DistilPose: Tokenized Pose Regression with Heatmap Distillation

Suhang Ye[1*], Yingyi Zhang[2*], Jie Hu[1*], Liujuan Cao[1], Shengchuan Zhang[1✉],
Lei Shen[2], Jun Wang[3], Shouhong Ding[2], Rongrong Ji[1]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, [2]Tencent Youtu Lab, [3]Tencent WeChat Pay Lab33

*Equal contribution

Tencent YouTu Lab

WeChat Pay

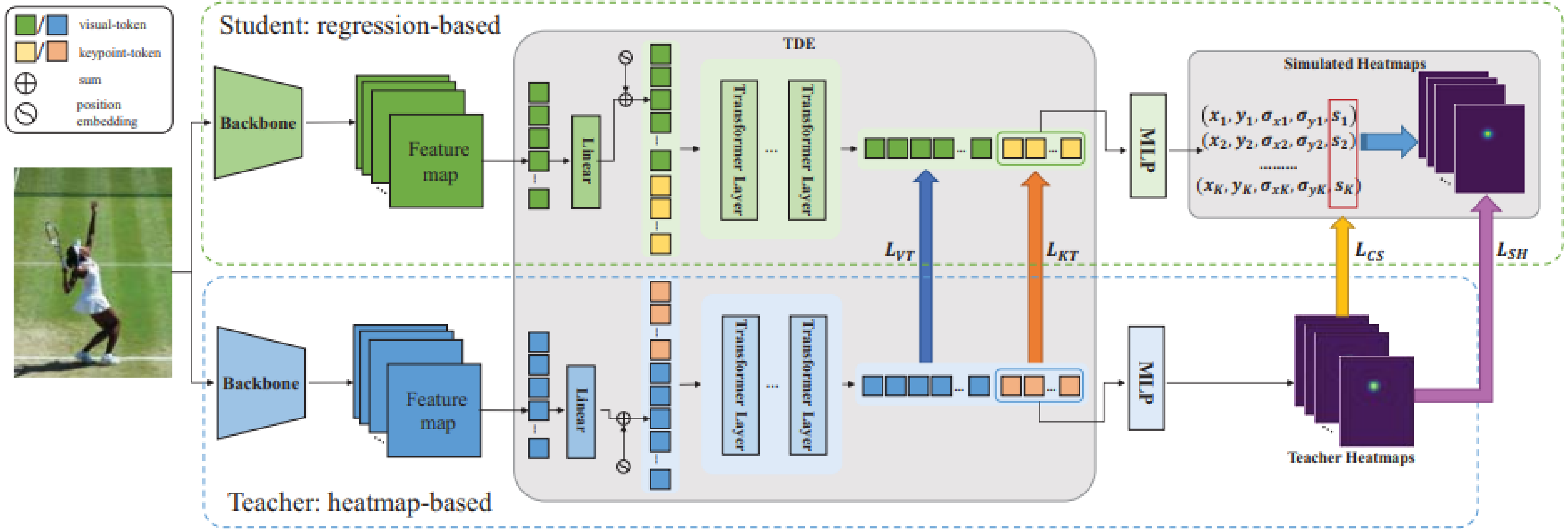
Paper Tag: TUE-AM-207

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Overview

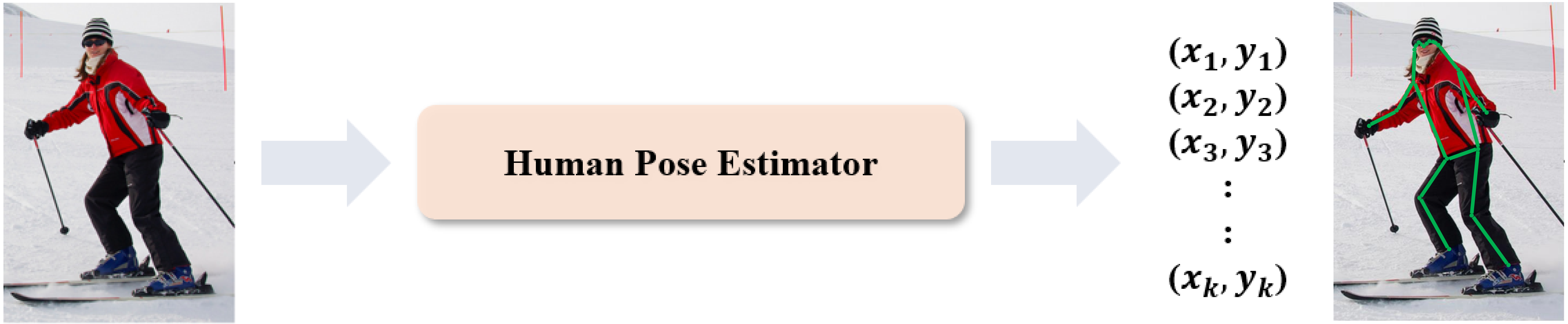
Code Paper

- Task: 2D Human Pose Estimation.
- Heatmap to Regression distillation !
- Greatly boost performance of regression-based student.



Human Pose Estimation

- 2D Human Pose Estimation
  - Aims to detect the ***coordinates*** of human's anatomical joints in a given image.
  - Multi-person : Top-Down.



Human Pose Estimator

(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
:
:
(x_k, y_k)

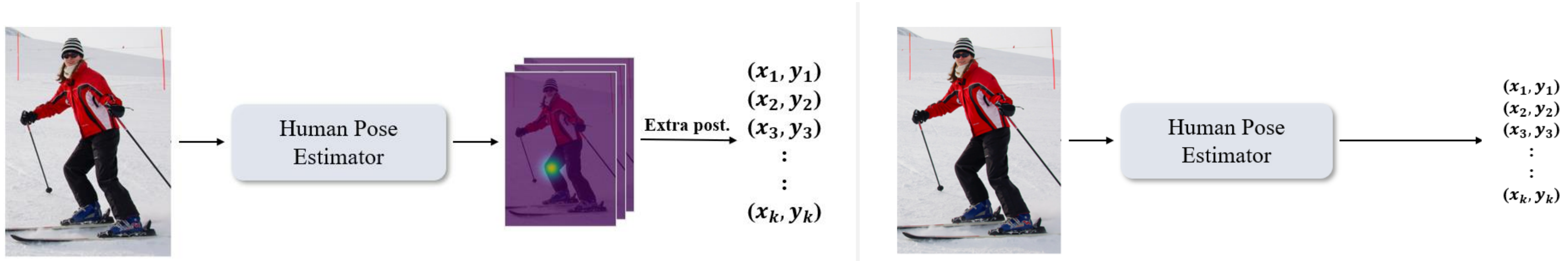
Input Image (Single Person, Fixed Resolution)

