arXiv:2301.00436v3 [cs.CV] 3 Apr 2023

面向视频动作识别的层次化解释

Sadaf Gulshad、Teng Long、Nanne vanNoord 阿
姆斯特丹大学{s.gulshad, t.long, n.j.e.vannoord}@uva.nl

摘要

解释深度神经网络的一种主要方法，是剖析视觉输入并找出导致分类结果的原型部分。然而，现有方法往往忽略这些原型之间的层级关系，因此无法同时解释较高层级（如水上运动）和较低层级（如游泳）的语义概念。受人类认知系统启发，本文利用层次化信息处理不确定性：当我们观察到水和人类活动，却无法确定具体动作时，可以将其识别为水上运动这一父类；只有观察到有人游泳后，才能将其明确细化为游泳动作。为此，我们提出层级原型解释器（*HIPE*），用于建立原型与类别之间的层级关系。*HIPE* 通过从类别层次结构的多个层级剖析输入视频帧，为视频动作分类提供推理过程；我们的方法也适用于其他视频任务。通过在 *ActivityNet* 和 *UCF-101* 上提供多层级解释并降低准确性与可解释性之间的权衡，我们验证了方法的忠实性。

1. 引言

在描述我们周围的世界时，我们可以根据可用信息或希望传达的细节程度，采用不同的粒度。例如，一段视频可能以欢呼的人群画面开场，让我们识别出这是一场体育赛事；随后镜头转向河流，我们便能推断这是一场水上运动赛事。然而，只有当木筏进入画面时，我们才能确定这段视频讲的是漂流。尽管如此，在描述这段视频时，我们仍可能只称其为体育赛事或水上运动赛事。我们的推理和描述过程建立在类别之间的层级关系之上，使我们能够在通用概念与具体概念之间切换。在这项工作中，我们通过学习层级概念来实现视频动作识别，并利用这些概念提升分类性能，同时生成多个粒度层次的解释，见图1。

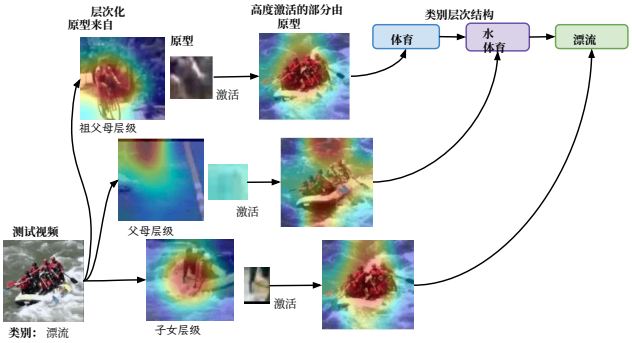


图1. 我们的方法通过学习不同层级的原型来生成解释：在祖父类层级，原型属于体育类别；在父类层级，原型属于水上运动类别；在类别层级，原型属于漂流类别。

尽管神经网络在视频理解任务中表现出色 [15, 18, 38, 42, 41, 54, 6, 4]，但要解释这些网络的决策仍然很困难，而这对实际应用至关重要。为此，越来越多的研究致力于在确保模型准确性的同时提升其可解释性 [16, 21, 14, 48, 25]。一种颇具前景的事后可解释性方法是概念瓶颈模型[22, 29, 20],，它通过在做出最终预测之前预测人类可理解的概念，来解释神经网络的决策。然而，这些方法需要密集的概念标注，或借助预训练语言模型等其他资源 [56, 36]。我们的方法能够提供内置解释，无需密集标注。

类似地，基于原型的模型 [24, 8, 10]着重于在训练期间学习原型，并在推理期间基于学到的原型进行预测。这使其能够提供这个看起来像那个式的解释。然而，先前基于案例的推理研究仅限于二维图像和模型。此外，它们只提供单一层级的解释；在存在不确定性的情况下，解释可能会非常随意，因为所有解释都被视为彼此同等相似。在这项工作中，我们着重捕捉动作之间的层级关系，以提供多层级

视频解释。

可解释模型（如概念瓶颈模型和基于原型的模型）面临的一项挑战是，它们引入了准确性与可解释性之间的权衡：可解释性是以准确率为代价的。本文旨在提出一种内置可解释性的模型，使其受这种权衡的影响较小。为实现这一目标，我们受到近期自然语言处理领域中关于学习双曲嵌入空间（而非欧几里得嵌入空间）的研究启发 [47, 9] ，以及计算机视觉任务中的相关研究 [1, 13, 28]。这些研究表明，以层次化知识引导嵌入空间有助于提升性能；我们认为，这也将有益于可解释性。这一观点源于人类的层次化认知过程：我们倾向于按照从具体到一般的顺序组织概念 [55, 32, 31], , 也源于类别在双曲空间中的表示方式。

本文的主要贡献如下：1）我们提出了层级原型解释器（HIPE），这是一种用于解释视频动作识别的推理模型。2）我们证明，即使存在不确定性或信息缺失，HIPE 也能通过提供多层级解释（即类别、父类或祖父类层级的解释）给出有意义的说明。3）我们进行了基准测试，并表明 HIPE 能够解决准确性与可解释性之间的权衡。

2. 相关工作

2.1. 视频的解释

神经网络的解释大致可分为两类：1）在网络训练完成后，针对其决策拟合解释，即事后解释 [16, 21, 14, 39, 30], 2）；2）构建内在于网络的解释机制，即内置 解释。[48, 25, 26, 49]。在这项工作中，我们专注于学习用于训练期间分类的语义表示，而不是对黑箱网络进行事后解释。

大量已有工作聚焦于视频动作识别、检测、分割等任务 [15, 18,38, 42, 41, 54, 6, 4], 然而，这些工作大多致力于为特定任务设计黑盒模型，并未解释模型为何做出某个决定。此外，视觉解释领域的大多数研究都聚焦于图像。只有少数工作关注如何解释用于视频的这些网络 [17, 2, 45, 25, 44],，而由于视频多了一个时间维度，基于图像的解释方法无法直接应用于视频。

[17] 和 [2] 专注于可视化循环神经网络中的时空注意力，而卷积神经网络仅用于提取特征。受图像类别激活图（CAM） [57] 的启发， [45]将其扩展到视频，通过找出对分类起作用的区域和帧。 [25]利用扰动来提取输入中对输出起决定作用、信息量最大的部分。 [45, 25] 都是事后方法，这意味着它们在预测过程中不会使用解释，因此可能无法忠实反映网络的计算过程 [10]。 [44] 提出了类别特征金字塔，这种方法遍历整个网络，搜索网络不同深度中对分类起作用的卷积核，因此计算开销很大。与之相反，我们实现了内置的多层级解释，不会增加任何计算复杂度。本文通过为每个类别学习层次原型，并将其追溯回输入视频，实现了视频的多层级解释。

2.2. 基于案例的推理模型

基于案例的推理模型主要分为两类：概念瓶颈模型，这类模型引入瓶颈层来学习人类可理解的概念；以及基于原型的模型，这类模型学习与训练集样本更接近的原型。概念瓶颈模型通过将神经网络的最终层替换为预测人类可理解概念的层，提供事后解释例如，对于红雀类别，概念可以是红色翅膀、红色喙和黑色眼睛 [22, 29, 20, 52]。随后，利用这些预测出的概念执行分类。然而，概念瓶颈模型需要密集概念标注来学习这些概念。此外，由于它们使用预测出的概念进行分类，因此面临可解释性与准确率的权衡。

[56, 36] 致力于解决这些局限，方法是将自然语言描述中的概念迁移过来，或通过 GPT 模型生成概念。相比之下，我们的工作通过为每个类别及其（祖）父类学习代表性样本，在兼顾可解释性与准确率的同时提供内置解释。我们的方法不需要大量标注，而是利用数据集中已有的层次结构，或者根据类别之间的关系轻松定义层次结构。

与我们的方法更相关的是， [52] 建立了神经元与层级概念之间的关系。不过，我们的方法与 [52]有三点显著区别：1）他们的方法属于事后解释（需单独训练一个概念分类器），而我们的方法是内置的；2）他们利用 ImageNet 的 WordNet 层次结构来学习层级概念，而我们没有；3）他们考虑神经元的贡献（人类难以理解），而我们通过原型层可视化学到的特征。

基于原型的模型旨在通过原型提供内置解释，这一想法最早在[24], 中得到探索，作者在具有编码器-解码器架构的网络中引入了一个原型层。该原型层存储与编码后的训练样本接近的权重——

样本，并使用解码器将其可视化。然而，他们的模型无法为自然图像生成逼真的可视化结果。因此 [8] 提出了为每个类别学习原型，并通过将其追溯至输入图像来进行可视化，而不使用解码器。我们从 [8]提供内置解释的思路中汲取灵感，但他们的工作仅限于二维图像，且只提供单层级解释；我们将其扩展为面向视频的多层级解释。

[40] 致力于通过在不同类别之间寻找共享原型，减少每个类别的原型数量。 [53]提出了一种不同的相似度度量，用于计算原型与图像块之间的相似度。Deformable ProtoPNet [10] 学习空间上灵活的原型，以捕捉输入中的姿态和上下文变化。此前所有基于原型的解释方法，都未考虑类别之间的层级关系，而是在定义明确的图像数据集 CUB 鸟类数据集 [51]和 Stanford Cars 数据集 [23] 上提供解释。相比之下，受人类解释方式的启发，我们在面向视频数据集学习各类别原型的同时，也考虑类别之间的层级关系。

