江西省研究生创新专项资金项目

申　　请　　表

项　目　名　称: 基于知识蒸馏的人体姿态估计

申　　请　　人: 谢唯嘉

指　导　教　师: 易见兵　副教授

培 养 单 位 (盖　章): 江西理工大学

填　报　时　间: 2022 年 6 月 10 日

一、项目申报人基本情况

姓名	谢唯嘉	性别	男	
出生年月	1997.07	籍贯	江西吉安	
在读学历层次	硕士研究生	入学日期	2021.09	
在读专业	信息与通信工程			
身份证号码	362432199707250O1X			
指导教师姓名	易见兵	研究方向	人体姿态估计	
本科 (硕士) 毕业学校	江西理工大学	专业	信息与通信工程	
所在院系	信息工程学院	E-mail	695449217@qq.com	
联系电话	18379716287	手机	18379716287	

二、项目基本情况

项目主要研究内容（2000 字以内。文科包括：研究的主要问题目的、意义、研究方法、对策建议、创新点等；理工科包括：主要问题、关键技术、解决方案、研究方法、创新点等）：

2.1 研究背景

人机交互 [1] 是一门系统与用户之间互动的学科，从上世纪七十年代末，随着人们对它的不断认识，逐渐形成了一种多模态形式的人机互动也就是模仿人生活中的交互方式，比如：手势、触觉、表情、语音等，让机器也能够获取外界的世界信息，获得视觉感知的能力也就是像人类一样能够拥有自己的“眼睛”，所以遇到问题时候是需要机器能够正确理解人类的行为，也是在这样的背景下，姿态估计被提出，让它成为了一种当下重要技术之一，而且人体姿态估计存在着潜在的应用价值，让学术界、工业界备受关注。

姿态估计 [2] [3] [4] 是计算机视觉中重要任务，也是计算机理解人体的动作、行为不可或缺的一部分。姿态估计是过去几十年计算机视觉界一直关注的一个重要问题。姿态估计 [5] 常常还和姿态识别也就是行为识别联系到一起，这是两个概念，行为识别实在最终输出的是图像或者视频的行为的类别，而姿态估计是在输入图像视频之后，能都定位到某人的某个身体部位出现的位置，也就

是能够重建到人体各个部位以及各个关节，估计到人的关节点的坐标，行为识别可以借助姿态估计的相关成果来实现。本创新项目主要研究姿态估计，姿态估计大多数是人体姿态估计，还有一些有手部姿态估计，手部姿态估计分为有标记和无标记的姿态估计，用于理解手部行为的意思。而本创新项目主要围绕人体姿态估计进行阐述。

目前人体姿态估计已经取得较为不错的效果，它在不断的从二维到三维，从图像到视频，从复杂网络到轻量化网络。紧接着又随着深度学习技术的不断发展，将深度学习的许多理论加入到了姿态估计。

2.1.1 基于深度学习的姿态估计

深度学习时机器学习领域近几年发展较为迅猛的分支。深度学习是自我解释型的学习方式，简单方便，功能强大，很多领域都在使用。姿态估计也陆续出现了许多方法，利用深度卷积神经网络来增强人体估计系统的性能，在网络架构方面，基于深度学习的姿态估计分为单级网络和多级网络，单级网络通常的难点在于后面的特征融合工作，多级网络一般就是重复叠加某个细小的小网络结构。单级网络都是会采用特征提取已经训练好的分类网络，微调后作为新网络的一个 backbone，常用到的网络比如:VGGNet [6] 或 ResNet [7]（残差网络）。模型主要有 Deeppose [2]（直接回归坐标）、CPM [3]（热力图回归坐标）。对于深度学习的人体姿态估计的算法来讲，大致分为自顶向下方法和自底向上的方法。

自顶向下（Top-Down Approaches）

自顶向下这种方法也就是先检测到整个人的存在，再具体检测每个关节点的位置，但是自顶向下的缺点有两点，一是对物体检测比较敏感，也就是如果行人未检测出来，关键点就不会检测到; 二是处理量会随着人数的增加而增加，成正比。

同时它的优点也是很突出:1、召回率较好，因为 top-down 是先做人体检测，人体往往会比部位更大，所以从检测角度来讲会更简单，相应找到的召回率也会更高。2、关键点的定位精度会更准，这部分原因是基于裁剪的框，对空间信息有一定的对齐处理，同时因为在做单人估计的时候，可以获得一些中间层的上下文信息，对于点的定位是很有帮助的。

2017 年 CVPR 采用的自顶向下后，先是使用 faster rcnn [8] 当人整体的检测器，其次用一个

残差网络结构对检测到的人做密集热力图并且预测坐标偏移量，最后将其二者结融合结果得到关键点的坐标。同年，MSCOCO 比赛的冠军采用级联金字塔网络改进过程，采用多尺度特征融合，还用损失函数来处理关键点。肖等人还提供了人体姿态估计的基准方法。自顶向下的检测算法有 RMPE [9]，Mask-RCNN [10]，HRNet [11]。

Regional Multi-Person Pose Estimation（RMPE）[9] 是上海交大和腾讯优图的论文，被 ICCV 2017 接收。它对于多人姿态估计的方法采用传统的自顶向下的方法，即先检测人，再识别人体姿态。检测使用的是 SSD-512，识别人体姿态使用的是达到最高水准的 Stacked Hourglass [4] 方法。致力于解决对于不完美的结果，通过调整，使得裁剪的单人能够被单人姿态估计方法很好的识别，从而克服检测带来的定位误差。

对于 Mask RCNN [10] 是一个非常流行的语义和实例分割架构。该模型可以同时预测图像中多个物体的候选框位置及分割其语义信息的掩膜。该模型的基础架构很容易被扩展到人体姿态估计上来。

对于 HRNet [11] 是目前人体姿态估计邻域中最主流的基准网络。为了这些任务位置信息更加精准，很容易想到的做法就是维持高分辨率的特征图，事实上 HRNet 之前几乎所有的网络都是这么做的，通过下采样得到强语义信息，然后再上采样恢复高分辨率恢复位置信息，然而这种做法，会导致大量的有效信息在不断的上下采样过程中丢失。而 HRNet 通过并行多个分辨率的分支，加上不断进行不同分支之间的信息交互，同时达到强语义信息和精准位置信息的目的。

自底向上（Bottom-Up Approaches）

这种方法就是先检测到每个人关节部位的关键点，然后再用图结构、条件随机场等算法组接成每个人的姿态，缺点就是将关键点与人关联起来，分组成人具有挑战, 而且受遮挡影响太大。代表算法：openpose [12]，DeepCut [13]，HigherHRNet [14]。

OpenPose 和很多自底向上的方法一样，OpenPose [12] 首先检测出图像中所有人的关节（关键点），然后将检出的关键点分配给每个对应的人。用 Part Affinity Fields（PAFs）来学习如何将身体关键和个人匹配起来。主要思想是利用贪心算法自下而上的解析步骤从而达到高准确率和实

