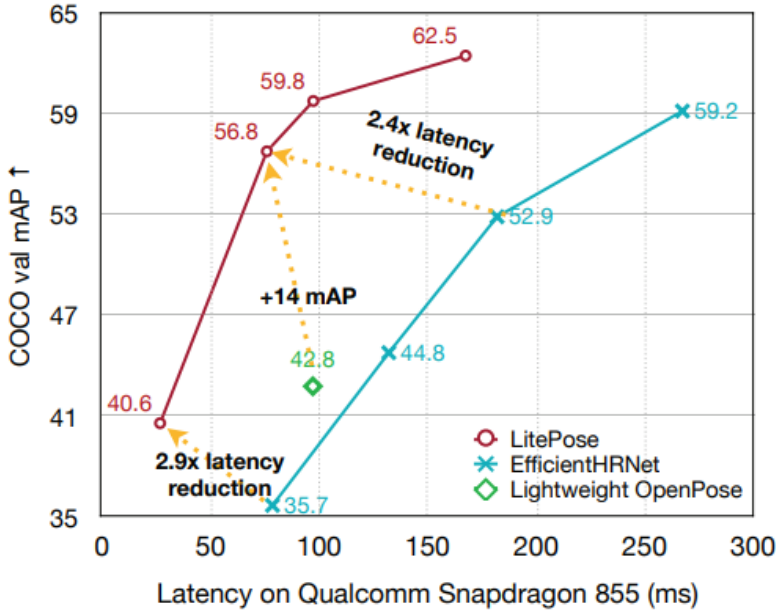


Figure 1. LitePose provides up to 2.9× latency reduction compared to EfficientHRNet [36] on Qualcomm Snapdragon 855 while achieving higher mAP on COCO. Compared with Lightweight OpenPose [39], LitePose obtains 14% higher mAP on COCO with lower latency.

- Wang, Yihan, et al. "Lite pose: Efficient architecture design for 2d human pose estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

谢谢！