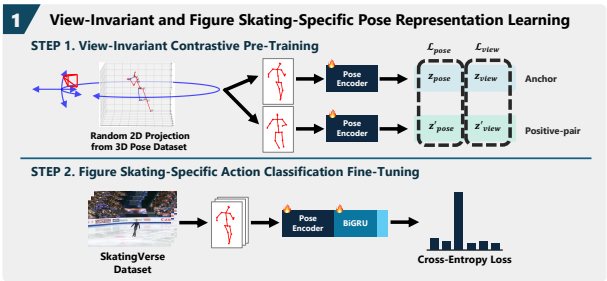
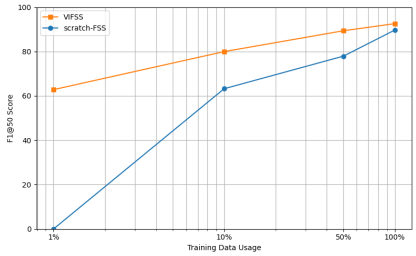
Graphical Abstract

VIFSS: 用于时间动作分割的视图不变和花样滑冰特定的姿态表示学习

Ryota Tanaka, Tomohiro Suzuki, Keisuke Fujii





Features	Acc	F1@50	F1@75	F1@90
2D Pose (baseline)	71.34	78.78	75.74	35.39
3D Pose	70.17	76.57	71.62	29.52
VIFSS (proposed)	**85.82**	**92.56**	**90.65**	**49.62**
scratch-FSS (ablation)	82.72	89.65	86.42	41.03

TAS Results in Element-level

Highlights

VIFSS: 用于时间动作分割的视图不变和花样滑冰特定的姿态表示学习

Ryota Tanaka, Tomohiro Suzuki, Keisuke Fujii

- We introduce a new framework for temporal action segmentation using VIFSS (View-Invariant and Figure Skating-Specific) pose representations.
- We construct FS-Jump3D, the first publicly available 3D pose dataset for figure skating jumps.
- We propose a fine-grained annotation strategy to capture precise jump procedures.
- Our VIFSS approach achieves superior performance on figure skating TAS tasks over 2D pose and 3D pose features.

VIFSS: 用于时间动作分割的视图不变和花样滑冰特定的姿态表示学习

Ryota Tanaka[a], Tomohiro Suzuki[a], Keisuke Fujii[a,b,*]

[a]*Graduate School of Informatics, Nagoya University, Chikusa-ku, Nagoya, 464-8601, Aichi, Japan*
[b]*RIKEN Center for Advanced Intelligence Project, 1-5 Yamadaoka, Suita, 565-0871, Osaka, Japan*

Abstract

从视频中理解人的动作在包括体育分析在内的各种领域中发挥着关键作用。在花样滑冰中，准确识别滑冰者执行的跳跃类型和时间点对于客观的表现评估至关重要。然而，由于跳跃程序的细致和复杂性，该任务通常需要专家级知识。尽管最近的方法尝试使用时间动作分割（TAS）来自动化这个任务，但对于花样滑冰来说，TAS 存在两个主要限制：标注数据不足，以及现有方法没有考虑跳跃动作固有的三维方面和程序结构。在这项工作中，我们提出了一个新的花样滑冰跳跃 TAS 框架，该框架明确结合了跳跃运动的三维性质和语义程序。首先，我们提出了一种新的视角不变、花样滑冰特定的姿态表示学习方法（VIFSS），结合了对比学习作为预训练和动作分类作为微调。为了实现视角不变的对比预训练，我们构建了 FS-Jump3D，这是第一个专门为花样滑冰跳跃公开可用的 3D 姿态数据集。其次，我们介绍了一种细致的注释方案，标记了“进入（准备）”和“着陆”阶段，使 TAS 模型能够学习跳跃的程序结构。大量实验表明我们框架的有效性。我们的方法在需要识别跳跃类型和旋转级别的元素级 TAS 上实现了超过 92 % F1@50。此外，我们证明视角不变的对比预训练在微调数据有限时特别有效，突出了我们方法在真实世界场景中的实用性。

Keywords: computer vision, temporal action segmentation, 3D human pose, contrastive learning, sports, dataset

*Corresponding author
Email addresses: tanaka.ryota@g.sp.m.is.nagoya-u.ac.jp (Ryota Tanaka), suzuki.tomohiro@g.sp.m.is.nagoya-u.ac.jp (Tomohiro Suzuki), fujii@i.nagoya-u.ac.jp (Keisuke Fujii)

理解视频中的人类动作是一项基础任务，具有广泛的应用范围，从体育到自动驾驶、监控和老年护理。在体育领域，视频数据在重放判决、自动判决和训练中提供反馈等应用中起着关键作用。特别是在花样滑冰中，由于跳跃元素的技术复杂性逐渐增加，使得视频的利用变得不可或缺。目前，花样滑冰的评分主要依赖于技术专家和重放操作员的人工标注，他们分别在表演中识别和记录跳跃类型和时机。这个过程需要专业知识和相当数量的时间和精力。为了解决这些挑战，本研究提出了一种自动化的流水线系统，通过时间动作分割（TAS）从广播视频中识别花样滑冰跳跃的类型和时机。TAS 的目标是将未修剪的视频分割成逐帧的动作标签。现有的 TAS 基准数据集主要由程序性活动的视频录音组成，比如烹饪、组装家具或玩具。这些数据集在帧级别标注明细动作，例如“拿起杯子”和“打开冰箱”，并训练 TAS 模型理解这些程序性转换。一个广泛使用并成功的方法是将 TAS 任务分为两个阶段：提取时间特征并基于这些特征分割动作。提取特征的质量大大影响分割性能，使得特征选择成为一个关键考虑因素。在标准的 TAS 任务中，来自物体或工具的视觉线索常常有助于识别动作。因此，通常采用利用 3D-CNNs 提取的基于图像的特征，如 I3D（Carreira and Zisserman, 2017）。然而，以往关于理解花样滑冰动作的研究（Liu et al., 2021; Hong et al., 2021）提出，二维姿态或嵌入式姿态表示可能比基于图像的特征更具信息性。这反映了花样滑冰作为一个以动作为中心的任务的本质，其中识别人类运动比识别物体或背景的外观更为重要。尽管如此，花样滑冰动作本质上是三维的，并在广播画面中从各种视点进行捕捉。利用三维姿态信息构建视角不变的姿态表示，为这一任务提供了令人期待的潜力。然而，先前没有研究探讨基于 3D 姿态表示对于 TAS 的有效性。此外，如图 **??** (a) 所示，现有的关于花样滑冰的 TAS 研究（Liu et al., 2021; Gan et al., 2024）在标注跳跃时未考虑进入（准备）或着陆阶段。由于 TAS 模型旨在学习序列动作过程，明确标注完整的跳跃过程，包括从进入到着陆的过渡，可能有助于有效学习。

在本文中，我们提出了一种新的 TAS 框架，该框架利用了一种视图不变和花样滑冰特定（VIFSS）的姿态表示学习方法，如图 **??** 所示，该方法捕捉到花样滑冰动作的三维特性。首先，本研究通过提出一种新的注释策略来解决上述问题，该策略针对花样滑冰 TAS 任务整合了详细的跳跃过程。我们还研究了视图不变和花样滑冰特定的 3D 姿态表示作为 TAS 模型输入的效用。为了解决在 TAS 中整合花样滑冰动作三维特性所遇到的挑战，我们引入了 VIFSS，该方法使用姿态嵌入特征而不是直接使用估算的 2D 或 3D 关节坐标。虽然原始关节坐标可以几何地表示动作，但它们往往缺乏对视点变化的鲁棒性，并可能阻碍在不同领域中的泛化。为了克服这些限制，我们提出了一种两阶段学习方法。首先，我们在 3D 姿态数据集上使用对比学习对姿态编码器进行预训练，以获取视图不变的表示。然后，

我们通过针对花样滑冰领域的动作分类对编码器进行微调。通过将视图不变的预训练与领域特定的微调结合，我们的方法在有限注释和多样视图的条件下，为花样滑冰中的 TAS 提供了更有效和灵活的表示学习。

尽管用于姿态表示的对比学习具有不需要显式 3D 姿态数据集的优势，只依赖于同步的多视角视频，但现有的 3D 姿态数据集可以进一步加速并增强学习过程。然而，大多数 3D 姿态数据集主要集中在日常活动，如走路或交谈（Ionescu et al., 2013; Joo et al., 2015; Von Marcard et al., 2018; Sigal et al., 2010），而体育特定的 3D 姿态数据集（Ingwersen et al., 2023; Nibali et al., 2021; Suzuki et al., 2025）仍然有限。此外，据我们所知，目前没有公开可用的 3D 姿态数据集能够捕捉花样滑冰运动。为了解决这一空白，我们构建了 FS-Jump3D，这是首个针对花样滑冰跳跃的开源 3D 姿态数据集，使用无标记动作捕捉技术。通过将 FS-Jump3D 与其他 3D 姿态数据集（Ionescu et al., 2013; Mehta et al., 2017; Li et al., 2021）集成到统一的训练框架中，我们能够有效学习适用于花样滑冰复杂而动态运动的姿态表示。本研究的目标是使花样滑冰跳跃的 TAS 能够精细检测起跳和落地时机。

本研究的主要贡献如下：

1. 我们提出了一种称为 VIFSS 的两阶段学习框架用于 TAS，该框架结合了对比预训练和动作分类微调，以实现花样滑冰的视角不变和领域专用嵌入。
2. 我们构建了两个数据集，以学习花样滑冰的特点：（i）FS-Jump3D，一个花样滑冰跳跃的三维姿态数据集，以及（ii）一个精细的 TAS 注释集，捕捉跳跃过程以及起跳和着陆的精确时间。
3. 我们进行了全面的实验，以验证所提出的方法在花样滑冰 TAS 任务中的有效性。