Human Keypoints Coordinates

Background

- Heatmap-based Methods
  - **Pros**
  - **Cons**
- Regress
  - **Pros**
  - **Cons**

Pros in both speed and accuracy,
Pros in both information and structure

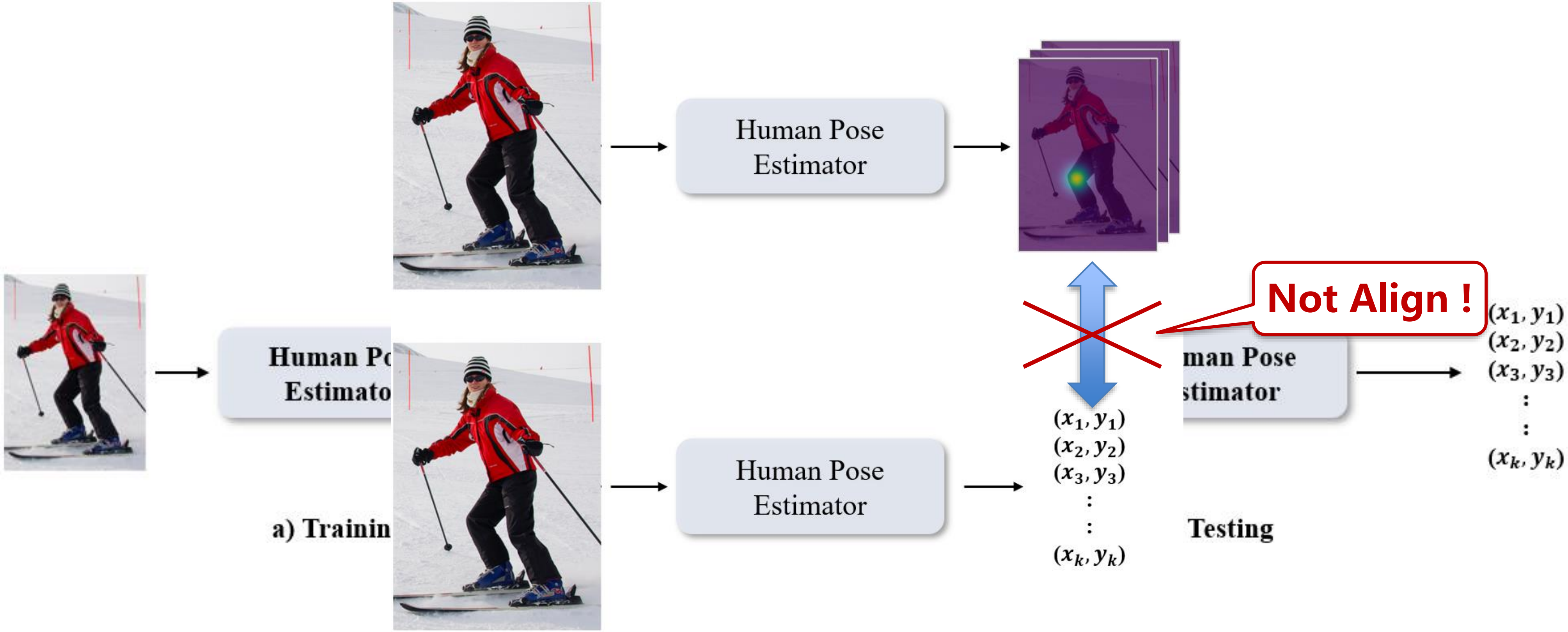


a) Heatmap-based Methods

b) Regression-based Methods

Previous Works

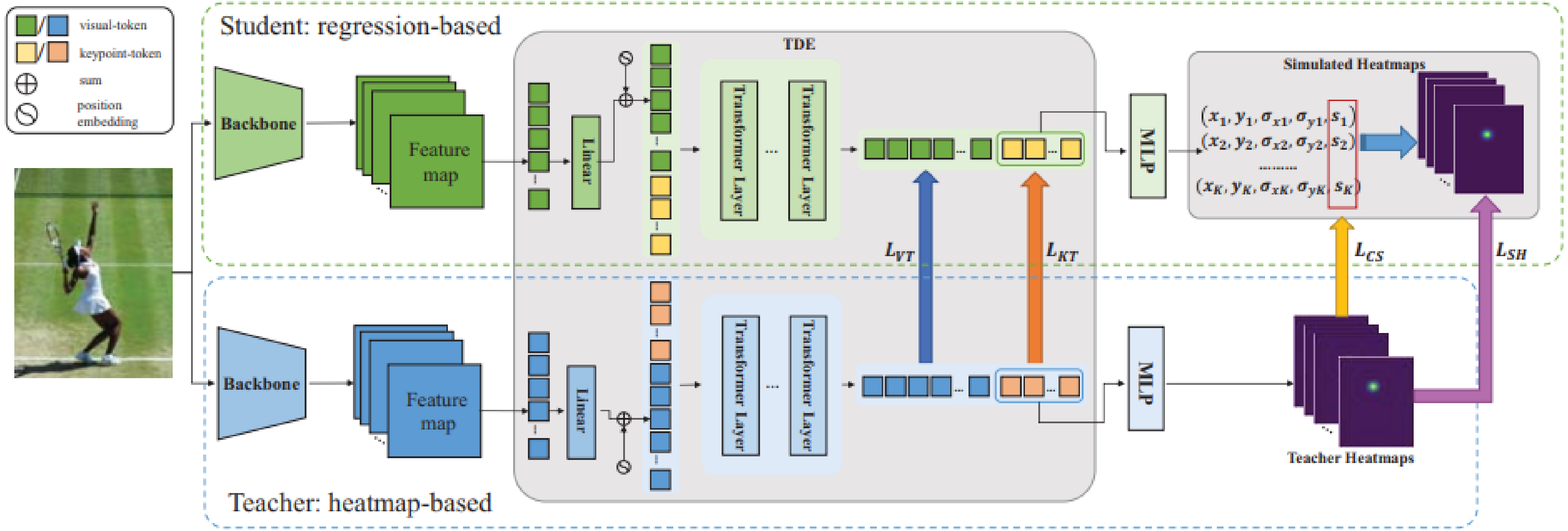
- Heatmap-based Pretrained Model[1]
- Heatmap Auxiliary Training[2]



