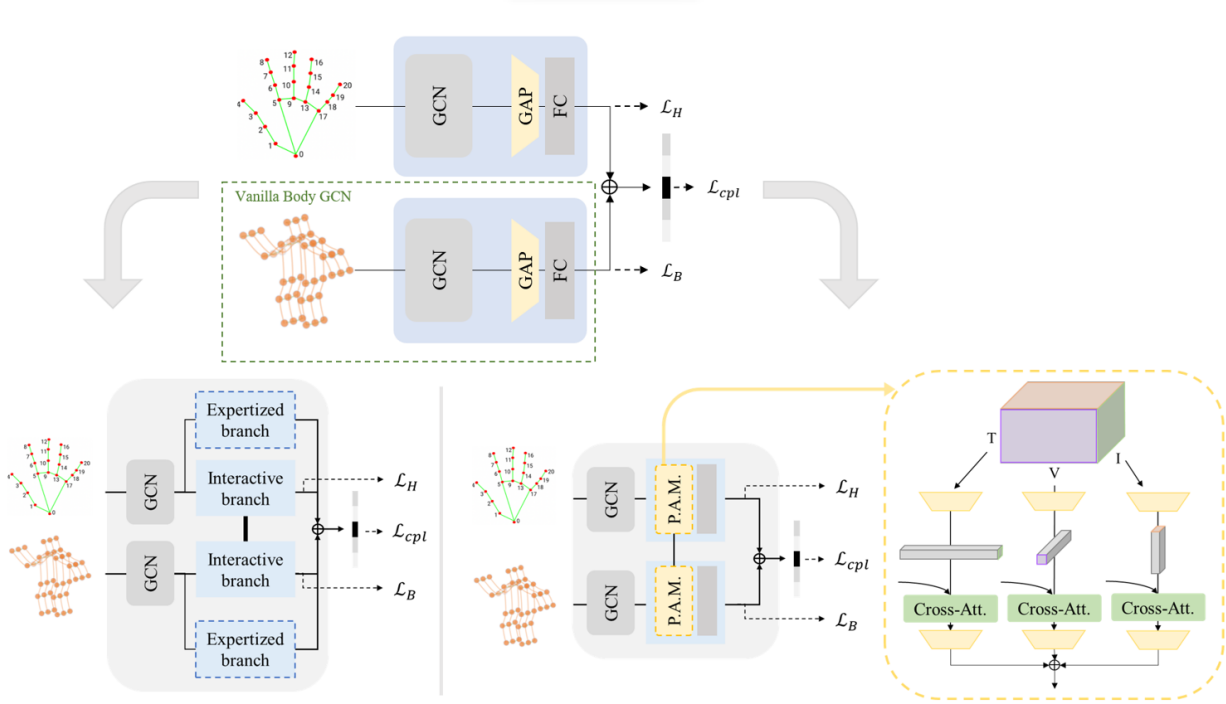
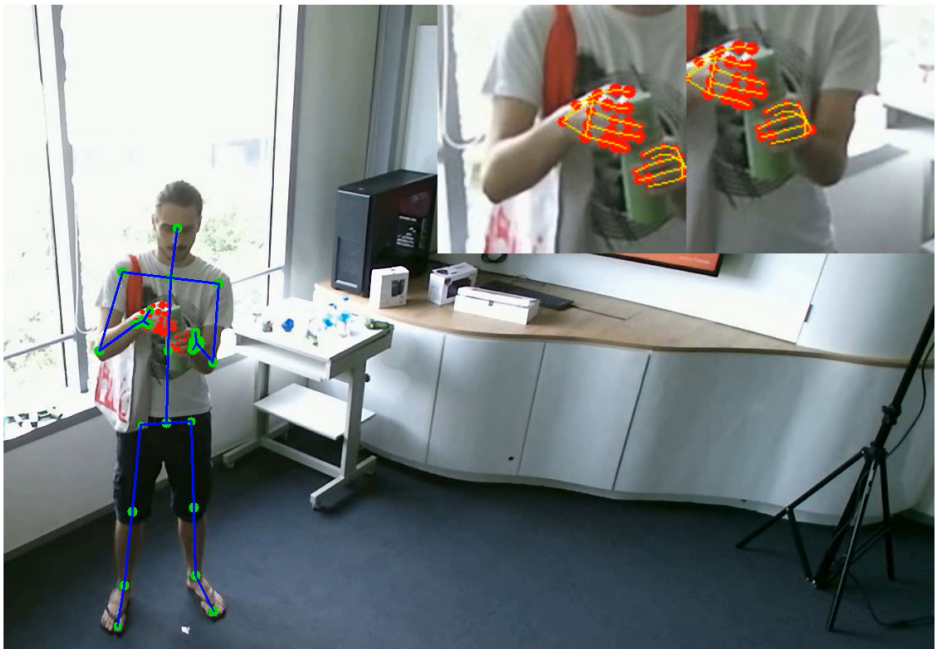
arXiv:2601.00369v1 [cs.CV] 1 Jan 2026

图文摘要

BHaRNet：用于细粒度骨架动作识别的可靠性感知的身体-手部模态专家网络

Seungyeon Cho、Tae-Kyun Kim





亮点

BHaRNet：面向细粒度骨架动作识别的可靠性感知身体-手部模态专家网络

Seungyeon Cho，Tae-Kyun Kim

- 基于概率的双流人体–手部框架，结合免校准骨架学习与可靠性感知融合。
- 基于 Noisy-OR 的融合损失，用于建模身体与手部可靠性的非对称性，并在关键点存在噪声时稳定双流学习。
- 统一的模态内至跨模态集成，将关节、骨骼和运动线索扩展至 RGB，以实现高效的骨架–RGB 动作识别。

BHaRNet：用于细粒度骨架动作识别的可靠性感知身体-手部模态专家网络⋆，⋆⋆

Seungyeon Cho[a],*,1, Tae-KyunKim[a],**,2

[a]*KAIST* 计算学院，大田，*34141*，大韩民国

文章信息

关键词：
骨架基础动作识别 多模态动作识别 细粒度动作识别 身体 - 手部双流架构 可靠性感知融合 Noisy-OR建模

摘 要

基于骨架的人体动作识别（HAR）借助基于图的架构取得了显著进展。然而，现有方法大多仍以身体为中心，关注大幅度运动，却忽视了对细粒度识别至关重要的细微手部关节运动。本文提出一种概率双流框架，融合可靠性建模与多模态整合，将不确定性条件下的专家化学习推广至骨架模态内和跨模态领域。该框架包含三个关键组件：（1）免校准预处理流程，移除规范空间变换，直接从原生坐标中学习；（2）概率 Noisy-OR 融合，无需显式置信度监督，即可稳定可靠性感知的双流学习；（3）模态内至跨模态集成，将四种骨架模态（关节、骨骼、关节运动和骨骼运动）与 RGB表征相结合，在统一的跨模态表述中衔接结构与视觉运动线索。在多个基准数据集（NTU RGB+D 60/120、PKU-MMD、N-UCLA）以及一个新定义的以手部为中心的基准数据集上的全面评估表明，该方法性能持续提升，并能在噪声和异质条件下保持稳健。

1. 引言

人体动作识别（HAR）是计算机视觉领域一项长期存在的挑战，应用于人机交互、视频理解和行为分析 [32, 34]。在 RGB、深度和骨架等多种感知模态中，基于骨架的表示已成为一种紧凑且易于解释的形式，能够实现高效的运动建模，并在不同对象和环境中具备出色的泛化能力。

时空图卷积网络（ST-GCN） [35] 及其后续方法 [4, 6, 14, 28,31, 38] 通过联合建模身体关节之间的时空依赖关系，显著推动了基于骨架的人体动作识别的发展。尽管取得了这些进展，大多数方法仍以身体为中心，强调大尺度的身体运动，却忽略了对细粒度识别至关重要的精细手部关节运动。尽管 SkeleT 等统一图模型 [36] 在整体拓扑结构中整合了身体关节和手部关节，但占主导地位的身体动态往往会掩盖细微的手部线索，这是由固有的尺度和特征不平衡造成的，从而限制了专门化程度和鲁棒性。

在实际应用中，手部骨架由小尺度关节和细粒度关节活动构成，极易受到遮挡和估计噪声的影响，从而导致身体与手部模态之间出现根本性的可靠性不对称：身体关节较为稳定，而手部关节则可能存在噪声或缺失。



图1：两种手部中心动作的可视化——“打哈欠”（左）和“安静”（右），展示了从 NTU RGB+D 中裁剪的帧及身体骨架。两种表示具有几乎相同的身体姿态，表明不同手势之间存在很强的全局姿态相似性。这凸显了仅凭身体骨架区分细粒度动作的难度，也说明了采用可靠性感知手部建模的必要性。

因此，鲁棒识别需要一种能够显式建模这种不对称性并自适应调节模态交互的方案。为此，我们引入三个组件，并将其统一到一个具备可靠性感知能力的身体–手部双流框架中，将其命名为**BHaRNet**（**B**ody–**H**and **a**ction **R**ecognition **Net**work，身体–手部动作识别网络）。

免校准学习。早期的手部姿态框架通常采用规范空间对齐来归一化手部坐标，并假设关节位置 [13, 16, 27] 准确无误。这类变换在存在遮挡或运动模糊时往往会放大噪声，使局部误差沿整个骨架传播。为此，我们采用免校准设计，取消规范空间变换，使模型能够直接在原生坐标空间中运行。

*通讯作者地址：大韩民国大田广域市儒城区大学路291号 KAIST E3-2楼2312室，邮编34141。电话号码： +82) 10-4898-1126。

**主要通讯作者vinny@kaist.ac.kr (S. Cho)；kimtaekyun@kaist.ac.kr (T. Kim)；https://github.com/VinnyCSY/BHaRNet(S. Cho)；https://sites.google.com/view/tkkim/home (T. Kim) ORCID：