与我们工作最密切相关的是 [48, 33]。 [48] 提出了动态原型网络（DPNet），用于发现深度伪造视频中的时间伪影和非自然运动。然而，DPNet 的目标与我们的不同；它只有两个目标类别——伪造/真实，因此任务较为简单。PrototypeTrees [33]采用具有预定义结构的决策树，并在每个决策节点学习独立的原型。这些原型经过优化，以提高沿树路径的纯度。然而，对于 PrototypeTrees，节点在树中的位置和顺序并不表示层级关系；也就是说，越接近根节点并不意味着语义层级越一般。此外，由于原型数量取决于树的大小，学习 ProtoTree 的计算开销很大。我们提出的多层级解释遵循非常明确的语义分布，其中（祖）父类原型更加通用，且不会增加任何额外的计算复杂度。

2.3. 双曲嵌入

近年来，双曲嵌入受到越来越多的关注，因为它们能够以连续表示的形式呈现层级知识 [34, 7]。这种连续性使得父类（及祖父类）的信息能够被隐含地——因此被纳入其中，使双曲训练仍保持单标签。它们在文本[47, 12, 58] 和视觉数据 [19, 1, 13, 28]中的有效性也已得到证明。双曲嵌入还被用于零样本学习 [27, 11]以及视频动作识别 [28, 46]。视频之间的层级关系，以及以层次化方式向人类解释决策的需求，都促使我们采用双曲空间。在此，我们利用双曲嵌入学习层次原型，为视频动作识别提供符合人类认知的

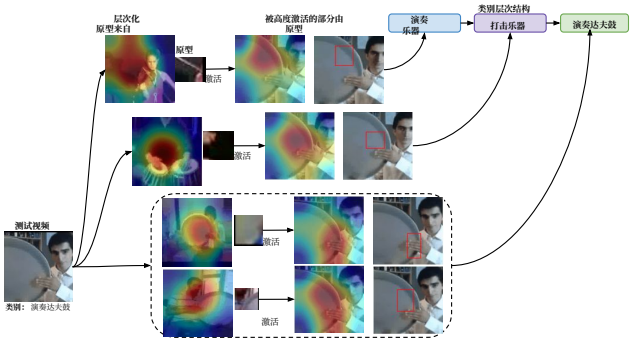


图 2。**层次化解释的重要性**。针对演奏达夫鼓的输入视频生成的层次化解释。即使有人不知道什么是“演奏达夫鼓”，其父类解释也表明这涉及“打击乐器”，祖父类解释则表明它是一种音乐“乐器”。

视频动作识别的解释。

3. 层次化解释

单层解释并不总是足以解释深度神经网络的决策，例如在图2中，单层解释只能说明预测为“演奏达夫鼓”的依据。然而，不熟悉达夫鼓的人可能无法理解这一解释。相反，当我们在更抽象的层次，即父类和祖父类层次，提供解释时，通过查看父类的解释，可以推断出这涉及“打击乐器”，而从祖父类可以得知它属于某种音乐“乐器”。我们提出将这种基于层次先验知识提供解释的强大机制融入用于视频动作识别的网络中。

3.1. 层级原型解释器

图 3概述了我们为视频动作识别提出的层次化原型解释器（HIPE）。HIPE 由一个 3D 卷积神经网络骨干网络 f 组成，用于从视频帧中提取特征；还包含一个层次原型层 g_p，用于为每帧学习原型。原型层后接一个全连接层 h，用于整合原型相似度分数，并通过指数映射将其映射到共享双曲空间中。我们将以动作层次结构形式表示的动作间关系先验知识，通过判别式嵌入投影到共享空间中。随后，我们采用双曲学习，获得能够实现多层次可解释性的层次原型。

作为骨干网络架构，我们使用视频动作分类网络 3D-Resnet [15]。对于每个输入视频$v \in \mathbb{R}^{W \times H \times T \times 3}$，包含 T 帧，它会提取视频特征$Z \in \mathbb{R}^{W_0 \times H_0 \times T_0 \times D}$，其空间分辨率为 $W_0 \times H_0$，帧数为 T_0，通道数为 D。该骨干网络的一个关键特点是

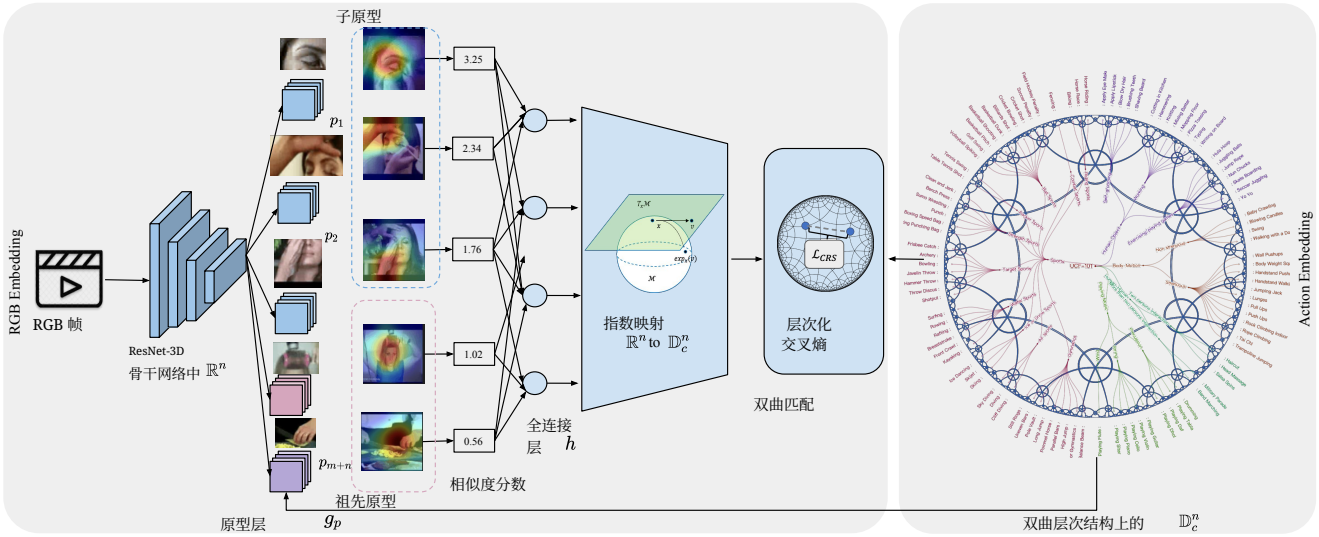


图 3. **层级原型解释器概览**。ResNet-3D 骨干网络提取视频特征，原型层为子类和父类学习原型，然后通过最大池化将这些原型转换为单个相似度分数。最后，分数经全连接层处理后，通过指数映射从 $\mathbb{R}^n$ 转换为 $\mathbb{D}^n$，映射到共享双曲空间，以进行层次化学习。通过在 $\mathbb{D}^n$上学习判别式嵌入，将动作映射到共享双曲空间。

这是因为 $T_0 < T$时间池化，因此特征Z 是针对片段而非单独帧提取的。由于采用了时间池化，HIPE 学到的原型具有时空特性，因此能够解释片段中的哪些部分表明视频中发生了该动作。

3.2. 层次原型层

给定从 3D-Resnet Z中提取的特征，我们添加两层带有 LeakyReLU 激活的 $1 \times 1 \times 1$ 卷积，以调整顶层的通道数。通过将动作的先验层级知识表示为嵌入并融入网络，学习层次原型。我们通过在双曲空间中使层次原型与动作嵌入对齐来优化它们。

给定动作类别集合 $\mathcal{A} = \{1, 2, ..., |\mathcal{A}|\}$，在层次化动作识别中，我们还会考虑它们的祖先类别 $\mathcal{H} = \{|\mathcal{A}|+1, |\mathcal{A}|+2, ..., |\mathcal{A}|+|\mathcal{H}|\}$，从而构建一个包含三个层级的层次树，即祖父类、父类和子类（见图3 右侧）。只需考虑类别之间的关系，就能轻松定义这些层次结构，无需对单个实例进行标注。每个数据集的层次结构嵌入过程只需离线执行一次。不过，也可以轻松地针对其他层次结构重复这一过程。更多细节请参阅补充材料。

对于每个子类 $\mathcal{A}$ 及其父类 $\mathcal{H}$ 动作，网络分别学习 m 和 n 原型 $P = \{p_j\}_{j=1}^{m+n}$，其形状为 $W_1 \times H_1 \times T_1 \times D$，包含 $W_1 \leq W_0$、$H_1 \leq H_0$ 和 $T_1 \leq T_0$。因此，每个原型都表示视频中的一个时空部分。给定卷积输出 $Z = f(v)$ 和原型 p，原型层 g_p会计算每个原型 p_j 与来自 Z 的图像块之间的距离，并使用以下方式将其转换为相似度分数

$$g_p(p_j, Z) = \max_{z \in Z} \log \frac{(||z - p_j||_2^2 + 1)}{(||z - p_j||_2^2 + \epsilon)}, \epsilon > 0 \quad (1)$$

每个原型与图像块之间的距离决定了该原型在输入中存在的程度。我们预期会分别为子类、父类及其祖父类学习不同的原型。例如，在图 3中，为 涂眼妆 子类学习到的原型是一只眼睛；对于父类，由于其类别是 自我修饰，学习到的原型聚焦于头发；而对于祖父类，原型则呈现人类与物体交互 类别中的手。随后，我们将相似度分数与全连接层的权重相乘 h，以获得用于投影到双曲联合空间中的嵌入，从而学习层次化原型。

3.3. 分层视频嵌入

从全连接层得到的嵌入 $\mathbf{h} = h(g_p(p, f(v)))$ 位于欧几里得空间中，无法直接映射到双曲嵌入空间，因此，我们使用指数映射 [12] 将视频嵌入映射到双曲空间中。

$$\exp_{\mathbf{x}}(\mathbf{h}) = \mathbf{x} \oplus \left(\tanh\left(\frac{||\mathbf{h}||}{1 - \|\mathbf{x}\|^2}\right) \frac{\mathbf{h}}{||\mathbf{h}||}\right) \quad (2)$$