时性。身体部位定位和关联是在两个分支上同时进行的。

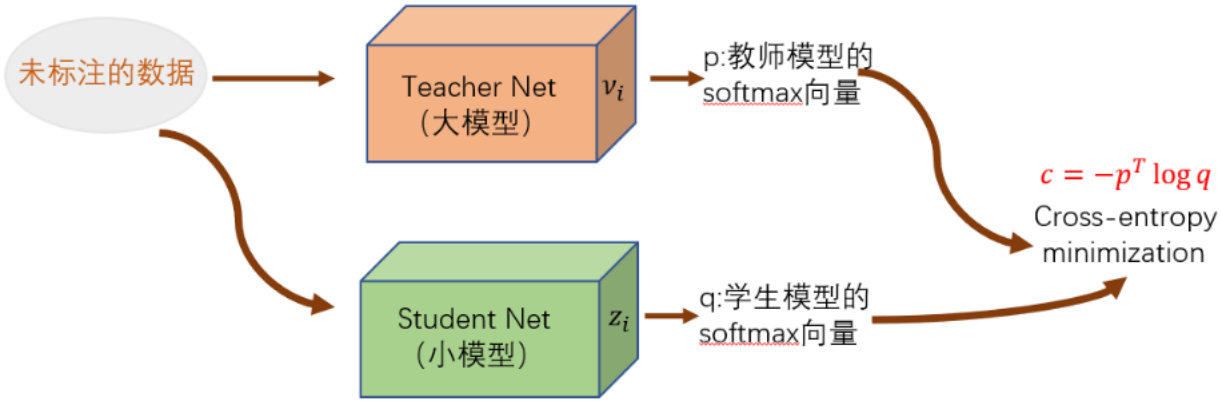
2015 年提出 DeepCut [13] 模型，深度卷积预测所有点，再聚类组成图。DeepCut 是一个自底向上的多人人体姿态估计方法。针对人体姿态估计任务，作者定义了以下问题：1. 生成一个由 D 个关节候选项组成的候选集合。该集合代表了图像中所有人的所有关节的可能位置。在上述关节候选集中选取一个子集。2. 为每个被选取的人体关节添加一个标签。标签是 C 个关节类中的一个。每个关节类代表一种关节，如“胳膊”“腿”“躯干”等。3. 将被标记的关节划分给每个对应的人。

CVPR 2020 录用的 HigherHRNet [14] 是一种新的自下而上的人体姿势估计方法，用于使用高分辨率特征金字塔学习尺度感知表示。该方法配备了用于训练的多分辨率监督和用于推理的多分辨率聚合，能够解决自下而上的多人姿势估计中的尺度变化挑战，并能更精确地定位关键点，尤其是对于小人物。HigherHRNet 中的特征金字塔包括 HRNet 的特征图输出和通过转置卷积进行上采样的高分辨率输出。

2.1.2 知识蒸馏

蒸馏是一种热力学的分离工艺，它利用混合液体或液-固体系中各组分沸点不同，使低沸点组分蒸发，再冷凝以分离整个组分的单元操作过程。具体在化学中，蒸馏是一个有效的分离沸点不同的组分的方法，大致步骤是先升温使低沸点的组分汽化，然后降温冷凝，达到分离出目标物质的目的。在深度学习中所提出的知识蒸馏就是通过引入超参数温度来提高对交叉熵损失中负对数的重视程度。

Geoffrey E. Hinton, Oriol Vinyals 和 Jeffrey Dean [15] 在 2015 年首次提出了知识蒸馏的概念，学生网络从教师提供的基础真实标签和软标签中学习, 所以知识蒸馏定义了一种学习方式，在这种学习方式中，一个更大的教师网络被用来指导一个更小的学生网络的训练，以完成许多任务。教师网络中蕴涵知识是通过教师的软标签传递给学生的。为了提高对负对数的重视，引入了超参数温度。



接下来的工作可以分为两类:logits 蒸馏和中间特征

以往 logit 精馏的工作主要是提出有效的正则化和优化方法，而不是新颖的方法。DML [16] 提出了一种相互学习的方式，同时培训学生和教师。TAKD [17] 引入了一个名为“教师助理”的中型网络，以弥合教师和学生之间的差距。

FitNet [18] 通过一个阶段的中间特征来提取知识。FitNet 的想法很简单，通过卷积层将学生的网络特征转换为和教师网络模型输出特征的相同形状。用 L2 距离用来测量它们之间的距离。PKT 将教师的知识建模为概率分布，用 KL 散度来测量距离。CRD [19] 将对比学习与知识提炼相结合，利用对比目标进行知识转移。

2.2 主要问题

1. 最近的研究 [20] 已经证明了长期空间依赖性对人体姿态估计的好处。长程空间相关性的一般概念是在大视场中对空间信息的全局理解。以前的高分辨率网络 [21] [22] [23] 主要依赖于并行分来构建空间相关性，但受宽度和深度的限制，轻量级网络的容量会受到严重影响。因此，其中一个问题是如何增强轻量级高分辨率网络，以更有效的方法建模长程空间依赖性。

2. 我们观察到，最先进的人体姿态估计基础网络 (如 HRNet [11]) 的基本 CNN 构建模块在建立大型网络时并不具有成本效益，因为每层有大量的通道，而且更难训练。并且在后续的边缘设备进行部署时有对于边缘设备较高的要求，不利于批量部署操作。在本研究中，我们考虑在不降低模型性能的情况下提高人体姿态估计效率，但保留可比较的精度结果的问题。

3. 目前还不清楚 PSAM [24] 如何使嵌入分类和置换的像素级回归得到最大的好处。在复杂的 DCNN 头部回归，如那些在，无锚的人体姿态检测任务。据我们所知，大多数现有的工作与自我注意块只插入块骨干网络。我们未来的工作是探索 PSAM 在 DCNN 头部的应用。

2.3 关键技术

2.3.1 极化自注意力模块 (PSAM)

注意力机制 (Attention Mechanism) 是人们在深度学习模型中嵌入的一种特殊结构，用来自动学习和计算输入数据对输出数据的贡献大小。注意力机制是上世纪九十年代，一些科学家在研究人类视觉时，发现的一种信号处理机制。人工智能领域的从业者把这种机制引入到一些模型里，并取得了成功。目前，注意力机制已经成为深度学习领域，尤其是自然语言处理领域，应用最广泛的“组件”之一。这两年曝光度极高的 BERT [25]、GPT [26]、Transformer [27] 等等模型或结构，都采用了注意力机制。

一些学者尝试让模型自己学习如何分配自己的注意力，即为输入信号加权。他们用注意力机制的直接目的，就是为输入的各个维度打分，然后按照得分对特征加权，以突出重要特征对下游模型或模块的影响。这也是注意力机制的基本思想。

自注意块作用于输入张量 X 来突出或抑制特征，这很像光学透镜过滤光线。在摄影中，横向方向上总是有随机的光线产生眩光/反射。偏振滤波，只允许光通过正交于横向方向，可以潜在地提高照片的对比度，由于总强度的损失，滤波后的光通常有一个小的动态范围，因此需要额外的增强。我们借用摄影的关键因素，提出了极化自我注意 (PSA) 机制:

