# Video Pose Distillation for Few-Shot, Fine-Grained Sports Action Recognition

James Hong Matthew Fisher Michaël Gharbi Kayvon Fatahalian

Stanford University Adobe Research

## 摘要

人体姿势对于细粒度的体育动作理解是一个非常有用的特征。然而，在运动视频上运行时，由于域移位和运动模糊、遮挡等因素，姿态估计器往往是不可靠的。这导致下游任务(如动作识别)在依赖于姿态时的精度较差。虽然端到端学习绕过了姿势，但需要更多的标签来泛化。

我们引入了视频姿态蒸馏(VPD)，这是一种弱监督技术，用于学习新视频领域的特征，比如对于姿态估计来说很有挑战性的单人运动。在VPD下，学生网络学习从体育视频RGB帧中提取稳健的姿势特征，这样，每当姿势被认为是可靠的，特征就匹配预训练的教师姿势检测器的输出。我们的策略保留了姿势和端到端的最佳策略，利用了原始视频帧中丰富的视觉模式，同时学习目标视频域中与运动员姿势和运动一致的特征，以避免对与运动员运动无关的模式进行过拟合。

VPD特征在四个真实的体育视频数据集中提高了在少镜头、细粒度动作识别、检索和检测任务上的性能，而不需要额外的地面真实姿势注释。

## 介绍

分析体育视频需要健壮的算法来自动化大规模视频集合中的细粒度动作识别、检索和检测。当体育运动以人为中心时，人体姿势是一个有用的特征。

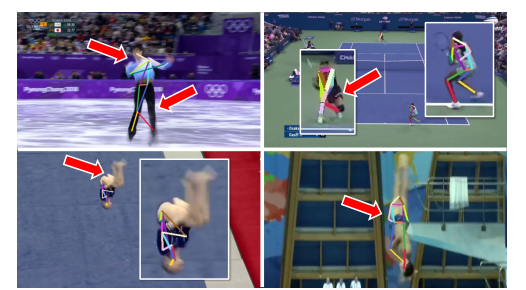


图1当前2D姿态检测器的局限性。最先进的姿态估计器[54]在具有挑战性的运动帧中产生噪声和不正确的结果，这在体育视频中很常见。这里有一些来自花样滑冰、网球、体操和跳水的例子，姿势估计的结果是不正确的。

最先进的基于骨骼的动作识别深度学习技术[36,68]依赖于精确的2D姿态检测来提取运动员的动作，但最好的姿态检测器[54,65]通常在具有复杂模糊和遮挡的快节奏运动视频中失败，通常在对动作至关重要的帧中(图1)。为了规避这些问题，端到端学习模型直接在视频流上操作[8,14,32,52,62,73]。然而，由于它们消耗的是像素而不是姿势输入，当用很少的标签进行训练时，它们往往会锁定特定的视觉模式[10,63]，而不是精细的动作(例如，运动员的衣服或球的出现)。因此，在具有挑战性的体育视频中，当标签稀缺时，先前的姿势和端到端方法往往不能很好地推广细粒度任务。虽然收集带有精细动作和姿势注释的大型数据集是可能的，但针对每一项新运动这样做是无法规模化的。

我们提出了视频姿态蒸馏(VPD)，这是一种弱监督技术，学生网络学会从一个新的视频领域(单个运动)的RGB视频帧中提取鲁棒姿态特征。VPD的设计是这样的，只要姿势是可靠的，特征就与预先训练过的教师姿势检测器的输出相匹配。我们的策略保留了姿态和端到端世界的最佳。首先，像直接监督的端到端方法一样，我们的学生可以利用原始帧中存在的丰富的视觉模式，包括但不限于运动员的姿势，并在姿势估计不成功时继续操作。其次，通过约束我们的描述符以在高置信度姿势可用时与姿态估计器一致，我们避免了与运动员动作无关的视觉模式过拟合的陷阱。第三，弱姿态监督允许我们加强一个额外的约束:我们要求学生不仅预测瞬时姿态，而且预测其时间导数。这鼓励我们的特征随着时间的推移拾取视觉上的相似性(例如，一个运动员从一个姿势发展到另一个姿势)。当我们用监督不力的方式训练学生观看一组没有标签的体育视频时，学生学会了“填补喧闹的姿势老师留下的空白”。这些特性一起导致了一个学生网络，在下游应用中使用时，该网络的特征优于教师的姿势输出。