其中，$\oplus$ 表示 1–曲线莫比乌斯加法，$\mathbf{x}$ 是连接切空间 $\mathcal{T}_0\mathbb{D}^n$ 与 $\mathbb{D}^n$的切点。x 的不同取值会导致不同的切空间；为避免任何

歧义，我们设定$\mathbf{x} = \mathbf{0}$，并将视频嵌入投影到双曲空间，以便与层级动作进行匹配。

3.4. 训练

我们的训练过程包含多个步骤：在最初的训练轮次中，我们对新添加的层进行预热。预热完成后，我们对整个网络进行端到端训练。每经过 10 个训练轮次，我们仅更新原型层，随后对原型层之后的各层进行微调。

双曲空间中的视频与动作匹配

我们的目标是学习一个潜在空间，使对分类至关重要的图像块聚集在相似的原型周围。为了学习层次原型，我们优化原型 $P = \{p_j\}_{j=1}^{m+n}$，使视频与双曲空间中的动作嵌入相匹配，因此我们的优化由 $\mathbf{\Phi} \in \mathbb{D}^{n\times(|\mathcal{A}|)}$ 监督。令 $\{(v_i,\ y_i)\}_{i=1}^{N}$ 表示训练集，其中 $v \in \mathbb{R}^{W\times H\times T\times 3}$ 和 $y_i \in \mathcal{A}$。我们的目标是求解：

$$\mathcal{L}_{crs} + \lambda_1\mathcal{L}_{cls} + \lambda_2\mathcal{L}_{sep} \tag{3}$$

层级交叉熵。 我们的损失函数中的第一项是层级交叉熵损失 $\mathcal{L}_{crs}$，它会对误分类进行惩罚，其定义如下：

$$\mathcal{L}_{crs} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log p(y=k|v) \tag{4}$$

交叉熵中的 Softmax 定义为双曲空间中视频嵌入与层次化动作嵌入之间的负距离：

$$p(y=k|v) = \frac{\exp(-d(\mathbf{h}_e), \mathbf{\Phi}_k)}{\sum_{k'}\exp(-d(\mathbf{h}_e), \mathbf{\Phi}_k))}, \tag{5}$$

其中 $\mathbf{h}_e = \exp\mathbf{0}(\mathbf{h})$ 表示对全连接层输出$\mathbf{h}$应用指数映射。

层级聚类。 为了在层次结构的每个层级上提供有意义的解释，我们的层级聚类损失鼓励输入帧中的特征至少有一个图像块更接近子类、父类或祖父类原型。

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=1}^{N}\min_{j:p_j\in P_{|\mathcal{A}|+|\mathcal{H}|}}\min_{z\in patches(f(v_i))}||z-p_j||_2^2 \tag{6}$$

层级分离。 我们的层级分离损失鼓励帧的潜在图像块远离不属于同一子类、父类或祖父类的原型。

$$\mathcal{L}_{sep} = -\frac{1}{N}\sum_{i=1}^{N}\min_{j:p_j\notin P_{|\mathcal{A}|+|\mathcal{H}|}}\min_{z\in patches(f(v_i))}||z-p_j||_2^2 \tag{7}$$

3.5. 更新层级原型层

我们将原型投影到训练视频中最近的视频特征上。我们针对子类、父类和祖父类动作类别执行此操作。数学上，对于来自子类、父类和祖父类的原型 p_j，即 $p_j \in P_{|\mathcal{A}|+|\mathcal{H}|}$，我们将原型层更新为：

$$p_j \leftarrow \underset{z\in\mathcal{Z}_j}{\operatorname{argmin}}||z-p_j||_2 \tag{8}$$

其中 $\mathcal{Z}_j = \{\tilde{z}:\ \tilde{z} \in patches(f(v_i))\ \forall i \text{ s.t. } y_i = |\mathcal{A}| + |\mathcal{H}|\}$。我们的原型层不仅使用属于子类的原型进行更新，也使用父类及（祖）父类的原型，从而能够学习类别之间的层次关系。

3.6. 层次原型可视化

为构建可视化结果，需要将学习到的原型映射到时空输入空间中。我们通过将输入 v 输入网络，并对原型层 $g_p(p_j,\ Z)$生成的激活图进行空间和时间上的上采样（针对视频），从而选取对原型 p_j 激活程度最高的图像块。我们分别可视化子类、父类和祖父类的 p_j，为层次结构的所有层级提供解释。

4. 实验设置

4.1. 数据集

为了评估 HIPE 在视频上的表现，我们在两个视频数据集上进行实验：UCF-101 [43] 和 Activity-Net1。3 [5]。

层次化 UCF-101。 UCF-101 [43]包含 13,320 个视频，属于 101 个动作类别，总时长为 27 小时。我们定义了两个额外的层次结构级别，一级、二级和三级的类别数量分别为 5、20 和 101。层次结构第三级的类别是该数据集原有的 101 个类别。完整的层次结构收录于补充材料中。

层次化 ActivityNet。 ActivityNet [5] 包含 14,950 个未剪辑视频，每个视频由一个或多个动作片段组成，这些动作片段属于 200 个动作类别，总时长约为 648 小时。我们使用 10,024 个视频进行训练，使用 4,926 个视频进行验证。我们遵循 [15]，在剪辑后的视频上训练和测试模型，以确定视频级准确率。我们遵循 [28]，使用数据集自带的层次结构来定义类别级动作层次结构。该数据集在一级、二级和三级分别包含 200、38 和 6 个类别。

4.2. 实现细节

层次化动作嵌入是通过使用黎曼 Adam 优化器 [3]进行训练生成的，该优化器由*geoopt* 和 PyTorch [37]实现。除此之外，

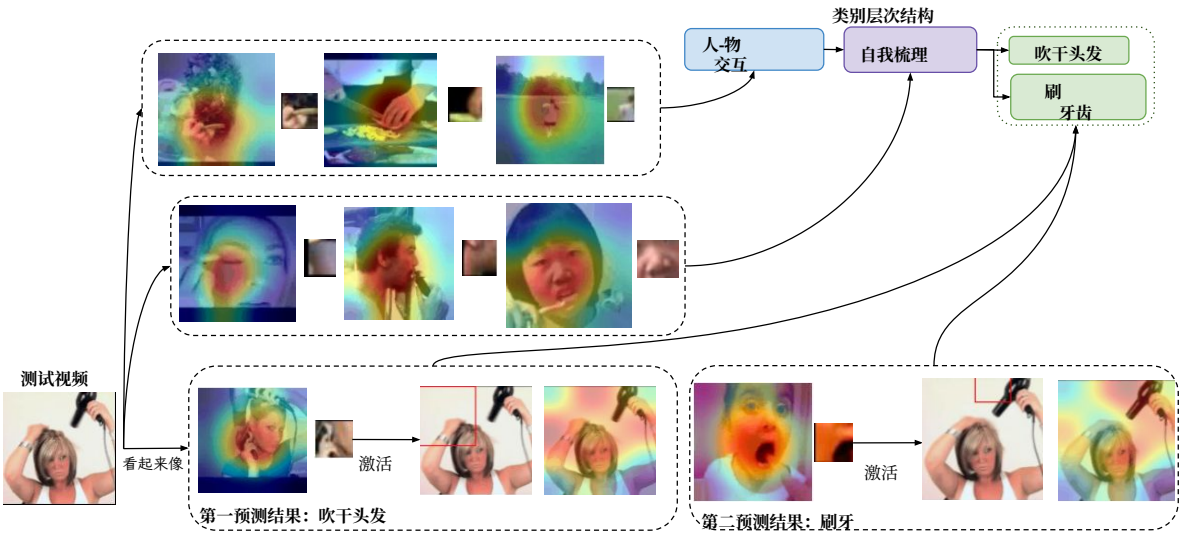


图4. **层次化解释**。 此示例展示了祖父类人-物交互类别、父类自我仪容整理类别和真实标签吹干头发类别中的原型。我们还观察到，模型的第二预测结果是其兄弟类别刷牙。这表明我们的模型正在学习类别之间的层级关系。

生成层次化动作嵌入这一离线的一次性步骤外，HIPE 以端到端的方式进行训练。对于特征提取，我们使用了在 Kinetics 上预训练的 ResNet-3D-18 [15] [6]，并添加了两个 1×1 卷积层（采用 LeakyReLU）、一个原型层以及最后的嵌入层。我们每 10 个训练轮次进行一次原型投影和可视化。

我们报告了 HIPE 两种变体上的结果：比较每个类别使用 5 个原型和 10 个原型进行原型投影的效果，以探究这种额外监督是否能够减少所需的原型数量。作为对比，我们将 ProtoPNet [8] 适配到视频数据，将 2D ResNet 骨干网络替换为 3D ResNet。

评估指标。我们报告片段级和视频级的性能。片段级准确率是每个片段预测正确的比率，而视频级准确率则是对视频中所有片段的预测进行多数投票所得的结果。此外，为了展示层次化学习的优势，我们报告了三项指标的准确率：类别准确率是在与真实标签相距 0 跳的层次空间中预测正确的比率；兄弟类别准确率是在与真实标签相距 2 跳时预测正确的比率；表亲类别准确率则是相距 4 跳时预测正确的比率。兄弟类别和堂表类别指标的表现越好，说明误分类结果在层次结构中越接近真实类别，因此语义上也越相似。

5. 视觉解释

层次化解释。图4展示了 HIPE 提供的多层级解释示例。我们的模型学习将视频片段特征表示为层次原型，这些原型分别属于祖父类、父类和子类。例如，在图4 中，我们的模型从祖父类人-物交互、父类 自我梳理和动作类别吹干头发 中学习到了原型（为便于展示，图中仅显示了十个原型中的三个）。图中还展示了其中一个原型及其在原始视频上的激活情况。我们还观察到，第二个最有可能的预测是其兄弟姐妹类别刷牙。