(1) 滤波: 在一个方向上完全折叠特征，同时保持其正交方向上的高分辨率;

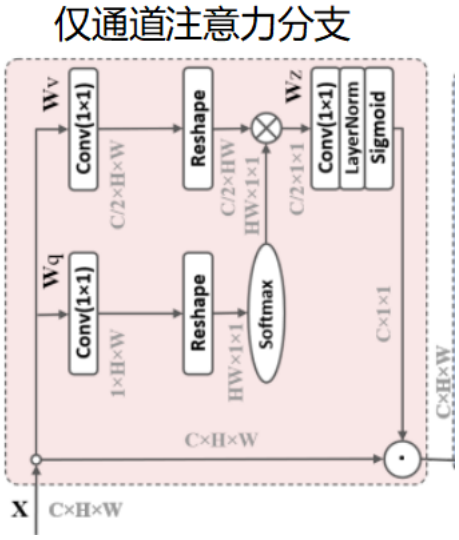
(2) 高动态范围: 在瓶颈张量 (注意块中最小的特征张量) 处进行 Softmax 归一化，增加注意的动态范围，然后进行色调映射 Sigmoid 函数。形式上，我们将 PSA 机制 [24] 实例化为下面的 PSA 块:

仅通道注意力分支： $A^{ch}(X) \in \mathcal{R}^{C\times 1\times 1}$

$$A^{ch}(X) = F_{SG}[W_{z|\theta_1}(\sigma_1(W_v(X)) \times F_{SM}(\sigma_2(W_q(X))))]$$

W_z, W_q, W_v 是三个 (1 x 1) 卷积核，它们学习不同通道之间空间特征的线性组合。σ_1 和 σ_2 两个张量变形算符。F_{SM} 是 SoftMax 操作而 $\times$ 是矩阵点积运算。

仅通道分支的输出为 $Z^{ch} = A^{ch}(X) \odot^{ch} X \in \mathcal{R}^{C\times H\times W}$,$\odot^{ch}$ 是一个基于通道的乘法运算符。

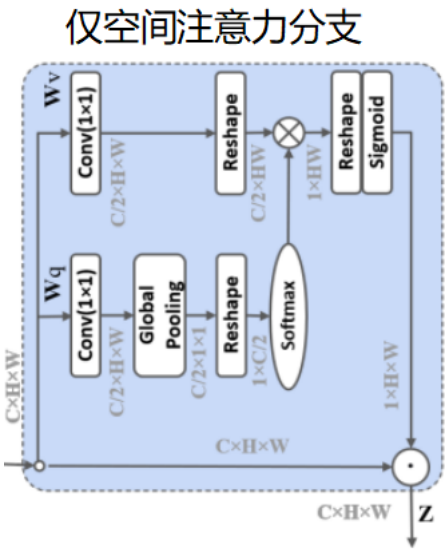


仅空间注意力分支： $A^{sp}(X) \in \mathcal{R}^{1\times H\times W}$

$$A^{sp}(X) = F_{SG}[\sigma_3(F_{SM}(\sigma_1(F_{GP}(W_q(X)))) \times \sigma_2(W_v(X)))]$$

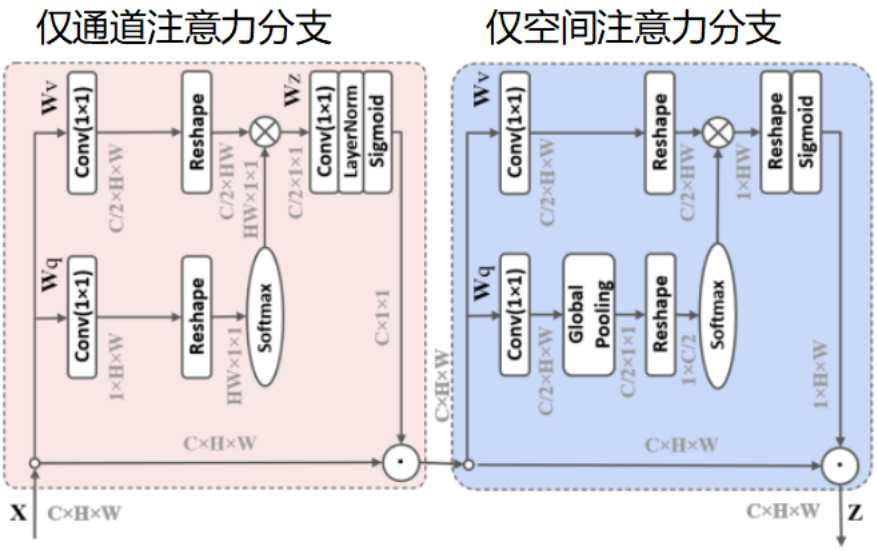
W_q 和 W_v 是分别标准的 1×1 卷积层。σ_1,σ_2 和 σ_3 是三个张量变形算符。F_{SM} 是 SoftMax 操作, F_{GP} 是一个全局池操作符。$\times$ 是矩阵点积运算。

仅空间分支的输出为 $Z^{sp} = A^{sp}(X) \odot^{sp} X \in \mathcal{R}^{C\times H\times W}$,$\odot^{sp}$ 是一个空间相关的乘法运算符。



整体注意力模块：

$$PSAM(X) = Z^{sp}(Z^{ch}) = A^{sp}(A^{ch}(X) \odot^{ch} X) \odot^{sp} A^{ch}(X) \odot^{ch} X$$



具体点我们在每个 Residual 块的第一个 3×3 卷积后添加 PSAM。但是目前还不清楚 PSAM 如何才能从嵌入分类的像素级回归中获益在复杂的 DCNN 头部。据我们所知，大多数现有的工作与自我注意块只插入块骨干网络。我们工作是还会探索 PSAM 在 DCNN 头部位置中的应用。

PSAM 与其他自我注意模块的关系: 我们将 PSAM 加入到表 1 中进行对比，PSA 模块相对现有的注意力模块更先进的原因:

(1) 内部分辨率 vs 复杂性: 与最高配置下的现有注意力块相比，PSA 在两个分支上都保持了最高的注意力分辨率 $(C/2)^3$ 和空间 ($[W, H]$) 维度。此外，在我们只关注通道的情况下，将 Softmax 重加权与挤压激励融合，利用 Softmax 作为瓶颈张量 $C/2 \times W \times h$ 的非线性激活。通道数 $C{-}C/2{-}C$ 遵循挤压激励模式，这对 GC 块和 SE 块都有利。我们的设计实现了高分辨率的压缩激励，同时计算复杂度可与 GC 块相当。

(2) 我们只关注空间的注意力不仅保持了饱满 $[W, H]$ 空间分辨率，而且内部保持 $2 \times \ \times /2$ 在 W_q 和 W_v 中可学习的参数，用于非线性 Softmax 重加权，这是比现有块更强大的结构。

