西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Effective Whole-body Pose Estimation with Two-stages Distillation

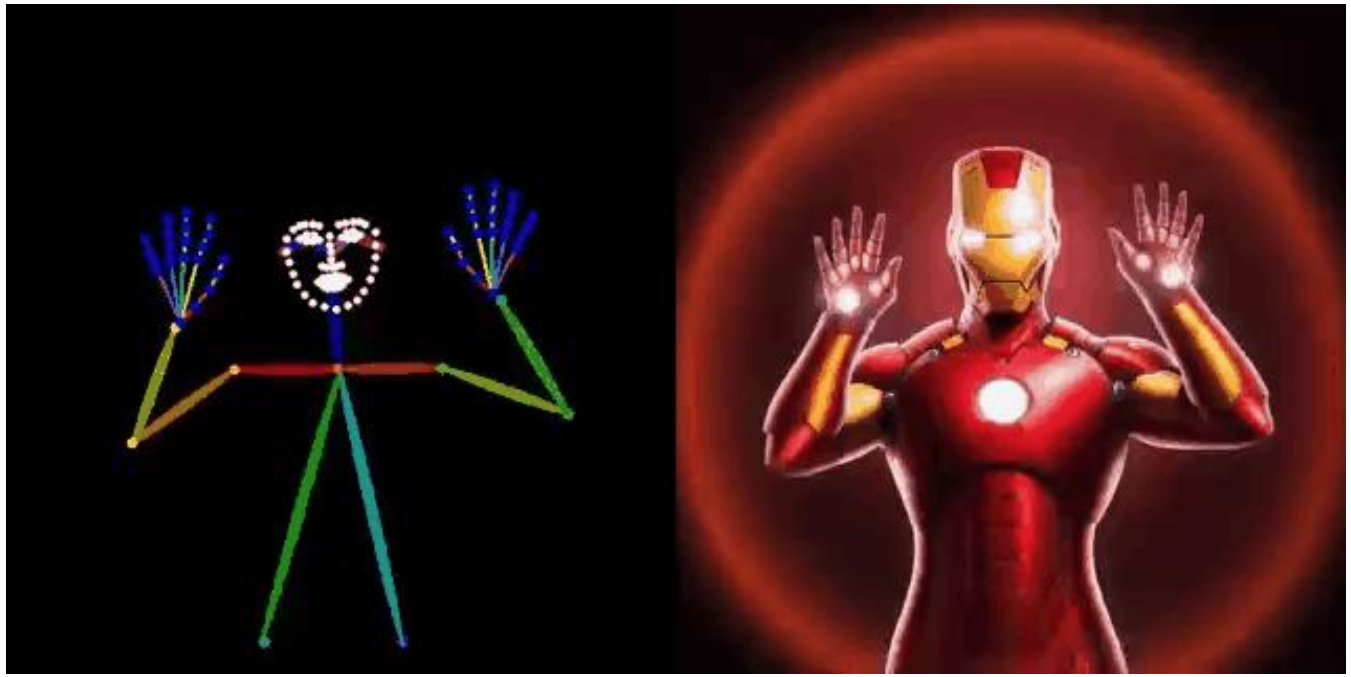
主讲人：赵行康

2024.3.6

研究背景及意义

人体全身姿态估计是指在图片中定位人体中的脸、手、腿等关键点。其在理解和生成任务中都起着重要作用，但其也面临着如数据限制，复杂的身体部位匹配，手势面部定位等问题。

为提升人体全身姿态估计的表现和效率，并且得到一系列可应用于下游任务的高效模型。作者提出了双阶段人体姿态蒸馏估计器（two-stage pose **D**istillation for **W**hole-body **P**ose estimators），也就是DWPose。



DWPose　　DWPose+ControlNet

基础知识介绍

知识蒸馏

知识蒸馏（Knowledge Distillation），简称KD，是一种模型压缩的方法，其目的是让一个小的模型（学生模型）学习一个大的模型（教师模型）的知识，从而提高小模型的性能和精度。其基本思想是利用教师模型的输出概率作为学生模型的额外监督信息，让学生模型去拟合教师模型的输出分布。由于知识蒸馏方法简单、有效，故其在工业界被广泛应用。

MMPose

MMPose 是一款基于 PyTorch 的姿态估计算法库，是 OpenMMLab 项目的成员之一。它支持人体、人手、人脸、动物、服装等多类物体的 2D/3D 姿态估计。