时空解释。图5 展示了一个视频的解释，该视频包含六个片段，每个片段有 16 帧。顶行展示了视频不同片段中的单帧。在第二行中，我们可以看到，模型未能正确预测第三和第四个片段的类别。底行展示了原型所来源的训练视频帧（为便于查看，每个片段展示一个原型），以及分类所依据的内容。我们可以看到，当网络关注自行车轮胎或车把等更具体的概念时，就能正确分类。而当网络关注草地和栅栏杆，并在原始片段中激活骑自行车者背景中的栏杆时，就会被误分类为骑马类别。类似地，当网络关注田野中的草地，并激活原始片段中骑自行车者背景里的绿植时，就会被误分类为赛马类别。所有片段的父类和祖父类仍然相同，即骑乘运动和体育，即使发生误分类，这些类别仍能提供有用的信息。

层次化解释在失败情况下的有效性。在图6 中，我们展示了另一个多层级解释发挥作用的场景。我们可以看到，原始的

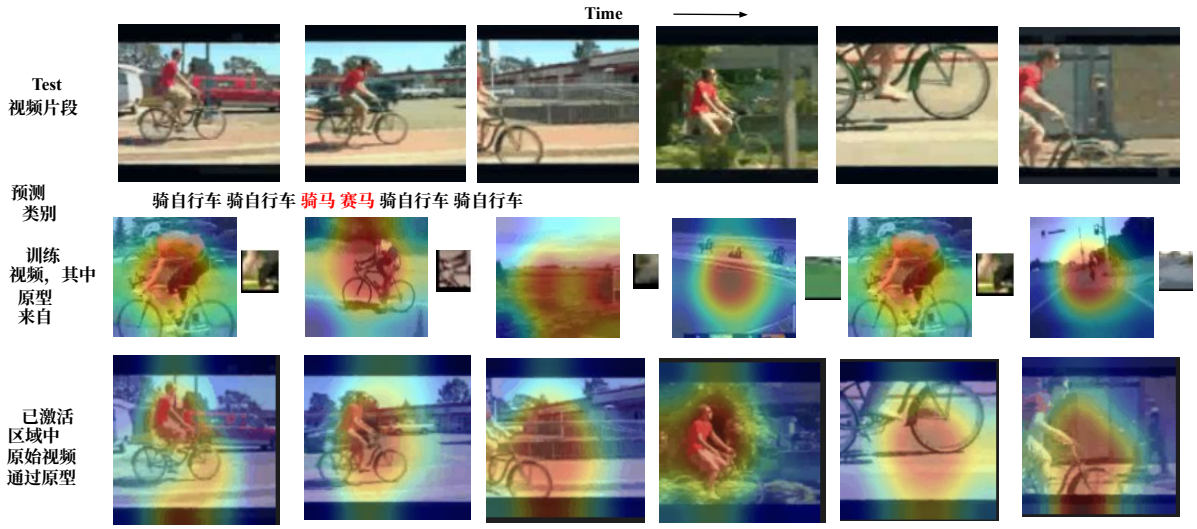


图5. **时空解释**。 包含六个片段的视频解释。顶行：每个片段的一帧。第二行：类别预测。底行：原型来源以及分类所依据的训练视频帧。当网络关注草地和栅栏杆等更抽象的概念时，视频会被误分类为骑马 类别。同样，当网络关注田野中的草地时，视频会被误分类为赛马 类别。而当网络关注自行车轮胎或车把等更具体的概念时，则能正确分类。

	网络	准确率	兄弟类别准确率	表亲类别准确率	原型数量 每个类别
不可解释 模型	3D-Resnet [15]	83.34	89.73	93.62	-
	双曲 ResNet [28]	82.64	89.99	93.28	-
可解释 模型	ProtoPNet [8]	78.30	85.92	90.98	10
	HIPE	79.45	88.88	92.73	5
	HIPE	**80.40**	**89.30**	**93.02**	10

表1. **UCF-101 视频上不同模型的片段级准确率比较**。我们观察到，HIPE 每类使用 10 个原型时，在提供多层级解释的同时，显著弥补了准确性与可解释性之间权衡造成的准确率下降。

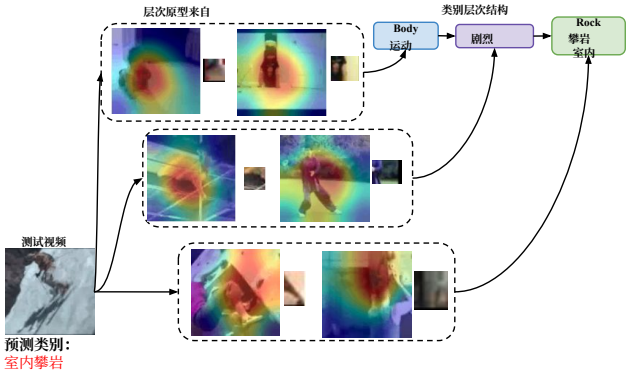


图6. **失败情况下的有效性**。即使发生分类错误，我们的多层级解释也能通过为父类和祖父类学习的原型提供有用信息。

滑雪视频被误分类为室内攀岩类别。然而，对于更抽象的解释，我们可以观察到，它的父类剧烈运动 和祖父类身体动作 都被正确识别。因此，我们的层次化解释即使在发生误分类时，也能为我们提供有用的信息。

6. 准确性与可解释性之间的权衡

为了说明 HIPE 能够降低准确性与可解释性之间的权衡，我们在此对比了我们的模型与可解释基线和不可解释基线的定量表现。

不可解释模型。 不可解释模型在 UCF-101 和 ActivityNet 上的表现见表格前两行 1、2和3。为公平比较这两种不可解释模型，普通 ResNet [15] 和双曲 ResNet [28], 均采用相同的数据增强，并以相同的训练轮次进行端到端训练。这两种不可解释模型的唯一区别在于，ResNet 模型通过欧几里得超平面分隔类别，而双曲 ResNet 则利用双曲嵌入空间分隔类别。我们在 UCF-101 上的结果表明，

	网络	准确率	兄弟类别准确率	表亲类别准确率	原型数量 每个类别
不可解释模型	3D-Resnet [15]	49.99	51.59	63.88	-
	双曲 Resnet [28]	49.95	52.03	65.44	-
可解释模型	ProtoPNet [8]	46.06	47.74	61.67	10
	HIPE	45.67	48.72	62.84	5
	HIPE	**46.26**	**48.93**	**62.97**	10

表2. **不同网络在 ActivityNet 视频上的片段级准确率比较**。我们观察到，每个类别使用 10 个原型的 HIPE 弥补了同级类和近邻类上的性能下降，在类别准确率方面表现与常规 ProtoPNet 相当，同时还能提供多层级解释。

	网络	UCF-101 准确率	ActivityNet 准确率	原型数量 每个类别
不可解释模型	3D-Resnet [15]	87.92	69.15	-
	双曲 ResNet [28]	87.15	70.19	-
可解释模型	ProtoPNet [8]	84.48	66.46	10
	HIPE	84.87	63.82	5
	HIPE	**86.22**	**66.48**	10

表3。**在 UCF-101 和 ActivityNet 视频上，不同模型的视频级准确率对比**。我们观察到，对于 UCF-101，采用每类 10 个原型的 HIPE 显著弥补了视频级准确率的下降；对于 ActivityNet，其性能与常规 ProtoPNet 相当，同时还能提供多层级解释。

普通 ResNet 和双曲 ResNet 在片段级（表 1）和视频级（表3）的表现相近。

对于 ActivityNet，片段级类别准确率（表2）在两个网络之间相当。然而，由于双曲 ResNet 采用了层次化学习，其兄弟姐妹类别准确率和表亲类别准确率更高；此外，它在视频级类别准确率方面也有所提升（见表3）。总体而言，在 UCF-101 上，非可解释网络的性能相当；而在 ActivityNet 上，双曲网络的性能有所提升。

可解释模型。可解释模型在 UCF-101 和 ActivityNet 上的表现如表1、2和3的倒数三行所示。我们报告了适配视频的常规 ProtoPNet [8] 以及 HIPE 各种变体的结果。

对于 UCF-101，使用每个类别有 10 个原型的常规 ProtoPNet 时，准确率大幅下降：片段级类别准确率降至 78.30，视频级准确率降至 84.48。这是因为可解释人工智能中普遍存在可解释性与准确率的权衡，[8] 中也报告了这一现象。相比之下，每个类别有 5 个原型的 HIPE 受影响小得多，并弥补了准确率下降：类别准确率提高 1.15，兄弟类别准确率和表亲类别准确率分别提高 2.96 和 1.75。将原型数量增加到每个类别 10 个后，类别准确率、兄弟类别准确率和表亲类别准确率分别进一步提高 2.10、3.38 和 2.04。此外，视频级表现达到 86.22（见表3）。因此，我们提出的 HIPE 的两种变体都减小了准确率降幅。

在 ActivityNet 上（见表2），我们观察到明显的对于常规 ProtoPNet，准确性与可解释性之间存在权衡，片段级和视频级准确率均有所下降。然而，尽管 HIPE 的类别准确率也出现了类似下降，但我们观察到，它在兄弟姐妹和堂表类别指标上部分弥补了这一降幅。对于采用 5 个原型的 HIPE 和每个类别采用 10 个原型的变体，这一表现都成立；后者在兄弟姐妹和堂表类别指标上甚至分别提升了 1.19 和 1.3 。尽管 ActivityNet 仍然具有挑战性，但兄弟姐妹和堂表类别准确率的提升直接有利于可解释性，如第 5 节所示。