这种设计有助于学习统一的骨架表示，同时减轻误差传播。

概率 Noisy-OR 融合。确定性融合方案 s 通常假设各模态具有相同的可靠性，因此当手部线索不确定或缺失时，聚合结果会不稳定。我们引入一种 Noisy-OR 融合，用于建模模态间的可靠性差异。它充当一道门控，在放大强而一致的证据的同时抑制不可靠信号，且无需依赖显式的置信度监督。这一概率机制使身体与手部流之间具备可靠性感知能力的聚合更加稳定，并提升了鲁棒性，尤其适用于关键点存在噪声或缺失时由手部驱动的细粒度动作识别。

超越骨架的多模态集成。传统的骨架内多模态集成——整合四种骨架模态（关节、骨骼、关节运动和骨骼运动）已被证明能有效建模基于骨架的识别中的时空线索。我们进一步将这一原则扩展至骨架域之外：受 MMNet 启发 [3], 我们采用身体引导的调制策略，以关节运动特征动态调节 RGB 特征学习，从而在统一的跨模态表述中连接结构表征与视觉表征。

本文基于我们的会议论文 [7] ，并将其扩展为统一的概率双流框架。第3 节重新审视确定性的身体–手部基线和预处理流程，第4 节则介绍广义概率框架。第 5 节报告全面的评估结果，包括新增的以手部为中心的基准、消融实验，以及鲁棒性和效率分析。实现细节和完整的定量结果见附录。

贡献。

- 广义概率双流框架：在统一表述下整合免校准骨架学习、可靠性感知融合和跨模态集成，并对身体与手部模态之间的不对称可靠性进行建模。
- Noisy-OR 融合：提出一种概率机制，无需显式的置信度监督即可稳定可靠性感知融合。
- 从骨架内集成扩展到跨模态集成：将骨架内运动线索整合（关节、骨骼、关节运动、骨骼运动）扩展至视觉模态，从而以跨模态表述实现统一的骨架–RGB 学习。
- 全面的验证与泛化能力分析：涵盖一项新的以手部为中心的基准（NTU-Hand 11/27）、针对免校准和 Noisy-OR 组件的扩展消融实验、帧丢失条件下的噪声鲁棒性，以及跨多个基准数据集的跨模态评估，结果表明模型在各种条件下均取得了稳定提升。

2. 相关工作

2.1. 基于骨架的动作识别

基于骨架的动作识别经历了多种架构范式的发展，从序列模型到图网络和基于注意力的网络。早期方法依赖循环神经网络 [9] 来捕捉时间动态，但在显式建模关节之间的空间关系方面存在局限。随后，研究者探索了卷积架构，将关节轨迹编码为类似图像的表示，并应用卷积神经网络 [11, 12], 。这提高了效率，但对底层骨骼结构的处理仍然只是隐式的。

时空图卷积网络（ST-GCN）的引入 [35] 标志着一个关键转变：它将关节表示为图节点，将运动学连接表示为边，从而能够联合建模空间和时间依赖关系。在这一方法的基础上，一系列基于图卷积网络的方法 [4, 6, 14, 19, 28, 38] 改进了图拓扑设计、聚合算子和多流表示，以增强鲁棒性和判别能力。此外，基于图的 Transformer框架 [1, 10, 23, 31] 引入了自注意力，以捕获长程依赖关系和全局运动模式。

尽管这些模型大多仍以身体为中心，但近期的研究（如SkeleT） [36] 将身体、手部和足部关键点整合到统一图中，以捕获全面的人体动态。然而，在这类统一表示中，全局身体运动仍可能掩盖细微的手部关节动作，因此有必要采用能够显式考虑模态特定可靠性和专门化的框架；本研究正是致力于此。

2.2. 多模态动作识别

早期基于图的方法 [20, 28] 主要依赖两种输入模态——关节表示和骨骼表示。后续研究 [4, 5, 36]通过引入关节运动（JM）和骨骼运动（BM）所体现的时间动态，将这一范式扩展到四种模态。进一步的拓展 [6, 19]探索了六流变体，以联合捕捉多种时空线索。近期，3MFormer [31]证明，基于超图的表述能够灵活聚合多尺度时空关系，验证了在基于骨架的识别中融合多种模态的重要性。

除纯骨架数据外，越来越多的研究开始整合视觉模态，以利用外观线索和上下文。近期方法 [2, 3, 11, 25] 通过架构对齐或特征融合，将姿态与 RGB 表征结合起来。其中，MMNet [3] 提出了一种身体引导的调制策略，利用身体特征对 RGB 激活进行加权，从而有效地将结构先验迁移到视觉表征中。这些进展促使我们提出一种概率双流集成方法，在考虑可靠性的框架下统一骨架模态与视觉线索。

2.3. 相关领域中的身体–手部协调

在手语识别、动作预测和第一人称视角理解等相关领域，身体与手部依赖关系建模一直备受关注。例如，UNI-SIGN [17] 将全身、手部和面部关键点嵌入共享潜在空间以进行手语识别，而 ExpForecastAI [8] 和 REWIND [15] 则侧重于未来姿态对齐和时间一致性预测。与这些方法不同，我们将身体和手部保留为独立的专家分支，并采用轻量级交叉注意力，无需共享嵌入或高精度定位。

2.4. 规范空间归一化与视角鲁棒性

规范空间归一化已广泛应用于手势识别和手部姿态估计 [13, 16, 27]，用于对齐局部手部坐标并减少视角变化，通常通过将手部关节变换到自我中心坐标系或规范坐标系来实现。研究表明，这类变换能够提升受控近距离场景中的识别效果 [16, 27],，但其前提是能够精确估计手部关节，并获得稳定的局部几何结构——这一假设在大规模第三人称人体动作识别数据集中很少成立，因为手部关节尺寸较小，常被遮挡或缺失。在这些情况下，规范映射可能会放大估计噪声，并使局部误差沿骨架传播，最终降低识别性能。

在人体动作识别领域，早期基于骨架的方法依靠规范映射来处理视角变化 [20, 35], ，而较新的架构则倾向于通过在训练期间对三维骨架进行序列级随机旋转，以数据驱动的方式增强鲁棒性 [4, 22, 37]。这些研究表明，无需固定的规范空间，也能实现出色的性能和视角不变性。我们的工作沿着后一方向，进一步研究更具挑战性的身体–手部场景，其中手部估计的噪声远高于身体关节：我们不采用规范对齐，而是在原生坐标空间中采用免校准设计，并显式建模身体与手部模态之间的可靠性不对称，详见第4节。

2.5. 概率融合与可靠性感知建模

Noisy-OR模型 [24] 长期以来一直是概率推理中的基本算子，用于建模由多个独立原因触发某个事件的可能性。其在深度学习中的现代改进 [29, 33] 使感知和融合任务能够实现具备可靠性感知的多实例聚合与不确定性建模。与确定性平均或拼接不同，Noisy-OR聚合能够自然地编码不同来源之间不对称的置信度，在抑制不可靠信号的同时保留有力证据。本文在双流设置下实现了这种概率推理，采用Noisy-OR融合机制，对身体–手部模态进行具备可靠性感知的聚合。

3. 背景与基础框架

本节将形式化介绍全文所用的符号，描述身体–手部骨架构建的预处理流程，并回顾构成我们概率框架基础的确定性双流架构。

3.1. 符号与问题定义

我们将身体-手部骨架序列表示为

$$X_B \in \mathbb{R}^{C\times T\times V_B}, \qquad X_H \in \mathbb{R}^{C\times T\times V_H}, \tag{1}$$

其中， C 表示坐标通道 (x,y, z)， T 表示时间长度， V_B, V_H 分别表示身体和手部的关节数量。给定 K 个动作类别，目标是在关键点存在噪声或缺失的情况下鲁棒地学习

$$f_\theta : (X_B, X_H) \to y, \qquad y \in \mathbb{R}^K, \tag{2}$$

在关键点存在噪声或缺失时仍具有鲁棒性。

身体-手部建模的一个典型挑战是不同模态之间存在可靠性不对称：身体关节通常较为稳定，而手部关节则容易受到遮挡、模糊和频繁的估计失败影响。这促使我们引入第 4 节所述的可靠性感知学习。

3.2. 预处理流程概述

我们采用基线框架中的预处理流程：手部关键点提取（MediaPipe [21]）、时间平滑、用于保持拓扑一致性的零掩码虚拟节点、规范变换、以髋部为中心的归一化，以及时间重采样。这些步骤构成了一个确定性预处理流程，与我们的会议基线 [7],完全相同；但在本研究中，我们移除了规范变换，同时保留其余步骤（第 4.2 节）。更多预处理细节见附录A。

3.3. 基础双流架构

我们的会议框架 [7]提出了一种确定性的双流设计，用于联合建模身体动态和手部关节运动。我们简要总结其关键组成部分，作为本文所提出的广义概率框架的基础。

3.3.1. 采用互补损失的双流训练

基线双流网络由两个时空 GCN 骨干网络组成，分别专用于身体模态或手部模态。给定它们的对数几率输出 $\hat{y}_B$ 和$\hat{y}_H$，训练目标由个体损失 $\mathcal{L}_{\text{idv}}$和互补损失 $\mathcal{L}_{\text{cpl}}$组成：

$$\mathcal{L}_{\text{idv}} = \text{CE}(\hat{y}_B, l) + \text{CE}(\hat{y}_H, l), \tag{3}$$

$$\mathcal{L}_{\text{cpl}} = \text{CE}\left(\tfrac{1}{2}(\hat{y}_B + \hat{y}_H),\, l\right), \tag{4}$$

$$\mathcal{L}_{\text{total}} = \lambda_{\text{idv}}\mathcal{L}_{\text{idv}} + \lambda_{\text{cpl}}\mathcal{L}_{\text{cpl}}. \tag{5}$$