(3) 输出为非线性分布。PSAM 仅通道和仅空间分支都使用 Softmax 和 Sigmoid 组合。将 Softmax 和 Sigmoid 进行组合作为概率分布函数，考虑关键点热图可以在线性变换上近似。因此，我们期望非线性可以充分利用保存在内部的高分辨率信息 PSAM 注意分支。

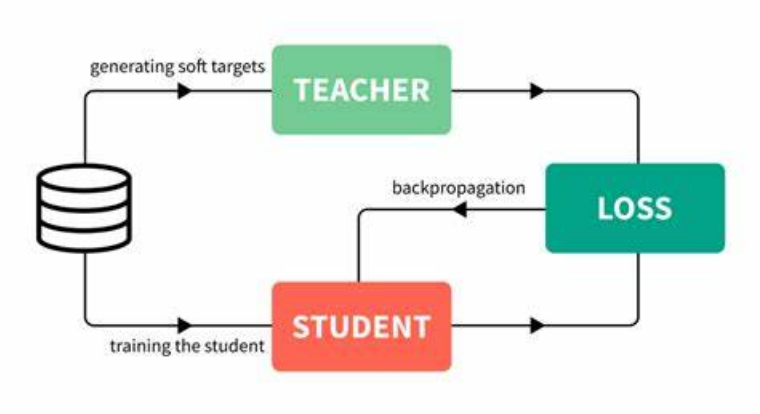
NL [28],GC [29],SE [30],CBAM [31],DA [32],EA [33]

方法	通道分辨率	空间分辨率	非线性激活	复杂度
NL[5]	C	[W,H]	SM	$C^2WH + CW^2H^2$
GC[27]	C/4	-	SM+ReLU	CWH
SE[15]	C/4	-	ReLU+SD	CWH
CBAM[28]	C/16	[W,H]	SD	CWH
DA[12]	C/8	[W,H]	SM	$C^2WH + CW^2H^2$
EA[29]	d_k($\ll$ C)	d_v($\ll$ min(W,H))	SM	CWH
PSA(our)	C/2	[W,H]	SM+SD	CWH

表 1　重新审视现有注意力块中的关键设计方面

2.3.2 知识蒸馏 (Distilling Knowledge)

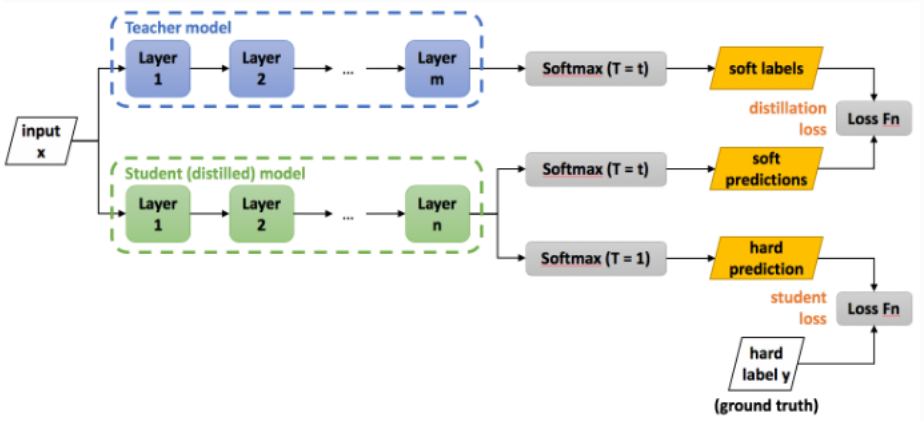
训练过程相当于一个有几十年经验的老师 (水平高，知识全面，学的快等特点) 相当于大模型，有个刚入学的学生 (知识储备能力小，经验不丰富) 相当于小模型。让教师网络指导学生网络进行训练。



知识蒸馏整体过程

一旦训练了繁琐的模型，我们就可以使用另一种训练方式，我们称之为“蒸馏”，将知识从繁琐的模型转移到更适合部署的小模型。一般认为，用于训练的目标函数应尽可能接近地反映使用者的真实目标。对于这个转移阶段称为知识蒸馏过程，我们可以使用相同的训练集，也可以使用单独的数据集。当繁琐的模型是简单模型的大集合时，我们可以使用它们各自预测分布的算术或几何平均值作为软目标。当软目标有很高的熵, 他们提供更多的信息/培训情况比硬目标和更少的方差之间的梯度训练情况, 所以小模型通常可以比原来繁琐的模型训练更少的数据和使用更高的学习速率 [34]。

繁琐的模型几乎总是以很高的置信度产生正确的答案，学习函数的很多信息都存在于软目标中非常小的概率的比率中。在最简单的精馏形式中，知识通过在一个转移集上对其进行训练，并对通过在其 soft-max 中使用高温的繁琐模型产生的转移集中的每个情况使用软目标分布来转移到精馏模型。训练蒸馏模



目标检测知识蒸馏整体过程

型时使用相同的高温，但训练后使用的温度为 1。

在目标识别领域中。原来我们需要让新模型的 softmax 分布与真实标签匹配，现在只需要让新模型与原模型在给定输入下的 softmax 分布匹配了。直观来看，使用知识蒸馏比前者具有这样一个优势：经过训练后的原模型，其 softmax 分布包含有一定的知识——真实标签只能告诉我们，某个图像样本是一辆宝马，不是一辆垃圾车，也不是一颗萝卜；而经过训练的 softmax 可能会告诉我们，它最可能是一辆宝马，不大可能是一辆垃圾车，但绝不可能是一颗萝卜。而这些信息对于学生网络模型的学习是具有正向的促进作用。

接续前面的讨论，我们的目标是让新模型与原模型的 softmax 输出的分布充分接近。直接这样做是有问题的：在一般的 softmax 函数中，自然指数 e 先拉大 logits 之间的差距，然后作归一化，最终得到的分布是一个 arg max 的近似，其输出是一个接近 one-hot 的向量，其中一个值很大，其他的都很小。这种情况下，前面说到的” 可能是垃圾车，但绝不是萝卜” 这种知识的体现是非常有限的。相较类似 one-hot 这样的硬性输出，我们更希望输出更” 软” 一些。

一种方法是直接比较 logits 来避免这个问题。具体地，对于每一条数据，记原模型产生的某个 logits v_i 是，新模型产生的 logits 是 z_i，我们需要最小化

$$D = \frac{1}{2}(z_i - v_i)^2$$

文献 [15] 中提出了更通用的一种做法。考虑一个广义的 softmax 函数:

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

其中 T 是温度，这是从统计力学中的玻尔兹曼分布中借用的概念。容易证明，当温度 T 趋向于 0 时，soft-

max 输出将收敛为一个 one-hot 向量，温度 T 趋向于无穷时，softmax 的输出则更“软”。具体地，在训练时我们需要最小化两个分布的交叉熵 (Cross-entropy)，记新模型利用 softmax 函数公式所产生的分布是 q，而原模型产生的分布是 p，则我们需要最小化：

$$C = -p^T logq$$

转移集中的每一种情况都贡献了教师模型和学生模型输出的交叉熵梯度其对某一个 logit z_i 的梯度：(z_i 相对于蒸馏模型的每个 logit, v_i 教师模型的 logits, p_i 产生软目标概率, T 是在有温度调节的情况下进行转移训练):

