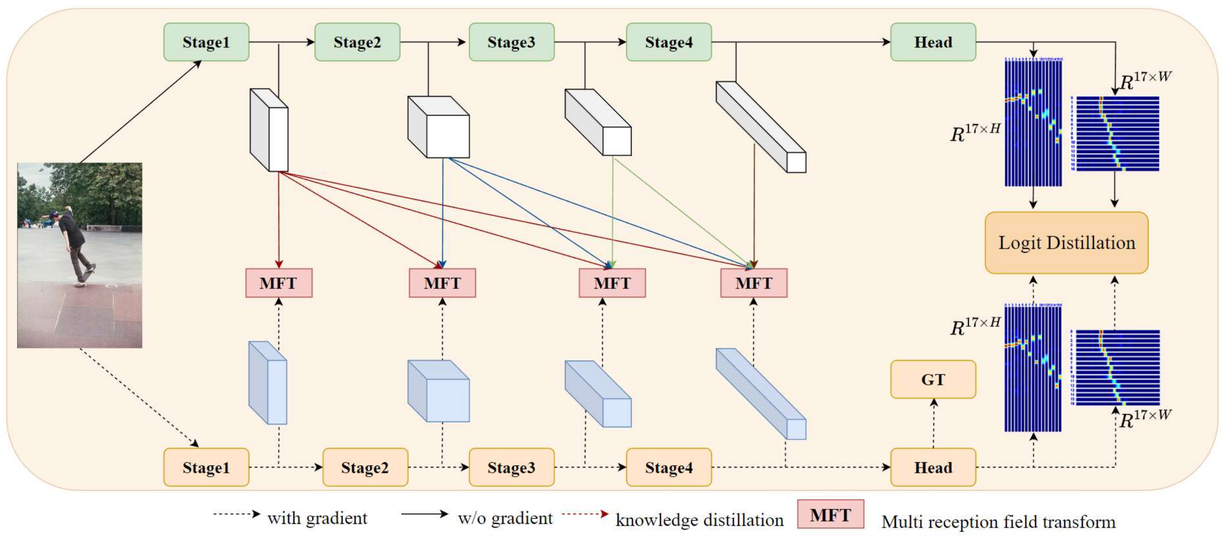
Lightweight Human Pose Estimation with Enhanced Knowledge Review

Hao Xu[1], Shengye Yan[1,2],Wei Zheng[2]
[1]Nanjing University of Information Science and Technology, 210044, Nanjing, China
[2]Shenzhen Youjia Innov Tech Co., Ltd., 518000, Shenzhen, China

Overview

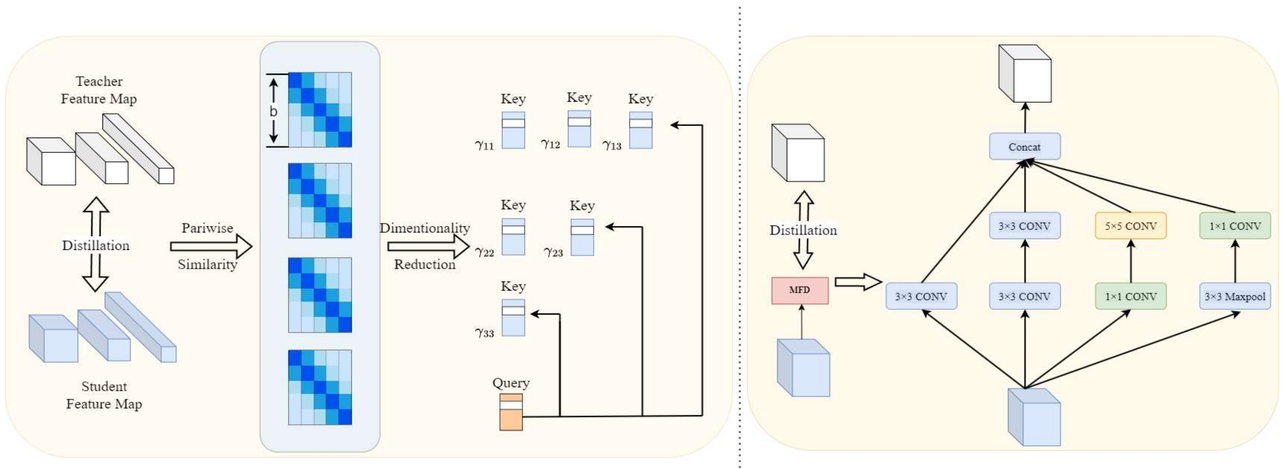


Compared with traditional knowledge distillation methods that typically involve layer-to-layer or block-to-block feature distillation between teacher and student networks, we present a novel approach to cross-stage knowledge distillation for enhancing human pose estimation.