FS-Jump3D 数据集是公开可用的，以支持花样滑冰研究和复杂体育运动的分析。我们还发布了对跳跃过程敏感的 TAS 注释以及所提出的两阶段姿势表示学习框架的完整项目代码。我们通过一种新的 VIFSS 方法增强 TAS 性能，并进行了详细分析的综合实验，扩展了我们之前的会议研讨会论文（Tanaka et al., 2024）。

1. 相关工作

1.1. 体育视频理解

在所有类型的体育运动中，定量评估运动员的表现是至关重要的。因此，各种研究任务被提出，以理解体育视频中的人类动作，包括动作预测（Honda et al., 2022; Kaneko et al., 2024; Ibh et al., 2024），动作识别（分

类） (Held et al., 2023; Zhu et al., 2022; Deyzel and Theart, 2023; Chappa et al., 2023; Askari et al., 2023, 2024; Ibh et al., 2023; Nonaka et al., 2024) ，时间动作检测 (Nonaka et al., 2022; Santra et al., 2025; Giancola et al., 2023; Cabado et al., 2024; Xarles et al., 2024) ，动作质量评估 (Pirsiavash et al., 2014; Parmar and Morris, 2019a,b; Okamoto and Parmar, 2024) ，以及 TAS (Liu et al., 2021; Gan et al., 2024) 。

传统方法在这些任务中通常利用通过卷积神经网络或光流 (Carreira and Zisserman, 2017; Giancola et al., 2018; Sanford et al., 2020; Piergiovanni and Ryoo, 2018; Shim et al., 2018) 提取的图像特征。然而，近年来，基于物体检测 (Redmon et al., 2016; Jocher et al., 2023) 、物体跟踪 (Aharon et al., 2022; Zhang et al., 2022) 和 2D/3D 姿态估计 (Yeung et al., 2024; Deyzel and Theart, 2023; Ibh et al., 2023, 2024) 的方法已成为从视频中提取时空信息的主流。这些方法与基于图像的特征不同，它们对拍摄条件和摄像机角度的变化更加稳健，在体育领域特别有用，因为在该领域中，收集大规模数据集通常具有挑战性。

在体育中理解人类动作涉及识别各种元素,包括群体层面的行为 (Chappa et al., 2023; Askari et al., 2023, 2024) 、个体层面的动作及关节动态 (Suzuki et al., 2024; Tanaka et al., 2023a) ，以及如球类或其他设备的物体轨迹 (Van Zandycke and De Vleeschouwer, 2022; Nakabayashi et al., 2024; Gossard et al., 2024; Vandeghen et al., 2022; Liu and Wang, 2022) 。因此，像物体检测、跟踪和姿态估计等技术的进步预计将显著提高元素识别的准确性，从而增强下游视频理解任务的整体性能。

在花样滑冰的背景下，已经研究了多种任务，包括动作识别、时间动作检测、动作质量评估和 TAS。由于花样滑冰是一项评估技术技能和艺术表现的裁判性运动，许多研究旨在自动化评分过程。一种常见的方法是训练深度学习模型以从表演视频和过去比赛中的相应裁判分数中回归整体分数。然而，这些端到端方法通常缺乏透明性和可解释性，使得难以理解模型如何评估技术熟练度或艺术质量。

因此，近期自动化花样滑冰评分的研究已将重点转向对单独的评分组件进行建模。这包括跳跃类型分类 (Liu et al., 2020; Hong et al., 2021) 、刃错误检测 [1] (Tanaka et al., 2023b,a) 和不足旋转检测 (HIROSAWA, 2020) 。此外，TAS，还对未剪辑的表演视频进行按帧动作标签分配研究，以识别花样滑冰中的跳跃和旋转动作。

与许多强调团队动态或物体轨迹的其他运动不同，花样滑冰在很大程度

[1]边缘错误检测评估滑冰者在起跳时是否根据规则使用了正确的刃（即刀刃倾斜）。这主要适用于翻转跳和鲁兹跳，并影响跳跃的基础分值和执行分数。

上依赖于分析滑冰者的细粒度、连续动作。FSD-10（Liu et al., 2020）采用光流和二维姿态估计解决了 10 种最常见跳跃和旋转类型的分类任务，研究表明，关注关节位移较大的帧可以改善模型性能。MCFS（Liu et al., 2021）通过使用二维姿态序列作为输入特征来提高跳跃和旋转的任务分割准确性。这些研究表明，基于姿态的特征，特别是那些对背景或服装不变的特征，有助于花样滑冰视频的理解。

虽然 2D 姿态特征的实用性已经得到验证，但基于视图不变的 3D 姿态嵌入的潜力仍未被充分探索。此外，诸如准备和落地阶段的跳跃过程元素对于识别跳跃类型和时间至关重要。然而，目前没有一个现有的花样滑冰 TAS 数据集明确标注了这些过程。在这项工作中，我们提出了一种新的标注方法，结合跳跃过程元素，并研究了使用视图不变的姿态表示用于花样滑冰 TAS 的有效性。

1.2. 3D 姿势估计与姿势表示学习

视图不变姿态表示可以通过两种主要方法实现：直接估计 3D 姿态或学习姿态的潜在嵌入。

对于 3D 姿态估计，主要有两种常见方法：一种是直接从图像中估计 3D 姿态，另一种是将估计的 2D 姿态提升到 3D，这种方法被称为 2D 到 3D 提升。随着最近 2D 姿态估计的进步，后一种方法备受关注，使研究人员能够仅专注于 2D 到 3D 的扩展。这种方法在最近的研究中表现出了强大的性能。像 SimpleBaseline（Martinez et al., 2017）、SemGCN（Zhao et al., 2019）和 JointFormer（Lutz et al., 2022）等方法提出了基于单帧 2D 到 3D 提升的 3D 姿态估计模型。相比之下，MotionBERT（Zhu et al., 2023）和 MotionAGFormer（Mehraban et al., 2024）整合了来自多个连续帧的时间建模用于 2D 到 3D 提升。虽然这些时间模型能够实现更准确的 3D 姿态估计，但它们往往会面临处理速度较慢的困境，并且仅限于固定长度的帧序列作为输入。

在姿态表示学习中，目标是将姿态嵌入到一个潜在空间中，大多数方法同样采用 2D 姿态作为输入。Pr-VIPE（Sun et al., 2020; Liu et al., 2022）引入了一种基于三元组损失的框架，以学习一个能够以视角不变的方式捕捉姿态相似性的潜在空间。由编码器从 2D 姿态输入生成的姿态嵌入对于动作识别和视频对齐等下游任务非常有效。CV-MIM（Zhao et al., 2021）提出了一种基于最大化互信息的对比学习方法，用于从 2D 姿态中解开与姿态相关和视图相关的表示。这种方法在捕捉姿态特定和视图特定特征方面表现出色，但一个限制是它在训练过程中没有明确利用视图标签。因此，尚不清楚学习到的视图相关特征是否能够准确反映不同视图之间的视点相似性。

大多数现有的三维人体姿态数据集，如 Human3.6M (Ionescu et al., 2013)、3DPW (Von Marcard et al., 2018)、MPI-INF-3DHP (Mehta et al., 2017) 和 HumanEva-I (Sigal et al., 2010)，依赖于光学标记的动作捕捉系统，并主要包含行走和挥手等日常活动。虽然这些系统提供高精度的追踪，但它们可能限制自然的身体动作，使其不太适合捕捉运动的复杂动态。

为了克服这一局限性，一些最新的数据集采用了无标记动作捕捉系统来记录更具动态的活动，包括足球踢球和棒球投球 (Ingwersen et al., 2023)、跳跃和接球 (Nibali et al., 2021)、以及舞蹈表演 (Li et al., 2021)。这些数据集扩展了 3D 姿态估计模型在传统数据集中未能表现的高动态场景中的适用性。

尽管取得了这些进展，目前还没有公开的花样滑冰数据集，而花样滑冰是一项专门在冰场这个独特环境中进行的运动。在冰上，滑冰者利用惯性和离心力来执行高度复杂和动态的动作，这些动作与其他运动中的动作有显著不同。因此，现有的数据集不足以建模或估算花样滑冰特有的姿势。

为了填补这一空白，我们介绍了 FS-Jump3D，这是第一个专注于花样滑冰跳跃的 3D 姿态数据集。该数据集是在冰上使用硬件同步的多摄像头设置和无标记动作捕捉系统拍摄，提供了来自专业滑冰运动员的高保真 3D 跳跃动作数据。有关数据集特征的详细比较，请参见之前会议论文中的表 1

2. 方法

本研究旨在使用 VIFSS（视图不变且特定于花样滑冰）姿势表示作为输入特征来进行花样滑冰跳跃的 TAS。先前的研究 (Liu et al., 2021) 已证明，基于姿势的方法使用 2D 关键点优于基于图像的方法，因为它们对外观变化（如服装或背景）不那么敏感。然而，花样滑冰动作本质上是三维的，而二维姿势是将三维身体配置投影到图像平面上，其结果可能因相机视点而变化很大。直接基于 3D 姿势的方法面临两个关键挑战：（1）TAS 表现高度依赖于 3D 姿势估计的质量，以及（2）简单的基于坐标的表示可能缺乏捕捉复杂运动动态的能力。

为了解决这些挑战，我们提出了一种方法，该方法利用对比学习和一个三维姿态数据集来预训练一个姿态编码器，以提取视角不变的姿态嵌入。预训练的编码器随后在一个花样滑冰特定的动作分类任务上进行微调，所得的姿态嵌入则作为下游 TAS 任务的输入特征。我们提出的姿态表示学习流程的概述如图 1 所示。

通过对比学习进行视角不变姿态表示学习需要一个合适的 3D 姿态数据集。然而，大多数现有数据集 (Ionescu et al., 2013; Mehta et al., 2017; Von Marcard et al., 2018; Sigal et al., 2010) 集中于日常活动，缺乏花样滑