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}(\frac{e^{z_i/T}}{\sum_j e^{z_i/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_i/T}})$$

如果温度高于对数的量级:

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(\frac{1+z_i/T}{N+\sum_j z_i/T} - \frac{1+v_i/T}{N+\sum_j v_i/T})$$

如果我们现在假设所有 logits 对每个样本都是零均值化的 (即 $\sum_j z_j = \sum_j v_j = 0$)，那么:

$$\frac{\partial C}{\partial z_i} = \frac{1}{NT^2}(z_i - v_i)$$

在较低的温度下，蒸馏很少注意匹配比平均值相差得多的对数。另一方面，但是比平均值相差得多的对数可能会传递由繁琐的模型获得的知识的有用信息。

知识蒸馏和最小化 logits 的平方差是等价的 (因为梯度大致是同一个形式)。实验表明，温度不能取太大，而应该使用某个适中的值，这表明忽略极负的 logits 对新模型的表现很有帮助 (较低的温度产生的分布比较“硬”，倾向于忽略 logits 中极小的负值)。

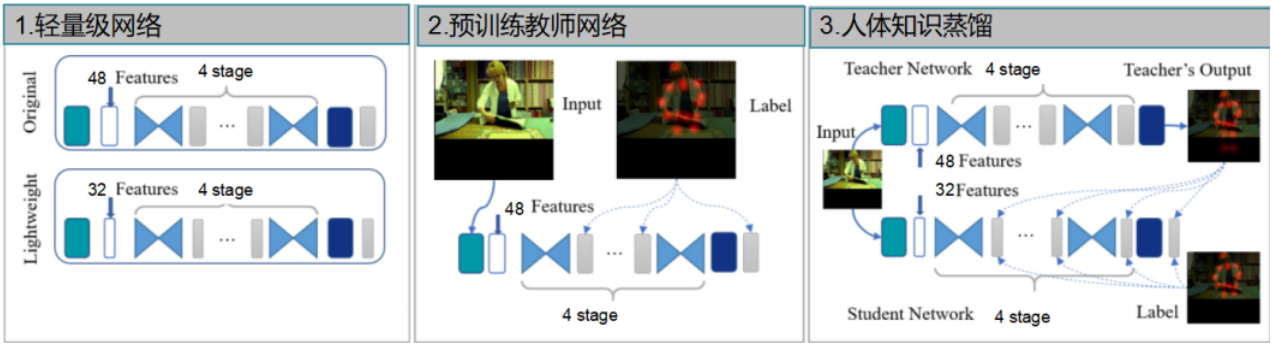
通过知识蒸馏加强监督

我们采用知识精馏的通用模型训练策略:

1. 我们先训练一个老师模型。在我们的实验中，默认我们选择了原始的 HigherHRNet 模型 [16]，因为它设计干净，模型鲁棒性强，可以不受任何限制地考虑其他更强的模型。它是目前人体姿态估计中精度最高的模型。

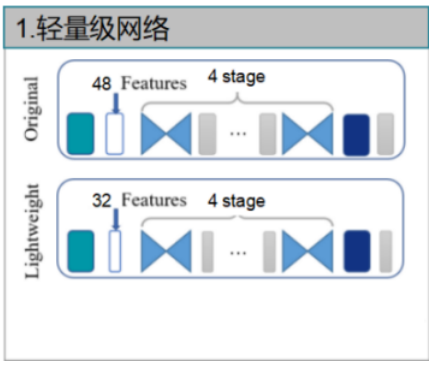
2. 然后，我们利用教师模型学习到的知识来训练目标学生模型。知识的提炼发生在这一步。

我们为了描述整个学生网络训练过程的概貌，将知识蒸馏主要步骤呈现如图 1 所示，并之后进行分阶段介绍。



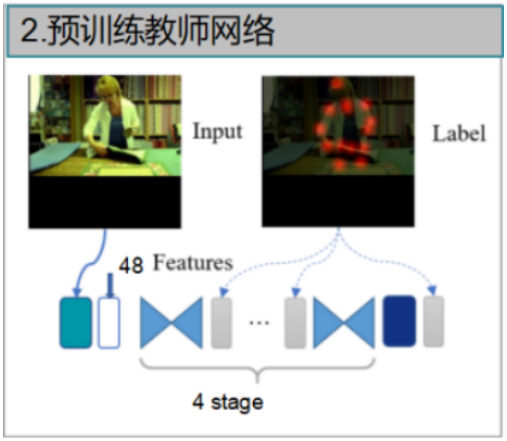
知识蒸馏主要步骤

阶段一：设置轻量级网络。



原始的网络结构通过 Stem 后变为通道数为 48 的特征图，之后下采样后通道数翻倍，长宽减半。而轻量级的网络结构通过后变为通道数为 32 的特征图，之后下采样后通道数翻倍，长宽减半。而两者的 stage 数都为 4。

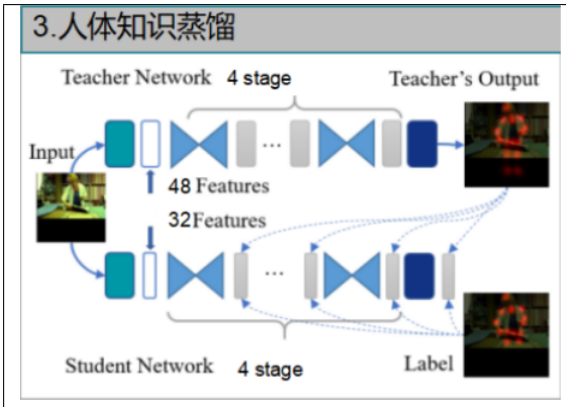
阶段二：预训练教师网络。



教师网络结构通过 Stem 后变为通道数为 48 的特征图，之后下采样后通道数翻倍，长宽减半。

训练过程是通过网络预测的关键点位置和真实关键点位置信息做 L_2 范数，并将其作为损失函数。在训练过程中实时对其进行反向传播，进而优化权重参数。

阶段三：人体知识蒸馏。



将输入的图片分别放入教师模型和学生模型中进行训练，教师的网络结构通过 Stem 后变为通道数为 48 的特征图，之后下采样后通道数翻倍，长宽减半。而学生的网络结构通过后变为通道数为 32 的特征图，之后下采样后通道数翻倍，长宽减半。

为了提炼老师原有的知识来对学生模型进行训练的关键点是设计一个适当的蒸馏损失函数，使得能够有效地提取和转移教师的知识到学生模型的训练过程中。我们将通过教师网络预测的关键点位置和真实关键点位置信息做 L2 范数和教师网络预测的关键点位置和学生网络预测的关键点位置做 L2 范数，并将两个损失函数作为损失函数。