因此，我们可以观察到，在两个数据集上，HIPE 受准确性与可解释性之间的权衡影响较小，同时还能提供多层级解释。

7. 结论

在这项工作中，我们提出了用于视频动作识别的层次化原型解释器。通过学习层次化原型，我们能够提供多个粒度层级的解释，不仅说明为何视频被分类为某个类别，还能指出哪些时空部分促使其归属于父类别。结果表明，HIPE在UCF-101上的表现优于先前的非层次化方法，同时在ActivityNet上的表现同样出色。此外，我们展示了多层级解释，使人们能够了解哪些时空部分有助于祖父类、父类和类别级分类。因此，我们的层次化方法在提供更丰富解释的同时，能以较小的性能损失换取可解释性。

参考文献

[1] Mina Ghadimi Atigh、Julian Schoep、Erman Acar、Nanne van Noord和Pascal Mettes。《双曲图像分割》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第4453–4462页，2022年。2, 3[2] Sarah Adel Bargal、Andrea Zunino、Donghyun Kim、Jian- ming Zhang、Vittorio Murino和Stan Sclaroff。《循环神经网络激励反向传播》。载于《IEEE计算机视觉与模式识别会议论文集》，第1440–1449页，2018年。2[3] Gary B´ecigneul和Octavian-Eugen Ganea。《黎曼自适应优化方法》。arXiv预印本 arXiv:1810.00760，2018年。5[4] Luca Bertinetto、Jack Valmadre、Joao F Henriques、Andrea Vedaldi和Philip HS Torr。《用于目标跟踪的全卷积孪生网络》。载于《欧洲计算机视觉会议论文集》，第850–865页。施普林格，2016年。1, 2[5] Fabian Caba Heilbron、Victor Escorcia、Bernard Ghanem和Juan Carlos Niebles。《ActivityNet：用于人体活动理解的大规模视频基准》。载于《IEEE计算机视觉与模式识别会议论文集》，第961–970页，2015年。5, 12[6] Joao Carreira和Andrew Zisserman。《Quo vadis，动作识别？一种新模型与Kinetics数据集》。载于《IEEE计算机视觉与模式识别会议论文集》，第6299–6308页，2017年。1, 2, 6[7] Ines Chami、Zhitao Ying、Christopher R´e和Jure Leskovec。《双曲图卷积神经网络》。《神经信息处理系统进展》，第32卷，2019年。3[8] Chaofan Chen、Oscar Li、Daniel Tao、Alina Barnett、Cynthia Rudin和Jonathan K Su。《这看起来像那个：用于可解释图像识别的深度学习》。《神经信息处理系统进展》，第32卷，2019年。1, 3, 6, 7, 8[9] Bhuwan Dhingra、Christopher J Shallue、Mohammad Norouzi、Andrew M Dai和George E Dahl。《在双曲空间中嵌入文本》。arXiv预印本 arXiv:1806 04313，2018年。2[10] Jon Donnelly、Alina Jade Barnett和Chaofan Chen。《可变形ProtoPNet：一种使用可变形原型的可解释图像分类器》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10265–10275页，2022年。1, 2, 3[11] Pengfei Fang、Mehrtash Harandi和Lars Petersson。《双曲空间中的核方法》。载于《IEEE/CVF 计算机视觉国际会议论文集》，第10665–10674页，2021年。3[12] Octavian Ganea、Gary B´ecigneul和Thomas Hofmann。《用于学习层次化嵌入的双曲蕴含锥》。载于《国际机器学习会议论文集》，第1646–1655页。PMLR，2018年。3, 4[13] Mina Ghadimi Atigh、Martin Keller-Ressel和Pascal Mettes。《使用理想原型的双曲布塞曼学习》。《神经信息处理系统进展》，第34卷：103–115，2021年。2, 3

[14] Sadaf Gulshad 和 Arnold Smeulders。通过反事实视觉属性和示例进行解释。载于《2020年国际多媒体检索会议论文集》，ICMR '20，2020年。1, 2[15] Kensho Hara、Hirokatsu Kataoka 和 Yutaka Satoh。时空3D卷积神经网络能否追溯2D卷积神经网络和ImageNet的发展历程？载于《IEEE计算机视觉与模式识别会议论文集》（CVPR），第6546–6555页，2018年。1, 2, 3, 5, 6, 7, 8[16] Lisa Anne Hendricks、Ronghang Hu、Trevor Darrell 和 Zeynep Akata。视觉解释的依据定位。载于 ECCV，2018年9月。1, 2[17] Andrej Karpathy、Justin Johnson 和李飞飞。循环网络的可视化与理解。arXiv 预印本 arXiv:1506.02078，2015年。2[18] Andrej Karpathy、George Toderici、Sanketh Shetty、Thomas Leung、Rahul Sukthankar 和李飞飞。使用卷积神经网络进行大规模视频分类。载于 CVPR，2014年。1, 2[19] Valentin Khrulkov、Leyla Mirvakhabova、Evgeniya Ustinova、Ivan Oseledets 和 Victor Lempitsky。双曲图像嵌入。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第6418–6428页，2020年。3[20] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、James Wexler、Fernanda Viegas 等。超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试。载于《国际机器学习会议论文集》，第2668–2677页。PMLR，2018年。1, 2[21] Jinkyu Kim、Anna Rohrbach、Trevor Darrell、John Canny 和 Zeynep Akata。自动驾驶车辆的文本解释。载于 ECCV，2018年9月。1, 2[22] Pang Wei Koh、Thao Nguyen、Yew Siang Tang、Stephen Mussmann、Emma Pierson、Been Kim 和 Percy Liang。概念瓶颈模型。载于《国际机器学习会议论文集》，第5338–5348页。PMLR，2020年。1, 2[23] Jonathan Krause、Michael Stark、Jia Deng 和李飞飞。用于细粒度分类的3D物体表征。载于第4届IEEE 3D表征与识别研讨会（3dRR-13），澳大利亚悉尼，2013年。3[24] Oscar Li、Hao Liu、Chaofan Chen 和 Cynthia Rudin。通过原型进行基于案例的推理的深度学习：一种能够解释其预测的神经网络。载于《AAAI人工智能会议论文集》，第32卷，2018年。1, 2。[25] Zhenqiang Li、Weimin Wang、Zuoyue Li、Yifei Huang 和 Yoichi Sato。通过扰动实现视频理解网络的视觉解释。载于《IEEE/CVF 计算机视觉应用冬季会议论文集》，第1120–1129页，2021年。1, 2[26] Xinmiao Lin、Wentao Bao、Matthew Wright 和 Yu Kong。用于视频理解模型视觉解释的梯度频率调制。arXiv 预印本 arXiv:2111 01215，2021年 2。.[27] Shaoteng Liu、Jingjing Chen、Liangming Pan、Chong-Wah Ngo、Tat-Seng Chua 和 Yu-Gang Jiang。用于零样本识别的双曲视觉嵌入学习。载于《

IEEE/CVF计算机视觉与模式识别会议论文集》，第9273–9281页，2020年。3[28] Teng Long、Pascal Mettes、Heng Tao Shen 和 Cees GM Snoek。在双曲空间中搜索动作。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1141–1150页，2020年。2, 3, 5, 7, 8, 12[29] Max Losch、Mario Fritz 和 Bernt Schiele。超越分类输出的可解释性：语义瓶颈网络。arXiv 预印本 arXiv:1907.10882，2019年。1, 2[30] Scott M Lundberg 和 Su-In Lee。解释模型预测的统一方法。《神经信息处理系统进展》，第30卷，2017年。2[31] James L McClelland 和 Timothy T Rogers。语义认知的并行分布式处理方法。《自然神经科学评论》，4(4):310–322，2003年。2
[32] 1982 2
[33]
Marvin Minsky。语义信息处理。。Meike Nauta、Ron van Bree 和 Christin Seifert。用于可解释细粒度图像识别的神经原型树。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第14933–14943页，2021年。3。[34] Maximillian Nickel 和 Douwe Kiela。用于学习层次表示的庞加莱嵌入。《神经信息处理系统进展》，
30, 2017 3。
[35]
Maximilian Nickel 和 Douwe Kiela。用于学习层次表示的庞加莱嵌入。载于 NeurIPS，2017年。12[36] Tuomas Oikarinen、Subhro Das、Lam M Nguyen 和 Tsui-Wei Weng。无标签概念瓶颈模型。载于《国际学习表征会议》。1, 2[37] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。PyTorch：一种命令式风格、高性能的深度学习库。《神经信息处理系统进展》，第32卷，2019年。5[38] Xiaojiang Peng 和 Cordelia Schmid。用于动作检测的多区域双流 R-CNN。载于《欧洲计算机视觉会议论文集》，第744–759页。施普林格，2016年。1, 2[39] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。“我为什么要相信你？”——解释任意分类器的预测。载于《第22届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集》，第1135–1144页，2016年。2[40] Dawid Rymarczyk、Łukasz Struski、Jacek Tabor 和 Bartosz Zieli´nski。ProtoPShare：用于可解释图像分类和相似性发现的原型共享。arXiv 预印本 arXiv:2011.14340，2020年。3[41] Karen Simonyan 和 Andrew Zisserman。用于视频动作识别的双流卷积网络。《神经信息处理系统进展》，第27卷，2014年。1, 2[42] Gurkirt Singh、Suman Saha、Michael Sapienza、Philip HS Torr 和 Fabio Cuzzolin。在线实时多时空动作定位与预测。载于《IEEE国际计算机视觉会议论文集》，第3637–3646页，2017年。1, 2
[43] Khurram Soomro、Amir Roshan Zamir 和 Mubarak Shah。UCF101：一个包含野外视频中 101 个人类动作类别的数据集。arXiv 预印本 arXiv:1212.0402，2012年。5，12[44] Alexandros Stergiou、Georgios Kapidis、Grigorios Kalli-atakis、Christos Chrysoulas、Ronald Poppe 和 Remco Veltkamp。用于视频解释的类别特征金字塔。载于 2019年 IEEE/CVF 国际计算机视觉会议研讨会（ICCVW），第 4255–4264 页。IEEE，2019年。2
[45] Alexandros Stergiou、Georgios Kapidis、Grigorios Kalli-atakis、Christos Chrysoulas、Remco Veltkamp 和 Ronald Poppe。显著性管：时空卷积的视觉解释。载于 2019年 IEEE 国际图像处理会议（ICIP），第1830–1834 页。IEEE，2019年。2[46] D´ıdac Sur´ıs、Ruoshi Liu 和 Carl Vondrick。学习未来的可预测性。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 12607–12617 页，2021年。3[47] Alexandru Tifrea、Gary B´ecigneul 和 Octavian-Eugen Ganea。Poincar\é 手套：双曲词嵌入。arXiv 预印本 arXiv:1810.06546，2018年。2, 3[48] Loc Trinh、Michael Tsang、Sirisha Rambhatla 和 Yan Liu。通过动态原型实现可解释且可信赖的深度伪造检测。载于 IEEE/CVF 计算机视觉应用冬季会议论文集，第1973–1983 页，2021年。1，2，3[49] Tomoki Uchiyama、Naoya Sogi、Koichiro Niinuma 和 Kazuhiro Fukui。通过自适应遮挡敏感性分析对视频分类中的 3D 卷积神经网络预测进行视觉解释。arXiv 预印本 arXiv:2207.12859，2022年。2[50] Abraham A Ungar。双曲正方形与莫比乌斯变换。《巴拿赫数学分析杂志》，2007年。12[51] Catherine Wah、Steve Branson、Peter Welinder、Pietro Perona 和 Serge Belongie。Caltech-UCSD Birds-200-2011 数据集。2011年。3[52] Andong Wang、Wei-Ning Lee 和 Xiaojuan Qi。HINT：层次化神经元概念解释器。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 10254–10264 页，2022年。2[53] Jiaqi Wang、Huafeng Liu、Xinyue Wang 和 Liping Jing。通过构建透明嵌入空间实现可解释图像识别。载于 IEEE/CVF国际计算机视觉会议论文集，第 895–904 页，2021年。3[54] Limin Wang、Yuanjun Xiong、Zhe Wang、Yu Qiao、Dahua Lin、Xiaoou Tang 和 Luc Van Gool。时间片段网络：迈向深度动作识别的良好实践。载于欧洲计算机视觉会议，第20–36 页。施普林格，2016年。1，2[55] Elizabeth K Warrington。语义记忆的选择性损伤。《实验心理学季刊》，27(4)：635–657，1975年。2[56] Mert Yuksekgonul、Maggie Wang 和 James Zou。事后概念瓶颈模型。arXiv 预印本 arXiv:2205.15480，2022年。1，2