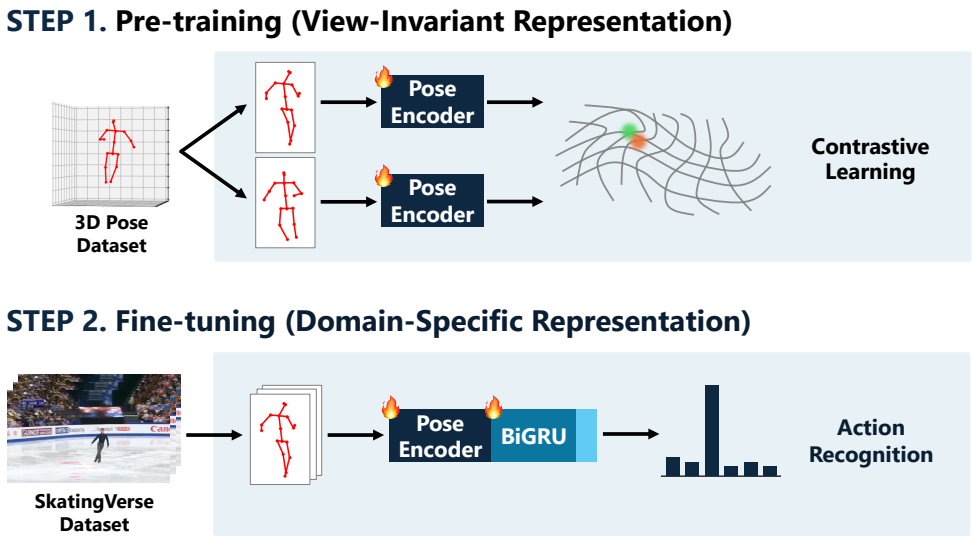


Figure 1: 针对花样滑冰定制的姿态表示学习流程概述。在第一阶段，将对 3D 姿态数据集（FS-Jump3D 和 （Ionescu et al., 2013; Mehta et al., 2017; Li et al., 2021））应用对比学习，以预训练一个姿态编码器，该编码器能够在潜在空间中将来自不同视点的相似姿态映射到接近的位置。在第二阶段，在滑冰动作分类数据集 SkatingVerse Gan et al. (2024) 上微调预训练的编码器，以便专门化姿态表示，用于花样滑冰动作。

冰中由惯性和离心力驱动的领域特定的复杂和动态动作。据我们所知，目前没有公开可用的专为花样滑冰定制的 3D 姿态数据集。为了解决这个限制，我们引入 FS-Jump3D，这是一个捕捉精英级花样滑冰跳跃动作的 3D 姿态数据集，包括使用无标记动作捕捉技术记录在冰上的三周跳跃。

我们进一步提出了一种针对 TAS 的新颖标注策略，该策略考虑了花样滑冰跳跃的程序结构。由于 TAS 模型通常旨在学习人类动作的程序，因此对花样滑冰跳跃的准备和着陆阶段进行标注可以引导模型更好地捕捉跳跃的运动序列。

提出的使用 VIFSS 姿势特征的 TAS 框架是基于新构建的 FS-Jump3D 数据集和一种新的标注方案，该方案考虑了花样滑冰跳跃的程序结构。因此，?? 节介绍了 FS-Jump3D 数据集。然后，?? 节给出了我们特定于跳跃的 TAS 标注方法。最后，2.1 节描述了通过由视角不变对比预训练和花样滑冰特定微调组成的两阶段框架学习到的 VIFSS 姿势特征。

为了捕捉花样滑冰跳跃复杂且动态的动作，我们使用无标记运动捕捉系

统（Theia3D, Theia）构建了一个 3D 姿态数据集，FS-Jump3D。数据收集是通过安装在滑冰场周围的 12 台硬件同步运动捕捉摄像机（Miqus Video, Qualisys）进行的。这种配置使得可以覆盖大范围区域，不仅捕捉跳跃的空中阶段，还能捕捉进入（准备）和着陆动作，这些对于 TAS 任务至关重要。无标记设置允许滑冰者进行高难度跳跃，包括三周跳，且不会有动作限制，同时保持毫米级的精确度。我们记录了四位经验丰富的花样滑冰者（称为滑冰者 A 到 D）跳跃的数据，每人对六种类型的跳跃进行 10 次试验。因此，FS-Jump3D 数据集包含了 253 个跳跃序列，每个序列包括来自 12 个摄像机视角的同步影像和相应的 3D 姿态数据。每个姿态由 83 个关节的 3D 坐标组成（头部：16，躯干：16，手臂：30，腿：34），能够进行详细的生物力学分析。关于 FS-Jump3D 的更多细节，请读者参考之前的论文。

对于 TAS 数据集，我们使用了来自奥运会和世界锦标赛的男子和女子短节目广播录像，这也是在先前有关花样滑冰跳跃分类和动作侦测的研究中采用的方法（Hong et al., 2021）和（Santra et al., 2025）。该数据集总共包含 371 个广播视频。

之前关于花样滑冰中的 TAS 研究的一个限制（Liu et al., 2021; Gan et al., 2024）是，他们的标注策略没有考虑跳跃的程序阶段（即准备和落地动作）。具体来说，在这些 TAS 数据集中，仅仅为起跳前和落地后的几个帧标注了单一动作标签（例如“2Axel”）。这样的标注模糊了起跳和落地的确切时机，缺乏信息来帮助 TAS 模型理解跳跃动作的顺序结构。在花样滑冰中，每种类型的跳跃都有独特的准备动作以产生旋转和升高。这使得准备阶段对跳跃分类至关重要。同样，落地动作为检测落地时刻提供了重要线索。

为了解决这个问题，我们提出了一种新的注释方案，将花样滑冰跳跃分为三个阶段：

1. 条目：起跳前的准备动作。
2. 跳跃：从起跳到落地的腾空阶段。
3. 着陆：着陆后的立即动作。

进入阶段包括从起飞前的三个步骤开始的帧，其中“步骤”被定义为转弯或滑行腿转换等变化。着陆阶段始于冰刀首次接触冰面的帧，并在滑冰者使用外侧后刃持续滑行时继续进行。进入标签对于六种跳跃类型中的每一种都是独特的（例如，“阿克塞尔进入”，“萨尔考进入”），而着陆标签在所有跳跃类型中是通用的。

根据之前的工作（Liu et al., 2021; Gan et al., 2024），我们定义了一个分层的标签结构，用以控制花样滑冰 TAS 中的任务难度，该结构包括集合级别和元素级别的标注。在集合级别，每个跳跃根据其类型（例如，”阿克

塞尔“,”萨霍夫“）被分配到六个标签之一。元素级别提供了更细的粒度，共有 23 个标签，考虑了跳跃类型和旋转次数（例如，”3 阿克塞尔“,”4 萨霍夫“）。元素级别的任务更具挑战性，因为它不仅需要识别跳跃类型，还需要识别旋转次数。

将跳跃阶段标签与分层标签结构相结合，在集合级别总共有 13 个动作标签，在元素级别有 30 个标签。与任何跳跃无关的帧（例如旋转、步序）被标注为“NONE”。

2.1. 视角不变性和花样滑冰特定姿态表示学习

之前在花样滑冰动作识别（Hong et al., 2021）和 TAS（Liu et al., 2021）的工作中已经证明，使用估计的二维人体姿态作为输入比使用 I3D（Carreira and Zisserman, 2017）提取的原始图像特征能获得更好的表现。虽然二维姿态对背景和服装的变化具有鲁棒性，但它们对相机视角和角度的变化较为敏感。为了缓解这一问题，之前的工作（Tanaka et al., 2024）提出了一种方法，利用估计的三维姿态坐标作为花样滑冰中 TAS 的输入。虽然这种方法在集级水平上显示出了一定的性能提升，但三维姿态估计的准确性仍是一个瓶颈，这意味着简单的基于三维坐标的表示不足以完全捕捉花样滑冰动作的复杂性。

为了克服这些挑战，我们提出了一种视角不变且领域特定的姿态编码器，使用三维姿态数据集进行预训练，并使用一个动作分类数据集进行微调，如图 1 所示。该姿态编码器的训练包括两个主要阶段：

1. 通过对比学习进行预训练以获得视角不变的姿态嵌入。
2. 对花样滑冰动作分类进行微调，以便调整姿态编码器来适应花样滑冰特定动作。

在预训练阶段，我们采用了一种对比学习框架，该框架处理从不同视角捕获的相同基础 3D 姿态的二维姿态对。其目的是学习反映多视角相似度的姿态嵌入。在微调阶段，我们针对花样滑冰动作分类任务优化预训练的姿态编码器，以增强编码器捕捉花样滑冰动作的能力。

以下几个小节详细描述了本研究中提出的学习视角不变和领域特定姿态嵌入的方法。

受先前工作的启发，如 Pr-VIPE（Sun et al., 2020; Liu et al., 2022）和 CV-MIM（Zhao et al., 2021），我们采用了一个对比学习框架来学习视图不变的姿势表示。在对比学习中，以 SimCLR（Chen et al., 2020）和 BYOL（Grill et al., 2020）为例，不同的增强方法被应用于同一个样本来创建一个“锚点”和一个共享语义相似性的相应“正例”。编码器被训练来将这些对映射到嵌入空间中的邻近点。虽然基于图像的任务通常使用诸如掩

蔽、旋转、翻转或抖动之类的增强策略，我们将这一策略应用于三维人体姿态数据。具体来说，我们使用虚拟相机在随机视点生成 3D 姿势的多个 2D 投影。这些多视图的 2D 姿势被视为训练的锚-正例对。

对于姿态编码器，我们采用了 Jointformer，这是一种基于 Transformer 的 3D 人体姿态估计器，专为每帧 2D 到 3D 的提升而设计。除了我们的 FS-Jump3D 数据集，我们还整合了几个主要的 3D 人体姿态数据集：Human3.6M、MPI-INF-3DHP 和 AIST++。Human3.6M 包含日常活动的室内录音，而 MPI-INF-3DHP 则包括室内日常动作和室外活动。AIST++ 是一个大型室内舞蹈动作数据集。花样滑冰不仅涉及跳跃，还包括各种各样的动作。为了提高姿态编码器的通用性并确保其适应花样滑冰，我们采用跨数据集的训练策略，使编码器接触各种人体姿态。