这里，CE 是采用 softmax 的标准交叉熵。

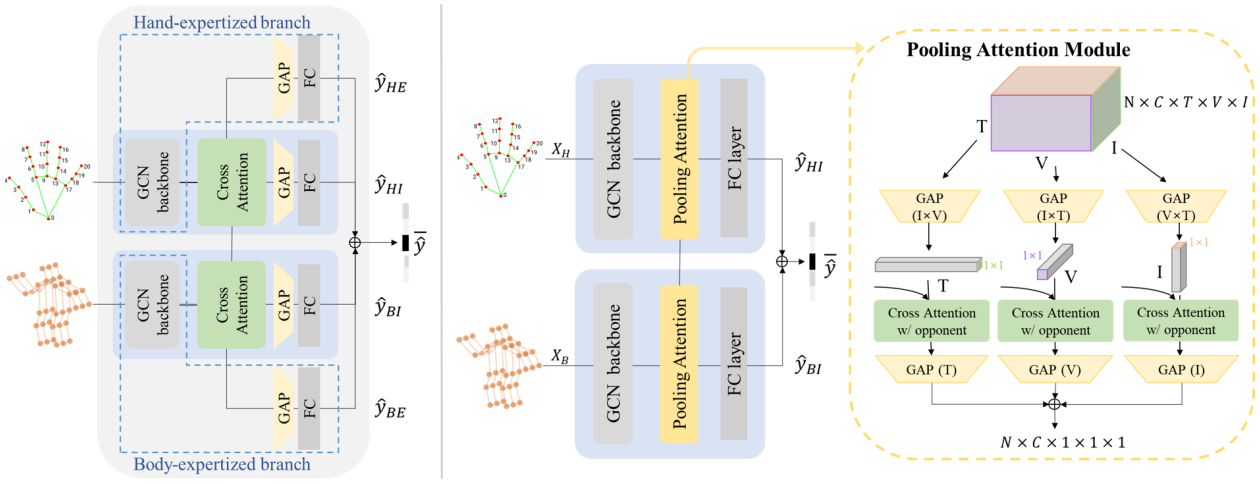


图 2： 双流架构概述。**左：** BHaRNet-P，由通过轻量级交叉注意力连接的身体交互（BI）分支和手部交互（HI）分支组成。**右：** BHaRNet-E，增加了专家分支（BE、HE），在通过交互分支共享上下文的同时保留特定于模态的线索。确定性基线和概率框架对应的分支—损失配置总结见表1。

模型	变体	分支	$\mathcal{L}_{\text{idv}}$	$\mathcal{L}_{\text{cpl}}$	$\mathcal{L}_{\text{nor}}$
确定性基线	P	BI、HI	CE(y_{BI})+交叉熵(y_{HI})	交叉熵(avg(BI,HI))	–
	E	BI,HI,BE,HE	CE(y_{BI})+交叉熵(y_{HI})	交叉熵(avg(all))	–
概率框架	P	身体交互分支，手部交互分支	交叉熵(y_{BI})+交叉熵(y_{HI})	交叉熵(avg(身体交互分支,手部交互分支))	交叉熵(Noisy-OR($\sigma(y_{BI})$, $\sigma(y_{HI})$))
	E	身体交互分支，手部交互分支，身体专家分支，手部专家分支	交叉熵(y_{BI})+交叉熵(y_{HI})	CE(avg(all))	交叉熵(Noisy-OR($\sigma(y_{BE})$, $\sigma(y_{HE})$))

表1确定性基线与概率框架的分支–损失–输出映射。BI/HI：身体/手部交互分支

;

BE/HE：身体/手部专家分支。 σ：Sigmoid 激活。Noisy-OR 按元素应用。

该联合训练方案促进了模态间协作，同时保留了模态特定的专业化能力。为缓解估计的手部骨架与身体关节之间的域差异，我们采用了规范空间变换来对手部坐标进行归一化。尽管这种预处理在某些情况下稳定了训练，但当手部关节估计不准确时，也会放大噪声传播，因此促成了第 4.2 节中的免校准设计。

3.3.2. 交叉注意力机制

为增强身体与手部流之间的信息交换，我们在特征层引入了交叉注意力机制。**BHaRNet-P**变体采用了两个交互分支——身体交互分支（BI）和手部交互分支（HI）——通过池化注意力模块交换特征（图2）。**BHaRNet-E** 变体进一步引入了专家分支——身体专家分支（BE）和手部专家分支（HE）——以平衡模态特定专门化与跨模态通信。训练期间，交互分支和专家分支均参与损失计算；推理时，仅使用专家分支的轻量级配置即可实现高效率。

3.3.3. 确定性骨架–*RGB*集成

除仅基于骨架的识别外，该确定性框架还扩展为纳入 RGB 线索，以进行基于外观的推理。遵循 MMNet [3]，的原则，来自关节模态的身体专家特征 f_{BE} 被用作 RGB 激活的空间加权信号：

$$f'_{\text{RGB}} = f_{\text{RGB}} \odot \text{weight}(f_{BE}), \tag{6}$$

其中， weight(·) 表示一种可学习变换，用于生成区域级重要性图。该调制将结构先验从骨架特征迁移至 RGB 表征，使空间显著性与关节运动相对齐。设 J 和 B 分别表示关节模态和骨骼模态，则最终预测通过确定性加权求和得到：

$$\hat{y}_{\text{final}} = w_J \hat{y}_J + w_B \hat{y}_B + w_{\text{RGB}} \hat{y}_{\text{RGB}}. \tag{7}$$

该集成在统一的确定性表述下，为整合结构线索和外观线索提供了强有力的基线，集成权重固定为 $(w_J : w_B : w_{\text{RGB}})$ =(1:1:1)。

上述架构构成了确定性基础，我们的广义概率双流框架便建立于此。在下一节中，我们将重新表述这些

确定性组件，将其转化为一种具备可靠性感知能力的概率学习框架，显式建模不确定性，并将多模态融合扩展至时间维度。

4. 广义概率双流框架

我们现在提出一种广义概率双流框架，通过引入可靠性感知学习和结构化多模态集成，对确定性基线（第 3 节）进行扩展。该框架在保留原始设计的高效性和确定性推理的同时，将概率建模融入训练目标，以提升模型在不确定骨架数据上的鲁棒性和专门化能力。图 2 展示了双流架构（BHaRNet-P/E），它们作为确定性基线和我们提出的概率扩展的骨干网络。

4.1. 概述

该框架包含三个关键组成部分：（1）不使用规范变换的表征学习；（2）基于 Noisy-OR 损失的可靠性感知概率学习；（3）融合时间线索的多模态集成。总训练目标结合确定性监督和概率监督，定义如下：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{idv}}\mathcal{L}_{\text{idv}} + \lambda_{\text{cpl}}\mathcal{L}_{\text{cpl}} + \lambda_{\text{nor}}\mathcal{L}_{\text{nor}}, \tag{8}$$

其中，$\mathcal{L}_{\text{idv}}$ 和 $\mathcal{L}_{\text{cpl}}$ 分别表示基线中的单项损失和互补损失，$\mathcal{L}_{\text{nor}}$ 表示所提出的概率Noisy-OR正则项。在推理期间，预测仍然是确定性的，通过对各分支的 logit 求和得到，从而确保效率与基线相同。

4.2. 无规范变换的表征学习

以往的身体–手部框架通常依赖规范空间归一化，以对齐不同受试者的手部坐标。尽管这种校准能够降低受试者间的差异，但它依赖于准确的关键点估计——这一条件在大规模第三人称HAR数据中经常无法满足，因为手部关节会受到遮挡和运动模糊的影响。因此，局部手部关节的误差会通过变换在全局范围内传播，从而放大噪声（图3）。

我们移除了这一规范步骤，直接从原生骨架坐标中进行学习。这种设计保留了局部空间一致性，避免了不可靠关节导致的误差传播，并使身体和手部处于同一坐标系中，从而允许模型隐式学习跨尺度对应关系。在实践中，我们发现，在手部遮挡和运动模糊的情况下，无需校准的设置能够带来更稳定的训练动态，相关分析见第5节。

4.3. 采用 Noisy-OR损失的可靠性感知概率学习

第3 节中的确定性损失公式隐含地假设所有分支的可靠性相同。实际上，

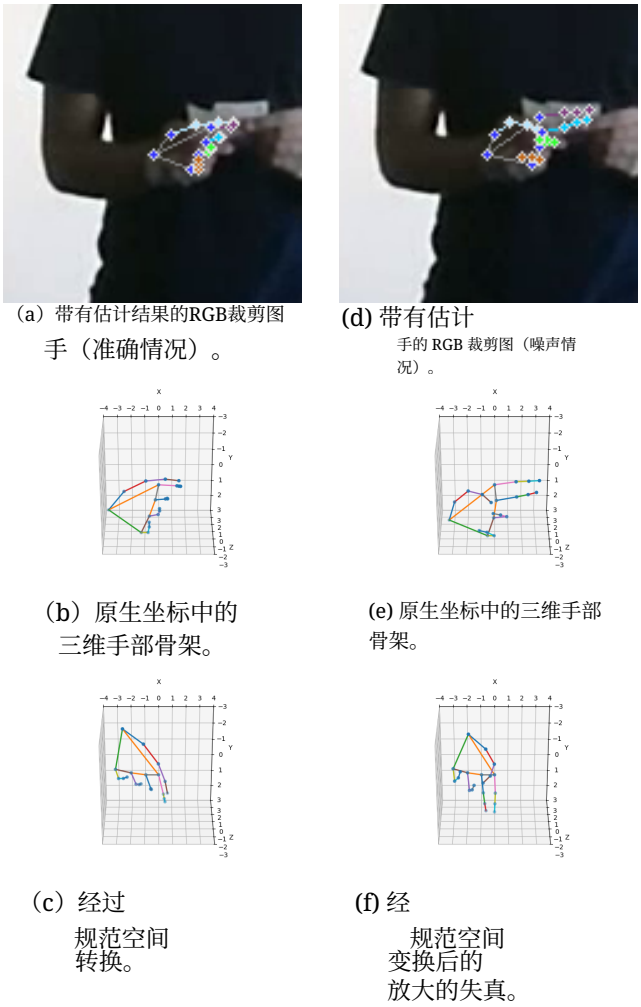


图3： 用于我们广义框架的免校准表示学习示例。我们展示了来自单个序列、帧率为30 fps的连续两帧。左：第17帧，手部估计准确。右：第18帧，其中食指和中指受到噪声干扰。各行分别显示：（上）带有手部估计的RGB裁剪图；（中）原生三维手部骨架；（下）经过规范空间变换后的三维手部骨架。为便于公平比较，所有三维视图均采用相同的视角和坐标轴尺度。在原生空间中，除受噪声影响的食指和中指关节附近外，整体手部构型保持稳定。经过规范空间变换后，局部噪声传播至整只手，导致较大的逐关节位移和明显的形状失真。