[1] Li J, Bian S, Zeng A, et al. Human pose regression with residual log-likelihood estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11025-11034.
[2] Tian Z, Chen H, Shen C. Directpose: Direct end-to-end multi-person pose estimation[J]. arXiv preprint arXiv:1911.07451, 2019.

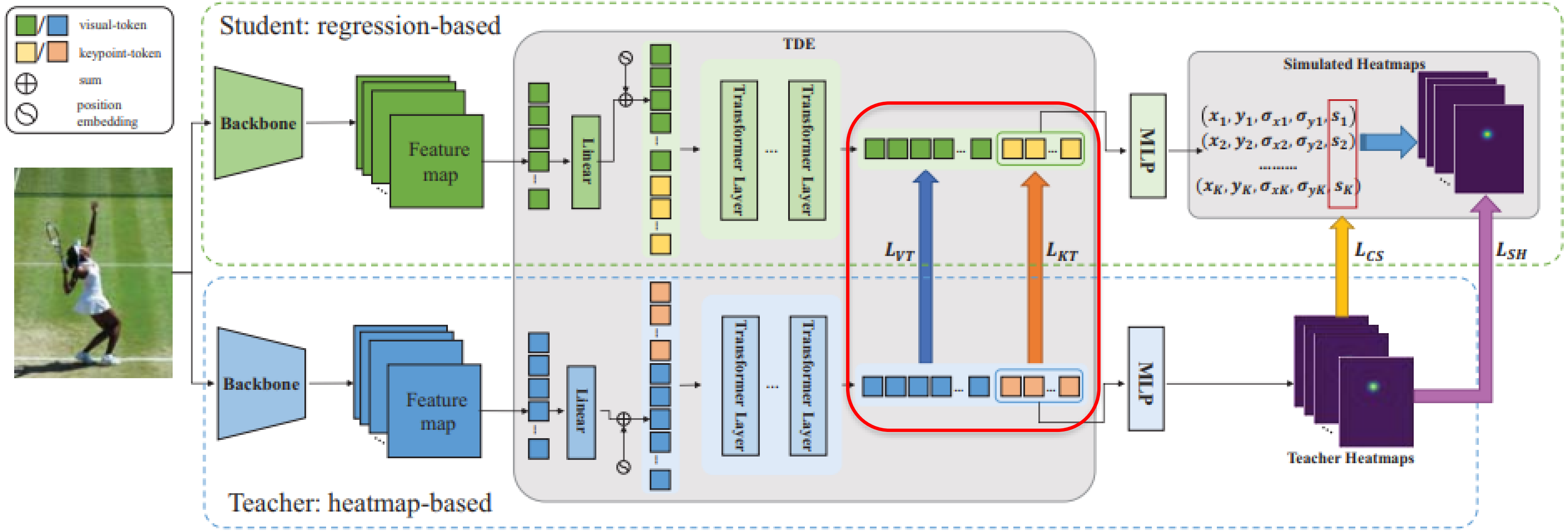
Framework

- **Teacher** : heatmap-based model
- **Student** : regression-based model
- **Distillation**:
    - ❑ Token-Distilling Encoder (TDE)
    - ❑ Simulated Heatmaps



DistilPose

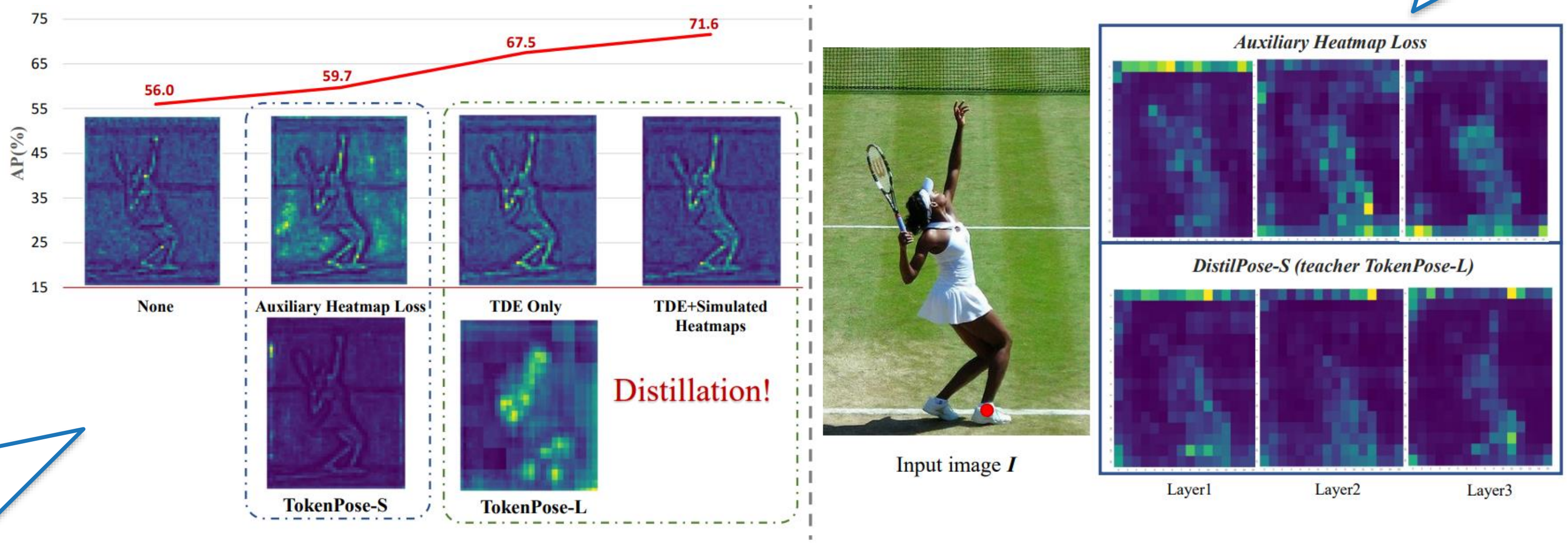
- Token-Distilling Encoder (TDE)
  - ViT-like structure.
  - Tokenization, align feature space.