对于 3D 姿态数据集的预处理步骤，我们首先通过对齐数据集中关节关键点的数量、空间位置和顺序，来标准化表示姿态的关键点定义。接下来，我们应用对齐和归一化程序，以定义每个 3D 姿态的一致虚拟摄像机方向。最后，在加载训练数据集时，我们进行随机增强以构建用于对比学习的 2D 姿态对。

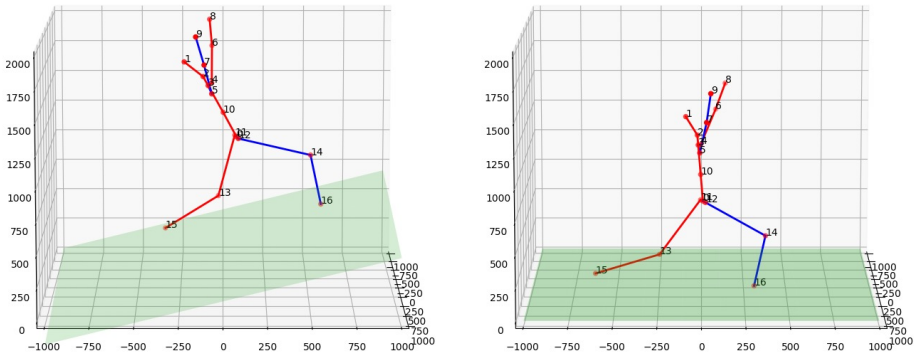
对于 3D 姿势对齐，我们首先通过 RANSAC (Fischler and Bolles, 1981) 如图 2 所示，估计地面平面，将 3D 姿势的重力方向对齐到全局 z 轴。我们在每个姿势中检测最低的 z 坐标，并假设这些最低点的 50 % 对应到实际接触点，而剩下的 50 % 被视为异常值。然后根据该假设使用 RANSAC 估计一个地面平面。将 3D 姿势旋转，使其估计的地面平面对齐全局 xy 平面。随后，我们进一步围绕 z 轴旋转每个姿势，以使所有姿势面向同一方向，如图 3 所示：左髋关节对齐到正 x 轴，右髋关节对齐到负 x 轴。

归一化的过程包括将每个姿势的中心定位于中间髋关节，并重新调整姿势的比例，使从中间髋关节到胸部以及从胸部到颈部的距离之和等于 0.4。这样，中心化的三维关节坐标就被缩放到几乎在绝对值为 1 的范围内。

数据增强包括使用虚拟相机进行随机二维投影、水平翻转、抖动和遮罩。虚拟相机相对于以原点为中心的三维姿态定位，方位角均匀地从 $\pm 180^\circ$ 中采样，仰角从 $\pm 30^\circ$ 中采样，距离从 $[5, 10]$ 范围中采样。

通过透视投影从这个虚拟视点生成一个二维姿态。水平翻转是在投影之前通过将三维关节的 x 坐标乘以 -1 来实现的。抖动通过加入方差为 0.01 的零均值高斯噪声到每个二维关节来实现。掩蔽随机设置每帧二维关节坐标的 1 % 为 $(0, 0)$ 。这些增强提高了数据的多样性，促进了鲁棒潜在表示的学习，增强了对实际二维姿态估计中的噪声的抵抗力。

目标函数. 受到 CV-MIM (Zhao et al., 2021) 的启发，该方法学习视图不变姿态表示，我们提出了一种简化的目标函数，不仅获取这种不变特征，还明确学习识别花样滑冰旋转所需的视图相关特征。



(a) Before ground plane alignment (b) After ground plane alignment

Figure 2: 使用 RANSAC 进行地面平面对齐。我们在所有 3D 姿态数据集中进行地面平面对齐，使得世界坐标系统的 z 轴与重力方向对齐，使用基于 RANSAC 的平面检测。

设二维姿态输入为 $x \in \mathbb{R}^{2\times N}$,具有 N 个关节,编码器输出为嵌入 $z \in \mathbb{R}^{d}$ 。该嵌入由姿态不变分量 $z_{\text{pose}} \in \mathbb{R}^{d_{\text{pose}}}$ 和视图相关分量 $z_{\text{view}} \in \mathbb{R}^{d_{\text{view}}}$ 组成，因此 $d = d_{\text{pose}} + d_{\text{view}}$ 。

给定一个锚定姿态及其正样对，令 z 和 z' 分别为它们的嵌入。为了加强视角不变的表示学习，我们使用 Barlow Twins 损失 (Zbontar et al., 2021) 定义姿态损失 $\mathcal{L}_{\text{pose}}$ ：

$$\mathcal{L}_{\text{pose}} = \text{BarlowTwins}(z_{\text{pose}}, z'_{\text{pose}})$$

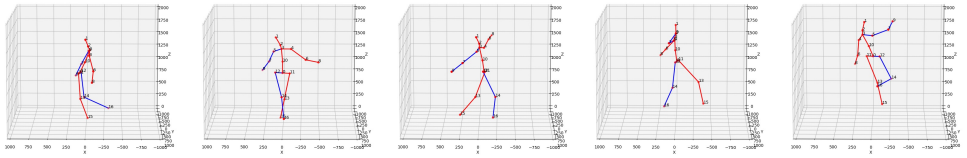
Barlow Twins 损失减少了锚定和正样嵌入之间的冗余，在不依赖负样本的情况下鼓励多样和不相关的特征，从而提高计算效率。

为了学习视角依赖的特征，我们定义了一个视角损失 $\mathcal{L}_{\text{view}}$ 。令 v_{c} 和 v'_{c} 分别表示从原点（髋部中点）到锚点和正视图的虚拟摄像机的单位向量。该损失使用余弦相似度和均方误差（MSE）定义为：

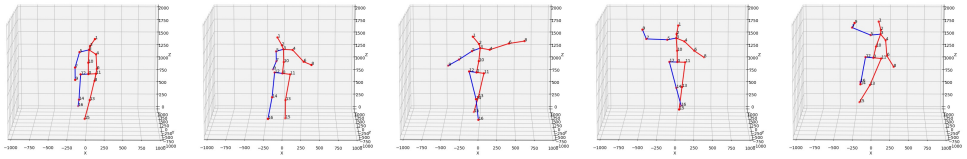
$$\mathcal{L}_{\text{view}} = \text{MSE}(\text{cossim}(z_{\text{view}}, z'_{\text{view}}), \text{cossim}(v_{\text{c}}, v'_{\text{c}}))$$

这个公式将相对于归一化和对齐的 3D 姿势一致定义的虚拟摄像机方向纳入学习目标。它鼓励模型捕捉对视点变化敏感的特征，这些特征在传统姿态表示学习中通常被忽视。

最后，为了防止学习到的嵌入崩溃到简单的解决方案并鼓励嵌入空间的多样性，我们引入了一个正则化项 $\mathcal{L}_{\text{R}}$ 。该项由两部分组成：一个调整输出



(a) Before pose alignment



(b) After pose alignment

Figure 3: 姿态对齐。对于所有的 3D 姿态数据集，我们将每一帧中的每一个姿态进行对齐，使其在整个序列中始终面向同一方向。

特征方差到目标值 σ^2_{target} 的方差损失，以及一个将输出分布正则化为均匀分布的 KL 散度损失：

$$\mathcal{L}_{\text{R}} = \text{VarianceLoss}(z) + \text{VarianceLoss}(z') + \text{KLUniformLoss}(z) + \text{KLUniformLoss}(z')$$

这两个损失定义如下：

$$\text{VarianceLoss}(z) = \frac{1}{d}\sum_{i=1}^{d}\left(\sigma_i^2(z) - \sigma^2_{\text{target}}\right)^2$$

$$\text{KLUniformLoss}(z) = \frac{1}{d}\sum_{i=1}^{d}\left(z_i \log z_i + (1 - z_i)\log(1 - z_i)\right)$$

在我们的实验中，我们为方差损失设置 $\sigma^2_{\text{target}} = 1.0$ 。

基于上述内容，姿态编码器的对比学习的整体损失函数定义为：

$$\mathcal{L}_{\text{total}} = w_{\text{pose}} \cdot \mathcal{L}_{\text{pose}} + w_{\text{view}} \cdot \mathcal{L}_{\text{view}} + w_{\text{R}} \cdot \mathcal{L}_{\text{R}}$$

，其中 w_{pose} 、w_{view} 和 w_{R} 是每个损失项的权重。在我们的实现中，我们设定 $w_{\text{pose}} = 1.0$ 、$w_{\text{view}} = 10.0$ 和 $w_{\text{R}} = 1.0$ 。

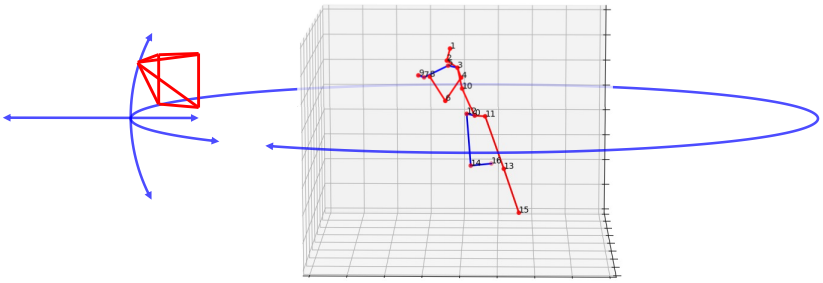


Figure 4: 随机虚拟摄像机排列。虚拟摄像机相对于对齐和归一化的 3D 姿势的根关节（即原点）定位。摄像机位置从一个球面坐标范围内随机采样：方位角来自 $\pm 180°$，俯仰角来自 $\pm 30°$，距离从范围 $[5, 10]$ 到原点。始终假设虚拟摄像机面向根关节。

