基于多粒度感知的音视视点预测

王国涛, 陈程立诏*, 范登平, 郝爱民, 秦洪

摘要—由于深度学习技术的快速发展和广泛可用的大规模训练集，视频显著性检测模型的性能一直在稳步提高。然而，基于深度学习的视听视点预测仍处于起步阶段。目前，在真实的音视环境中收集的真实视点数据集，只提供了少数音视视频序列。因此，在相同的视听环境下重新收集真实的视点既没有效率也没有必要。为了解决这个问题，本文提出了一种新的弱监督方法，以缓解视听模型训练对大规模训练集的需求。通过仅使用视频类别标签，本文提出了选择性类别激活映射机制（***s**elective* **c**lass **a**ctivation **m**apping，SCAM）及其升级版本（SCAM+）。在时空音频环境中，前者遵循从粗到精的策略来选择最具辨别力的区域，并且这些区域通常能够与真实人眼视点表现出高度一致性。后者为SCAM配备了额外的多粒度感知机制，使整个过程更符合真实人类视觉系统的处理过程。此外，本文从这些区域提取知识，以获得完整的空间-时序-音频视点预测网络（**s**patial-**t**emporal-**a**udio （STA） **f**ixation **p**rediction （FP）），从而在视频标签不可用的情况下实现广泛的应用。在不依赖任何真实人眼视点的情况下，这些STA-FP网络的性能与完全监督网络的性能相当。代码和结果可在https://github.com/guotaowang/STANet公开获取。

Index Terms—弱监督学习，音视视点预测，多粒度感知

1 摘要和意图

在深度学习时代，我们见证了视频显著性检测技术的不断发展 [3]， [4]， [5]， [6]， [7]， [8]，其主要任务是定位一系列视频序列中最独特的区域。目前，这一研究领域由两个平行的研究方向组成，视频显著性物体检测（**v**ideo **s**alient **o**bject **d**etection，VSOD）和视频视点预测（**v**ideo **f**ixation **p**rediction，VFP）。在实践中，前者 [9]， [10]， [11]， [12]， [13]， [14]， [15]， [16]， [17] 旨在**分割**具有清晰物体边界的最显著物体（例如，图 *1-B*）。后者是本文的主要主题，它预测人眼的视点——散布在整个场景中的分散坐标，没有任何清晰的边界（例如，图 *1-A*）。事实上，近几十年来，这一主题已被广泛研究。

与之前的方法 [18]， [19]， [5]， [20]不同，本文的兴趣在于利用深度学习技术来预测视觉和音频环境下的视点，也称为音视视点预测，该主题仍处于早期探索阶段。

人类视觉和音频视点预测广泛应用于各个领域。代表性应用包括运动学 [21]、犯罪心理学 [22]、飞行员技能培训 [23]、测谎 [24]、手术风险评估 [25]、360度视频监控 [26]， [27]。例如，当人们撒谎时，他们会产生一系列生理反应，如瞳孔扩张、眨眼次数增加、声音颤抖等。这种“外在现象”已经被用于测谎。然而，在某些情况下，我们将从内部观察受试者。通过分析受试者眼睛视点的细微变化，可以提高测谎准确性。在实践中，**受试者通常处在音视环境中**，因为传统的基于视觉的视点方法无法很好地处理。

目前，几乎所有的 **s**tate-**o**f-**t**he-**a**rt （SOTA）音视视点预测方法都是在深度学习技术的帮助下开发的，使用的是简单的编码器-解码器结构，有各种注意机制，并以完全监督的方式进行训练。尽管取得了进展，但这些完全监督的方法仍受到一个关键限制的困扰。

众所周知，深度模型的性能在很大程度上取决于所采用的训练集，目前，本文的研究领域已经可以使用大规模视觉相关训练集。然而，与视听环境相关的训练数据相当稀缺，因为在这种多模态环境中收集真实的人眼视点是一个费时费力的过程。据本文所知，只有少数配备了真实视点数据的视听序列可用于视听视点预测任务，只有一小部分被用于网络训练，这使得数据短缺的困境更加严重。因此，根据本文进行的广泛定量评估，大多数现有的基于深度学习的视听视点预测模型 [28]， [29]，在本质上可能是过拟合的。