DistilPose

- Token-Distilling Encoder (TDE)
  - ViT-like structure.
  - Tokenization, align feature space.

TDE can learn the relationship between keypoint-tokens and visual-tokens of the corresponding position



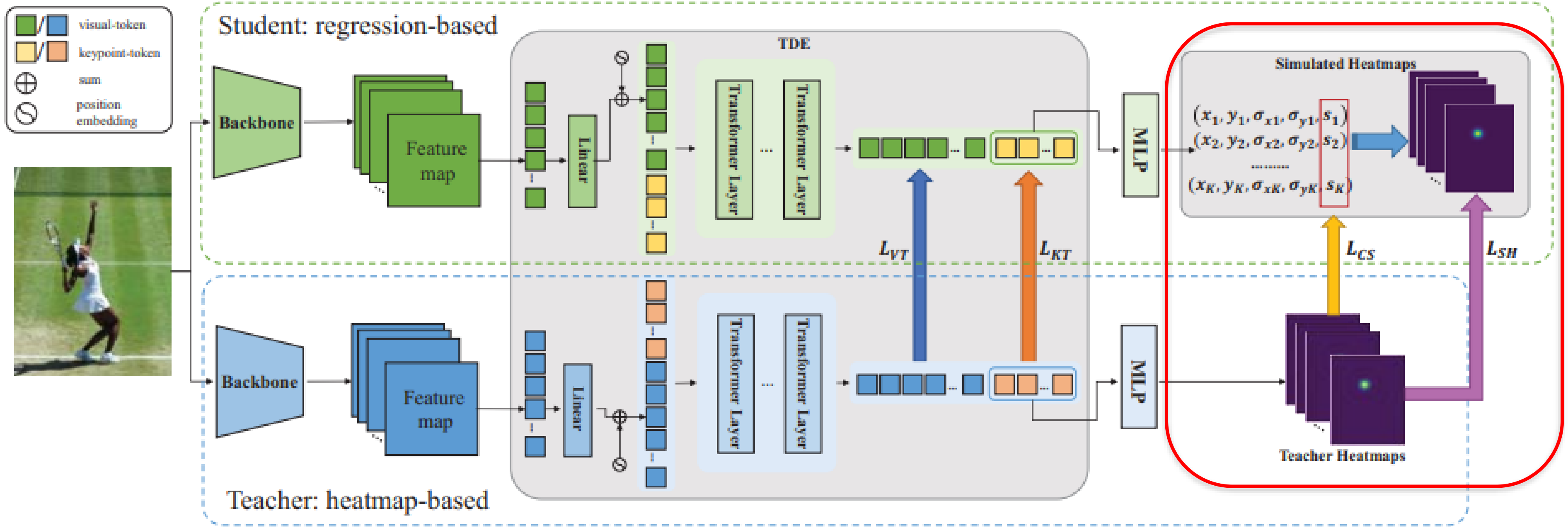
(a) Backbone Feature Map

(b) Attention Matrix

Student learns information more focused on human body itself, and achieves higher performance

DistilPose

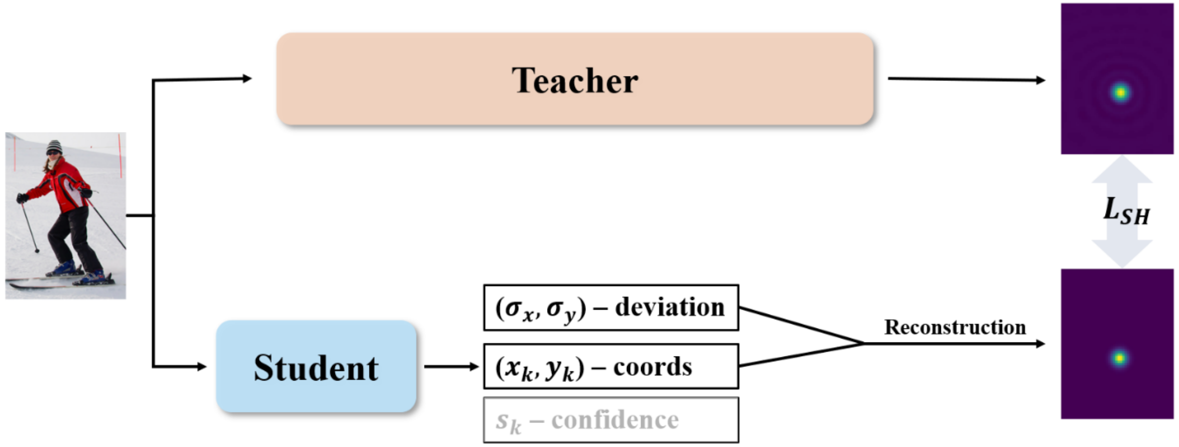
- Simulated Heatmaps
  - Basic Distribution Simulation
  - Confidence Distillation