VPD功能提高了在目标运动领域中少镜头、细粒度动作识别、检索和检测任务的性能，而不需要额外的地面真实动作或姿势标签。我们在四个具有细粒度动作标签的不同体育视频数据集上演示了VPD的好处:跳水[31]、自由体操[49]、网球[69]和一个用于花样滑冰的新数据集。在一个镜头有限的监督设置中，使用提取的VPD特征训练的动作识别模型可以显著优于直接根据来自教师的特征以及来自先前基于骨骼和端到端学习工作的基线训练的模型。例如，当从跳水和自由体操(两个最具挑战性的姿势数据集)中限制在每节课8到64个训练示例时，VPD特征比次优方法分别提高了6.8到22.8%和5.0到10.5%的细粒度分类准确性。即使在标签充足的情况下，VPD仍然具有竞争力，在四个测试数据集中的三个上实现了优越的准确性。总而言之，VPD在利用姿势至关重要的情况下(例如，少镜头)超越了它的老师，在端到端方法占主导地位的情况下(例如，不可靠的姿势和高数据/充分监督设置)也具有竞争力。最后，我们展示了VPD特征在细粒度动作检索和少镜头时间检测任务中的应用。

本文做出了以下贡献:

1. 一种弱监督方法VPD，用于适应新的视频域的姿势特征，在2D姿势估计不可靠的情况下，它显著提高了动作识别、检索和检测等下游任务的性能。

2. 在使用VPD特征的少镜头、细粒度动作理解任务中具有最先进的准确性，适用于各种运动。在动作识别方面，VPD特征在每个类只有8个例子的情况下表现良好，即使训练数据增加，也仍然具有竞争力或最先进。

3.一个新的数据集(花样滑冰)和对现实世界体育视频的三个数据集的扩展，以包括对表演者的跟踪，以促进未来对细粒度体育动作理解的研究。

## 相关工作

**姿态表示**为人类行为理解提供了强大的抽象。尽管在2D和3D姿势估计方面取得了显著进展[44,45,54]，但依赖姿势的下游算法在体育视频中仍然受到不可靠估计的困扰。由于可用的标签很少，对于像细粒度动作识别这样的任务，模型必须同时学习动作和应对噪声输入。

VIPE[53]和CV -MIM[72]表明，学习过的姿态嵌入(将摄像机视图分解并放弃显式3D姿态估计)是有用的;它们在域外3D姿态数据上进行训练，以嵌入2D姿态输入，当2D姿态可靠时，它们是有效的。VPD扩展了这些工作，使用蒸馏取代不可靠的2D姿态估计步骤，用一个直接从像素嵌入到姿态嵌入的模型。[25, 45, 70]从视频中学习人体动作，但产生3D姿势而不是嵌入。

**视频动作识别**以端到端模型为主[3,8,14,32,52,59,59,62,73]，这些模型通常在不同但粗粒度的分类任务(例如，' golf '， ' tennis '等)上进行评估[26,29,42,51,71]。体育领域的细粒度动作识别是最近的发展[31,49]。

除了对体育视频分析是必要的外，单个运动项目内的细粒度分类也很有趣，因为它避免了粗粒度任务中的许多上下文偏差[10,31,63]。[2, 12, 17, 61]也是细粒度数据集，但不同于体育运动中以身体为中心的动作。

基于姿态或骨骼的方法[11,36,68]似乎非常适合于以人为中心的运动中的动作识别。它们依赖于可靠的2D或3D姿势，这些姿势存在于受控设置中捕获的数据集中[35,48]，但不适用于公共体育视频，因为在公共体育视频中没有地面真相，自动检测器通常表现不佳(例如，[31,49])。

在以人为中心的运动数据集中，VPD改进了基于姿态和端到端方法，特别是在姿态不可靠的情况下。像VIPE[53]一样，VPD产生了有效的姿势特征，在某种程度上，相对简单的下游模型，如最近邻搜索[53]或通用BiGRU[16]网络，可以与最先进的动作识别相竞争——在少镜头和高数据的情况下。为了证明这一点，我们将其与第4.1节中的几种最近的动作识别方法[36,52]进行比较。

VPD特征可以用于任何预先训练的姿势特征可能有用的任务，如动作检索和时间细粒度检测(例如，以200毫秒粒度识别网球拍摆动)。后者很有趣，因为先前的基线[13,24]关注的是更一般的类别，而不是单一运动中的以人为中心的动作，而且很少有论文[1,67]涉及少镜头设置。

**少镜头动作识别**文献遵循许多范式，包括元学习、度量学习和数据增强方法[1,7,30,41]。这些作品关注粗粒度数据集[13,26,29,51]，采用各种协议将数据集划分为可见/不可见类和/或执行简化的N-way, K-shot分类(例如，5-way, 1或5- shot)。VPD的不同之处在于，当训练特性时，它对动作标签完全不可知，对下游任务(如动作识别)不需要特定的体系结构。与试图推广到不可见类的“少镜头”学习不同，我们在标准分类任务上进行评估，所有类都是已知的，但在训练时每个类只限制k个示例。我们的评估类似于[50,72]，它们在有限的监督下执行动作和图像识别，并且，像[50,72]一样，我们在不同的监督级别上进行测试。