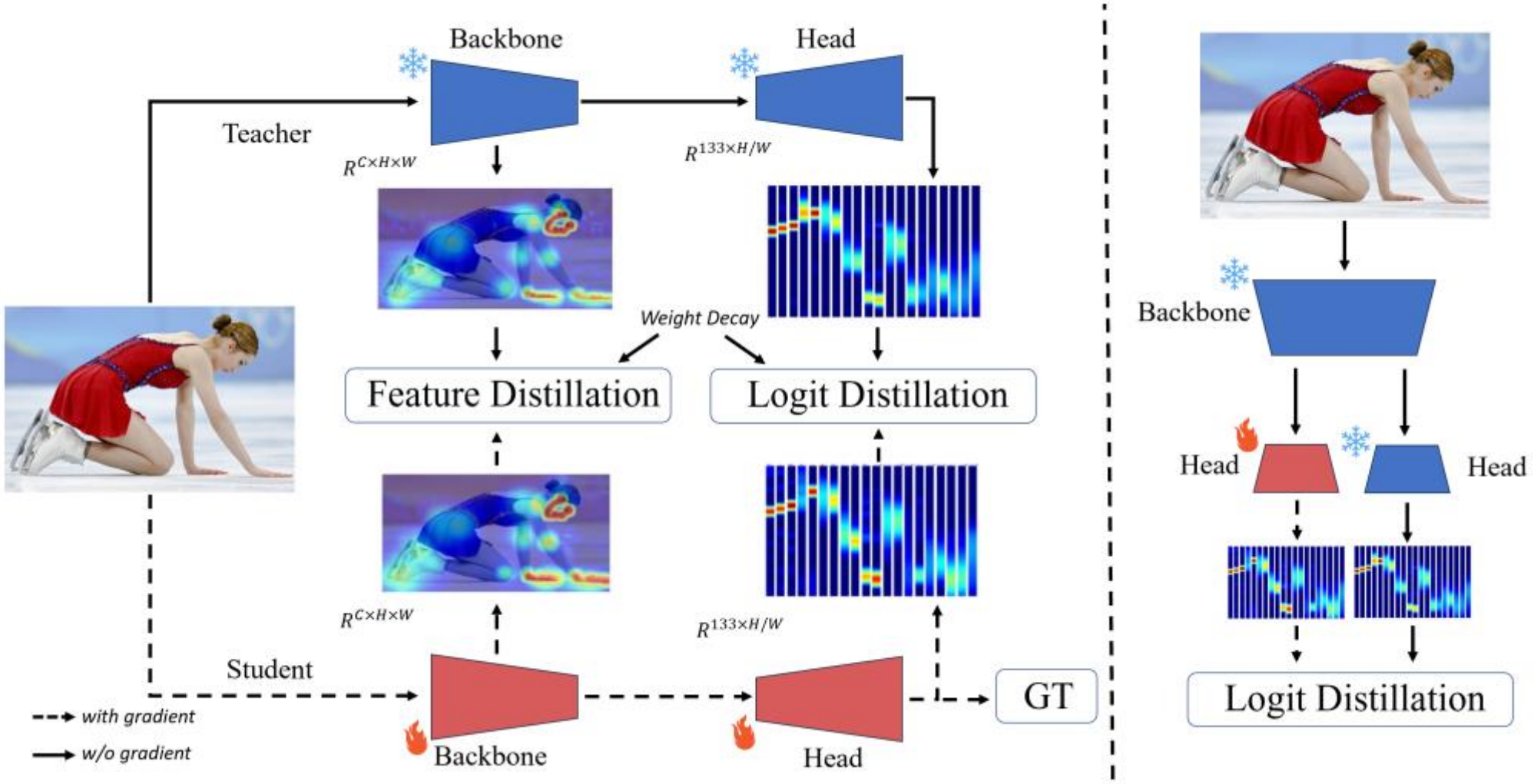
MMPose有较好的适用性，其支持多种姿态估计任务、模型结构、数据集以及评估指标。

基础知识介绍

RTMPose

RTMPose（Real-Time Multi-Person Pose Estimation based on MMPose）是一种基于 MMPose 的实时多人姿态估计框架。其使用 CSPNeXt 作为主干网络，使用 SimCC 作为关键点定位算法，使用自注意力机制来捕获关键点相关性。其中CSPNeXt为ResNet的变体，其相比于ResNet，在保持甚至是提高精度的同时，大幅降低了参数计算量。

双阶段蒸馏（Two-stages Pose Distillation，TPD）



first-stage

second-stage

一阶段蒸馏

一阶段蒸馏为传统知识蒸馏形式，其中教师模型规模大于学生模型，该阶段蒸馏包含了backbone的输出feature和head的输出logits两个层面，即Feature-based distillation和 Logit-based distillation。

在一阶段蒸馏中，作者令教师和学生的backbone特征输出分别为 $\boldsymbol{F^t}$ 和 $\boldsymbol{F^s}$ ，令教师和学生的logit输出为 $\boldsymbol{T_i}$ 和 $\boldsymbol{S_i}$ 。一阶段训练的目标就是使学生的特征和输出尽量贴近 $\boldsymbol{F^t}$ 和 $\boldsymbol{T_i}$ 。