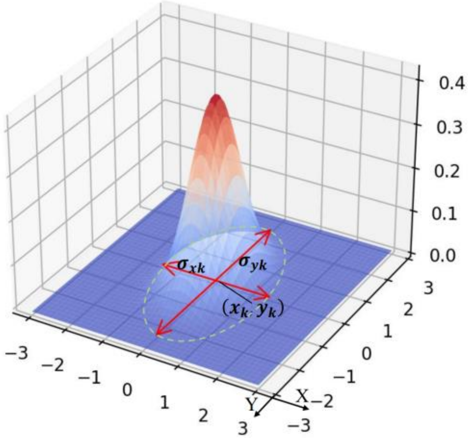


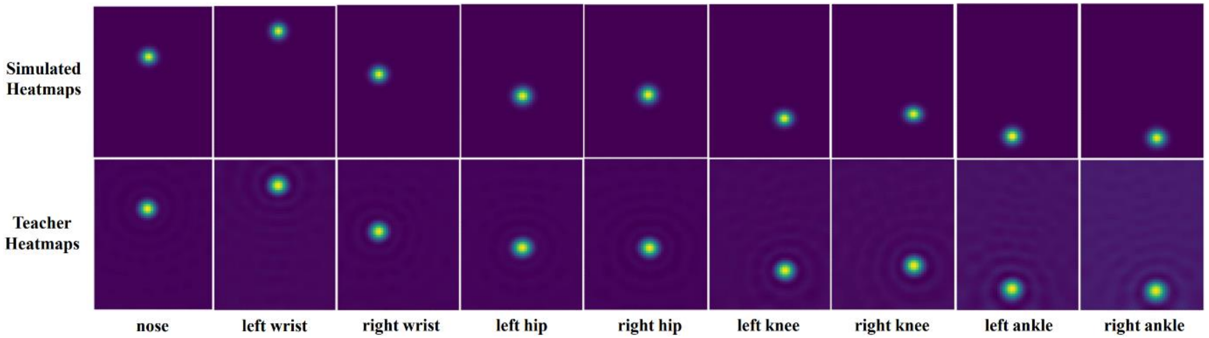
DistilPose-Simulated Heatmaps

Basic Distribution Simulation

- Aims to learn the distribution information contained in teacher heatmaps.



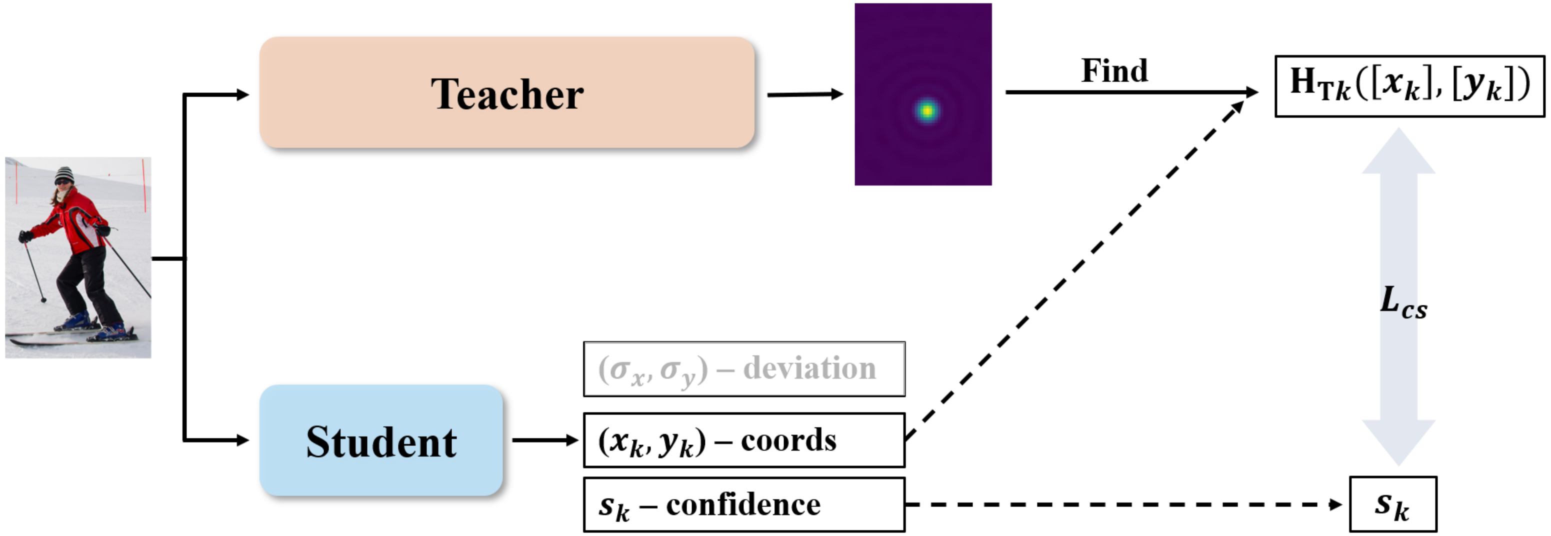




$$H_k(x,y) = e^{-\frac{1}{2}\left(\frac{(x-\mu_{xk})^2}{\sigma_{xk}^2}+\frac{(y-\mu_{yk})^2}{\sigma_{yk}^2}\right)}$$

DistilPose-Simulated Heatmaps

- Confidence Distillation
  - Most of previous regression-based methods (except RLE[3]) can’t provide a available confidence score.
  - DistilPose provide a novel method to predict achieve this goal.



[3] Li J, Bian S, Zeng A, et al. Human pose regression with residual log-likelihood estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11025-11034.