对于传统的人体姿态模型训练过程中，我们通常使用基于损失函数的均方误差 (Mean-Squared Error, MSE)[27][28]。为了表示真实人体关节标签，我们为每个关节 k 生成一个置信图 $m_k(k \in 1, \cdots, k)$，通过将高斯核核以标记位置 $z_k = (x_k, y_k)$ 为中心。更具体地说，第 k 个关节标号的高斯置信映射 m_K 可以写成:

$$m_k(x,y) = \frac{1}{2\pi\sigma^2} exp(\frac{-[(x-x_k)^2+(y-y_k)]^2}{2\sigma^2})$$

其中 (x, y) 表示像素位置，超参数 σ 表示预先确定的空间方差。则得到 MSE 损失函数为:

$$\mathcal{L}_{mse} = \frac{1}{K}\sum_{k=1}^{K} \left\| m_k - \hat{m}_k \right\|_2^2$$

其中 m_k 表示第 k 个关节的预测置信图，$\hat{m}_k$ 表示第 k 个关节的真实位置。

在此之前的蒸馏函数设计都是用于对象分类环境下的单标签 softmax 交叉熵损失 [3,13] 中，不适合在 2D 图像空间中传递结构化的人体姿态知识。为了解决上述问题，我们设计了一个联合置信度图专用姿态蒸馏损失函数，公式为:

$$L_{kd} = \frac{1}{K}\sum_{k=1}^{K} \left\| m_k^s - m_k^t \right\|_2^2$$

式中，m_k^s 和 m_k^t 分别为预先训练的教师模型和受训的学生目标模型预测的对第 K 个关节信息的置信图。我们选择 MSE 函数作为蒸馏的损失函数，用来衡量学生和教师模型之间的分歧，以最大化与姿势监督学习损失的可比性。

对于学生模型的损失函数还是使用传统的 MSE 函数作为损失函数。

我们将训练过程中人体姿态估计使用知识精馏的整体损失函数表示为:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{kd} + (1-\alpha)\mathcal{L}_{mse}$$

其中 α 为两个损失项之间的平衡权重，通过交叉验证估计。

2.4 解决方案

该项目用 Python 语言进行编程，将项目的数据集调用进项目网络结构中进行各个深度卷积神经网络权重学习操作。

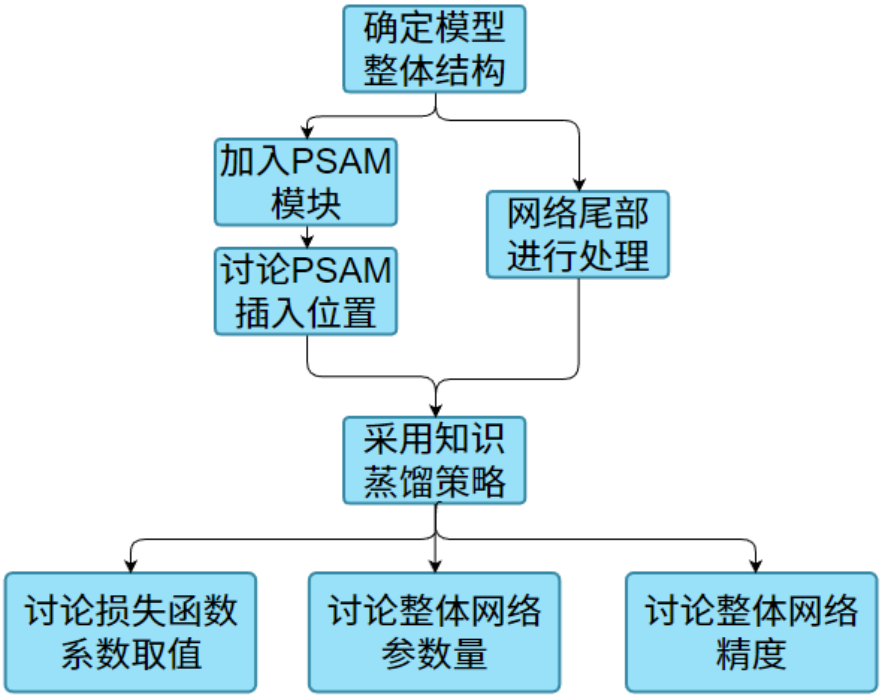
(1) 针对如何更有效的使用人体姿态知识蒸馏这一训练策略，可通过现阶段关于目标识别中的做法类比于人体姿态估计这一具体的任务中。更重要的是理解人体姿态估计这一领域中已经采用的蒸馏方法，具体这一训练策略的可行性要通过实验数据和理论方面进行验证确定。对人体姿态知识蒸馏中的损失函数设置做相关的消融实验。

(2) 在现阶段轻量级模型的基础上，对于轻量级模型，添加先阶段泛化能力很好的注意力模块，提高轻量级网络整体的准确性。对于要加入的注意力模块，将有意的进行讨论 PSAM 如何才能从嵌入分类的像素级回归中获益在复杂的 DCNN 头部。据我们所知，大多数现有的工作与自我注意块只插入块骨干网络。我们工作是还会探索 PSAM 在 DCNN 头部位置中的应用。

(3) 对于整体的网络结构尾部，加入现阶段泛化能力很好的反卷积模块，用于生成比输入特征图分辨率高的高质量特征图，重新组合各个阶段生成的特征图，从各个特征图中学习到不同维度的信息，进而提高轻量级网络整体的准确性。

2.5 研究方法

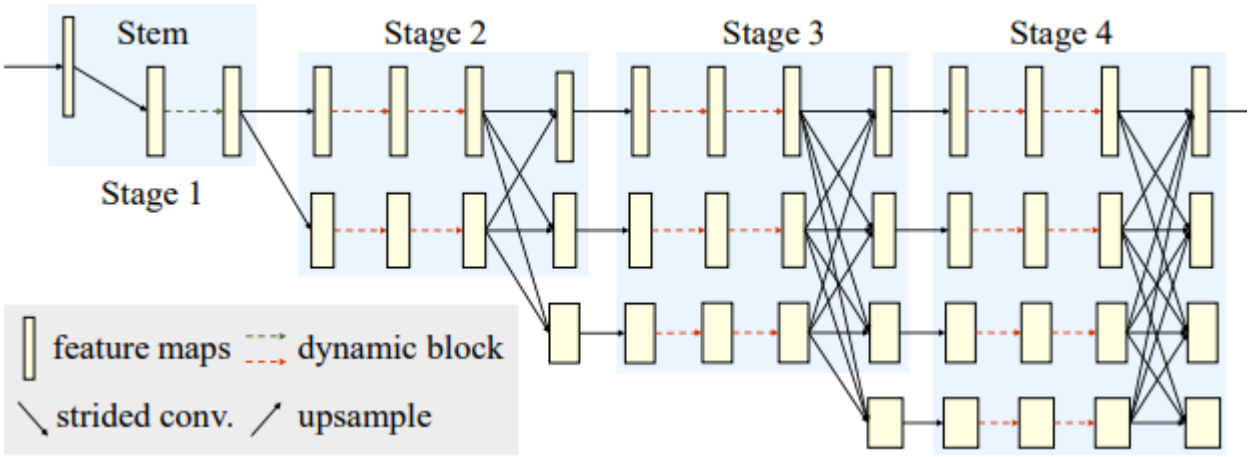
首先我们建立适合程序运行的实验平台和环境，Ubuntu 系统，Pytorch 深度学习框架等。在这个平台和环境上进行方法和技术的研究，对课题研究内容进行整体框架规划分析和设计，合理安排各阶段任务和分配有限资源，部分设计各个击破。该项目的研究方法及过程如下图所示。