**自监督/蒸馏。**VPD仅依靠机器生成的位姿注释进行弱监督和蒸馏。VPD类似于[66]，因为蒸馏的主要目标是提高学生的鲁棒性和准确性，而不是提高模型效率。大多数自我监督工作集中在预训练和联合训练场景中，在这些场景中，自我监督损失是次要的，最终任务损失是次要的，后续或并发微调是获得竞争性结果的必要条件[9,18,23,27,33]。相比之下，我们的VPD学生在蒸馏后是固定的。

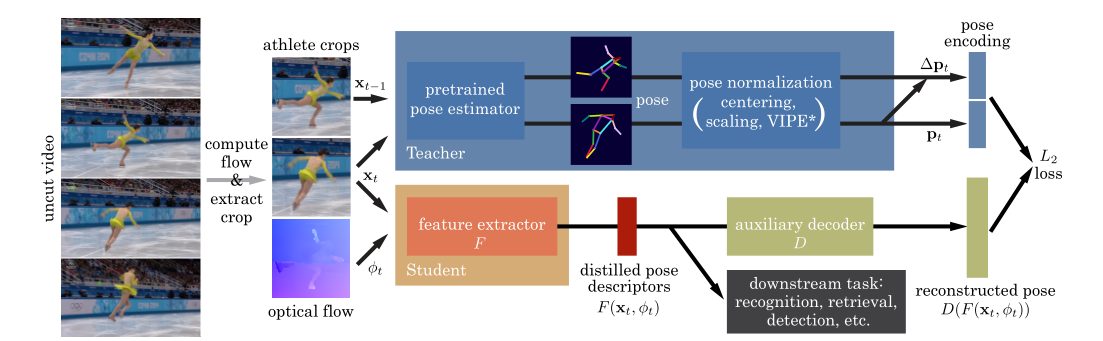


图2方法概述。VPD有两个数据路径:一个是生成监督的教师，一个是学习在目标(运动)领域嵌入姿势和动作的学生。当在帧t上进行训练时，教师应用现成的2D姿态估计器，然后进行姿态规范化步骤，以获得弱姿态特征:pt和∆pt。学生路径接收到局域RGB xt和光流ϕt，并计算一个描述子F (xt， ϕt)∈Rd，全连接网络D由此回归(pt，∆pt)。训练结束后，只保留F在完整测试数据集上提取嵌入。

## 视频姿态蒸馏

我们的策略是从现有的现成的姿势检测器(教师)中提取不准确的姿势估计，通过通用的姿势数据集进行训练，将其提取到一个学生网络中，该网络专门为特定目标运动领域的视频生成健壮的姿势描述符(图2)。学生(章节3.2)将运动员周围裁剪的RGB像素和光流作为输入。它生成一个描述符，从中我们回归由老师发出的运动员的姿势(章节3.1)。我们使用高置信度教师输出的稀疏集作为对学生的弱监督，在一个大型、未切割和未标记的目标域视频语料库(第3.3节)上运行这个蒸馏过程。

由于教师已经接受过培训，VPD在目标视频域中不需要新的姿势注释。同样，学习姿势特征也不需要下游特定于应用程序的标签(如用于识别的动作标签)。然而，VPD确实要求在每个输入帧中识别运动员，因此我们假设在每一帧中为运动员提供一个近似的边界框，作为数据集的一部分。有关讨论和限制，请参阅第5节。

### 3.1教师网络

为了强调VPD是一种可以应用于不同教师模型的通用方法，我们提出了VPD的两种教师变体。第一种使用现成的姿势估计器[54]从xt(第t帧的RGB像素)估计2D关节位置。我们像[53]那样通过缩放和定心对2D关节位置进行归一化，并将关节坐标收集到一个向量pt∈Rd中。我们将其称为2D- vpd，因为老师生成了2D关节位置。

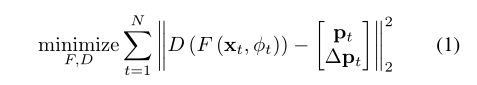
我们的第二个老师变量进一步将2D关节位置处理为一个视图不变的姿势描述符，发出为pt。我们的实现使用VIPE\*来生成这个描述符。VIPE\*是对来自PrVIPE[53]的概念的重新实现，该概念被扩展为在额外的合成3D姿势数据上进行训练[39,46,74]，以实现更好的泛化。我们将这种变体称为VI-VPD，因为老师会生成一个视图不变的姿势表示。(关于VIPE\*及其与Pr-VIPE相比的质量的详细信息，请参阅附录。)