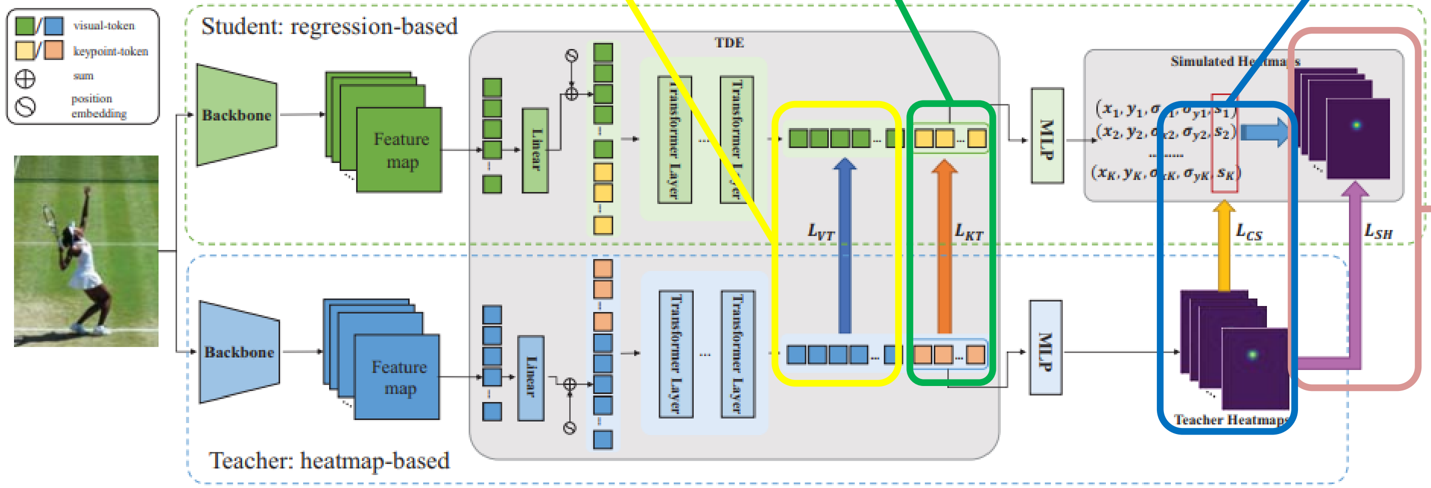
DistilPose

- Loss

$$L_{CS} = \sum_{i=1}^{K} |\mathbf{H}_{\mathbf{T}k}([x_k],[y_k]) - s_k|$$

$$L_{VT} = MSE(\boldsymbol{VT}_h, \boldsymbol{VT}_r)$$

$$L_{KT} = MSE(\boldsymbol{KT}_h, \boldsymbol{KT}_r)$$



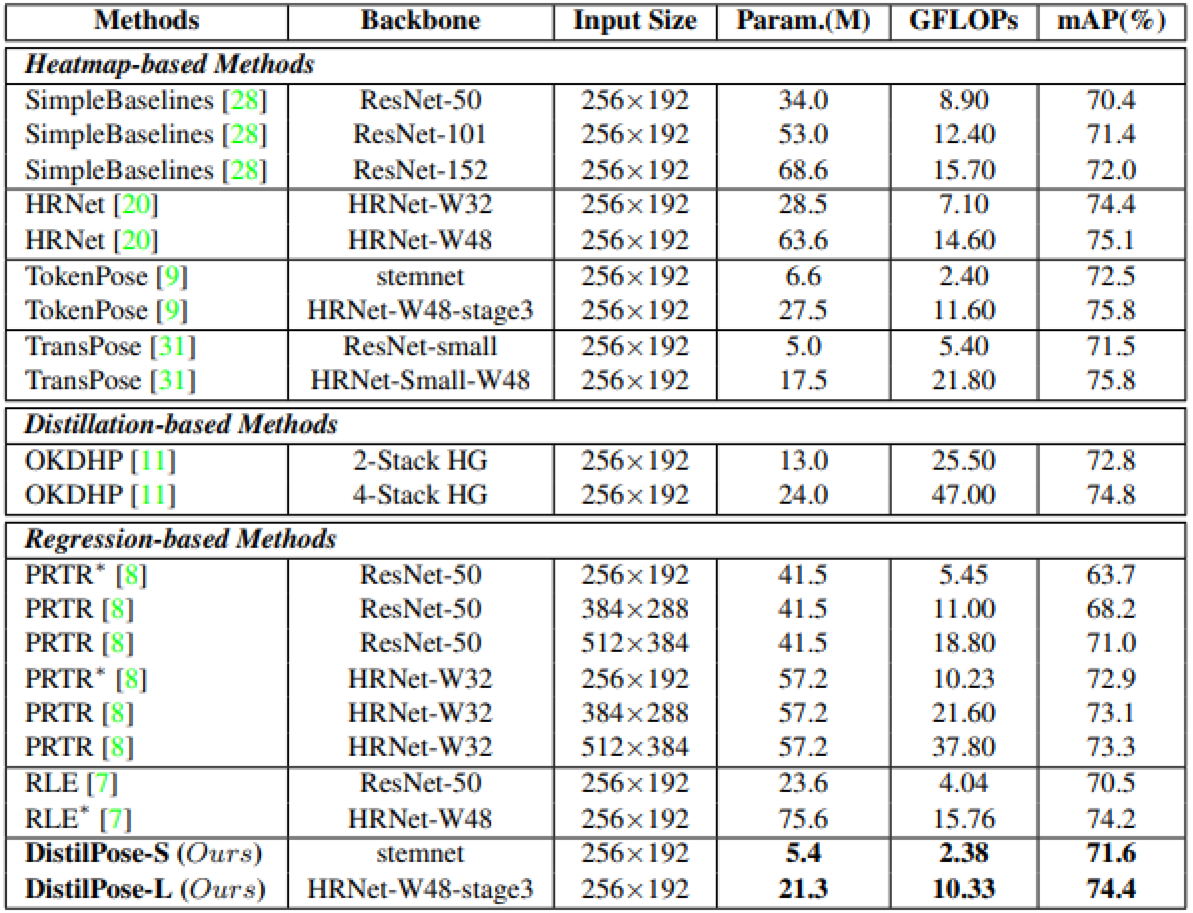
$$L_{SH} = \sum_{i=1}^{K} MSE(\mathbf{H}_k, \mathbf{H}_{\mathbf{T}k})$$

Total Loss:

$$L = L_{reg} + \alpha_1 L_{KT} + \alpha_2 L_{VT} + \alpha_3 L_{SH} + \alpha_4 L_{CS}$$