项目的整体研究过程

1、确定模型整体结构。

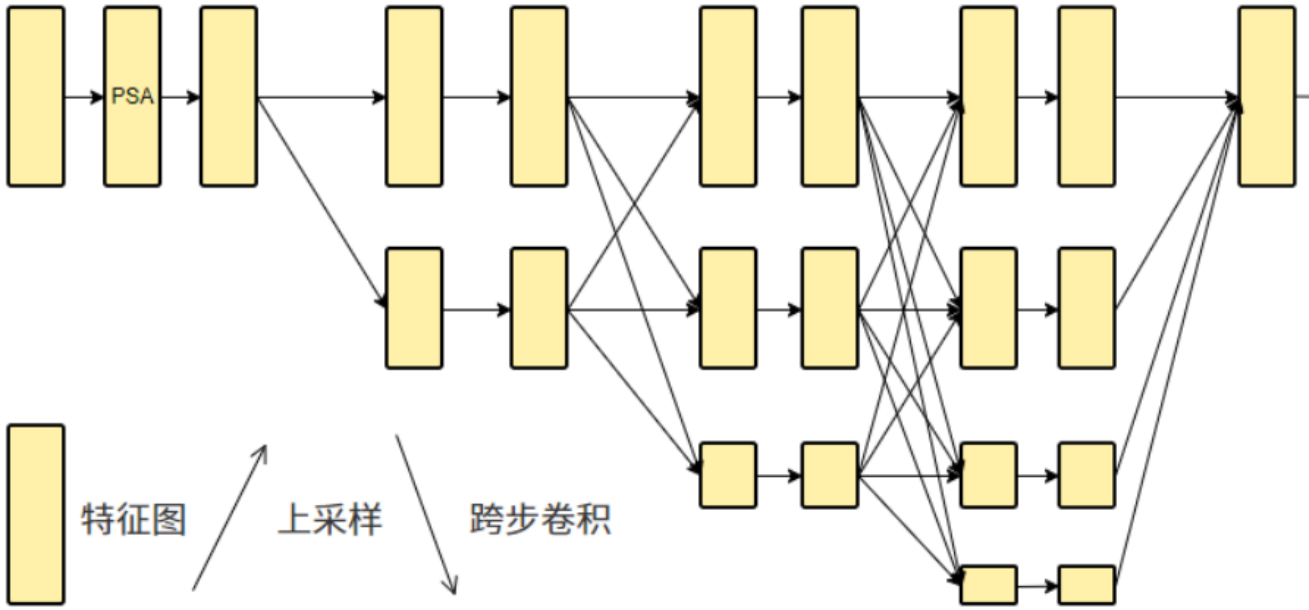
我们采用在 IJCAI 上发表的最新动态轻量级的高分辨率网络 (Dite-HRNet [35])，可以高效地提取多尺度上下文信息和建立人体姿态估计的长期空间依赖性模型。实验结果表明，该网络在 COCO 和 MPII 人体姿态估计数据集上都取得了优异的性能，超过了目前最先进的轻量级网络。Dite-HRNet 的整体网络结构如下图所示。



Dite-HRNet 整体网络结构

2. 加入 PSAM 模块。

为了提高网络结构对于特征图信息的学习能力，而注意机制在深卷积神经网络 (DCNN) 已经成为流行推动特征图信息中远程依赖关系, 特定于像素级别的关注。而 PSAM 模块似乎已经耗尽了其仅限通道和仅限空间分支的表示能力。实验结果表明，在二维人体姿态估计基准上，加入 PSAM 模块有所提高。模型的整体网络结构如下图所示。目前还不清楚 PSAM 如何才能使嵌入在复杂的 DCNN 头部像素级回归得到最大的好处。据我们所知，大多数现有的自注意块工作只在骨干网中插入自注意块。我们将讨论 PSAM 插入进 DCNN 其他位置的使用。



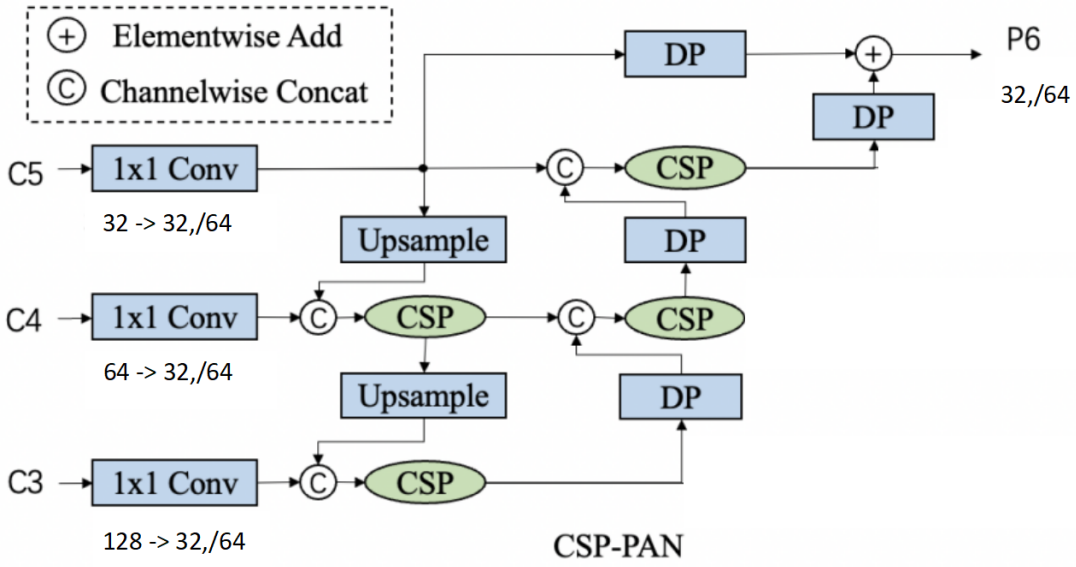
模型整体网络结构

3. 模型网络尾部进行处理。

在整体网络结构的尾部采用 PP-PicoDet [36] 提出了 CSP-PAN 结构，使用 1*1 的卷积将特征的通道数与 BackBone 输出的最小通道数进行统一，从而减少计算量，并保证特征融合性能不受影响。此外，PP-PicoDet 还在 CSP-PAN 的基础上再下采样一次，添加一个更小的特征尺度来提升大物体的检测效果。