### 3.2学生特征提取器

由于理解运动员的运动，而不仅仅是他们当前的姿势，是许多运动分析任务的关键方面，我们设计了一个学生特征提取器，它对运动员当前姿势pt和姿势变化速率∆pt:= pt−pt−1的信息进行编码。

学生是一个神经网络F，它消耗一个围绕运动员裁剪的彩色视频帧xt∈R3hw，以及它的光流∈R2hw，来自前一帧。H和w为作物的空间维度，t为帧指数。学生生成一个描述子F (xt， ϕt)∈Rd，其维数与教师的输出相同。我们将F实现为具有5个输入通道的标准ResNet-34[19]，并将输入作物的大小调整为128×128。

在蒸馏过程中，F发出的特征通过辅助译码器D，预测当前姿态pt和时间导数∆pt。利用视频的时间方面，∆pt提供了一个额外的监督信号，迫使我们的描述符捕捉当前姿势之外的运动。D被实现为一个完全连接的网络，我们使用以下目标训练组合学生路径D◦F:



由于在推理过程中只需要F来产生描述符，我们在训练结束时丢弃D。

它的老师接受的训练是识别姿势和人体外表的一般分布，而学生F则不同，他专门研究新目标领域的帧和光流(例如，网球场上的运动员)。通过蒸馏的专门化可以让F专注于运动数据中解释姿势的模式。我们不期望，也不要求下游任务，F编码的姿势或人看不到的目标领域(例如，坐在长凳上，芭蕾舞者)，尽管他们可能是教师的培训分布的一部分。第4节中的实验表明，我们的姿态描述符F (xt， ϕt)提高了一些应用(包括少镜头、细粒度动作识别)的精度。

### 3.3训练数据的选择与扩充

**数据的选择。**教师的输出可能会有噪声，这是由于诸如运动模糊和闭塞等挑战，或者由于我们的目标视频和教师受训的数据之间的域转移。为了提高学生的学习能力和防止记忆老师的噪音，我们从老师的弱监督集合中排除了低姿势自信分数(具体来说，平均估计联合分数)的框架。默认情况下，阈值是0.5，但网球使用0.7。调优这个阈值对提取特征的质量有影响(详见补充)。我们还统一随机地保留固定比例的帧(20%)作为学生的验证集。

**数据扩充。**我们应用标准的图像增强技术，如随机调整大小和裁剪;水平翻转;和当训练学生F时进行颜色和噪声抖动。为了确保在水平增大xt和ϕt时保持身体左右方向，我们还必须翻转老师的输出pt。对于2D关节位置和2D- vpd，这是很简单的。为了翻转用于训练VI-VPD的VIPE\*(本身就是一种手性姿势嵌入)，我们必须将二维姿势输入翻转到VIPE\*，然后重新嵌入它们。

## 结果

我们评估了VPD在四个细粒度运动数据集上产生的特征，这些数据集展示了广泛的运动范围。

**花样滑冰**由371名男子和女子单打短节目表演组成，分别来自冬奥会(2010-18年)和世锦赛(2017-19年)，总计17个视频小时。在分类任务FSJump6中，ISU[22]定义了六种跳跃类型。2018年的所有视频(134个动作，520个跳跃)都被拿出来测试。其余的跳转被分割为743/183用于训练/验证。

**网球**包括两个锦标赛(温布尔登和美国网球公开赛)中的9场单打比赛，在击球的框架上标注了挥拍[69]。网球中有七节挥杆课。训练/验证集包含来自5个匹配的4,592/1,142个例子，测试集包含来自其余4个匹配的2,509个例子。根据比赛视频进行分割，由于服装的多样性有限和个人(10名职业球员)的独特性，该数据集具有挑战性。

**自由体操**。我们使用FineGym99数据集[49]中的女子自由体操项目(FX35)，包含1214个例程(34小时)。有35个类和7634个动作。

**跳水48**[31]包含16,997个注释实例，48个跳水序列由国际泳联[15]定义。我们对作者发布的修正的V2标签进行评估，并重新训练现有的最先进的方法GSM[52]，以进行比较。

所有这四个数据集都包含姿态不能很好地估计或不确定的帧，尽管它们的分布是不同的(详情见补充)。如前所述，在快速运动的帧中，由于运动模糊和非典型的运动姿势(如翻转或跳水)，姿态估计通常更差;参见图1中的示例。这些数据集的一个共同挑战是，快速运动帧通常是区分感兴趣的细粒度动作所必需的。

我们假设行为的主体已被识别和跟踪。如果画面中有多个人，快速移动的具有挑战性姿势的运动员通常会被忽略:例如，与静止的观众或裁判相比，他们被检测到的信心更低，或者根本没有被检测到。在公平的比较中，我们通过提供与我们的方法相同的输入来提高基线，这将显著改善它们的结果。