在使用对比学习对姿态编码器进行预训练以获得视角不变的姿态嵌入之后，我们对编码器进行微调，使其专门用于识别花样滑冰动作。这是通过在一个特定于花样滑冰的动作分类任务上进行训练来实现的。

对于动作分类任务，我们使用了 SkatingVerse 数据集，该数据集包含注释过的花样滑冰跳跃和旋转的视频剪辑，并分为 28 个动作类别。其中，23 个类别对应于六种跳跃类型（Axel、Salchow、Toe Loop、Loop、Flip 和 Lutz）和四个旋转等级（单、双、三、四）。剩下的五个类别包括四种旋转类型（Camel Spin、Sit Spin、Upright Spin 和其他旋转），以及一个用于不包含跳跃或旋转段的“NONE”类别。

该数据集提供了 1,687 个官方花样滑冰视频的标注，生成了 19,993 个训练片段和 8,586 个测试片段，每个片段都根据其特定动作进行了剪辑和标记。

为了微调用于花样滑冰动作分类的姿态编码器，我们采用了一种受之前研究启发的时间建模架构 (Hong et al., 2021)，该架构将基于 BiGRU 的序列模型连接到预训练的姿态编码器。整体架构如图 5 所示。

给定从每个视频帧估计的 2D 姿态序列，姿态编码器输出一个姿态嵌入的序列。这些嵌入被输入到一个两层的 BiGRU 中，其输出通过时间最大池化进行聚合。池化后的特征然后通过一个带有 dropout 的全连接层，接着

是一个 ReLU 激活，另一个 dropout 层，最后是一个全连接层，以预测 28 个动作类别的概率。

姿态编码器使用通过视图不变对比学习预训练的权重初始化。整个模型通过在动作分类任务上的交叉熵损失进行微调。此预训练使模型能够通过微调有效地学习适合花样滑冰动作理解的表示。

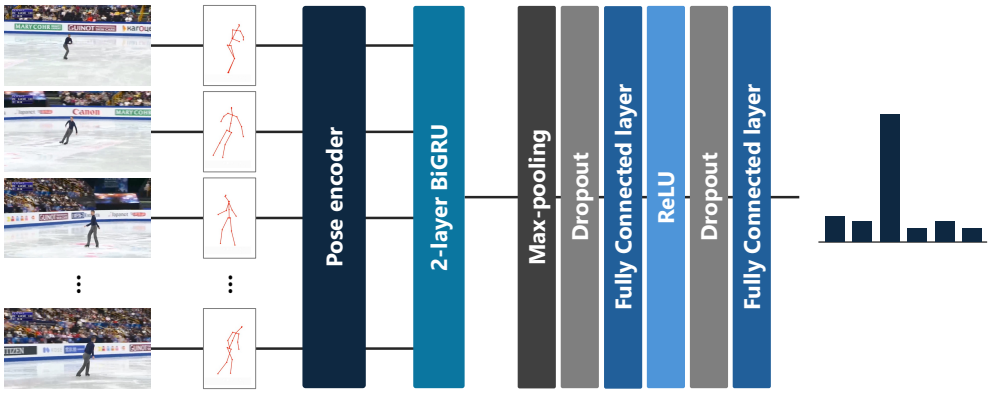


Figure 5: 使用两层 BiGRU 的动作分类模型概述。该模型通过在花样滑冰动作分类任务上进行训练来微调姿态编码器。

3. 实验

3.1. 数据集

3.1.1. FS-Jump3D 数据集

FS-Jump3D 的最显著特点是其记录环境。现有的数据集主要是在受控的室内条件下收集的，比如实验室环境，并且只有少数是在户外录制，但 FS-Jump3D 是在溜冰场中录制的，这是一种高度专业化的环境。与标准的地面条件不同，溜冰场使滑冰者能够执行利用惯性和离心力的动态和复杂动作。

另一个关键特征是使用无标记运动捕捉来记录高难度动作，包括三周跳。花样滑冰跳跃对环境条件极为敏感，并且具有跌倒的风险；因此，为了避免限制滑冰者的自然动作，我们采用了一种无标记系统。这种方法使我们能够捕捉到真实的跳跃数据，包括错误和跌倒，而不降低跳跃的难度水平。

此外，FS-Jump3D 是使用十二台专用的运动捕捉摄像机记录的，这些摄像机具备硬件同步功能。鉴于花样滑冰中跳跃动作极其迅速，硬件同步对于减少摄像机间的时间错位和实现精确的空间重建至关重要。使用十二台摄像机还提供了足够的多视角覆盖，这有可能缓解花样滑冰二维姿态估计中常遇到的严重遮挡问题。

我们构建了一个花样滑冰跳跃的 TAS 数据集，并使用我们提出的程序感知方法进行了注释。该数据集包含 371 段来自国际比赛短节目项目的广播视频，包括冬季奥运会（2010、2014、2018）和世界花样滑冰锦标赛（2017–2019），涵盖了男子和女子比赛项目。

平均每个视频包含 4,265 帧，其中大约 382 帧（占 8.96 % ）带有动作标签，对应从跳跃进入到着陆的时间间隔。这种稀疏性突出表明了花样滑冰转播中 TAS 的固有挑战，其中带注释的片段仅占整个视频的一小部分。

图 6 展示了数据集中跳跃类型的分布。在集级别 (a) 中，勾手跳是最常见的，这可能是因为它们既可以作为单独的跳跃出现，也可以作为组合中的第二跳出现。阿克塞尔跳是第二常见的，反映了它们在男女短节目中作为规定元素的地位。

在动作元素层面，最常见的跳跃是三周 Toe Loop（脚尖点冰跳），随后是三周 Lutz（卢茨跳），两周 Axel（阿克塞尔跳），和三周 Axel。Axel 跳的高频率与其在比赛中的强制性包含一致，绝大多数女性尝试两周 Axel，而绝大多数男性尝试三周 Axel。三周 Lutz 也因其较高的基础分值而受欢迎。相比之下，难度更高的四周跳以及降级跳（例如，由三周或四周降为单周或双周跳）出现的频率较低。这些统计数据显示在动作元素层面上的显著类别不平衡，并突出展示了花样滑冰技术动作管理（TAS）的挑战。

为了评估我们的程序感知标注方法的有效性，我们定义了一个名为“集合级粗标注”的基线。在这个基线中，所有的进入和降落标签都被替换为“无”标签，只有跳跃动作标签（例如，“Salchow”，“Axel”）逐帧标注。此设置模仿了 MCFS (Liu et al., 2021) 的标注风格，允许进行公平比较，同时展示了引入详细跳跃程序阶段的影响。

3.2. 评估协议

3.2.1. 姿态特征在 TAS 输入中的比较

为了验证所提出方法的有效性，我们进行了一系列的评估实验。首先，我们通过比较各种类型的输入特征来评估所提出的 VIFSS 姿态特征对 TAS 性能的影响。此外，我们通过在不同的实验设置下评估 TAS 性能来研究 FS-Jump3D 数据集、所提出的标注方法以及我们的两阶段姿态表示学习框架的个别贡献。

为了识别花样滑冰跳跃中 TAS 的有效输入表示，我们比较以下四种姿态特征，包括我们提出的方法：

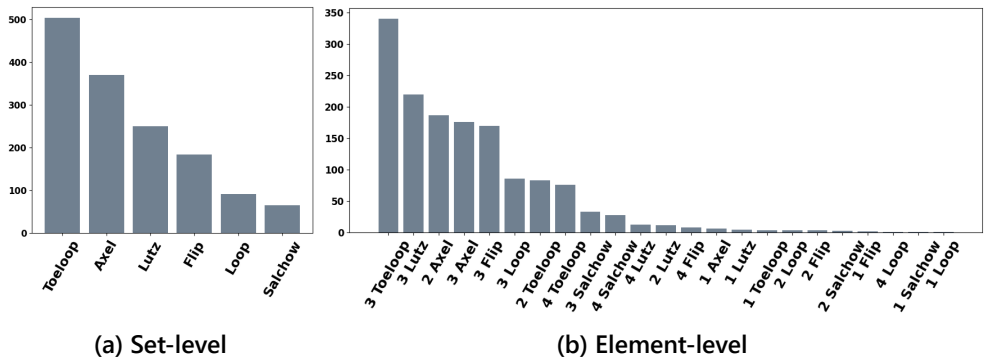


Figure 6: 我们 TAS 数据集中花样滑冰跳跃的标注统计。无论是集合级标签还是元素级标签都存在显著的类别不平衡。

1. 二维姿态特征（基线）
2. 三维姿势特征（Tanaka et al., 2024）
3. VIFSS 特征（提出）
4. scratch-FSS 特性（消融）

作为基线，我们采用 2D 姿态特征，这在花样滑冰动作识别中常被使用（Liu et al., 2021）。这些是通过对每个视频帧进行 2D 姿态估计直接获得的 2D 坐标。在我们的实现中，我们使用了 DWPose（Yang et al., 2023），该模型在 COCO-Wholebody 数据集上进行了预训练，并由作者公开发布，并对获得的 2D 关节坐标进行归一化处理。

3D 姿态特征是通过将估计的 2D 姿态提升到 3D 并应用视图不变的归一化生成的，使所有姿态面向一致的方向。我们利用 MotionAGFormer（Mehraban et al., 2024）进行 3D 姿态估计，并在用于生成建议的 VIFSS 姿态特征的相同数据集上训练。此外，在姿态对齐过程中使用的 3D 旋转角度被归一化并与对齐后的姿态坐标级联为辅助特征，使得表示能保留有助于动作识别的全局方向线索。有关 3D 姿态特征处理流程的更多细节，请参阅初步会议论文（Tanaka et al., 2024）。