手部相关分支常因关键点缺失或不准确而不稳定，导致学习信号不一致。为了解决这一问题，我们引入了一种基于 Noisy-OR 算子的可靠性感知项，用于聚合逐分支证据，并倾向于选择至少有一个分支能够有把握地支持真实类别的配置。

概率 *formul*化。给定分支级逻辑值 $y_i \in \mathbb{R}^K$，我们首先通过 Sigmoid函数计算逐类别分数：

$$p_{i,k} = \sigma(y_{i,k}), \quad k = 1, \ldots, K, \tag{9}$$

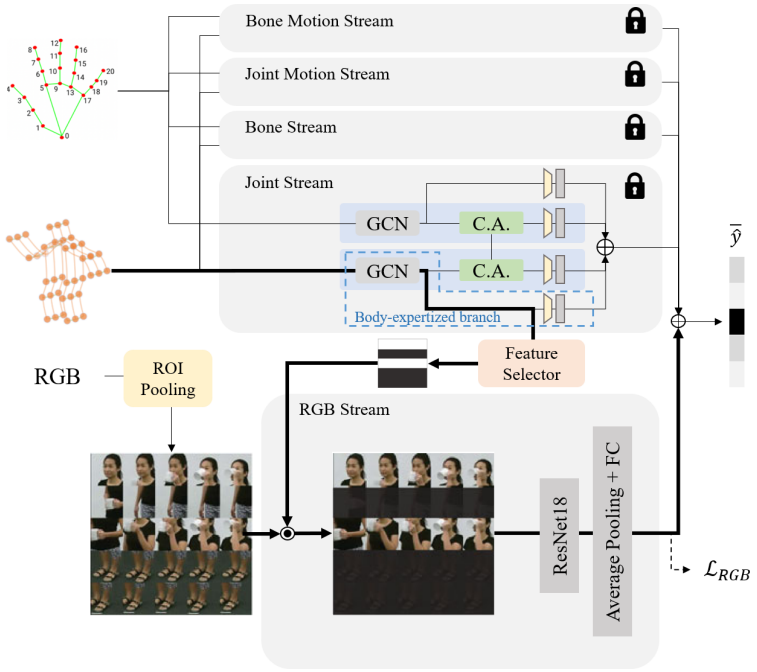


图4： BHaRNet-M 示意图。我们将用于骨骼流的 BHaRNet-E 集成进来，并添加了一条具有独立训练路径的 RGB 流（粗线）。RGB分支接收来自身体专家化分支的身体关节引导，使视觉特征提取器聚焦于相关的时空区域。

该函数将每个 logit 映射到 $[0,1]$，并稳定分支间的尺度差异。随后，我们使用逐元素 Noisy-OR算子对各分支的这些分数进行聚合：

$$p_{\mathrm{nor},k} = 1 - \prod_i \left(1 - p_{i,k}\right). \tag{10}$$

在这里，如果至少有一个分支对类别 k赋予较高置信度，那么 $p_{\mathrm{nor},\,k}$ 就会变大；而得分较低的不可靠分支则贡献甚微。

我们不将 $p_{\mathrm{nor}} = (p_{\mathrm{nor},1}, \ldots, p_{\mathrm{nor},K})$解释为归一化的概率分布，而是将其视为逐类别证据分数。随后，我们将 p_{nor} 输入基于 softmax 的交叉熵：

$$\mathcal{L}_{\mathrm{nor}} = \mathrm{CE}(p_{\mathrm{nor}}, l).$$

换言之，Noisy-OR 会在计算标准的单标签交叉熵损失之前，对逐分支证据进行可微池化。我们将 $\mathcal{L}_{\mathrm{nor}}$ 用作训练时正则项，以鼓励至少一个分支对正确类别提供有把握的支持，同时不改变确定性推理规则。

变体特定的应用。 表1总结了确定性变体和概率变体的分支—损失映射。在 BHaRNet-P 中，Noisy-OR 项定义在交互分支 BI 和 HI 上：

$$p_{\mathrm{nor}}^{(P)} = 1 - \left(1 - \sigma(y_{BI})\right) \odot \left(1 - \sigma(y_{HI})\right), \tag{11}$$

并且 $\mathcal{L}_{\mathrm{nor}}$ 由经 softmax 归一化的 Noisy-OR 分数计算得出。这促使模型依赖其中可靠的交互分支，同时减轻受损分支的影响。

在 BHaRNet-E 中，专家分支 BE 和 HE 旨在专攻模态特定推理，而交互分支 BI 和 HI 则通过互补损失维持跨模态上下文。在这里，Noisy-OR 项仅定义在专家分支上：

$$p_{\mathrm{nor}}^{(E)} = 1 - \left(1 - \sigma(y_{BE})\right) \odot \left(1 - \sigma(y_{HE})\right), \tag{12}$$

从而在身体专家或手部专家有把握时，任一方都能主导决策。这一设计符合轻量级推理模式，该模式仅使用 BE 和 HE。

解释与独立性。 在我们的设定中，无法保证分支之间严格的统计独立性，因为这些分支共享上游编码器和监督信号，且 BI/HI 通过交叉注意力交换信息。然而，在专家化配置（BHaRNet-E）中，BE/HE 不会直接相互关注，并且被鼓励专注于模态特定线索。因此，我们将 Noisy-OR 算子视为对部分独立专家的近似汇聚，而非完全生成式概率模型，并将其用作 $\mathcal{L}_{\mathrm{nor}}$ 训练时正则项，使学习倾向于至少有一个专家有把握地支持正确类别的配置。

4.4. 带时间线索的多模态集成

我们进一步通过在相同的 logit 求和规则下纳入时间运动模态，推广确定性集成（第 3 节）。该集成仍具有确定性且无需参数，并将空间骨骼线索扩展至

表 2基于骨架的动作识别中，在 NTU RGB+D 60/120 和 N-UCLA 基准数据集上的准确率（%）、浮点运算次数和参数量与最先进方法的比较。“-”表示参考文献未提供实验结果，“*”表示基于公开代码得到的结果。最佳和次佳结果分别以**粗体** 和下划线标出。BHaRNet-*†表示我们的会议基线，而“Ours”行对应所提出的概率变体。

方法	NTU 60		NTU 120		N-UCLA	十亿次浮点运算	参数量（百万）
	跨主体	跨视角	跨主体	跨设置	跨视角		
CTR-GCN [4]	92.4	96.8	88.9	90.6	96.5	7.9	5.8
InfoGCN [6]	93.0	97.1	89.8	91.2	97.0	10.0*	9.4
PoseConv3D [11]	94.1	97.1	86.9	90.3	-	31.8	4.0
BlockGCN [38]	93.1	97.0	90.3	91.5	96.9	6.5	5.2
DeGCN [22]	93.6	97.4	91.0	92.1	97.2	6.9	5.6
3MFormer [31]	94.8	98.7	92.0	93.8	**97.8**	58.5	6.7
ProtoGCN [19]	93.8	97.8	90.9	92.2	-	43.4	24.9
SkeleT [36]	**97.0**	**99.6**	94.6	**96.4**	97.6	9.6	5.2
BHaRNet-B† [7]	96.1	98.7	94.0	94.9	94.6	**3.3**	**2.8**
BHaRNet-E† [7]	96.2	98.8	94.3	95.0	94.6	5.4	5.5
BHaRNet-P† [7]	96.3	98.8	94.3	95.2	95.3	3.4	4.9
我们的方法（BHaRNet-B）	96.6	99.1	94.5	95.5	95.9	6.6	5.5
我们的方法（BHaRNet-E）	96.7	99.2	94.7	95.7	96.3	10.9	11.0
我们的方法（BHaRNet-P）	96.8	99.2	**94.8**	95.8	95.9	6.8	9.7

（关节、骨骼）与时间动态（关节运动、骨骼运动），并在统一的 logit 空间公式中整合 RGB 外观线索。

骨架内集成。 在骨架域中，将四种模态——关节（J）、骨骼（B）、关节运动（JM）和骨骼运动（BM）——聚合如下：

$$\hat{y}_{\text{skel}} = 2(\hat{y}_J + \hat{y}_B) + (\hat{y}_{JM} + \hat{y}_{BM}), \tag{13}$$

空间模态获得较高权重，以保持结构稳定性；运动模态则提供互补的时间敏感性。

骨架-*RGB*集成。 在 BHaRNet-M（图 4）中，RGB 流采用基于 MMNet 的调制策略 [3], 其中，骨架特征引导注意力聚焦于与运动相关的区域。最终预测将确定性集成扩展至包含外观线索：

$$\hat{y}_{\text{final}} = 2(\hat{y}_J + \hat{y}_B) + (\hat{y}_{JM} + \hat{y}_{BM}) + 3\hat{y}_{\text{RGB}}. \tag{14}$$