首先是Feature-based distillation，在这里作者直接让学生去模仿教师backbone中的网络层，接着用均方差损失衡量学生特征 $\boldsymbol{F^s}$ 和教师特征 $\boldsymbol{F^s}$ 之间的距离，其具体损失为：

$$L_{fea} = \frac{1}{CHW} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(F_{c,h,w}^{t} - f(F_{c,h,w}^{s}) \right)^2$$

其中C、H、W分别是教师输出特征的高度、宽度以及通道数；f 是一个1×1的卷积参，其目的是改变 $\boldsymbol{F^s}$ 的形状为 $\boldsymbol{F^t}$ 。

一阶段蒸馏

在Logit-based distillation中，RTMPose的原始分类损失同样可以应用到一阶段的损失中，其公式为：

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

其中N为一批样本的数量（batch）；K为关键点数量；$W_{n,k}$ 为目标权重被掩盖的关键点；L为定位箱的长度；V_i 是标签值。

同时，在不对目标权重进行掩盖时，与标签值不同，其中不可见的关键点（图片中被其它物品遮挡的地方）也可以由教师分配给学生一个合理的值。作者认为这个值很有帮助，并在后续的实验中对其进行了验证，这样Logit-based distillation的损失就可以表示为：

$$L_{logit} = -\frac{1}{N} \cdot \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{L} T_i log(S_i)$$

一阶段蒸馏

故一阶段蒸馏的损失就可以表示为：

$$L = L_{ori} + \alpha L_{fea} + \beta L_{logit},$$

同时作者对蒸馏采用权重衰减对策略时学生更多地关注标签，并取得更好的表现，一阶段的最终损失就可以表示为：

$$L_{s1} = L_{ori} + r(t) \cdot \alpha L_{fea} + r(t) \cdot \beta L_{logit},$$

$$r(t) = 1 - (t-1)/t_{max},$$

二阶段蒸馏

二阶段蒸馏则更接近一种自蒸馏的形式，它不需要一个额外的教师模型，仅需要利用训练过学生模型以一种自学的方式来获得更好的表现。

该阶段蒸馏，作者称之为头部感知蒸馏方法（head-aware distillation）。在训练模型的基础上，作者使用训练过的backbone和未训练过的head构建学生模型，使用都训练过的backbone和head构建教师模型。因为教师和学生具有相同的结构，所以只需要从backbone提取一次特征，再将特征输入教师和学生的head进行训练即可，其损失函数如下：

$$L_{s2} = \gamma L_{logit}$$

与以往的自蒸馏方法不同，最终头部感知蒸馏方法可以在20%的训练时间内有效地从头部提取知识，进一步提高定位能力。

实验

	Method	Input Size	GFLOPs	whole-body		body		foot		face		hand	
				AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
Whole-body	SN† [13]	N/A	N/A	32.7	45.6	42.7	58.3	9.9	36.9	64.9	69.7	40.8	58.0
	OpenPose [2]	N/A	N/A	44.2	52.3	56.3	61.2	53.2	64.5	76.5	84.0	38.6	43.3
Bottom-up	PAF† [3]	512×512	329.1	29.5	40.5	38.1	52.6	5.3	27.8	65.6	70.1	35.9	52.8
	AE [34]	512×512	212.4	44.0	54.5	58.0	66.1	57.7	72.5	58.8	65.4	48.1	57.4
Top-down	DeepPose [43]	384×288	17.3	33.5	48.4	44.4	56.8	36.8	53.7	49.3	66.3	23.5	41.0
	SimpleBaseline [47]	384×288	20.4	57.3	67.1	66.6	74.7	63.5	76.3	73.2	81.2	53.7	64.7
	HRNet [40]	384×288	16.0	58.6	67.4	70.1	77.3	58.6	69.2	72.7	78.3	51.6	60.4
	PVT [44]	384×288	19.7	58.9	68.9	67.3	76.1	66.0	79.4	74.5	82.2	54.5	65.4
	FastPose50-dcn-si [9]	256×192	6.1	59.2	66.5	70.6	75.6	70.2	77.5	77.5	82.5	45.7	53.9
	ZoomNet [19]	384×288	28.5	63.0	74.2	74.5	81.0	60.9	70.8	88.0	92.4	57.9	73.4
	ZoomNAS [48]	384×288	18.0	65.4	74.4	74.0	80.7	61.7	71.8	88.9	93.0	62.5	74.0
	ViTPose+-S [51]	256×192	5.4	54.4	-	71.6	-	72.1	-	55.9	-	45.3	-
	ViTPose+-H [51]	256×192	122.9	61.2	-	75.9	-	77.9	-	63.3	-	54.7	-
	RTMPose-m	256×192	2.2	58.2	67.4	67.3	75.0	61.5	75.2	81.3	87.1	47.5	58.9
	RTMPose-l	256×192	4.5	61.1	70.0	69.5	76.9	65.8	78.5	83.3	88.7	51.9	62.8
	RTMPose-l	384×288	10.1	64.8	73.0	71.2	78.1	69.3	81.1	88.2	91.9	57.9	67.7
	RTMPose-x	384×288	18.1	65.3	73.3	71.4	78.4	69.2	81.0	88.8	92.2	59.0	68.5
	RTMPose-l + UBody	256×192	4.5	62.1	70.6	69.7	76.9	65.5	78.1	84.1	89.3	55.1	65.4
	RTMPose-l + UBody	384×288	10.1	65.4	73.2	71.0	77.9	68.6	80.2	88.5	92.2	60.6	69.9
	DWPose-t	256×192	0.5	48.5	58.4	58.5	67.0	46.5	63.6	73.5	80.7	35.7	49.0
	DWPose-s	256×192	0.9	53.8	63.2	63.3	71.3	53.3	69.0	77.6	84.1	42.7	54.9
	DWPose-m	256×192	2.2	60.6	69.5	68.5	76.1	63.6	77.2	82.8	88.1	52.7	63.4
	DWPose-l	256×192	4.5	63.1	71.7	70.4	77.7	66.2	79.0	84.3	89.4	56.6	66.5
	DWPose-l	384×288	10.1	66.5	74.3	72.2	78.9	70.4	81.7	88.7	92.1	62.1	71.0