我们将提出的 VIFSS 特征与我们消融研究中使用的 scratch-FSS 特征进行了比较。这里，“VIFSS”表示姿势编码器首先通过对比学习进行预训练，以获得视图不变的表示，然后在花样滑冰特定动作分类任务上进行微调。相比之下，scratch-FSS 特征是从一个仅在动作分类任务上从头训练的编码器中提取的，没有经过视图不变的预训练阶段。此比较使我们能够量

化所提出的预训练策略对学习的姿势表示质量的贡献。

我们评估了各种输入特征表示对于 TAS 的有效性，使用了我们提出的细粒度注释，该注释结合了详细的跳跃步骤。

按照之前工作的实验方案 (Hong et al., 2021)，2018 年冬季奥运会和世界花样滑冰锦标赛的所有视频素材被指定为测试集。其余视频用于训练，其中 20 % 的训练数据保留用于验证。这个划分确保相同比赛年份的素材不会同时出现在训练集和测试集中，从而能够评估模型对未见过的视频条件的泛化能力，例如摄像机位置、观看角度和视频质量的变化。

为了评估所提出的基于姿态的特征的效用，我们采用 FACT (Lu and Elhamifar, 2024) 作为 TAS 模型。FACT 是一种基于 Transformer 的架构，利用交叉注意力机制共同学习帧级和动作级表示。在多个基准 TAS 数据集 (Kuehne et al., 2014; Stein and McKenna, 2013; Fathi et al., 2011; Damen et al., 2022) 上已展示出最先进的表现。

我们遵循 TAS 文献中的标准评估协议 (Ding et al., 2022)，计算逐帧准确率和 F1@ { 10, 25, 50, 75, 90 } 分数。这些指标在所有动作段上计算，排除“入场”，“着陆”和“无”标签。逐帧准确率量化正确标记帧的比例。F1@k 指标认为如果预测段与对应的真实段至少重叠其长度的 k % ，则该预测段是正确的，并在每个重叠阈值处计算精确率和召回率的调和平均值。

4. 结果

本节对我们提出的花样滑冰跳跃 TAS 方法在各种实验设置下进行了全面评估。在第 4.1 节中，我们对比了两种标注粒度下的 TAS 性能：集合级别和元素级别。然后，我们在两种标注级别上评估了我们提出的 VIFSS 姿态特征相对于其他特征类型的有效性。每个级别的详细结果在第 **??** 节和第 4.2 节中进行讨论。第 4.3 节通过比较包含和不包含 FS-Jump3D 数据集的视图不变预训练，研究了该数据集的影响。在第 **??** 节中，我们通过与粗略标注对比，评估了我们精细的、跳跃过程感知标注方案的好处，突出其在提高模型对跳跃动态理解中的作用。最后，第 4.4 节展示了一项消融研究，该研究量化了在不同监督水平下视图不变预训练的贡献，特别强调了其在低数据情境下的优势。

4.1. 集合级别和元素级别 TAS 注释的比较

表 1 报告了花样滑冰跳跃的 TAS 结果，使用不同的输入特征比较了集合级和元素级标注之间的表现。表 1a 和 1b 的比较显示，元素级的分割比集合级更具挑战性。在这两种设置下，F1@90 的分数都明显低于 F1@10 到 F1@75 的分数。在我们的广播视频数据集中，跳跃从起跳到落地的平均持续时间为 16.25 帧。在 F1@90 度量下，预测的段只有在与真实段重叠约

15 帧时才被视为正确，仅允许 1 到 2 帧的误差。相比之下，F1@75 的度量要求约 12 帧的重叠，允许大约四帧的误差（即在跳跃前后各两帧）。在这种 F1@75 度量下，我们提出的使用两阶段姿态表示学习的 VIFSS 特征在集合级和元素级任务中都达到了超过 90 %。

我们的注释方案能够精确评估起飞和着陆的时间。当 F1@90 作为模型精度的上限时，始终如一的高 F1@75 得分展示了所提议表示法的实际有效性。

Feature	Acc	F1@10	F1@25	F1@50	F1@75	F1@90
2D pose (Baseline)	78.55	85.12	84.93	84.17	81.52	35.83
3D pose (Tanaka et al., 2024)	79.89	87.13	86.94	86.56	82.36	33.36
VIFSS	89.91	95.44	95.44	94.68	93.16	51.71
scratch-FSS	86.38	92.48	92.29	91.72	88.87	42.44

(a) 集合级别的 TAS 结果。

Feature	Acc	F1@10	F1@25	F1@50	F1@75	F1@90
2D pose (Baseline)	71.34	78.97	78.97	78.78	75.74	35.39
3D pose (Tanaka et al., 2024)	70.17	77.71	77.33	76.57	71.62	29.52
VIFSS	85.82	92.75	92.75	92.56	90.65	49.62
scratch-FSS	82.72	89.65	89.65	89.65	86.42	41.03

(b) 元素级 TAS 结果。

Table 1: 在不同标注粒度水平下的 TAS 性能比较。

如表 1a 所示，VIFSS 和 scratch-FSS 特征在集合级别的 TAS 上表现优于基线方法。其中，VIFSS 特征实现了最高的整体性能。这些结果表明，所提出的视图不变对比预训练有效地缓解了二维姿态特征的视图依赖性质，从而增强了针对花样滑冰特定动作的下游动作分类性能。尽管三维姿态特征落后于所提出的姿态嵌入，但它们仍超过了基线的二维姿态特征。总体而言，这些发现表明，学习视图不变的姿态表示，特别是所提议的 VIFSS 特征，对于识别集合级别的跳跃类型和时机是有效的。

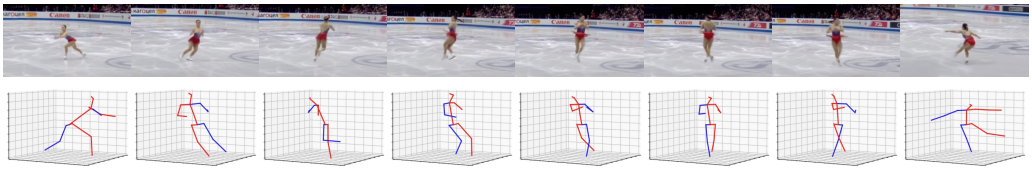
4.2. 元素级 TAS 性能

表 1b 展示了每种输入特征类型在元素级别的 TAS 结果。与集合级别的发现一致，姿态嵌入特征、VIFSS 和 scratch-FSS 特征均表现出色。尤其是，VIFSS 特征在 F1@50 上超过了 92 %。相比之下，3D 姿态特征在元素级别的表现不如基线的 2D 姿态特征。这表明，所提出的用于姿态嵌入的两阶段学习框架同样适用于细粒度的元素级别分割。同时，3D 姿态特征似乎不太适合需要精确识别复杂旋转运动的任务。

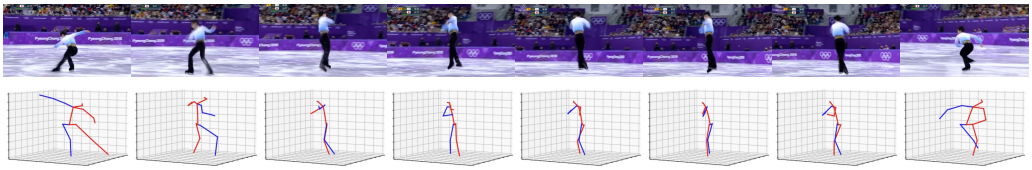
3D 姿态特征在元素级别上的性能下降可能是由于 3D 姿态估计模型本身的局限性。具体来说，本研究中使用的 3D 姿态估计器 MotionAGFormer (Mehraban et al., 2024) 是一个从 2D 姿态序列预测 3D 姿态的时间模型。然而，训练和推理数据的时间特性差异很可能影响了其性能。FS-Jump3D 训练数据集主要由 B 和 C 受试者进行的双跳和 A 受试者进行的三跳组成。相反，用于推理的广播视频主要包含三跳甚至四跳。此外，虽然 FS-Jump3D 数据集从 60fps 下采样到 30fps 进行训练，但广播视频是以 25fps 录制的。这些在旋转次数和训练与推理中的帧间隔方面的差异可能导致模型在识别旋转运动能力上的下降。

图 7a 和 7b 分别展示了在广播视频中成功和失败的跳跃旋转 3D 姿态估计的例子。图 7a 展示的是一个双轴跳，这在 FS-Jump3D 训练集中得到了良好表示（由主体 A 和 C 完成），并展示了旋转运动的准确 3D 姿态估计。相比之下，图 7b 展示的是一个四重跳，这在训练数据中没有包含，导致无法正确捕捉旋转运动。同样，对于一些跳跃动作中手臂处于不寻常位置的情况，例如双臂举过头部完成的三重跳跃，常常观察到失败。这些观察表明，时间 3D 姿态估计模型的性能在很大程度上依赖于训练 3D 姿态数据集的多样性，而模型的局限性成为后续 TAS 任务的瓶颈。

相比之下，姿态嵌入方法对这样的变化表现出稳健性。由于对比预训练阶段旨在学习独立于时间序列的每帧姿态表示，因此它较少受到跳跃旋转类内变异性的影响。此外，微调阶段不依赖于通常难以获得的 3D 姿态标注，这使得对于特定领域的应用更为实用。因此，即使对于要求更高的元素级 TAS 任务，基于嵌入的方法依然始终取得了优异的性能。



(a) 在广播视频中成功实现双轴三维姿态估计。旋转运动被准确地重建为一个三维姿态序列。



(b) 在广播视频中，对四周旋转跳跃的 3D 姿势估计失败。模型未能捕捉到旋转运动，这可能是由于训练数据中缺乏类似的高速旋转。

Figure 7: 广播视频中跳跃旋转的三维姿态估计算例。(a) 展示了一个成功的案例，而 (b) 则显示了一个失败的情况，可能是由于训练数据多样性的限制所致。

4.3. FS-Jump3D 对 TAS 性能的影响

表 2 比较了在训练过程中是否包含 FS-Jump3D 数据集时，使用 3D 姿态特征和 VIFSS 特征的集合级 TAS 表现。结果显示，对于两种特征类型，排除 FS-Jump3D 数据集会导致分割性能的下降。FS-Jump3D 由在独特环境下（冰场）捕获的花样滑冰跳跃动作组成，这与传统的 3D 姿态数据集有很大不同。这些结果表明，FS-Jump3D 在准确捕获花样滑冰中的跳跃动作方面起着关键作用，强调了其对 TAS 任务的重要性。