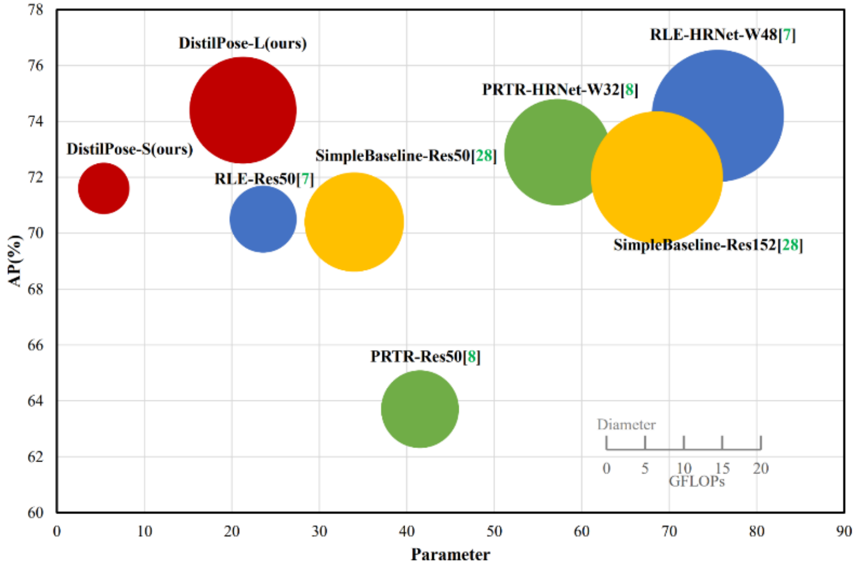
Experiments

Main Results on MSCOCO

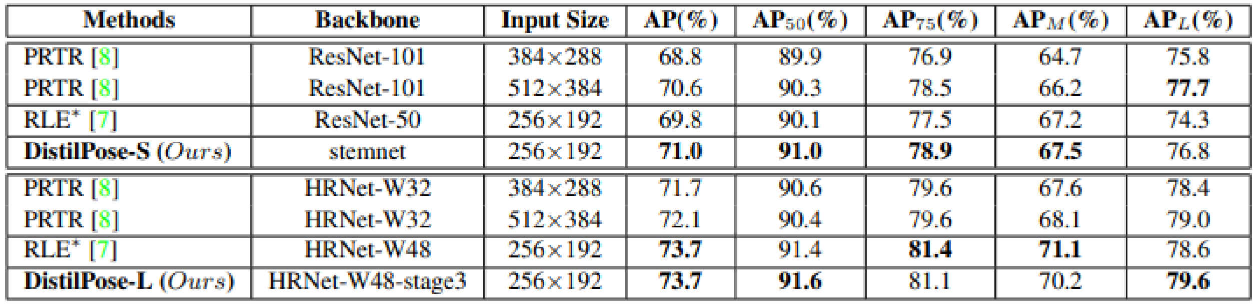


Methods	Backbone	Input Size	Param.(M)	GFLOPs	mAP(%)
Heatmap-based Methods					
SimpleBaselines [28]	ResNet-50	256×192	34.0	8.90	70.4
SimpleBaselines [28]	ResNet-101	256×192	53.0	12.40	71.4
SimpleBaselines [28]	ResNet-152	256×192	68.6	15.70	72.0
HRNet [20]	HRNet-W32	256×192	28.5	7.10	74.4
HRNet [20]	HRNet-W48	256×192	63.6	14.60	75.1
TokenPose [9]	stemnet	256×192	6.6	2.40	72.5
TokenPose [9]	HRNet-W48-stage3	256×192	27.5	11.60	75.8
TransPose [31]	ResNet-small	256×192	5.0	5.40	71.5
TransPose [31]	HRNet-Small-W48	256×192	17.5	21.80	75.8
Distillation-based Methods					
OKDHP [11]	2-Stack HG	256×192	13.0	25.50	72.8
OKDHP [11]	4-Stack HG	256×192	24.0	47.00	74.8
Regression-based Methods					
PRTR* [8]	ResNet-50	256×192	41.5	5.45	63.7
PRTR [8]	ResNet-50	384×288	41.5	11.00	68.2
PRTR [8]	ResNet-50	512×384	41.5	18.80	71.0
PRTR* [8]	HRNet-W32	256×192	57.2	10.23	72.9
PRTR [8]	HRNet-W32	384×288	57.2	21.60	73.1
PRTR [8]	HRNet-W32	512×384	57.2	37.80	73.3
RLE [7]	ResNet-50	256×192	23.6	4.04	70.5
RLE* [7]	HRNet-W48	256×192	75.6	15.76	74.2
DistilPose-S (*Ours*)	stemnet	256×192	**5.4**	**2.38**	**71.6**
DistilPose-L (*Ours*)	HRNet-W48-stage3	256×192	**21.3**	**10.33**	**74.4**

MSCOCO val2017



MSCOCO val2017

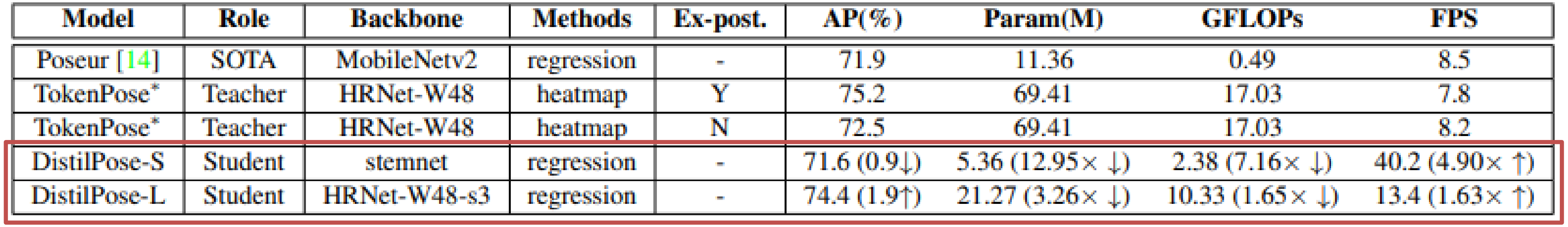


Methods	Backbone	Input Size	AP(%)	AP$_{50}$(%)	AP$_{75}$(%)	AP$_M$(%)	AP$_L$(%)
PRTR [8]	ResNet-101	384×288	68.8	89.9	76.9	64.7	75.8
PRTR [8]	ResNet-101	512×384	70.6	90.3	78.5	66.2	**77.7**
RLE* [7]	ResNet-50	256×192	69.8	90.1	77.5	67.2	74.3
DistilPose-S (*Ours*)	stemnet	256×192	**71.0**	**91.0**	**78.9**	**67.5**	76.8
PRTR [8]	HRNet-W32	384×288	71.7	90.6	79.6	67.6	78.4
PRTR [8]	HRNet-W32	512×384	72.1	90.4	79.6	68.1	79.0
RLE* [7]	HRNet-W48	256×192	**73.7**	91.4	**81.4**	**71.1**	78.6
DistilPose-L (*Ours*)	HRNet-W48-stage3	256×192	**73.7**	**91.6**	81.1	70.2	**79.6**