[57] Bolei Zhou、Aditya Khosla、Agata Lapedriza、Aude Oliva 和 Antonio Torralba。学习用于判别性定位的深度特征。载于 *IEEE*计算机视觉与模式识别会议论文集，第 2921–2929 页，2016年2。[58] Yudong Zhu、Di Zhou、Jinghui Xiao、Xin Jiang、Xiao Chen 和 Qun Liu。Hypertext：赋予 fastText 双曲几何特性。*arXiv* 预印本 *arXiv:2010.16143*，2020年。3

8. 补充材料

8.1. 层次化动作嵌入

将关于动作的先验层级知识纳入网络，需要我们将其表示为嵌入。本节将详细介绍如何在双曲空间中学习这些动作嵌入；下一节将说明如何学习层次原型，并通过将其与双曲空间中的动作嵌入对齐来进行优化。给定动作类别集合$\mathcal{A} = \{1, 2, ..., |\mathcal{A}|\}$，在层次动作识别中，我们还会考虑它们的祖先类别 $\mathcal{H} = \{|\mathcal{A}| + 1, |\mathcal{A}| + 2, ..., |\mathcal{A}| + |\mathcal{H}|\}$，从而构建一个包含三个层级的层次树，即祖父类、父类和子类（见论文中的图 2 右侧）。这一层级嵌入过程针对每个数据集离线执行一次。不过，也可以轻松地针对其他层次结构重复这一过程。

学习动作嵌入。我们将动作层次结构 $\mathcal{A} \cup \mathcal{H}$ 映射到共享双曲空间 $\mathbb{D}^n$ 中，以获得层次化动作嵌入，并在下一节中将其用作动作类别模板。令 $\mathcal{P} = \{(u, v)|u = \phi(v)\}$ 为 v 及其父类 $\phi(v)$ 构成的正样本对；而$\mathcal{N} = \{(u', v')|u' \neq \phi(v')\}$ 则为负样本对。判别损失与 [28]类似：

$$\mathcal{L}(\mathcal{P}, \mathcal{N}, \mathbf{\Phi}) = \mathcal{L}_H(\mathcal{P}, \mathcal{N}) + \lambda \cdot \mathcal{L}_S(\mathbf{\Phi}), \tag{9}$$

其中， $\mathbf{\Phi}$ 表示类别模板矩阵，其第 $c - th$列 $\mathbf{\Phi}_c$ 是 $\mathbb{D}^n$ 中类别 c 的模板向量。对于损失函数， $\mathcal{L}_H$ 鼓励保留父类与子类之间的关系，而 $\mathcal{L}_S$ 则促使不同子层次结构彼此分离。第一部分 $\mathcal{L}_H$ 与 [35], 类似，其中位置错误的子类—父类对将受到惩罚：

$$\mathcal{L}_H(\mathcal{P}, \mathcal{N}) = \sum_{(\mathbf{u}, \mathbf{v}) \in \mathcal{P}} \log \left(\frac{e^{-d(\mathbf{u}, \mathbf{v})}}{\sum_{(\mathbf{u}, \mathbf{v}') \in \mathcal{N}} e^{-d(\mathbf{u}, \mathbf{v}')}} \right), \tag{10}$$

其中， $-d(\mathbf{u},\mathbf{v})$表示两个动作嵌入$\mathbf{u}$ 和$\mathbf{v}$之间的双曲距离，可简写为：

$$d(\mathbf{v}, \mathbf{u}) := 2 \operatorname{arctanh} \left(\| -\mathbf{v} \oplus \mathbf{u} \| \right), \tag{11}$$

其中 $\oplus$ 表示 1–曲面双曲空间 $\mathbb{D}^n$中的莫比乌斯加法 [50] 。

在第二部分中，我们通过分离损失鼓励兄弟关系之间彼此分离，并更新 $\mathbf{\Phi}$ ：

$$\mathcal{L}_S(\mathbf{\Phi}) = - \sum_{i \in |\mathcal{A}|} ||\tilde{\mathbf{\Phi}}_i^T \hat{\mathbf{\Phi}}_i||_F + \gamma ||(\hat{\mathbf{\Phi}}_i \hat{\mathbf{\Phi}}_i^T - \mathbf{I})||_F, \tag{12}$$

ˆ其中， $\mathbf{\Phi}$ 由相对于˜动作类别 i的非兄弟向量组成，而 $\mathbf{\Phi}$ 由 i的兄弟向量组成。

通过上述目标进行学习后，我们得到 $\mathbf{\Phi}$，这是一个动作模板向量矩阵，其中包括动作 $\mathcal{A}$以及祖先（父类和祖父类）动作 $\mathcal{H}$。

8.2. UCF-101 和 ActivityNet 的层次结构

图7 和8 分别展示了 UCF-101 [43]和 ActivityNet 的层次结构 [5]。我们定义了三个层级；对于 UCF-101，一级、二级和三级分别包含 5、20 和 101 个类别。例如，祖级类别之一是演奏音乐，父类包括木管乐器、弦乐器、打击乐器，子类包括吹长笛、弹吉他、击鼓等。层次结构第三级（即子类层级）中的类别是该数据集原有的 101 个类别。

ActivityNet 在一级、二级和三级分别包含 200、38 和 6 个类别（见图8）。例如，祖级类别之一是个人护理，父类包括穿衣打扮、仪容整理、洗漱，子类包括穿鞋、理发、刮胡子等。