- **Novel Cross-Stage Knowledge Distillation Technique:** Our method leverages multi-layer feature distillation combined with cross-attention weight allocation. By expanding the receptive field of the student network's feature maps, we aim to minimize the negative regularization effects that arise from semantic mismatches during the distillation process.

- **Efficient SimDR-based Head:** We replace the traditional Heatmap-based head with a SimDR-based head, decoupling coordinate representations to reduce computational and memory overhead, thereby improving efficiency.

Method

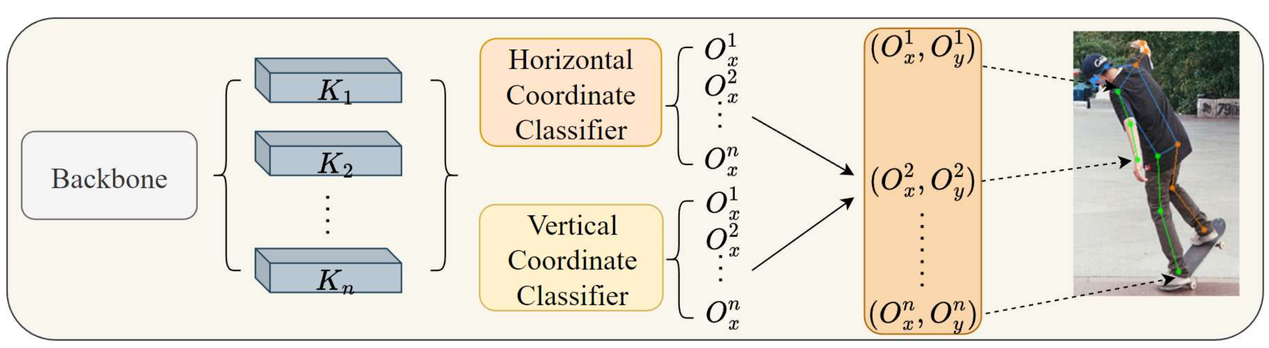


KD with Adaptable Cross-Stage Learning Weight in HPE

To address the negative regularization effects caused by feature mismatches in traditional knowledge distillation methods, we compute similarity matrices Sm and Tl for each pair of corresponding feature maps from the student and teacher networks. These matrices are then projected through linear layers, normalized using the L2 norm, and refined to obtain feature vectors. Using these refined vectors, we calculate attention weights that reflect the similarity between the student and teacher feature maps.

$$\gamma^i(t_l, s_m) = \frac{e^{S_m[i]\cdot T_l[i]^{Transpose}}}{\sum_{l=1}^{L}\sum_{m=l}^{M} e^{S_m[i]\cdot T_l[i]^{Transpose}}}$$

$$L_{fkd} = \sum_{i}^{b}\sum_{l=1}^{L}\sum_{m=l}^{M}\gamma^i(t_l, s_m)D(F_{t_l}[i] - f_{align}(F_{s_m}[i]))$$



Vanilla KD for the Output Horizontal and Vertical Coordinates

We use SimDR to convert the backbone network's output into a classification task for horizontal and vertical keypoints, reducing costs and accelerating inference. By replacing heatmap-based representations with SimDR in both networks, we apply vanilla knowledge distillation with KL divergence to guide the student model's output. This approach defines a task-specific loss function L_{task} based on coordinate classification, enhancing keypoint detection efficiency and accuracy.

$$L_{task} = -\sum_{n=1}^{N}\sum_{k=1}^{K} M_{n,k}\cdot\sum_{i=1}^{L}\frac{1}{L}\cdot V_i log(S_i)$$

$$L_{vkd} = -\frac{1}{N}\cdot\sum_{n=1}^{N}\sum_{k=1}^{K} KL(T_i, S_i)$$

Results

Table1. Results on MPII with student having different architectures.

Distillation Mechanism	Vanilla	Single Layer		Multiple Layers			
Method	KD [11]	FitNet [25]	RKD [22]	AF [39]	SP [27]	ReviewKD [4]	ours
PCKh@0.5	86.74	86.53	86.94	87.12	87.37	87.56	87.73

- We compare our model with state-of-the-art methods and summarize results using different distillation techniques, employing both vanilla and feature-based distillation methods.