Feature	Acc	F1@10	F1@25	F1@50	F1@75	F1@90
3D pose (Tanaka et al., 2024)	79.89	87.13	86.94	86.56	82.36	33.36
3D pose (w/o FS-Jump3D)	75.76	83.08	82.89	82.51	77.57	31.18
VIFSS	89.91	95.44	95.44	94.68	93.16	51.71
VIFSS (w/o FS-Jump3D)	87.53	93.16	92.97	92.78	91.25	46.01

Table 2: 在有和没有 FS-Jump3D 数据集条件下的 Set 级别 TAS 结果。

此研究中提出的注释方法旨在将跳跃程序步骤（进入和着陆阶段）的知识融入到 TAS 模型中，从而帮助模型更好地识别跳跃的类型和时机。表 3 显示了验证所提议注释在设定级别 TAS 任务上的有效性的结果。为了比较，通过将所提注释中与进入和着陆阶段相关的所有标签替换为 “None” 标签，创建了一个粗略注释，只保留跳跃标签作为有效的动作类。使用的评估指标是 F1@50，这是 TAS 中最常用的指标。

结果表明，通过提出的标注引入程序步骤可以提高 TAS 性能。除了为模型提供对跳跃过程的更深入理解外，提出的标注还可能通过为更多帧分配有意义的标签来促进性能提升。具体来说，在粗略标注中，只有 1.50 % 的所有帧被分配了动作标签（仅跳跃），而在提出的标注中，8.96 % 的帧被标记为动作标签（进入、跳跃或着陆）。动作标签比例的这种差异表明，在更长的时间范围内提供有意义的帧级标注是提高花样滑冰跳跃 TAS 性能的一个重要因素。

4.4. 视角不变预训练的有效性

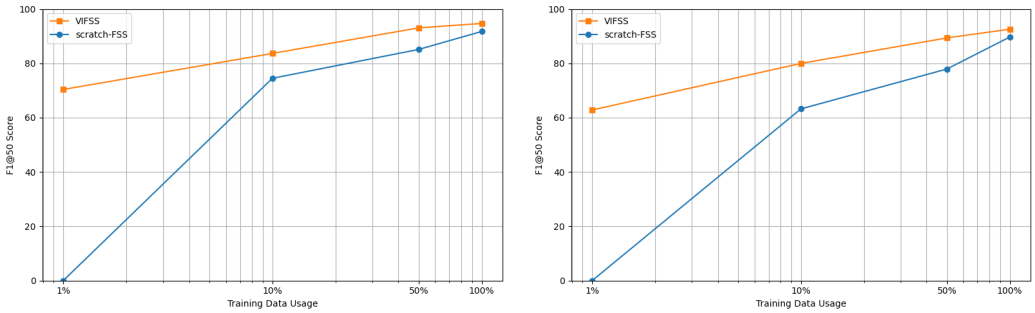
图 8 显示了一项消融研究的结果，该研究评估了我们用于 VIFSS 特征的两阶段学习框架中预训练的影响。我们比较了在集合级别和元素级别 TAS 任务中有无对比预训练的模型。在第二阶段微调时，我们将动作分类数据量分别设为 100 % 、50 % 、10 % 和 1 % ，并使用 F1@50 评估性能。

Feature	Proposed	Coarse
2D pose (baseline)	84.17	76.42
3D pose (Tanaka et al., 2024)	86.56	72.78
VIFSS	94.68	93.93

Table 3: 基于 F1@50 的集合级 TAS 中所提出的标注与粗标注之间的比较。

结果表明，结合视图不变对比预训练可以在两个层面上稳定地提高 TAS 表现。值得注意的是，随着微调数据量的减少，预训练的好处变得更加明显。在仅使用 1 % 数据的极端情况下，没有预训练的模型无法学习有意义的姿态嵌入，导致 TAS 表现几乎为零。相比之下，经过预训练的模型在集合层面实现了超过 70 % F1@50 的表现，在元素层面则超过了 60 %。

这些发现强调了视图不变预训练在低数据环境中特别有效。即使在注释的微调数据极为有限的情况下，它也能够高效地学习花样滑冰的领域自适应姿态嵌入。



(a) 在集合级别上，使用和不使用视角不变的预训练的 F1@50。

(b) 元素级别的 F1@50，包含和不包含视图不变的预训练。

Figure 8: 视图不变的人物姿态表示预训练对不同比例微调数据下 TAS 性能的影响。在两个标注级别中，预训练持续提高 TAS 性能，尤其是在低数据条件下观察到显著的提升。

5. 结论

在这项工作中，我们提出了 VIFSS 姿态特征，并展示了其在花样滑冰跳跃应用中的有效性。我们还引入了一种新的标注策略，该策略结合了跳跃程序，并构建了 FS-Jump3D，这是首个公开可用的花样滑冰 3D 姿态数据集，包括三周跳。

大量实验表明，通过对比学习预训练的姿态编码器导出的姿态嵌入，并在花样滑冰动作分类任务上进行微调，可以显著提升 TAS 在集合级和元素级的性能。这些结果表明，与直接回归 3D 关节点坐标相比，使用特定领域监督学习潜在姿态表示提供了更大的灵活性和准确性。此外，与粗略标注的比较证实了我们的程序感知标注在为 TAS 提供更丰富和更具信息量的监督信号方面的优势。消融研究还证实了我们的两阶段 VIFSS 学习方法强有力地支持领域适应，特别是在注释有限的场景中。

基于在花样滑冰中已证明的视角不变对比预训练和特定任务微调的效果，未来的工作将拓展我们的方法至其他花样滑冰动作如旋转和踏步序列的技术动作考核。我们还计划探索其在其他体育领域技术动作考核任务中的普遍适用性，旨在进一步验证我们所提方法的稳健性和多功能性。

6.

作者贡献 TS 支持 FS-Jump3D 数据集的创建和处理，并参与了稿件的修订。KF 监督了本研究并对稿件修订做出了贡献。所有作者都阅读并批准了最终稿件。

7.

致谢 本工作得到了日本学术振兴会（JSPS）项目编号 21H05300 和 23H03282，以及 JST PRESTO 项目编号 JPMJPR20CA 的资助。资助方在研究设计、数据收集、分析、数据解释或本文手稿的撰写中没有发挥作用。

8.

数据和代码可用性 我们的 VIFSS 姿态表示学习方法的 TAS 数据集标注和源代码将在发表时在 `https://github.com/ryota-skating/VIFSS` 公开提供。

9.

竞争利益 所有作者声明没有利益冲突。

References

Aharon, N., Orfaig, R., Bobrovsky, B.Z., 2022. Bot-sort: Robust associations multi-pedestrian tracking. ArXiv abs/2206.14651. URL: `https://api.semanticscholar.org/CorpusID:250113384`.

Askari, F., Jiang, R., Li, Z., Niu, J., Shi, Y., Clark, J.J., 2023. Self-supervised video interaction classification using image representation of skeleton data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 5229–5238.

Askari, F., Yared, C., Ramaprasad, R., Garg, D., Hu, A., Clark, J.J., 2024. Video interaction recognition using an attention augmented relational network and skeleton data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3225–3234.

Cabado, B., Cioppa, A., Giancola, S., Villa, A., Guijarro-Berdiñas, B., Padrón, E.J., Ghanem, B., Van Droogenbroeck, M., 2024. Beyond the premier: Assessing action spotting transfer capability across diverse domains, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3386–3398.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308.

Chappa, N.V.R., Nguyen, P., Nelson, A.H., Seo, H.S., Li, X., Dobbs, P.D., Luu, K., 2023. Spartan: Self-supervised spatiotemporal transformers approach to group activity recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 5158–5168.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, JMLR.org. pp. 1597–1607.

Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al., 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision , 1–23.

Deyzel, M., Theart, R.P., 2023. One-shot skeleton-based action recognition on strength and conditioning exercises, in: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 5169–5178.

Ding, G., Sener, F., Yao, A., 2022. Temporal action segmentation: An analysis of modern techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 1011–1030. URL: https://api.semanticscholar.org/CorpusID:252992530.

Fathi, A., Ren, X., Rehg, J.M., 2011. Learning to recognize objects in egocentric activities, in: CVPR 2011, pp. 3281–3288. doi:10.1109/CVPR.2011.5995444.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 381–395. URL: https://doi.org/10.1145/358669.358692, doi:10.1145/358669.358692.

Gan, Z., Jin, L., Cheng, Y., Cheng, Y., Teng, Y., Li, Z., Li, Y., Yang, W., Zhu, Z., Xing, J., Zhao, J., 2024. Skatingverse: A large-scale benchmark for comprehensive evaluation on human action understanding. IET Computer Vision n/a. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cvi2.12287, doi:https://doi.org/10.1049/cvi2.12287, arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cvi2.12287.

Giancola, S., Amine, M., Dghaily, T., Ghanem, B., 2018. Soccernet: A scalable dataset for action spotting in soccer videos, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1792–179210. doi:10.1109/CVPRW.2018.00223.

Giancola, S., Cioppa, A., Georgieva, J., Billingham, J., Serner, A., Peek, K., Ghanem, B., Van Droogenbroeck, M., 2023. Towards active learning for action spotting in association football videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 5098–5108.

Gossard, T., Krismer, J., Ziegler, A., Tebbe, J., Zell, A., 2024. Table tennis ball spin estimation with an event camera, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3347–3356.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271–21284.