8.3. 视觉解释

层次化解释。图9、10 和11 展示了我们多层级解释的定量结果。

失败情况下的有效性。图12 和13展示了我们的方法在失败情况下的有效性。

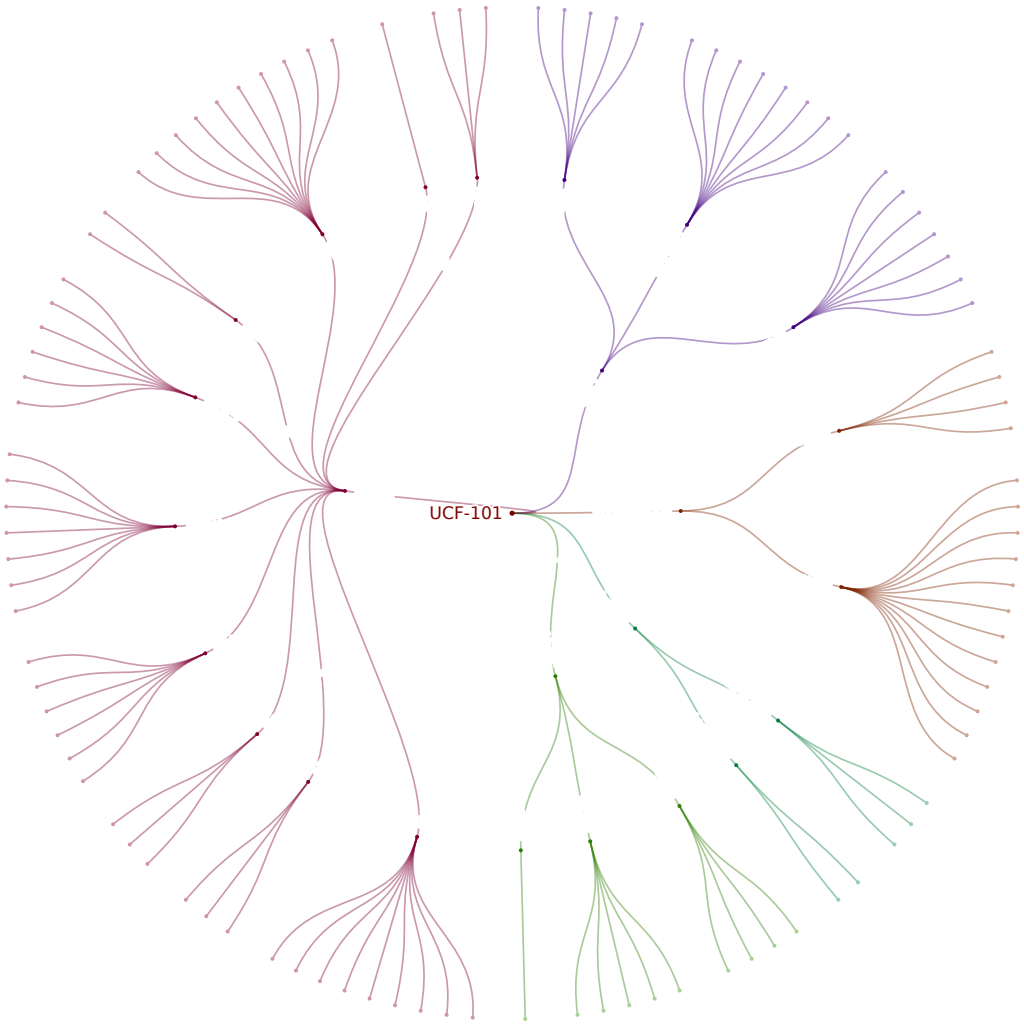


图7. **UCF-101 的层次结构**。UCF-101 数据集所定义层次结构的示意图。层次结构分为三个层级：祖级（演奏音乐）、父类（木管乐器、弦乐器、打击乐器）和子类（吹长笛、弹吉他、击鼓等）类别。

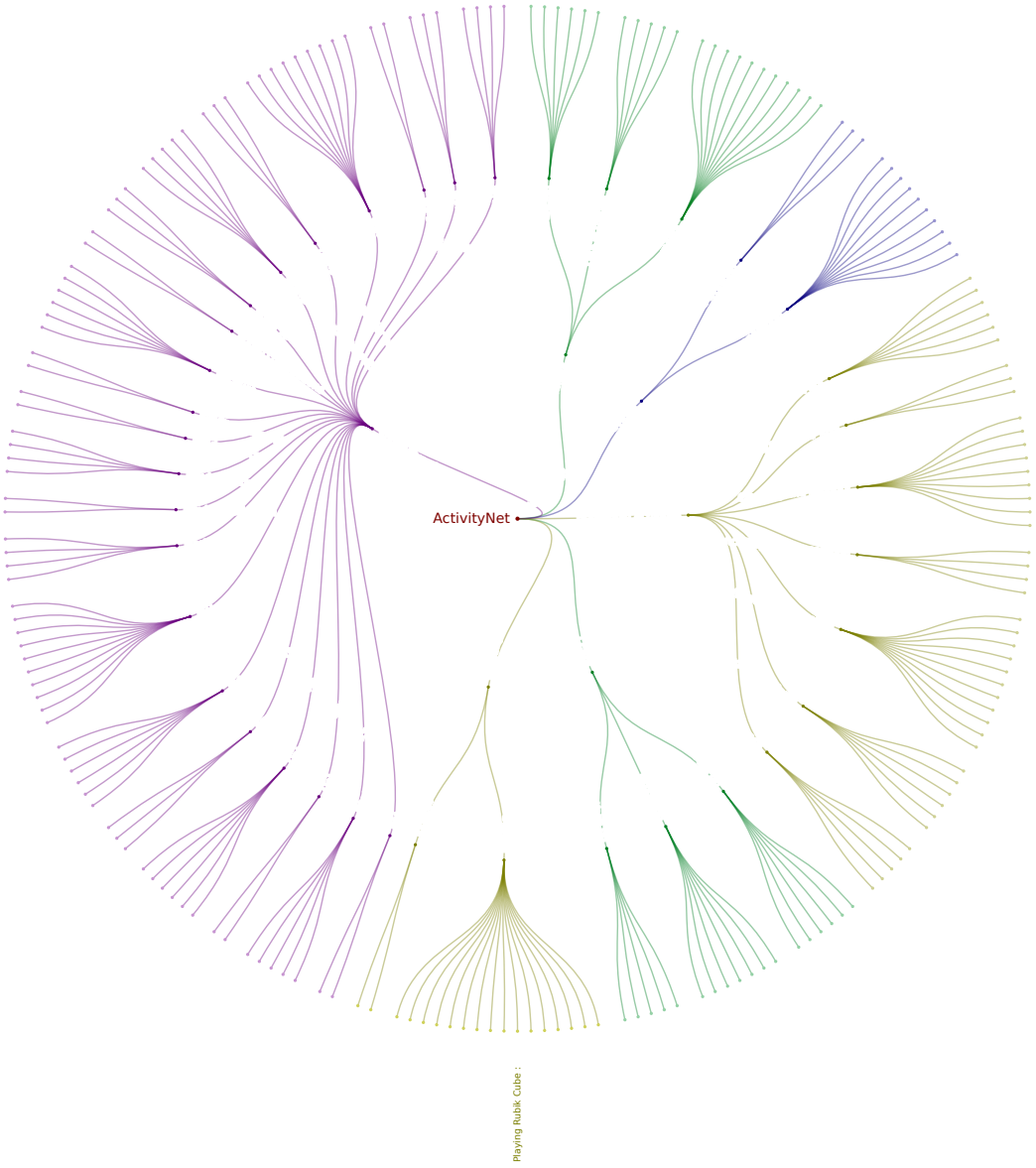


图8. **ActivityNet 的层次结构**。Activity 数据集所定义层次结构的示意图。层次结构分为三级：祖级（个人护理）、父类（穿衣打扮、仪容整理、洗漱）和子类（穿鞋、理发、刮胡子 等类别）。

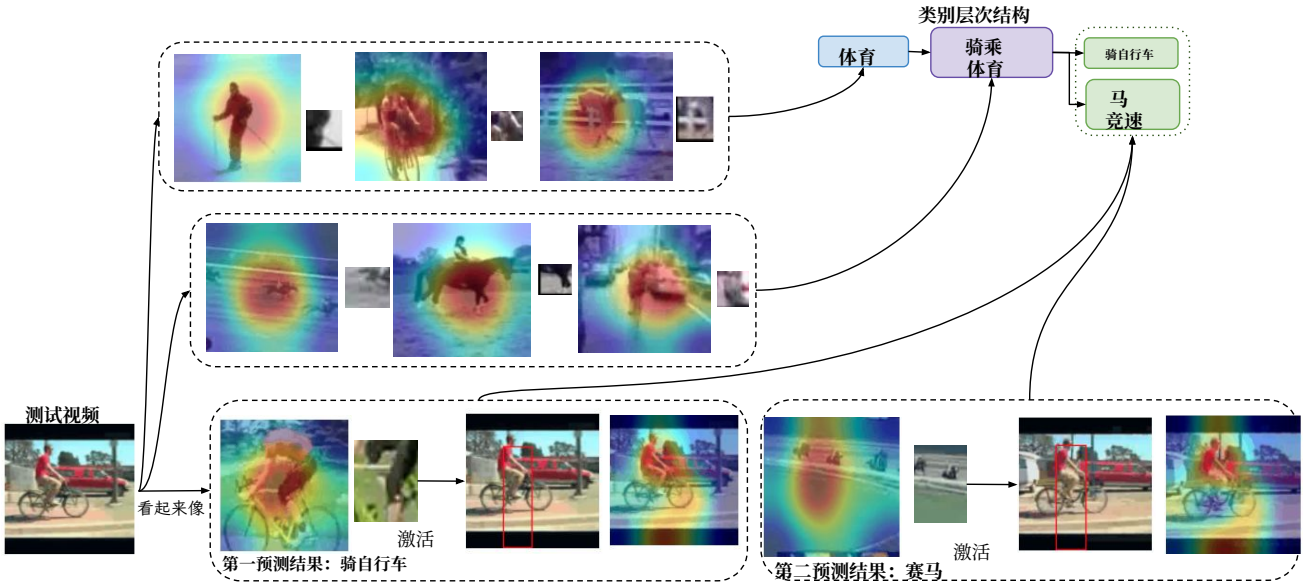
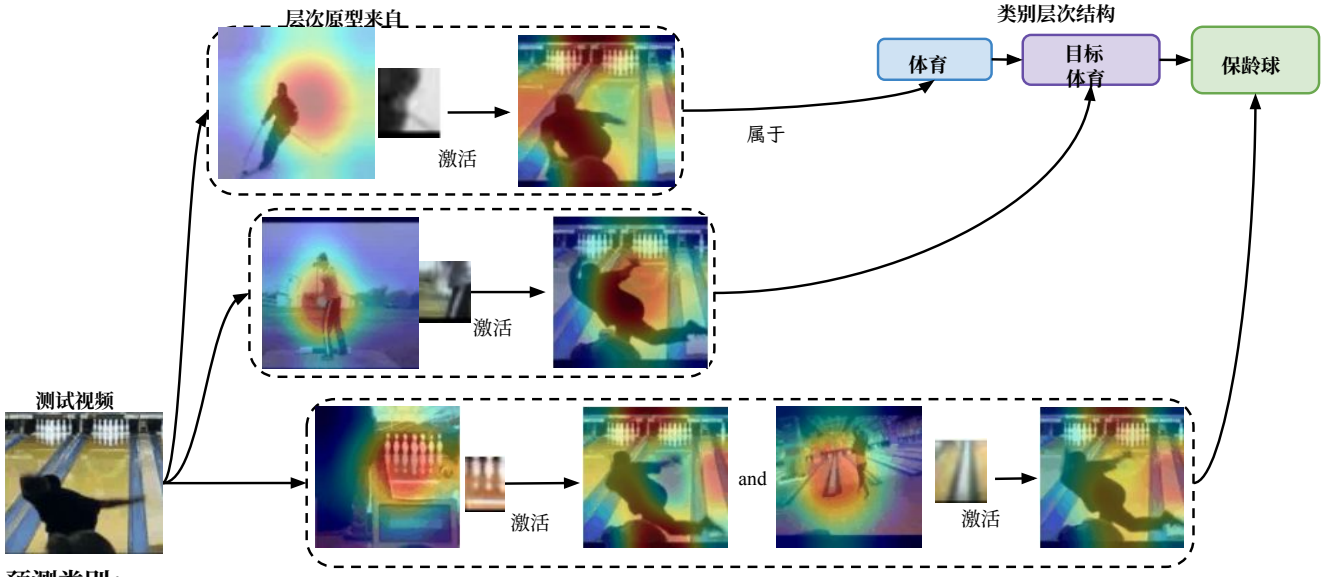