这种五模态融合（J、B、JM、BM、RGB）在统一的确定性规则下，同时丰富了时空表征和视觉表征。固定权重（2 : 2 : 1 : 1 : 3）是通过在验证划分上进行网格搜索选定的。

效果。添加时间模态能够持续提升对细粒度动作和手部中心动作的识别能力，同时保持推理过程简单。固定权重集成降低了模态间的方差，在多种噪声和视角条件下提供可靠的性能，并进一步表明，与异构模态集成时，具备可靠性感知能力的骨架特征仍然有益。

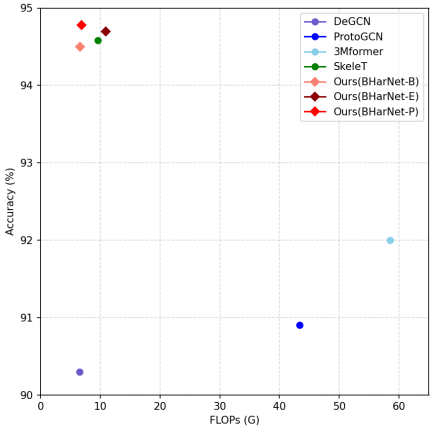


图 5： NTU 120跨主体骨架动作识别的准确率-GFLOPs权衡。红色调标记表示我们的概率模型（BHaRNet-B/E/P），绿至蓝色标记表示以往基于骨架的最先进方法（DeGCN、ProtoGCN、3MFormer、SkeleT）。

5. 实验

我们在主要的基于骨架的动作识别基准上评估所提出的概率双流框架，包括 NTU RGB+D 60/120 [26],PKU-MMD [18] ，以及 N-UCLA [30]。此外，我们还定义了一个新的以手为中心的基准 NTU-Hand 11/27，该基准取自 NTU 60/120 中以手部动作为主的类别，用于评估涉及细微手部关节运动的细粒度动作。为清晰起见，我们报告两个基线参考版本：(i) BHaRNet†：已发表的会议版本，作为所有主要结果表中的官方对比基线；(ii)BHaRNet‡：

表 3NTU RGB+D 60/120 和 PKU-MMD 基准数据集上的多模态动作识别。“*”表示参考论文中提供的实验结果。最佳和次佳结果分别以**粗体** 和下划线标出。BHaRNet-M† 表示会议基线，“我们的方法”表示所提出的概率变体。

方法	模态	NTU 60		NTU 120		PKU-MMD		十亿次浮点运算	参数量（百万）
		跨主体	跨视角	跨主体	跨设置	跨主体	跨视角		
MMNet[3]	J+B+RGB	96.6	99.1	92.9	94.4	97.4	98.6	89.2	34.2
PoseConv3D[11]	B+RGB	**97.0**	**99.6**	95.3	96.4	-	-	41.8*	31.6*
π-ViT[25]	J+RGB	94.0	97.9	91.9	92.9	-	-	590.0	121.4
EPAM-Net[2]	J+RGB	96.1	99.0	92.4	94.3	96.2	98.4	8.1	**2.5**
BHaRNet-M† [7]	J+B+RGB	96.3	99.0	95.1	96.0	96.9	97.9	**7.6**	16.7
我们的方法（BHaRNet-M）	J+B+RGB	96.8	99.3	95.3	96.2	97.3	98.5	**7.6**	16.7
	J+B+关节运动+骨骼运动+RGB	**97.0**	99.4	**95.5**	**96.5**	**97.5**	**98.7**	13.0	22.2

采用更新后的免校准预处理进行训练的复现版本。在消融实验（第 5.5 节）中，我们将 BHaRNet‡ 视为“+免校准”步骤，并在附录 C 中提供完整结果。

5.1. 实现细节

除非另有说明，训练配置均遵循会议版本 [7]。对于骨架模型（BHaRNet-B/E/P），我们采用两阶段训练方案：(1) 分别预训练身体与手部流（DeGCN [22] 骨干网络）；(2) 使用预训练权重微调整个双流模型。对于多模态模型（BHaRNet-M），我们额外引入一个阶段来训练 RGB流。所有实验均在 1 块 RTX 3090 GPU 上进行。骨架预处理步骤遵循第 4.2节中的无标定设计，变体命名见附录 B。

5.2. 基于骨架的识别主要结果

表 2 总结了我们的方法与 NTU RGB+D 60/120 和 N-UCLA 上最先进的（SOTA）基于骨架的模型的比较结果。我们的方法在 NTU 120 X-Sub 上取得了 SOTA 准确率，在其余所有评测协议上也取得了接近 SOTA 的结果，同时计算成本具有竞争力（6.6–10.9 十亿次浮点运算）。尽管在 N-UCLA 上的表现相对一般，但该数据集中的手部中心动作较少，训练样本也有限。尽管如此，我们的框架相较会议版本仍有显著提升，表明其在低数据量场景下具有更强的泛化能力。

观察结果。与先前的大规模SOTA模型（3MFormer [31], SkeleT [36]）相比，我们的框架仅使用约少30–90%的浮点运算次数，便实现了接近SOTA的性能，同时保持了良好的效率与准确率的权衡（图5）。

5.3. 融合RGB的多模态识别

表3 比较了我们的多模态框架（BHaRNet-M）与近期的姿态–RGB方法相比。所提出的模型在 NTU 120 和 PKU-MMD 上实现了最先进的准确率，在 NTU 60 上也取得了极具竞争力的结果，同时保持较低的计算开销。在相同模态

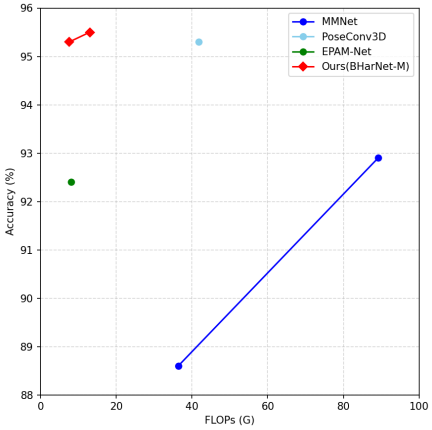


图6： NTU 120跨主体数据集上多模态（骨骼+RGB）动作识别的准确率–十亿次浮点运算权衡。红色标记表示我们的多模态框架（BHaRNet-M，左：J+B+RGB / 右：J+B+JM+BM+RGB），绿色至蓝色标记表示先前的姿态–RGB模型（MMNet、PoseConv3D、EPAM-Net）。

设置（J+B+RGB）下，我们更新后的 BHaRNet-M 已经超越其会议版本（BHaRNet-M†），这表明仅凭概率学习和免校准设计，就能稳定提升 0.2–0.6 个百分点（pp）。将骨架内运动线索（关节运动、骨骼运动）扩展至 RGB 后，性能进一步小幅但稳定地提升了 0.1–0.3pp，这表明运动线索可通过身体引导的调制有效迁移至 RGB，从而强化互补的时序–外观交互。

效率。图6 显示，在姿态–RGB 方法中，我们的五模态 BHaRNet-M (J+B+JM+BM+RGB) 在 NTU120 跨主体评测中取得了最高准确率，同时所需的浮点运算次数远少于以往高准确率模型。该模型能够平稳扩展，且没有明显过拟合，验证了概率 formul化在异构输入上的稳定性。

表4NTU-Hand11/27 上基于骨架的动作识别与最先进方法的性能比较。“NTU-Hand 11”指 NTU RGB+D 60 的 60 个类别中的 11 个以手部为中心的类别，“NTU-Hand 27”指 NTU RGB+D 120 的 120 个类别中的 27 个类别。

方法	NTU-Hand 11		NTU-Hand 27	
	跨主体	跨视角	跨主体	跨设置
ProtoGCN [19]	87.4	94.4	82.0	84.7
SkeleT [36]	94.7	**98.6**	89.9	92.6
BHaRNet-P†	95.4	97.8	90.8	92.2
我们的方法 (BHaRNet-E)	95.7	98.5	91.4	**93.1**
我们的方法 (BHaRNet-P)	**96.1**	98.5	**91.7**	**93.1**

表 5在 NTU RGB+D 120 和 NTU-Hand 27 跨主体数据集上进行模态扩展的消融研究。

方法	模态	NTU 120	NTU-Hand 27
我们的方法 (BHaRNet-E)	J+B	94.5	91.3
	J+B+关节运动+骨骼运动	94.7	91.4
我们的方法 (BHaRNet-M)	J+B+关节运动+骨骼运动+RGB	**95.5**	**92.4**

5.4. 以手部为中心的基准：NTU-Hand 11/27

为评估细粒度动作，我们引入 NTU-Hand 11/27，其中包含来自 NTU 60 的 11 个以手部为中心的类别，以及来自 NTU 120 的另外 16 个类别。以手部为中心的子集包含手部关节运动细微的动作，而这些动作在全类别评估中的贡献极小。评测采用标准的 X-Sub/X-View（NTU-Hand 11）和 X-Sub/跨设置（NTU-Hand 27）评测协议。表4 显示，我们的方法在两个划分上均取得最佳或相当的表现，在以手部为中心的子集上优于 SkeleT（最高达 2 个百分点）和 ProtoGCN（高出 4–9 个百分点）。这些结果表明，可靠性建模能够有效捕捉以身体为中心的模型往往忽略的细微关节活动。详细的类别列表、各类别准确率和完整的变体比较见附录C。

5.5. 消融实验

我们进行了一系列消融实验，以量化各组件的贡献。我们评估了三种逐步增强的配置：(i) 基线†，(ii) +无需校准的预处理（相当于复现的基线 BHaRNet‡），(iii) +Noisy-OR损失（我们的完整概率模型）。值得注意的是，这些步骤不会引入额外参数或浮点运算次数，从而能够单独评估可靠性感知学习的效果，而不将性能提升与架构扩展混为一谈。附录C提供了全面结果，包括各变体的详细分析以及跨数据集的模态扩展；结果显示的趋势与主表中报告的改进一致。