MSCOCO test-dev2017

Experiments

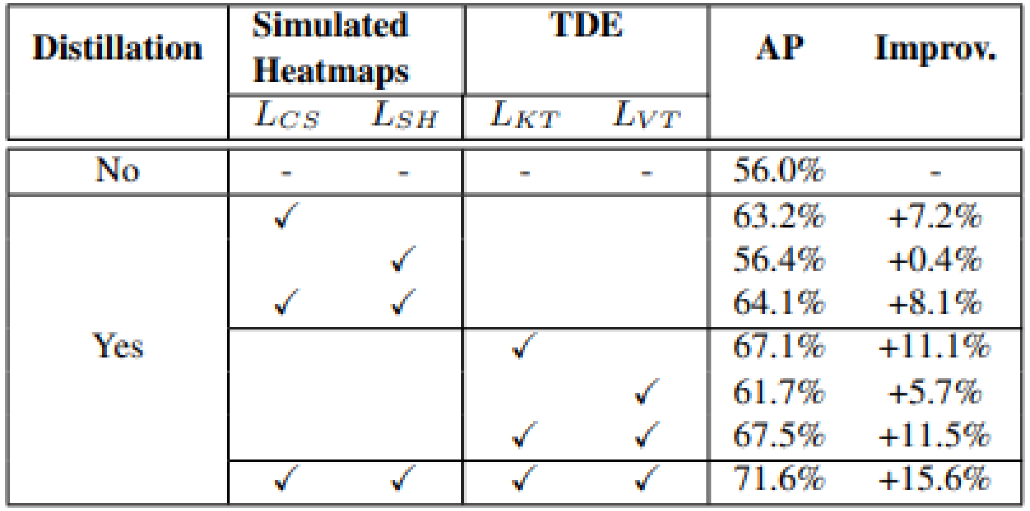
- Comparison between student and teacher.



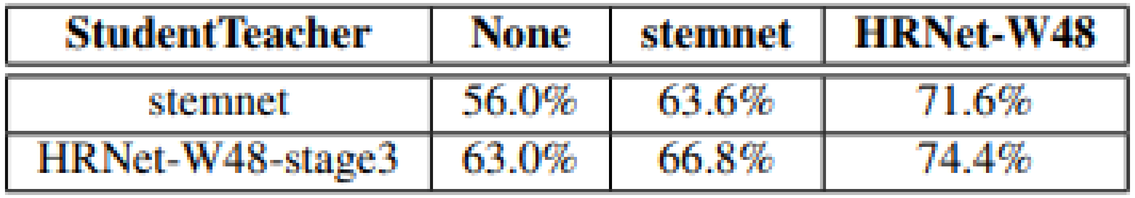
Model	Role	Backbone	Methods	Ex-post.	AP(%)	Param(M)	GFLOPs	FPS
Poseur [14]	SOTA	MobileNetv2	regression	-	71.9	11.36	0.49	8.5
TokenPose*	Teacher	HRNet-W48	heatmap	Y	75.2	69.41	17.03	7.8
TokenPose*	Teacher	HRNet-W48	heatmap	N	72.5	69.41	17.03	8.2
DistilPose-S	Student	stemnet	regression	-	71.6 (0.9↓)	5.36 (12.95× ↓)	2.38 (7.16× ↓)	40.2 (4.90× ↑)
DistilPose-L	Student	HRNet-W48-s3	regression	-	74.4 (1.9↑)	21.27 (3.26× ↓)	10.33 (1.65× ↓)	13.4 (1.63× ↑)

Experiments

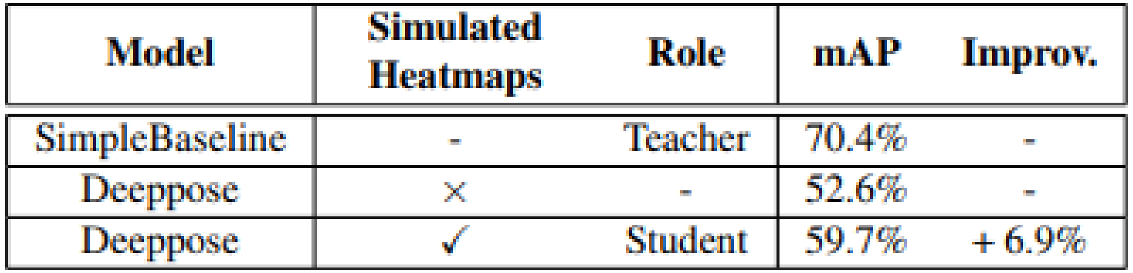
- Ablation Studies.



Distillation	Simulated Heatmaps		TDE		AP	Improv.
	L_{CS}	L_{SH}	L_{KT}	L_{VT}		
No	-	-	-	-	56.0%	-
Yes	✓				63.2%	+7.2%
		✓			56.4%	+0.4%
	✓	✓			64.1%	+8.1%
			✓		67.1%	+11.1%
				✓	61.7%	+5.7%
			✓	✓	67.5%	+11.5%
	✓	✓	✓	✓	71.6%	+15.6%



StudentTeacher	None	stemnet	HRNet-W48
stemnet	56.0%	63.6%	71.6%
HRNet-W48-stage3	63.0%	66.8%	74.4%



Model	Simulated Heatmaps	Role	mAP	Improv.
SimpleBaseline	-	Teacher	70.4%	-
Deeppose	×	-	52.6%	-
Deeppose	✓	Student	59.7%	+ 6.9%

Conclusion

- We propose a novel HPE framework, termed DistilPose.
- DistilPose includes a Token-Distilling Encoder and a Simulated Heatmaps to perform **heatmap-to-regression** knowledge distillation.
- DistilPose achieved state-of-the-art performance among regression-based methods with a much lower computational cost.

- Code: *https://github.com/yshMars/DistilPose*
- Paper: *https://arxiv.org/abs/2303.02455*

Code Paper

Thank you for your attention !