图9。**层次化解释**。此示例展示了祖父类体育类别、父类骑乘运动类别和真实标签骑自行车类别的原型。我们还观察到，模型的第二预测结果是其兄弟姐妹赛马类别。这证实了我们的模型正在学习类别之间的层级关系。



预测类别：
保龄球

图10。**层次化解释**。我们提出的层级原型解释器所采用的基于层级原型的推理过程示意图。

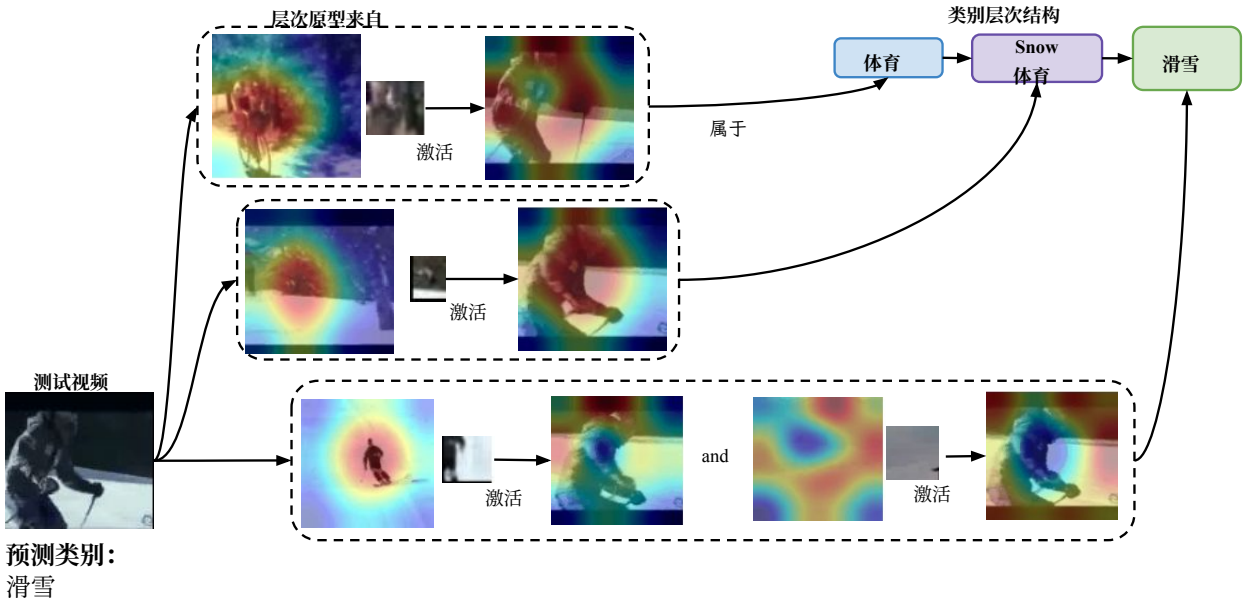


图11. **层次化解释**。 示意图展示了我们提出的层级原型解释器的基于层级原型的推理过程。

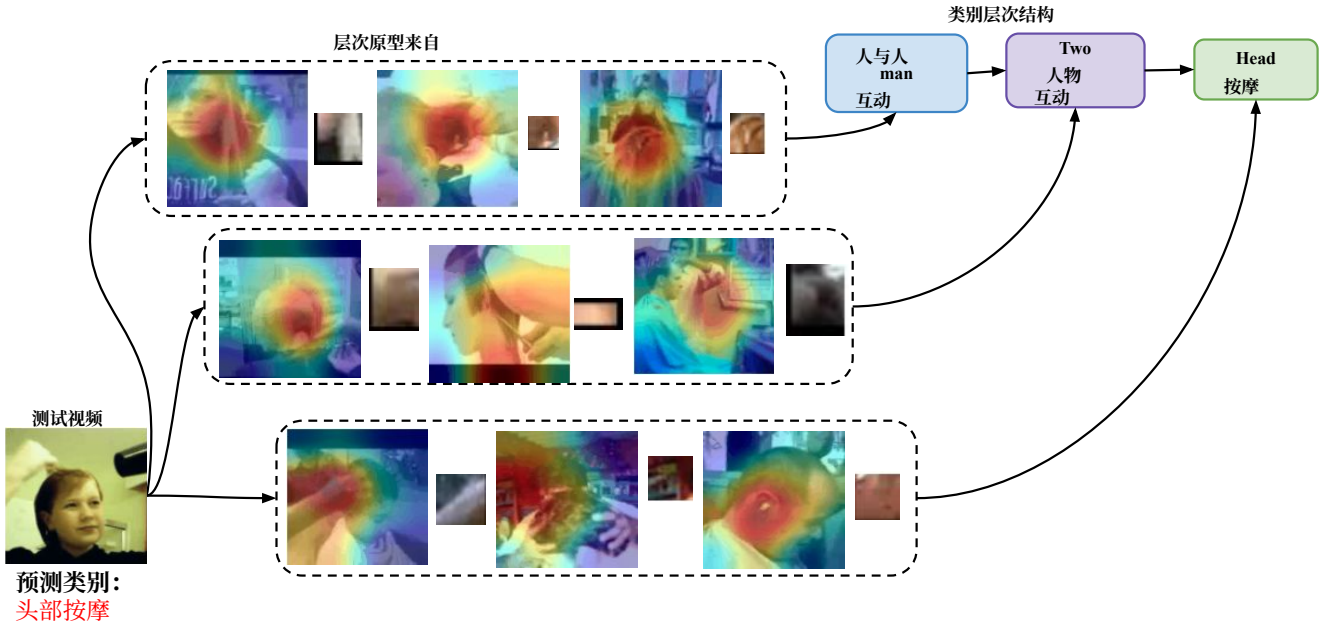


图12。**失败情况下的有效性**。即使发生误分类，我们的多层级解释也能通过为父类和祖父类学习的原型提供有用信息。

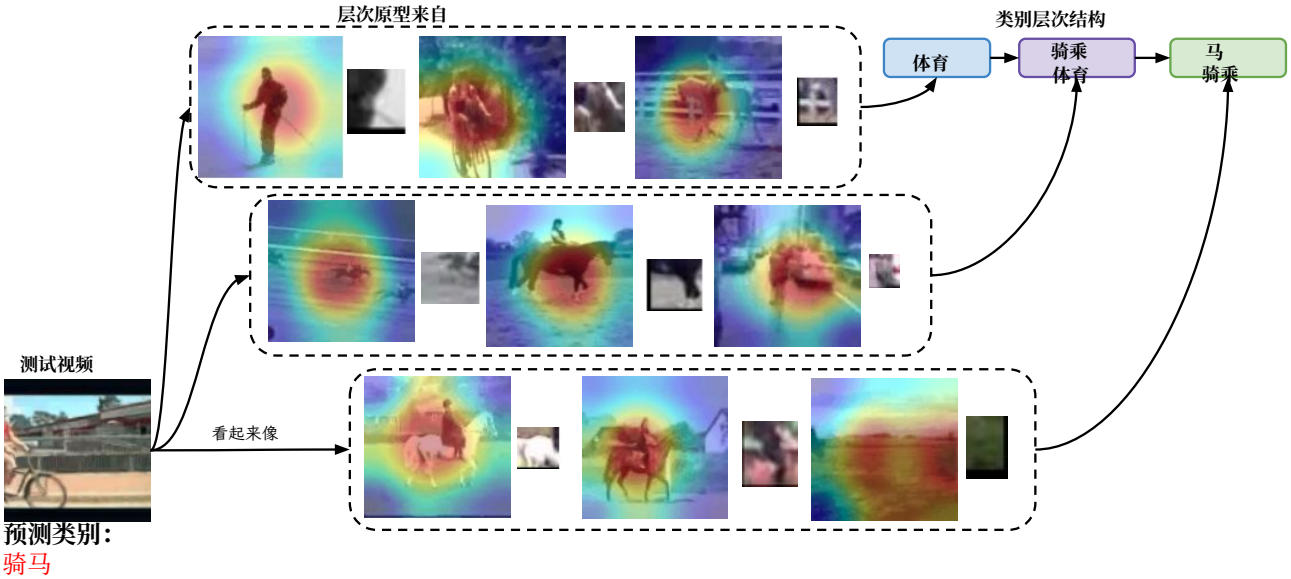


图13。**失败情况下的有效性**。即使发生误分类，我们的多层级解释也能通过为父类和祖父类学习的原型提供有用信息。