与此同时，PP-PicoDet 在 Neck 和 Head 部分均采用深度可分离卷积，将 $3x3$ 卷积核增大至 5×5，来增大感受野，还保持了速度不变。并且 PP-PicoDet 采用了通道数和 Neck 一致的“耦合头”，相比于小通道数的“解耦头”有更快的预测速度。

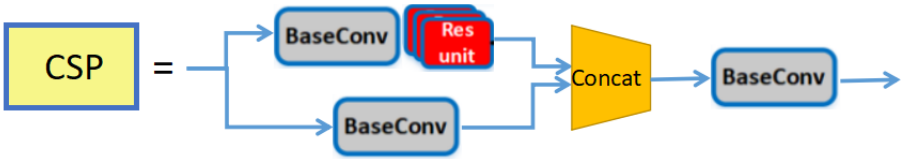
我们使用 PAN[41] 结构得到多级特征映射，使用 CSP 结构对相邻特征映射进行特征拼接和融合。整体网络的尾部使用 CSP-PAN 模块进行特征的融合，在原有的 CSP-PAN 中，每个输出特征图中的通道号与骨干输入的通道号保持一致。对于移动设备来说，信道数大的结构具有昂贵的计算成本。我们通过使所有特征图中的所有信道数等于 1×1 卷积的最小信道数来解决这个问题。整体修改后的 CSP 网络结构插入整体网络结构的尾部。



轻量级的 CSP-PAN 模块插入网络尾部

DP 模块采用了深度卷积和点卷积

该轻量级的 CSP 模块的整体结构如下图所示。



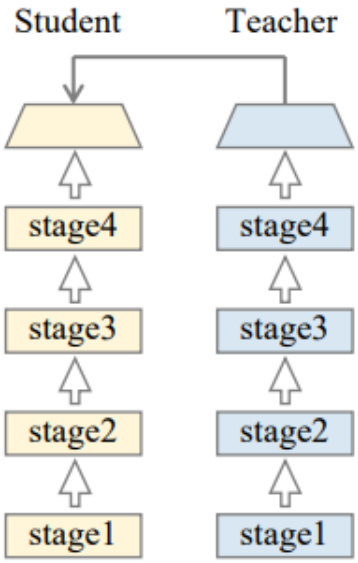
CSP 模块采用了深度卷积和点卷积

我们修改了现阶段泛化性较好的轻量级物体探测器，在移动设备的物体检测方面具有卓越的性能。将该网络模块融合进我们自己的网络结构中，进一步能够进行各个空间维度的特征融合且所使用的网络开销较少。进而使整个网络兼顾检测速度又兼顾检测的精度。

4. 采取知识蒸馏策略。

为了提炼老师原有的知识来对学生模型进行训练的关键点是设计一个适当的蒸馏损失函数，使得能够有效地提取和转移教师的知识到学生模型的训练过程中。我们将通过教师网络预测的关键点位置和真实关键点位置信息做 L2 范数和教师网络预测的关键点位置和学生网络预测的关键点位置做 L2 范数，并将两个损失函数作为损失函数。

下图是采取知识蒸馏训练策略的整体网络结构的训练过程。根据最后一层老师模型的输出结果对学生模型整体的训练过程进行监督，使学生能够学习到更多网络结构中深层隐含的特征，使学生模型整体的泛化性更好。



知识蒸馏训练策略的整体网络结构

2.6 创新点

1. 我们研究了研究不足的人体姿态模型效率问题，而现有的尝试主要集中在提高准确性性能，在部署时模型推断的成本很高。这是一个关键的问题，要解决扩大现有的深度姿态估计方法到实际应用。

2. 我们提出了一种人体关键点蒸馏 (HKD) 模型训练方法，使之能够更有效地训练极小的人体姿态检测 CNN 网络。这是基于一种知识蒸馏的思想，这种思想已经成功地应用于指导物体图像分类的深度模型。特别地，我们推导了一个姿势知识精馏学习目标，将预先训练的较大的教师模型的潜在知识转移到微小的目标姿势估计模型中。

3. 将目前泛化性较强的注意力模块 (PSAM) 加入进目前较为先进的人体姿态模型，提出了新型的人体姿态模型。与现有的偏好特定布局的通道道-空间组合相比，PSAM 布局之间只有微小的度量差异。这表明 PSAM 可能已经耗尽了其仅限通道和仅限空间分支中的表示能力。

4. 我们对于网络结构尾部模型增加了现泛化性好且轻量级的目标检测邻域的特征检测头，探究 PP-PicoDet 与检测下游任务的一体化的可能性。采用的检测头整体在所有模型尺寸上比其他模型结构实现了更好的速度和准确性之间的权衡，进一步提高现有的深度姿态估计方法到实际应用。

参考文献

[1] J. C. R. Licklider, "Man-computer symbiosis," *Ire Transactions on Human Factors in Electronics*, vol. 1, pp. 4–11, 1960.

[2] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1653–1660.

[3] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4724–4732.

[4] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9912. Springer, 2016, pp. 483–499.

[5] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 7103–7112.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.

[8] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 91–99.

[9] H. Fang, S. Xie, Y. Tai, and C. Lu, "RMPE: regional multi-person pose estimation," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2353–2362.

[10] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2980–2988.

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*. Computer Vision Foundation / IEEE, 2019, pp. 5693–5703.

[12] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1812.08008, 2018.

[13] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4929–4937.

[14] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 5385–5394.

[15] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[16] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 4320–4328.

[17] S. Mirzadeh, M. Farajtabar, A. Li, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher," *CoRR*, vol. abs/1902.03393, 2019.

[18] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *3rd International Conference on Learning Representations,*

ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015.

[19] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[20] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 5385–5394.

[21] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 5693–5703.

[22] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.

[23] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 10 440–10 450.

[24] H. Liu, F. Liu, X. Fan, and D. Huang, "Polarized self-attention: Towards high-quality pixel-wise regression," *CoRR*, vol. abs/2107.00782, 2021.

[25] S. Alaparthi and M. Mishra, "Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey," *CoRR*, vol. abs/2007.01127, 2020.

[26] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[27] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[28] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 7794–7803.

[29] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. IEEE, 2019, pp. 1971–1980.

[30] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 9423–9433.

[31] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11211. Springer, 2018, pp. 3–19.

[32] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 3146–3154.

[33] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*. IEEE, 2021, pp. 3530–3538.

[34] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, Eds. ACM, 2006, pp. 535–541.

[35] Q. Li, Z. Zhang, F. Xiao, F. Zhang, and B. Bhanu, "Dite-hrnet: Dynamic lightweight high-resolution network for human pose estimation," *CoRR*, vol. abs/2204.10762, 2022.

[36] G. Yu, Q. Chang, W. Lv, C. Xu, C. Cui, W. Ji, Q. Dang, K. Deng, G. Wang, Y. Du, B. Lai, Q. Liu, X. Hu, D. Yu, and Y. Ma, "Pp-picodet: A better real-time object detector on mobile devices," *CoRR*, vol. abs/2111.00902, 2021.