5.5.1. 模态扩展。

表 5 报告了模态扩展模块带来的逐步改进。从我们的概率 J+B 配置出发，我们进一步加入 JM/BM 和 RGB。值得注意的是，这些改进在 NTU 120 和 NTU-Hand 27 上始终如一，表明概率 f ormul化对大规模动作和细粒度动作均有益。整合时序运动线索（JM、BM）可进一步提升性能，而加入 RGB 则取得最高准确率，表明各组件均带来了互补的稳健性。

表6针对 NTU RGB+D 60 和 NTU-Hand 11 的跨视角评估消融研究。关节-骨骼集成模型的比较。

方法	NTU 60	NTU-Hand 11
	跨视角	跨视角
BHaRNet-E†	98.8	97.9
我们的方法（BHaRNet-E）	**99.0**	**98.2**

5.5.2. 跨视角鲁棒性。

表 6 评估了视角变化下的性能。我们更新后的框架持续提升跨视角泛化能力，这表明会议版本中的规范空间对齐可能已经对特定的相机几何结构产生了过拟合。

5.5.3. 新颖性细分。

表 7 分解了免校准预处理和 Noisy-OR 损失对各变体（B、E、P）的影响。在 NTU 120 和 NTU-Hand 27 上，所有变体均表现出一致的提升，证实所提出的模块能够适应不同的变体风格。唯一的例外是 BHaRNet-B 在 NTU-Hand 27 上的表现：免校准变体略优于加入 Noisy-OR 正则化的变体。我们将此归因于一种保守的转变，即转向

表7NTU 120 和 NTU-Hand 27 跨主体基准数据集上各项创新步骤的消融研究。采用关节模态进行处理。

方法	NTU 120	NTU-Hand 27
BHaRNet-B†	92.7	88.6
+ 免校准	93.0	**89.4**
+Noisy-OR损失	**93.2**	89.2
BHaRNet-E†	93.0	88.8
+ 免校准	93.3	89.4
+Noisy-OR损失	**93.5**	**89.5**
BHaRNet-P†	92.9	88.7
+ 免校准	93.3	89.4
+Noisy-OR损失	**93.4**	**89.5**

表 8NTU RGB+D 120 和 NTU-Hand 27 跨主体基准上的帧丢失鲁棒性。我们报告不同帧缺失率（0、0.25、0.5）下的准确率。

方法	NTU 120			NTU-Hand 27		
	0	0.25	0.5	0	0.25	0.5
BHaRNet-B†	92.7	90.1	89.8	88.6	88.6	88.0
+ 免校准	93.0	92.8	92.6	**89.4**	88.7	88.4
+Noisy-OR损失	**93.2**	**93.0**	**92.8**	89.2	**89.1**	**88.6**
BHaRNet-E†	93.0	90.1	90.1	88.8	88.9	88.3
+ 免校准	93.3	93.2	93.0	89.4	88.8	88.9
+Noisy-OR损失	**93.5**	**93.3**	**93.1**	**89.5**	**89.3**	**89.1**

仅专家设置下的身体线索，并在鲁棒性研究（第 5.5.4 节）中进行详细分析。

5.5.4. 噪声鲁棒性。

表 8分析了在帧丢弃条件（0、0.25、0.5）下，专家分支专用（BHaRNet-B）和专家+交互（BHaRNet-E）配置的鲁棒性。对于规范空间基线（BHaRNet-B†/E†），随着帧丢弃率的增加，准确率急剧下降，在 NTU 120 上下降超过 2–3 个百分点。移除规范变换后，曲线已明显变得平缓，尤其是对于 BHaRNet-E，这表明免校准预处理减轻了受损手部帧所导致的时间噪声放大。有趣的是，基线模型在 25% 帧丢弃条件下略有提升。这是因为规范空间对齐会放大手部关节噪声；移除一部分严重受损的帧实际上减少了这种传播的噪声，从而产生轻微的去噪效果。相比之下，我们的概率 formul化已经抑制了不可靠线索，因此无需移除帧即可获得稳定的性能（图7）。

采用 Noisy-OR损失后，我们的模型在 NTU-Hand 27 上同样变得更加鲁棒，展现出稳定的可靠性感知融合能力。对于 BHaRNet-B 和 BHaRNet-E，在 NTU-Hand 27 上，0 % 与 50% 帧丢弃之间的差距沿平滑曲线缩小，而免校准基线则遭受明显更大的性能下降。此外，对于仅包含专家分支的 BHaRNet-B 变体，加入 Noisy-OR损失后

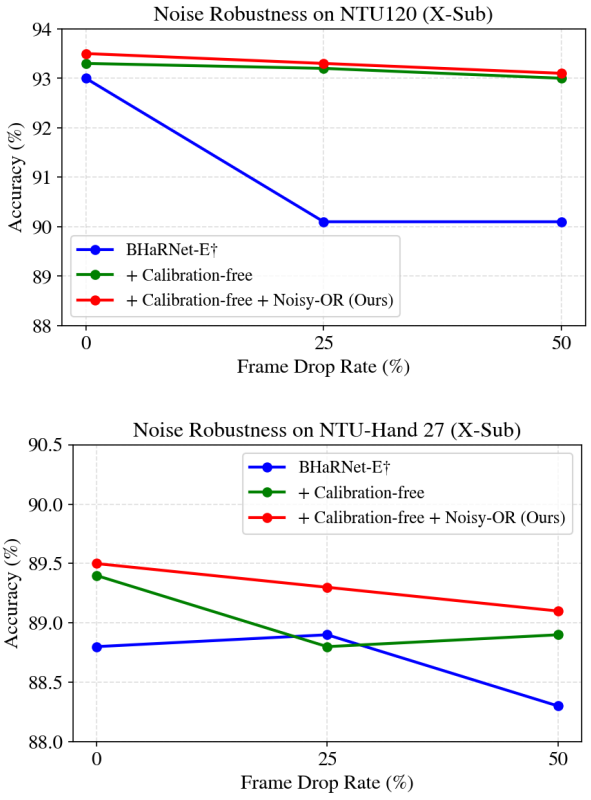


图 7： BHaRNet-E模型在NTU 120（上）和NTU-Hand 27（下）跨主体基准数据集的帧丢失条件下的噪声鲁棒性。我们评估了会议基线BHaRNet-E† （蓝线）、采用免校准预处理的概率模型（绿线），以及采用免校准预处理和Noisy-OR损失的概率模型（红线）在帧缺失率为0、25和50%时的表现。

在丢帧扰动下的准确率更加稳定，但与免校准基线相比，以手部为中心的性能略有下降。我们推测，在没有交互分支的情况下，概率正则项会使专家分支偏向更加保守的身体线索；而在完整的 BHaRNet-E 配置中，这一影响能够得到补偿，因为交互分支可以恢复细粒度的手部信息。

5.6. 效率与权衡分析

在仅使用骨架和多模态设置下，与大多数既有方法相比，我们的框架在浮点运算次数相近或更少的情况下，取得了更高或相当的准确率；相较于以身体为中心的模型，提升尤为显著。图8 进一步展示了在 1/2/4 模态设置（J/J+B/J+B+JM+BM）下，我们的方法与 SkeleT 在准确率-GFLOPs权衡方面的对比，表明我们的模型位于或接近比 SkeleT 更优的准确率-效率前沿。在验证数据上进行网格搜索后发现，固定的集成权重（2:2:1:1:3）在 ±50%扰动（<0准确率差异为 0.1 个百分点）范围内保持稳定，这支持了无参数推理，并表明模型在不同配置下具有鲁棒性。这些结果验证了所提出的

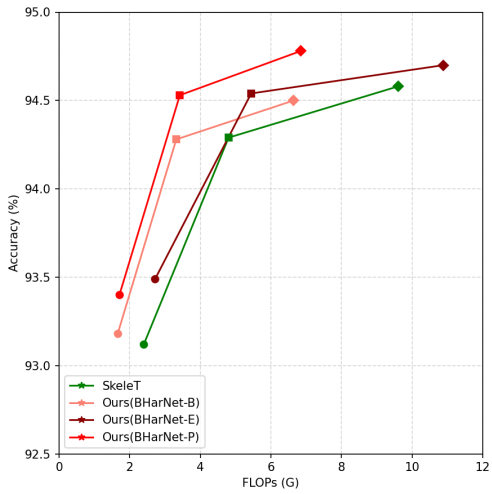


图8： 在 NTU 120 跨主体基准测试中，SkeleT 与我们的概率框架在单模态（线左侧，标记为 ●）、双模态（中间， ■）和四模态（右侧， ◆）配置下的准确率–GFLOPs 对比。绿色边框标记表示 SkeleT 变体，红色边框标记表示我们对应的 BHaRNet 变体（B/E/P）。

方案在不增加推理复杂度的情况下提升了鲁棒性，符合可靠性感知建模的设计目标。

6. 结论

我们提出了一个用于身体-手部动作识别的广义概率双流框架，统一了免校准骨架学习、可靠性感知融合和多模态整合。大量实验表明，所提出的框架以相近的计算成本实现了最先进或具有竞争力的性能，并持续优于我们的会议基线。新的以手部为中心的基准（NTU-Hand 11/27）进一步展现了在细粒度、手部主导动作上的显著提升。消融实验证实，免校准学习与Noisy-OR损失能够带来互补收益，在不增加推理复杂度的情况下，提高了模型对视角变化和丢帧扰动的鲁棒性。