实验

(a) OpenPose (b) MediaPipe (c) Ours

(a) OpenPose (b) MediaPipe (c) Ours

实验

Method	RTMPose* x-l			
First-stage	-	✓	-	✓
Second-stage	-	-	✓	✓
body	69.7	**70.4**	69.7	**70.4**
foot	65.5	65.8	65.9	**66.2**
face	84.1	84.1	84.2	**84.3**
hand	55.1	56.4	55.4	**56.6**
whole-body	62.1	62.9	62.2	**63.1**

Table 4. Ablation study of the two distillation stages. The teacher and student are RTMPose-x and RTMPose-l. '*' denotes the model is trained on COCO + UBody.

	body	foot	face	hand	whole-body
RTMPose-m	69.1	64.8	81.8	49.8	60.0
RTMPose-m + S2	69.4	65.1	81.9	50.3	60.4
RTMPose-l	69.5	65.8	83.3	51.9	61.1
RTMPose-l + S2	69.6	66.1	83.2	52.3	61.3
RTMPose-m*	68.6	63.6	82.5	52.3	60.4
RTMPose-m* + S2	68.5	63.6	82.8	52.7	60.6
RTMPose-l*	69.7	65.5	84.1	55.1	62.1
RTMPose-l* + S2	69.7	65.9	84.2	55.4	62.2
RTMPose-x*	70.3	65.3	84.9	56.4	63.0
RTMPose-x* + S2	70.4	65.3	84.9	56.6	63.2

Table 5. The impact of the proposed head-aware self-KD in the second-stage distillation (S2) on existing estimator RTMPose. '*' denotes the model is trained on COCO + UBody. All results are reported with AP on COCO-WholeBody.

Num. of persons	1	2	3	4	5	6	7	8	9
OpenPose	5.78	6.28	7.55	8.90	9.16	10.4	15.2	16.86	18.91
Ours	0.068	0.070	0.077	0.082	0.084	0.088	0.094	0.098	0.108

Table 8. Comparison of pose estimation speed with OpenPose. The table displays the average time cost (in seconds) for inferring an image on an Nvidia RTX 3090 with varying numbers of persons present. We use YOLOX-l and DWPose-l to test the speed.

实验

GT	Fea	Logit	Decay	whole-body
✓	-	-	-	60.4
✓	✓	-	-	61.8
-	-	✓	-	60.9
-	✓	✓	-	61.4
✓	✓	✓	-	62.0
✓	✓	✓	✓	**62.3**

此外，通过实验，作者发现不使用GT标签（ground truth label）而仅仅使用教师的logit输出来训练学生会使平均精度更高。这表明或许可以让教师模型对新数据进行标签而非人工标签，这样会节省人力成本并且取得更好地效果。但是如果结合GT标签和Fea、Logit层面再进行训练则会取得更好的效果。

Logit	Mask	whole-body
✓	-	60.9
✓	✓	59.8

在基于logit的蒸馏中，作者故意省略了目标权重掩模，它被用于区分可见的和不可见的关键点。通过实验结果，很明显可以看出目标权重掩模的存在显著地阻碍了蒸馏性能。结果导致学生的表现显著下降了1.1%。这些结果强调了教师对无形关键点的输入的重要性，肯定了其对学生训练过程的积极影响。

西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Questions and Discussions