### 4.1细粒度动作识别

细粒度动作识别测试VPD捕捉运动员姿势和动作精确细节的能力。我们考虑两种few-shot的设定，一个是只提供有限数量的动作例子，还有一个是传统的完全监督设定，其中训练集中的所有动作示例都可用。

我们的VPD功能是从体育语料库中的训练视频中提取出来的，未经切割，没有标签。为了提取测试集上的特征，我们使用固定的VPD学生F。VI-VPD和2D- vpd特性保持相同的维度d，它们的老师:d = 64对于VIPE\*来说，d = 26对于规范化的2D来说。对于Diving48, VIPE\*的d = 128，因为我们还提取了垂直翻转姿势的姿势嵌入并将它们连接起来。这种数据扩充对于VIPE\*来说是有益的，因为潜水姿势通常是倒置的，而在VIPE\*训练的超域3D姿势数据集中，潜水姿势的表现不太好。

**动作识别模型。**为了使用VPD进行动作识别，我们首先将每个动作表示为一个姿势特征序列。然后，我们使用双向门控循环单元网络(BiGRU)[16]在学生F生成的(固定)特征上进行训练，对动作进行分类。由于我们的特征是手性的，许多动作可以在左右方向下执行，我们将规则和水平翻转框架嵌入到学生中。参见附录了解实现细节。

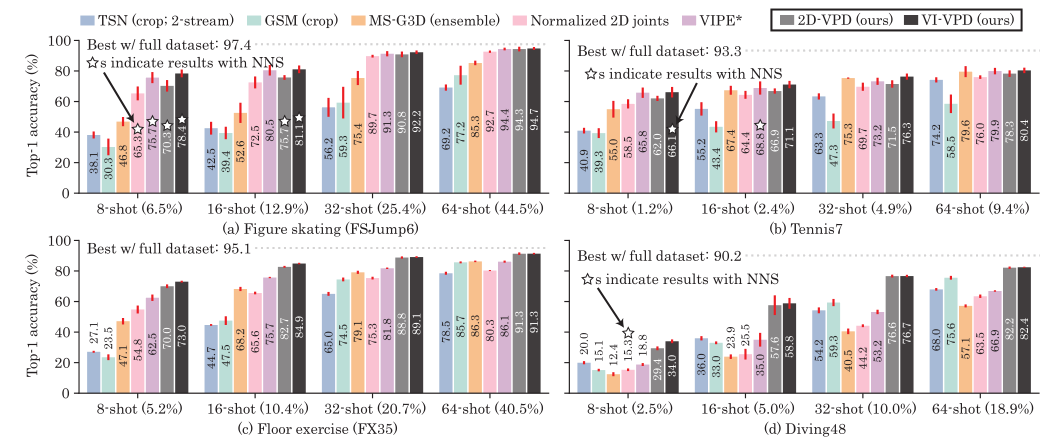


图3few-shot细粒度动作识别的精度。百分比给出了整个训练集的百分比。使用完整数据集作为监督的最先进的准确性用虚线表示(参见表1)。在除Diving48之外的每个数据集上，基于姿势的基线(MS-G3D [36]， 2D关节和VIPE\*)在几次快照设置中超过了端到端基线(GSM[52]和TSN[62])，这表明了姿势在标签稀缺时的重要性，以及姿势不可靠时的挑战。VI-VPD在FX35和Diving48上的性能明显优于基线和之前的方法;在FSJump6和Tennis7上的准确性也略有提高，但与VIPE\*相似。上面的星号结果使用最近邻搜索(NNS)代替BiGRU架构(NNS在这些情况下表现更好;完整结果见补充)。

先前的姿态嵌入工作已经探索了使用序列对齐后的最近邻检索[53]。我们还测试了一种最近邻搜索(NNS)方法，该方法使用动态时间扭曲来计算姿态特征序列之间的匹配代价。对于网络神经网络，每个测试例与所有训练例进行搜索，并预测最佳对齐匹配的标签。在大多数情况下，BiGRU是优越的，尽管NNS可以在少镜头的情况下有效，当这种情况下，我们指出。

**基线。**我们将提取的2D-VPD和VI-VPD功能与几个基线进行比较。

1. 来自老师的特性:VIPE\*或归一化的2D关节位置，使用相同的下游动作识别模型和数据增强。

2. 基于骨架的:MS-G3D集合[36]和STGCN[68]。两个基线接收相同的跟踪2D姿态，用于监督VPD。

3.端到端:GSM [52]， TSN [62]， TRNms[73](多尺度)。我们用裁剪过的运动员和全帧(w/o裁剪)作为输入进行测试，我们发现裁剪显著提高了四个数据集上的少镜头设置和除Diving48之外的所有数据集上的完全监督设置的准确性。适用时，结合RGB和光流模型的结果表示为2-stream。