尽管具有这些优势，我们的框架最终仍受到底层动作信号质量的限制，并且在 N-UCLA 等小型数据集上仍表现不佳。在第三人称人类动作识别视频中，手部通常较小、被遮挡或模糊，因此提取出的手部骨架天然带有噪声，有时也不可靠。在本研究中，我们有意聚焦于标准基准数据集上的单一 RGB 骨架提取流程，以便单独考察所提出的可靠性感知学习的效果。将该框架扩展到多样化的传感配置（例如 RGB-D摄像机、可穿戴传感器或自我中心传感器），并研究传感器选择与可靠性建模之间的相互作用，是未来值得探索的方向。

7. 致谢

我们谨此感谢 Hye-Yeon Kim 在本研究期间给予的宝贵帮助与支持。

A. 预处理与以手部为中心的基准详情

A.1. 身体–手部骨架构建

我们遵循第 4.2 节所述的身体–手部预处理流程。给定 RGB 视频帧，使用 MediaPipe 估计身体和手部关键点 [21]。随后，我们构建身体骨架序列 $X_B \in \mathbb{R}^{C\times T\times V_B}$ 和手部骨架序列 $X_H \in \mathbb{R}^{C\times T\times V_H}$，其中 C 表示坐标通道 (x, y, z)，T 表示时间长度，而 V_B，V_H 分别表示身体关节数和手部关节数。

时间处理。 我们首先识别手部检测结果可靠的有效帧，并通过筛选这些帧以及对较短间隔进行线性插值来执行时间平滑。随后，将得到的序列重采样至固定长度；必要时，按照近期研究中的常见做法，通过堆叠边界帧进行填充。

空间处理。 对于手部流，我们采用包含原生运动学边的 21 关节 MediaPipe 手部图。对于具有 25 个身体关节的数据集（NTU RGB+D 60/120、PKU-MMD），我们添加 21 个手部关节，并保留 4 个坐标经零掩码处理的虚拟关节，以维持拓扑结构一致。对于提供 20 个身体关节的 N-UCLA，我们从手部中移除 THUMB_IP 关节，以匹配精简后的身体关节配置。

手部中心策略。 在我们的会议框架中，手部中心节点是相对于场景中活动最频繁的节点定义的，并用于相对坐标归一化，与身体预处理的做法相同。相比之下，在手部建模中，我们使用每只手的腕关节分别定义一个局部中心，从而避免大幅度的全局身体运动主导局部手部运动。这种设计在保留主模型所用共享相机坐标系的同时，减少了手部流中的明显运动模糊。

A.2. NTU-Hand 11/27 类别定义

NTU-Hand 基准由 NTU RGB+D 60/120 [26] 中以手部为中心的子集构成。我们选取了手部关节运动细微且起主导作用的动作，并定义了两个子集：

NTU-Hand 11。该子集包含 NTU RGB+D 60 中的 11 个以手部为中心的类别：拍手、阅读、书写、撕纸、打电话、玩手机/平板电脑、键盘打字、指向某物、自拍、看时间（查看手表）、双手互搓。

NTU-Hand 27。该子集在 NTU-Hand 11 的基础上，新增了 NTU RGB+D 120 中的 16 个以手部为中心的类别：竖大拇指，竖倒拇指，做 *OK* 手势，做胜利手势，

装订书本、数钱、剪指甲、剪纸、打响指、开瓶、嗅闻、蹲下、抛硬币、折纸、揉纸团、玩魔方。NTU-Hand 11 和 NTU-Hand 27 均采用原始 NTU RGB+D 60/120 跨主体和跨视角/设置评测协议进行评估，评估范围限定为选定的以手部为中心的类别。

B. 实现细节

B.1. 网络变体与分支配置

我们的双流架构基于 DeGCN 骨干网络 [22]，适用于身体与手部流。我们定义了三种纯骨架变体和一种多模态变体：

- **BHaRNet-B**：一种纯专家配置，包含身体专家分支和手部专家分支（身体专家分支、手部专家分支）。
- **BHaRNet-E**：一种专家+交互配置，结合专家分支（身体专家分支、手部专家分支）与身体/手部交互分支（身体交互分支、手部交互分支），以共享上下文。
- **BHaRNet-P**：一种参数高效配置，采用通道宽度较小的轻量级交互分支。
- **BHaRNet-M**：一种多模态变体，在 BHaRNet-E 的基础上扩展了 RGB 流，并采用 MMNet 风格的身体引导调制，如第 3.3 节所述。

B.2. 训练方案

除非另有说明，我们遵循 DeGCN 和我们的会议框架 [7] 的训练设置。对于骨架模型（BHaRNet-B/E/P），我们采用两阶段流程：(1) 分别使用各自的骨架输入对身体与手部流进行独立预训练；(2) 在启用所有损失的情况下，对完整的双流架构进行微调。（$\lambda_{\text{idv}}, \lambda_{\text{cpl}}, \lambda_{\text{nor}}$）损失权重在所有第二阶段实验中保持不变，推理采用确定性的 logit 求和，不包含任何随机成分或测试时增强。

完成 4 种骨架内模态的处理步骤后，我们为多模态模型（BHaRNet-M）引入第三阶段。在冻结预训练骨架主干网络的同时，我们使用骨架引导调制训练 RGB 流。

C. 补充定量结果

本节提供扩展的定量结果，以补充第5节中的分析。表C.1汇总并扩展了模态消融实验和变体消融实验（表4、5和6）。表C.2 综合扩展了对帧丢弃鲁棒性和新颖性细分的分析（表7 和8）。表C.3扩展了与基于骨架的基线方法的比较（表2）。这些通用化表格并未改变主要结论，但更全面地展示了我们的概率框架在不同数据集、模态和扰动设置下的表现。

D. 补充定性可视化

我们提供了定性样例作为补充视频材料（视频 S1）。该视频展示了我们的双流人体–手部框架的表现，包括各分支的预测（身体与手部流）及其集成结果，以及来自 NTU 数据集样本序列中相应的身体和手部关键点估计。每个关键点都渲染为一个节点，其半径根据局部关节运动幅度进行缩放，因此关节运动幅度较大的关节在可视化中会更加醒目。

CRediT 作者贡献声明

SeungyeonCho: 概念化、方法、软件、验证、形式分析、调查、数据整理、初稿书写、可视化、项目管理。**Tae-Kyun Kim:** 监督、资源、经费获取、审阅与编辑。

参考文献

[1] Abdelfattah, M., Hassan, M., Alahi, A., 2024年. Maskclr：注意力-引导式对比学习，用于鲁棒的动作表征学习，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第18678–18687页。[2] Abdelkawy, A., Ali, A., Farag, A., 2025年。Epam-net：一种用于视频动作识别的高效姿态驱动注意力引导多模态网络。Neurocomputing 633，129781。[3] Bruce, X.B.等，2022年。Mmnet：一种用于RGB-D视频中人体动作识别的基于模型的多模态网络。IEEE模式分析与机器智能汇刊 45，3522–3538。[4] Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W., 2021年。用于基于骨架的动作识别的通道级拓扑细化图卷积，载于：IEEE/CVF国际计算机视觉会议论文集，第13359–13368页。[5] Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H., 2020年。采用移位图卷积网络的基于骨架的动作识别，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第183–192页。[6] Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K., 2022年。InfoGCN：面向基于人体骨架的动作识别的表征学习，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第20186–20196页。[7] Cho, S., Kim, T.K., 2025年。采用交叉注意力的身体-手部模态专家网络，用于细粒度骨架动作识别，载于：2025年IEEE/RSJ智能机器人与系统国际会议（IROS），第11614–11621页。doi：10.1109/IROS60139.2025.11246883。[8] Ding, P., Cui, Q., Wang, H., Zhang, M., Liu, M., Wang, D., 2024年。三维全身人体运动的表现力预测，载于：AAAI人工智能会议论文集，第1537–1545页。[9] Du, Y., Wang, W., Wang, L., 2015年。用于基于骨架的动作识别的分层循环神经网络，载于：IEEE计算机视觉与模式识别会议论文集，第1110–1118页。[10] Duan, H., Xu, M., Shuai, B., Modolo, D., Tu, Z., Tighe, J., Bergamo, A., 2023年。Skeletr：迈向野外基于骨架的动作识别，载于：IEEE/CVF国际计算机视觉会议论文集，第13634–13644页。[11] Duan, H.等，2022年。重新审视基于骨架的动作识别，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第2969–2978页。

表 C.1在 NTU RGB+D 60/120 和 NTU-Hand 11/27 上，对基线与概率变体及模态配置（J+B、J+B+JM+BM、J+B+JM+BM+RGB）的完整消融实验。

方法	模态	NTU 60		NTU 120		NTU-Hand 11		NTU-Hand 27	
		跨主体	跨视角	跨主体	跨设置	跨主体	跨视角	跨主体	跨设置
BHaRNet-B†	J+B	96.1	98.7	94.0	94.9	95.4	97.9	90.6	91.8
BHaRNet-E†	J+B	96.2	98.8	94.3	95.0	95.6	97.9	90.9	92.0
BHaRNet-P†	J+B	96.3	98.8	94.3	95.2	95.4	97.8	90.8	92.2
我们的方法（BHaRNet-B）	J+B	96.4	99.1	94.3	95.2	95.6	98.2	91.3	92.7
我们的方法（BHaRNet-E）	J+B	96.6	99.0	94.5	95.4	95.6	98.2	91.3	92.8
我们的方法（BHaRNet-P）	J+B	96.7	99.1	94.5	95.5	95.8	98.1	91.4	92.9
我们的方法（BHaRNet-B）	J+B+关节运动+骨骼运动	96.6	99.1	94.5	95.5	95.7	98.4	91.4	93.0
我们的方法（BHaRNet-E）	J+B+关节运动+骨骼运动	96.7	99.2	94.7	95.7	95.7	98.5	91.4	93.1
我们的方法 (BHaRNet-P)	J+B+关节运动+骨骼运动	96.8	99.2	94.8	95.8	96.1	98.5	91.7	93.1
我们的方法 (BHaRNet-M)	J+B+关节运动+骨骼运动+RGB	97.0	99.4	95.5	96.5	96.1	98.8	92.4	94.1