Held, J., Cioppa, A., Giancola, S., Hamdi, A., Ghanem, B., Van Droogenbroeck, M., 2023. Vars: Video assistant referee system for automated soccer decision making from multiple views, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 5086–5097.

HIROSAWA, S., 2020. Action quality assessment of figure skating jumps using machine learning. ICE SPORTS RESEARCH 1.

Honda, Y., Kawakami, R., Yoshihashi, R., Kato, K., Naemura, T., 2022. Pass receiver prediction in soccer using video and players' trajectories, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3503–3512.

Hong, J., Fisher, M., Gharbi, M., Fatahalian, K., 2021. Video pose distillation for few-shot, fine-grained sports action recognition, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9234–9243. doi:10.1109/ICCV48922.2021.00912.

Ibh, M., Graßhof, S., Hansen, D.W., 2024. A stroke of genius: Predicting the next move in badminton, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3376–3385.

Ibh, M., Grasshof, S., Witzner, D., Madeleine, P., 2023. Tempose: A new skeleton-based transformer model designed for fine-grained motion recognition in badminton, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 5199–5208.

Ingwersen, C.K., Mikkelstrup, C.M., Jensen, J.N., Hannemose, M.R., Dahl, A.B., 2023. Sportspose-a dynamic 3d sports pose dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5219–5228.

Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence 36, 1325–1339.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics YOLO. https://github.com/ultralytics/ultralytics. Version 8.0.0. Licensed under AGPL-3.0.

Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y., 2015. Panoptic studio: A massively multiview system for social motion capture, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3334–3342.

Kaneko, T., Kawakami, R., Naemura, T., Inoue, N., 2024. Augmenting pass prediction via imitation learning in soccer simulations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3194–3203.

Kuehne, H., Arslan, A., Serre, T., 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 780–787.

Li, R., Yang, S., Ross, D.A., Kanazawa, A., 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) , 13381–13392URL: https://api.semanticscholar.org/CorpusID:236882798.

Liu, P., Wang, J.H., 2022. Monotrack: Shuttle trajectory reconstruction from monocular badminton video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3513–3522.

Liu, S., Liu, X., Huang, G., Qiao, H., Hu, L., Jiang, D., Zhang, A., Liu, Y., Guo, G., 2020. Fsd-10: A fine-grained classification dataset for figure skating. Neurocomputing 413, 360–367. URL: https://www.sciencedirect.com/science/article/pii/S0925231220310973, doi:https://doi.org/10.1016/j.neucom.2020.06.108.

Liu, S., Zhang, A., Li, Y., Zhou, J., Xu, L., Dong, Z., Zhang, R., 2021. Temporal segmentation of fine-gained semantic action: A motion-centered figure skating dataset. Proceedings of the AAAI Conference on Artificial Intelligence 35, 2163–2171. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16314, doi:10.1609/aaai.v35i3.16314.

Liu, T., Sun, J.J., Zhao, L., Zhao, J., Yuan, L., Wang, Y., Chen, L.C., Schroff, F., Adam, H., 2022. View-invariant, occlusion-robust probabilistic embedding for human pose. International Journal of Computer Vision 130, 111–135. doi:10.1007/s11263-021-01529-w.

Lu, Z., Elhamifar, E., 2024. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18175–18185.

Lutz, S., Blythman, R., Ghosal, K., Moynihan, M., Simms, C., Smolic, A., 2022. Jointformer: Single-Frame Lifting Transformer with Error Prediction and Refinement for 3D Human Pose Estimation , in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 1156–1163. URL: https://doi.ieeecomputersociety.org/10.1109/ICPR56361.2022.9956366, doi:10.1109/ICPR56361.2022.9956366.

Martinez, J., Hossain, R., Romero, J., Little, J.J., 2017. A simple yet effective baseline for 3d human pose estimation, in: Proceedings of the IEEE international conference on computer vision, pp. 2640–2649.

Mehraban, S., Adeli, V., Taati, B., 2024. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 6920–6930.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C., 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision, in: 2017 International Conference on 3D Vision (3DV), pp. 506–516. doi:10.1109/3DV.2017.00064.

Nakabayashi, T., Higa, K., Yamaguchi, M., Fujiwara, R., Saito, H., 2024. Event-based ball spin estimation in sports, in: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3367–3375.

Nibali, A., Millward, J., He, Z., Morgan, S., 2021. ASPset: An outdoor sports pose video dataset with 3D keypoint annotations. Image and Vision Computing , 104196URL: https://www.sciencedirect.com/science/article/pii/S0262885621001013, doi:https://doi.org/10.1016/j.imavis.2021.104196.

Nonaka, N., Fujihira, R., Koshiba, T., Maeda, A., Seita, J., 2024. Rugby scene classification enhanced by vision language model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3256–3266.

Nonaka, N., Fujihira, R., Nishio, M., Murakami, H., Tajima, T., Yamada, M., Maeda, A., Seita, J., 2022. End-to-end high-risk tackle detection system for rugby, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3550–3559.

Okamoto, L., Parmar, P., 2024. Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3204–3213.

Parmar, P., Morris, B., 2019a. Action quality assessment across multiple actions, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1468–1476. doi:10.1109/WACV.2019.00161.

Parmar, P., Morris, B.T., 2019b. What and how well you performed? a multitask learning approach to action quality assessment, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 304–313. doi:10.1109/CVPR.2019.00039.

Piergiovanni, A., Ryoo, M.S., 2018. Fine-grained activity recognition in baseball videos, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1821–18218. doi:10.1109/CVPRW.2018.00226.

Pirsiavash, H., Vondrick, C., Torralba, A., 2014. Assessing the quality of actions, in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Com-

puter Vision – ECCV 2014, Springer International Publishing, Cham. pp. 556–571.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Sanford, R., Gorji, S., Hafemann, L.G., Pourbabaee, B., Javan, M., 2020. Group activity detection from trajectory and video data in soccer, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3932–3940. doi:10.1109/CVPRW50498.2020.00457.

Santra, S., Chudasama, V., Wasnik, P., Balasubramanian, V.N., 2025. Precise event spotting in sports videos: Solving long-range dependency and class imbalance, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 3163–3172.

Shim, M., Kim, Y.H., Kim, K., Kim, S.J., 2018. Teaching machines to understand baseball games: Large-scale baseball video database for multiple video understanding tasks, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham. pp. 420–437.

Sigal, L., Balan, A.O., Black, M.J., 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision 87, 4–27.

Stein, S., McKenna, S.J., 2013. Combining embedded accelerometers with computer vision for recognizing food preparation activities, in: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, pp. 729–738.

Sun, J.J., Zhao, J., Chen, L.C., Schroff, F., Adam, H., Liu, T., 2020. View-invariant probabilistic embedding for human pose, in: European Conference on Computer Vision, Springer. pp. 53–70.

Suzuki, T., Takeda, K., Fujii, K., 2024. Automatic detection of faults in simulated race walking from a fixed smartphone camera. International Journal of Computer Science in Sport 23, 22–36. URL: https://doi.org/10.2478/ijcss-2024-0002, doi:doi:10.2478/ijcss-2024-0002.

Suzuki, T., Tanaka, R., Yeung, C., Fujii, K., 2025. Athleticspose: Authentic sports motion dataset on athletic field and evaluation of monocular 3d pose estimation ability. URL: https://arxiv.org/abs/2507.12905, arXiv:2507.12905.

Tanaka, R., Suzuki, T., Fujii, K., 2024. 3d pose-based temporal action segmentation for figure skating: A fine-grained and jump procedure-aware annotation approach, in: Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports, Association for Computing Machinery, New York, NY, USA. p. 17–26. URL: https://doi.org/10.1145/3689061.3689077, doi:10.1145/3689061.3689077.

Tanaka, R., Suzuki, T., Takeda, K., Fujii, K., 2023a. Automatic edge error judgment in figure skating using 3d pose estimation from a monocular camera and imus, in: Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports, pp. 41–48.

Tanaka, R., Suzuki, T., Takeda, K., Fujii, K., 2023b. Automatic edge error judgment in figure skating using 3d pose estimation from inertial sensors, in: 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), IEEE. pp. 1099–1100.

Van Zandycke, G., De Vleeschouwer, C., 2022. 3d ball localization from a single calibrated image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3472–3480.

Vandeghen, R., Cioppa, A., Van Droogenbroeck, M., 2022. Semi-supervised training to improve player and ball detection in soccer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3481–3490.

Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G., 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera, in: Proceedings of the European conference on computer vision (ECCV), pp. 601–617.

Xarles, A., Escalera, S., Moeslund, T.B., Clapés, A., 2024. T-deed: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos, in: Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3410–3419.

Yang, Z., Zeng, A., Yuan, C., Li, Y., 2023. Effective whole-body pose estimation with two-stages distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4210–4220.

Yeung, C., Ide, K., Fujii, K., 2024. Autosoccerpose: Automated 3d posture analysis of soccer shot movements, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3214–3224.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S., 2021. Barlow twins: Self-supervised learning via redundancy reduction, in: International conference on machine learning, PMLR. pp. 12310–12320.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X., 2022. Bytetrack: Multi-object tracking by associating every detection box, in: Computer Vision –ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII, Springer-Verlag, Berlin, Heidelberg. p. 1–21. doi:10.1007/978-3-031-20047-2_1.

Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N., 2019. Semantic graph convolutional networks for 3d human pose regression, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3420–3430. doi:10.1109/CVPR.2019.00354.

Zhao, L., Wang, Y., Zhao, J., Yuan, L., Sun, J.J., Schroff, F., Adam, H., Peng, X., Metaxas, D., Liu, T., 2021. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12788–12797. doi:10.1109/CVPR46437.2021.01260.

Zhu, K., Wong, A., McPhee, J., 2022. Fencenet: Fine-grained footwork recognition in fencing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3589–3598.

Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y., 2023. Motionbert: A unified perspective on learning human motion representations, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 15085–15099.