#### 4.1.1few-shot和有限监督的设定

实验协议。每个模型给出了每个动作类的k个例子，但可能会使用未标记的数据或者从其他数据集中获取的知识。例如，基于骨骼的方法依赖于2D姿势检测;VIPE\*利用域外3D姿势数据;而VPD功能则是在未切割、未标记的训练视频中提取出来的。这个实验设置反映了现实世界的情况，其中标签很少，但未标记和域外数据大量存在。我们的评估指标是完整测试集上的top-1精度。为了控制为每个少镜头实验所选择的训练示例的变化，我们在数据的五个随机抽样和固定子集上运行每种算法，对于每个k，并报告平均精度。

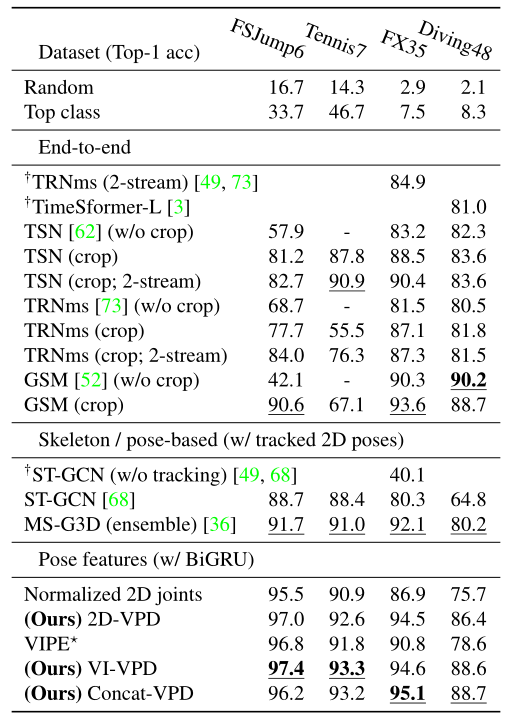


表1:使用所有训练数据的细粒度动作识别的准确性。顶部结果被加粗，每个方法类别被下划线。†表示先前工作的最佳结果。VI-VPD在FSJump6、Tennis7和FX35上实现SOTA精度，即使基线通过跟踪和裁剪输入得到改进。在Diving48上，VIVPD落后端到端GSM(w/o 裁剪)1.6%。VI-VPD和2D-VPD功能都具有竞争力;连接它们(Concat-VPD)可能会略微提高准确性。

**结果。**图3将2D-VPD和VI-VPD特性与它们的老师(和其他基线)进行了比较。在FSJump6和Tennis7上，VI-VPD比其最先进的VIPE\*有轻微的改进，精确度在几个百分点之内。在k≤32的情况下，FX35和VI-VPD的准确率分别比VIPE\*提高了10.5%和5%，在k = 64的情况下，比MS-G3D集合提高了5%。同样，在Diving48上，端到端GSM和2流TSN比非vpd基于姿态的方法更好，VI-VPD提高了6.8到22.8%的精确度。我们在FX35和Diving48上的结果表明，VI-VPD有助于将姿势的好处转移到最不可靠的数据集。

虽然视不变(VI)特征通常比二维相似物表现更好，但在摄像机角度不同的运动(如花样滑冰和自由体操)和小k时，VI-VPD和2D- vpd之间的精度差异更明显，在训练过程中，动作识别模型只能观察到少量视图。

#### 4.1.2传统的全训练集设置

即使在高数据体系中，VPD特征也具有竞争力(表1)。在所有四个数据集上，VI-VPD和2D-VPD都比它们的老师显著提高了准确性。VI-VPD在FSJump6(在VIPE\*上为0.6%)、Tennis7(在VIPE\*上为1.5%)和FX35(在GSM\*上为1.0%，有截断输入)数据集上也达到了最先进的准确性。

Diving48对于基于姿势的方法来说尤其具有挑战性，VI-VPD在没有裁剪的情况下比GSM表现更差，差了1.6%。带有裁剪的GSM也差了1.5%，可能是由于我们跟踪的错误和限制。然而，VIVPD的表现明显优于基于顶部姿势的基线(比MS-G3D, ensemble的8.4%)。

我们的结果表明，VPD的成功并不局限于少镜头。然而，由于表1中的许多方法在给定充足的数据时可以产生较高的准确率，达到或超过90%，因此我们将提高标签效率视为VPD和未来工作的更重要目标。