表 C.2NTU RGB+D 120 和 NTU-Hand 27（跨主体、关节模态）上 canonical、免校准及 Noisy-OR 变体的帧丢失鲁棒性完整消融实验。

方法	NTU 120			NTU-Hand 27		
	0	0.25	0.5	0	0.25	0.5
BHaRNet-B†	92.7	90.1	89.8	88.6	88.6	88.0
+ 免校准	93.0	92.8	92.6	89.4	88.7	88.4
+ Noisy-OR损失	93.2	93.0	92.8	89.2	89.1	88.6
BHaRNet-E†	93.0	90.1	90.1	88.8	88.9	88.3
+ 免校准	93.3	93.2	93.0	89.4	88.8	88.9
+ Noisy-OR损失	93.5	93.3	93.1	89.5	89.3	89.1
BHaRNet-P†	92.9	91.0	90.9	88.7	89.0	88.3
+ 免校准	93.3	93.2	93.0	89.4	89.1	88.7
+ Noisy-OR损失	93.4	93.3	93.0	89.5	89.5	88.7

表 C.3在 NTU RGB+D 60/120 和 N-UCLA 上对基于骨架的方法进行扩展比较：准确率（%）、十亿次浮点运算和参数数量。“–”表示原始文献未报告相关结果，“*”表示使用公开代码复现的结果。

方法	NTU 60		NTU 120		N-UCLA	十亿次浮点运算	参数量（百万）
	跨主体	跨视角	跨主体	跨设置	跨视角		
CTR-GCN [4]	92.4	96.8	88.9	90.6	96.5	7.9	5.8
InfoGCN [6]	93.0	97.1	89.8	91.2	97.0	10.0*	9.4
PoseConv3D [11]	94.1	97.1	86.9	90.3	-	31.8	4.0
BlockGCN [38]	93.1	97.0	90.3	91.5	96.9	6.5	5.2
DeGCN [22]	93.6	97.4	91.0	92.1	97.2	6.9	5.6
3MFormer [31]	94.8	98.7	92.0	93.8	**97.8**	58.5	6.7
ProtoGCN [19]	93.8	97.8	90.9	92.2	-	43.4	24.9
SkeleT [36]	**97.0**	**99.6**	94.6	**96.4**	97.6	9.6	5.2
BHaRNet-B† [7]	96.1	98.7	94.0	94.9	94.6	**3.3**	**2.8**
BHaRNet-E† [7]	96.2	98.8	94.3	95.0	94.6	5.4	5.5
BHaRNet-P† [7]	96.3	98.8	94.3	95.2	95.3	3.4	4.9
BHaRNet-B‡	96.5	99.0	94.2	95.2	95.7	9.6	5.2
BHaRNet-E‡	96.6	99.0	94.4	95.5	95.5	**3.3**	**2.8**
BHaRNet-P‡	96.6	99.1	94.5	95.5	95.0	3.4	4.9
我们的方法 (BHaRNet-B)	96.6	99.1	94.5	95.5	95.9	6.6	5.5
我们的方法 (BHaRNet-E)	96.7	99.2	94.7	95.7	96.3	10.9	11.0
我们的方法 (BHaRNet-P)	96.8	99.2	**94.8**	95.8	95.9	6.8	9.7

[12] Hachiuma, R., Sato, F., Sekii, T., 2023. 通过结构化关键点池化实现的统一基于关键点的动作识别框架，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第22962–22971页。[13] Lee, H., Ryu, J., 2025. 通过规范域方法实现高效的三维人体姿态估计泛化。IEEE Access。[14] Lee, J., Lee, M., Lee, D., Lee, S., 2023. 用于基于骨架的动作识别的分层分解图卷积网络，载于：IEEE/CVF国际计算机视觉会议论文集，第10444–10453页。[15] Lee, J., Xu, W., Richard, A., Wei, S.E., Saito, S., Bai, S., Wang, T.L., Sung, M., Kim, T.K., Saragih, J., 2025. REWIND：通过基于样例的身份条件实现实时第一人称全身动作扩散，载于：计算机视觉与模式识别会议论文集，第7095–7104页。[16] Li, Y., He, Z., Ye, X., He, Z., Han, K., 2019. 用于基于骨架的动态手势识别的时空图卷积网络。EURASIP图像与视频处理期刊 2019，78。[17] Li, Z., Zhou, W., Zhao, W., Wu, K., Hu, H., Li, H., 2025. Uni-sign：迈向大规模统一手语理解，载于：第十三届国际学习表征会议。网址：https://openreview.net/forum?id=0Xt7uT04cQ。[18] Liu, C., Hu, Y., Li, Y., Song, S., Liu, J., 2017. PKU-MMD：用于基于骨架的人体动作理解的大规模基准，载于：智能互联社区视觉分析研讨会论文集，第1–8页。[19] Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y., Sun, Z., 2025. 揭示关键细节以辨别差异：一种用于基于骨架的动作识别的新型原型视角，载于：计算机视觉与模式识别会议论文集，第29248–29257页。[20] Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W., 2020. 解耦并统一用于基于骨架的动作识别的图卷积，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第143–152页。[21] Lugaresi, C., et al., 2019. MediaPipe：用于构建感知流水线的框架。arXiv预印本 arXiv:1906.08172。[22] Myung, W., et al., 2024. DeGCN：用于基于骨架的动作识别的可变形图卷积网络。IEEE图像处理汇刊 33，2477–2490。[23] Pang, Y., et al., 2022. IGFormer：用于基于骨架的人体交互识别的交互图变换器，载于：欧洲计算机视觉会议，Springer，第1–17页。[24] Pearl, J., 1988. 智能系统中的概率推理：可信推理网络。Morgan Kaufmann。[25] Reilly, D., Das, S., 2024. 只需添加？！用于理解日常生活活动的姿态诱导视频变换器，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第18340–18350页。[26] Shahroudy, A., et al., 2016. NTU RGB+D：用于三维人体活动分析的大规模数据集，载于：IEEE计算机视觉与模式识别会议论文集，第1010–1019页。[27] Shamil, M.S., Chatterjee, D., Sener, F., Ma, S., Yao, A., 2024. 关于三维手部姿态对动作识别的效用，载于：欧洲计算机视觉会议，Springer，第436–454页。[28] Shi, L., et al., 2019. 用于基于骨架的动作识别的双流自适应图卷积网络，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第12026–12035页。[29] Tian, J., Cheung, W., Glaser, N., Liu, Y.C., Kira, Z., 2019. UNO：用于意外输入退化的不确定性感知 noisy-or 多模态融合，载于：arXiv预印本 arXiv:1911.05611。[30] Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C., 2014. 跨视角动作建模、学习与识别，载于：IEEE计算机视觉与模式识别会议论文集，第2649–2656页。[31] Wang, L., Koniusz, P., 2023. 3MFormer：用于骨架动作识别的多阶多模态变换器，载于：IEEE/CVF计算机视觉与模式识别会议，第5620–5631页。[32] Wen, Y., Tang, Z., Pang, Y., Ding, B., Liu, M., 2023。交互式时空令牌注意力网络用于基于骨架的通用交互动作识别，载于：2023年IEEE/RSJ智能机器人与系统国际会议（IROS），IEEE。第7886–7892页。[33] Wu, J., Yu, Y., Huang, C., Yu, K., 2015。用于图像分类和自动标注的深度多实例学习，载于：IEEE计算机视觉与模式识别会议论文集（CVPR），第3460–3469页。[34] Xu, Y., Peng, K., Wen, D., Liu, R., Zheng, J., Chen, Y., Zhang, J., Roitberg, A., Yang, K., Stiefelhagen, R., 2024。带噪声标签的基于骨架的人体动作识别，载于：2024年IEEE/RSJ智能机器人与系统国际会议（IROS），第4716–4723页。doi：10.1109/IROS58592.2024.10801681。[35] Yan, S., Xiong, Y., Lin, D., 2018。用于基于骨架的动作识别的时空图卷积网络，载于：AAAI人工智能会议论文集。[36] Yang, Y., Zhang, J., Zhang, J., Tu, Z., 2024。通过骨架变换实现基于骨架的动作识别的表达性关键点。CoRR abs/2406.18011。URL：https://doi.org/10.48550/arXiv.2406.18011。[37] Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N., 2020。语义引导神经网络用于高效的基于骨架的人体动作识别，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第1112–1121页。[38] Zhou, Y., Yan, X., Cheng, Z.Q., Yan, Y., Dai, Q., Hua, X.S., 2024。BlockGCN：重新定义基于骨架的动作识别的拓扑感知，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第2049–2058页。



Seungyeon Cho 在大韩民国大田的韩国科学技术院（KAIST）获得理学学士和理学硕士学位。他的研究兴趣包括基于骨架的人体动作识别、身体与手部协调、可靠性感知学习、高效多模态建模，以及将机器学习进一步应用于光子学和逆向设计的光学器件。



Tae-Kyun(T-K) Kim自2020年起担任KAIST计算学院教授兼计算机视觉与学习实验室主任，并于2010年至2024年在英国伦敦帝国理工学院担任副教授。他于2008年获得剑桥大学博士学位，并于2007年至2010年担任剑桥大学悉尼·萨塞克斯学院初级研究员（学院管理机构成员）。他在KAIST获得理学学士和理学硕士学位。他的研究兴趣包括面向三维视觉和生成式人工智能的机器（深度）学习，并已在该领域顶级会议和期刊上合著发表100余篇学术论文。他曾担任在伦敦举办的BMVC17大会主席、BMVC23程序委员会共同主席，并担任TPAMI、《模式识别》期刊和《图像与视觉计算》期刊的副编辑。他定期担任顶级人工智能与视觉会议的领域主席。