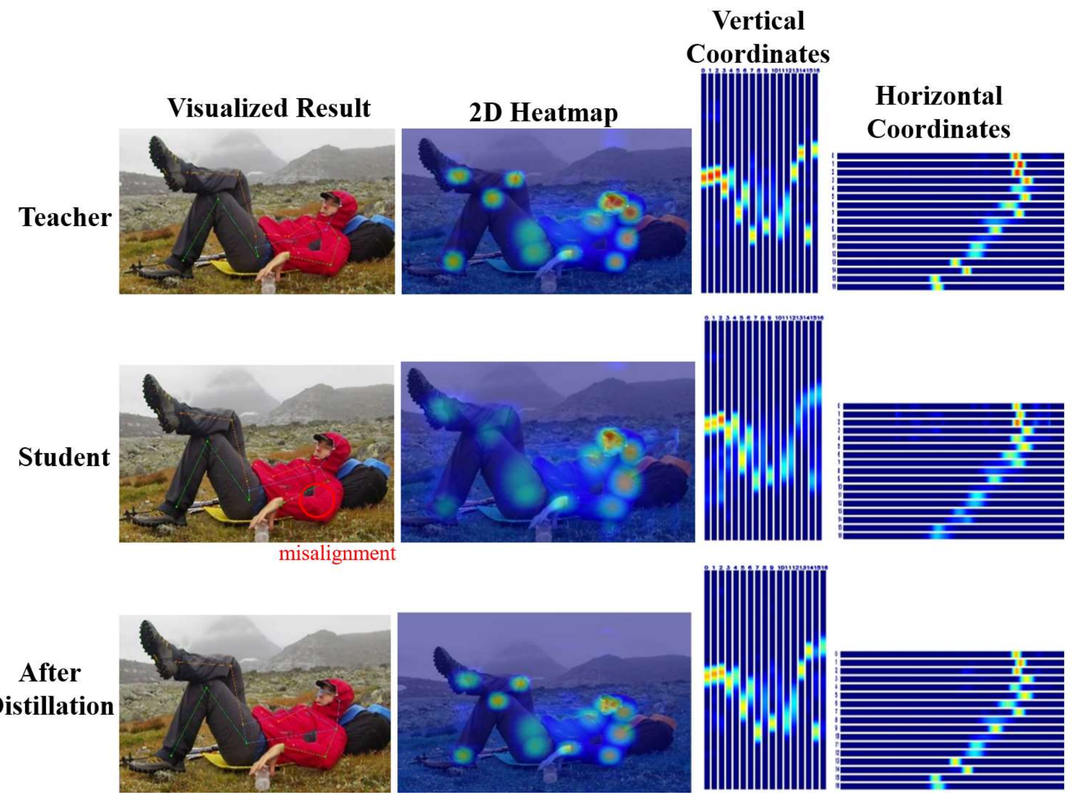
Table2. Experimental Results on the MPII validation set.

Model	Input size	Params	Infer memory	Speed	PCKh@0.5
ViTPose-B([32]	256 × 256	86M	14090MB	——	92.3
HRNET48([1]	256 × 256	64M	——	3.14s	91.4
HRNET32([28])	256 × 256	28.5M	8194MB	2.35s	90.8
8-Stack Hourglass([20])	256 × 256	25.1M	5174MB	1.20s	90.2
SimpleBaseline([31])	256 × 256	34.0M	6354MB	1.53s	88.5
MobileNetV2([12])	256 × 256	9.6M	3615MB	1.21s	85.4
ShuffleNetV2([41])	256 × 256	7.6M	1383MB	0.69s	82.8
2-Stack Hourglass([20])	256 × 256	18.6M	——	——	88.6
OKDHP([17])	256 × 256	18.6M	——	——	89.2
LiteHRNet18([38])	256 × 256	1.1M	2900MB	0.25s	86.1
LiteHRNet30([38])	256 × 256	1.8M	3334MB	0.28s	86.9
LiteHRNet18⋆(Multi-KD)	256 × 256	1.1M	2473MB	0.22s	87.7
LiteHRNet30⋆(Multi-KD)	256 × 256	1.7M	2886MB	0.23s	88.9

- On the MPII dataset, the SimDR-based pose estimation model significantly reduces computational overhead, with LiteHRNet18 achieving a 1.6% improvement in PCKh@0.5 and a 20% reduction in inference memory.

Table3. Ablation result on MPII validation set.

Cross-stage	MFT	WA	VKD	LiteHRNet18(Multi-KD)	LiteHRNet30(Multi-KD)
✗	✗	✗	✗	85.9	86.7
✓	✗	✗	✗	86.8	87.6
✓	✓	✗	✗	87.3	88.4
✓	✗	✓	✗	87.1	88.1
✓	✓	✓	✗	87.5	88.7
✓	✓	✓	✓	87.7	88.9



Acknowledgement

This work is supported by the Shenzhen Science and Technology Program (Grant No. KQTD20180411143338837).