#### 4.1.3消融和额外实验

我们强调VPD的两个重要的消融来理解VPD改进的来源:(1)分析蒸馏方法的部分和(2)只用视频的动作片段进行蒸馏。我们还考虑(3)一个未标记的设置，其中VPD是在整个视频语料库中提取的。其他实验请参考补充。

**蒸馏方法分析。**表2(a)显示了当我们在没有流量输入ϕt和没有运动预测1的情况下提取消融2DVPD和VI-VPD特征时，动作识别精度的提高。增量改进通常在少镜头设置中最为明显，在FX35和Diving48数据集上，VPD产生了最大的好处(见章节4.1.1)。

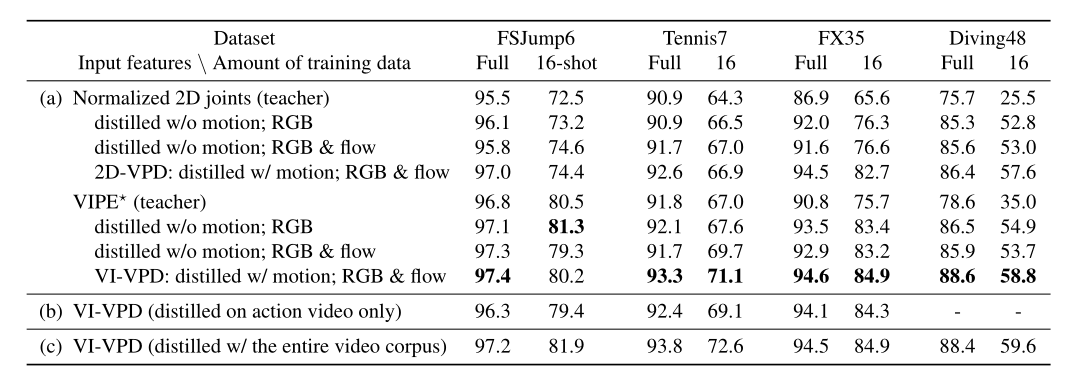


表2：动作识别实验。在全训练集和16个镜头场景下(a)对蒸馏方法进行消融，(b)只使用数据集的动作部分进行蒸馏，(c)在整个视频语料库(包括测试视频，没有标签)上蒸馏VI-VPD特征时，精确度为第一。结果由BiGRU提供。

有了VIPE\*，仅从RGB中蒸馏就能产生很大的效果(2.7%和7.7%，在FX35的全速和16次设置下;跳水是7.9%和19.9% 48)。在RGB之外添加流，而不添加运动，会产生混合的结果。最后，添加运动预测和解码器D，进一步提高了结果(FX35上1.1%和1.5%，全镜头和16镜头;在跳水48数据集上2.1%和3.9% )。蒸馏运动对FSJump6和Tennis7的影响在16杆设置下是混合的，尽管完全设置显示出改进。

2D-VPD可以被看作是视图不变量(VIPE\*)的消融，当进一步消融时，会显示出类似的模式。

**仅对视频的动作部分进行VPD训练。**细粒度操作类在FSJump6、FX35和Tennis7中占视频的比例分别不到7%、8%和28%。通过只提取训练视频的动作部分的VI-VPD特征，我们评估了在未切割视频上蒸馏VI-VPD是否能提高动作识别的泛化能力。

表2(b)总结了结果，并表明只提取动作视频在我们的数据集上表现较差。这是很有前途的，因为(1)未切割的表现比检测到动作的表现更容易获得，(2)在低监督设置下，VI-VPD提高了准确性，即使在训练语料库的其余部分中没有检测到动作。这也表明提炼更多视频可以提高功能的质量。

**用整个视频语料库进行蒸馏。**在一个新领域用视频构建真实世界的应用程序时(例如，[69])，未标记的语料库通常是起点。由于VPD仅由机器从未标记的视频中生成的姿态估计来监督，所以VPD特征可以从所有可用的视频中提取出来，而不仅仅是训练数据表2(c)显示了将VI-VPD与训练和测试视频(未切割和无标签)联合提取的结果。如果有改进的话，在所有四个数据集上都是很小的(≤1.5%，在Tennis7上获得16个镜头)，并且证明了在大数据集上蒸馏的VI-VPD能够在不查看测试视频的情况下进行泛化。

### 4.2动作检索

动作检索衡量VPD特征在搜索相似的未标记动作时的效果。在这里，VPD特征是在整个未标记的语料库上提取的。

**实验协议。**给定一个查询动作，表示为一系列姿势特征，我们使用姿势特征和动态时间扭曲之间的L2距离对语料库中的所有其他动作进行排名，以计算对齐得分。如果一个结果具有与查询相同的细粒度操作标签，则该结果被认为是相关的，我们通过k个结果的精度(所有查询的平均值)来评估相关性。

**结果。**在表3和所有四个数据集的所有截断点上，VPD特征的表现都优于它们的老师。在FX35和Diving48上可以看到相当大的改进。如果相机角度的数量是有限的(例如，网球7和跳水48)，视图不变量并不总是导致最高的精度，尽管它可能有助于检索更多样化的结果。

### 4.3few-shot动作检测的姿态特征

在时间粒度较细且标签较少的情况下，检测细粒度的动作，可以实现少镜头识别和检索等任务。我们评估了花样滑冰和网球数据集上的VPD特征，以分别对跳跃和摆动进行时间定位。平均跳跃长度为1.6秒(≈40帧)，而挥杆定义为球接触帧周围的200毫秒(≈5帧)。

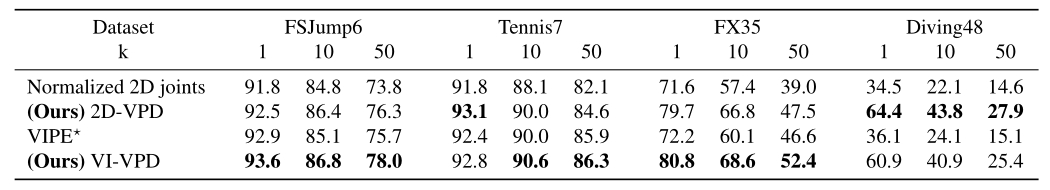


表3：动作检索:Precision@k结果(%)根据动态时间扭曲的对齐评分排序。VPD可以在所有四个数据集上得到更多相关的结果。在FSJump6和Tennis7上的进展不大，而在FX35和Diving48上的巨大改进表明，在姿势估计最不可靠的情况下，VPD特征更优越。

**实验协议。**我们遵循与FSJump6和Tennis7相同的视频级训练/测试分割，并且只在训练视频上提取特征。作为一个简单的基线方法，我们训练一个输出每帧预测的BiGRU，并将其合并生成预测的动作间隔(详见补充)。BiGRU根据五种动作(花样滑冰)和五种点(网球)的时间标签进行训练。为了得到更一致的结果，我们进行了五次交叉验证，并集成了每帧预测。在表4中，我们报告了在不同水平的时间交与并(tIoU)上的平均精度(AP)。

**结果。**VPD在这两个任务上都提高了AP。网球挥杆持续时间短，意味着每帧姿势估计中的噪声有很大的影响，并且VI-VPD在每个tIoU阈值(高达7.4比VIPE百科(tIoU = 0.5)提高AP。

## 局限和讨论

VPD需要**跟踪目标**，以确保姿势是正确的人。现实世界的体育视频中，除了主体之外，往往还会包含很多人，比如观众、裁判等。4.1节数据集中的跟踪注释是使用现成模型和启发式自动计算的(详情见补充)。

这是可能的，因为运动员在外表、空间和时间上都是突出的——运动视频是跟踪[4,64]和检测突出区域[6]的自然应用。我们观察到，在其他之前的方法(如[52,62,68])上跟踪和非跟踪输入之间的精度差异可以是惊人的(对于GSM [52]， FSJump6为48%，对于ST-GCN为40%，FX35为40% [68];见表1)。

为了评估我们的姿势特征的质量，我们专注于单个运动员或同步运动员的动作(包含在跳水48)。涉及许多人的任务和动作需要一个更复杂的下游模型，可以处理每帧的多个描述符或姿势。

**未来工作。**首先，用于监控VPD的2D姿态估计相对于摄像机视图本身具有模糊性，而诸如深度或行为先验等附加信息可以帮助缓解这种模糊性。除了运动和VIPE之外，其他监管不力的来源也可能有所帮助。第二，我们的蒸馏过程是线下的;在姿势特征提取阶段支持在线训练，类似于[43,56]，在时间演化的数据集中可能是有益的。蒸馏用于显式的2D或3D姿态估计是另一种可能。虽然VPD特征可以在有限的数据下提高准确性，但在少镜头和半监督任务上的表现仍有很大的提高空间，我们希望未来的工作继续探索这些主题。

## 结论

姿态特征对于研究新的体育视频数据集中以人为主的动作是很有用的。然而，对于现成的模型来说，这样的数据集往往具有挑战性。我们的方法VPD，通过从现有的姿势估计器中提取知识，在困难和少标签的背景下，提高了姿势特征的可靠性。VPD可以学习一些特征，这些特征能够提高目标(运动)领域中传统动作理解任务和少镜头动作理解任务准确性。我们相信，我们的基于蒸馏的方法是解决新视频领域应用面临的挑战的一个有用的范例。

**声明。**本工作得到III-1908727下的美国国家科学基金会(NSF)和Adobe Research的支持。我们也感谢匿名审稿人。