为了解决这个问题，本文尝试使用弱监督策略来实现音视视点预测。本文设计了一种新方案，通过仅使用视频类别标记作为监督，而不是使用劳动密集型的逐帧音视**g**round **t**ruths（GTs），来产生类似GT的音视伪视点。事实上，已经存在大量具有语义类别标记的音视序列（例如，AVE数据集 [30]），其中大多数最初是为音视分类任务收集的。请注意，从成本角度来看，手动为视频分配语义标记显然比收集真实的人眼视点更便捷。因此，关键是如何将视频标签转换为真实的视点数据。

本文的方法受到了各种物体定位和物体检测任务 [31]， [32]， [33]， [34]， [35]， [36]中使用的**C**lass **A**ctivation **M**apping （CAM， [37]）的启发。本文的理论依据是，对

- 这项工作的初期版本已经在 *CVPR 2021 [1]* 上发表。
- 通讯作者: 陈程立诏 *(E-mail: cclz123@163.com)*。
- 本文是 *Weakly Supervised Visual-Auditory Fixation Prediction with Multigranularity Perception [2]*的中文翻译版本，由王国涛翻译，由陈程立诏、范登平校稿。

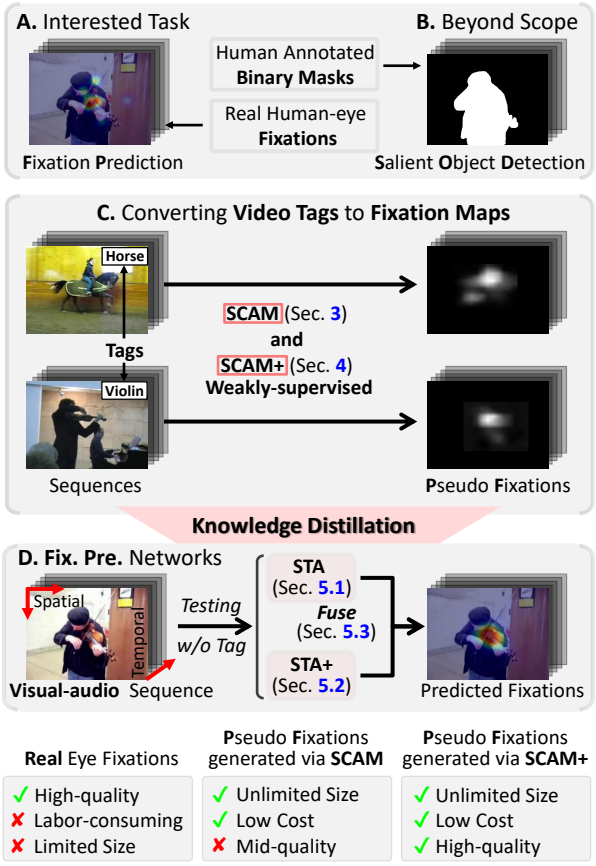


图 1. 本文主要致力于设计一种用于时空-时间-音频（STA）视点预测任务的**弱监督**方法，其关键创新之处在于，本文通过新提出的***selective* class activation mapping**（SCAM，Sec. 3）和升级版SCAM+（Sec. 4）自动将语义类别标记转换为伪视点，该版本还配备了多粒度感知能力。获得的伪视点可以用作学习目标，以**知识蒸馏**的方式指导两个单独的视点预测网络（即STA和STA+，Sec. 5）训练，这两个网络联合生成**通用的**视频视点预测，而不需要任何视频标签。在下侧，本文分别使用✓和✗表示每个标签的优点和缺点。

于分类任务，具有最强判别能力的图像区域应该是最显著的区域，这些区域通常比其它区域具有更大的分类置信度。

如图 1-C所示，考虑到本文的目标是视觉和音频环境中的视点预测，本文在生成高质量视点图方面提出了两项实用创新： 1）*selective* class activation mapping（SCAM）和2）SCAM+——SCAM的升级版本。 SCAM的主要原理依赖于“**数据源**”方面，该方面执行从粗到精的策略，以揭示来自多个数据源（即空间、时间、音频及其组合）的最具辨别力的区域，这些区域与真实人眼视点具有高度一致性。这种从粗到精的方法确保了前面提到的差别较小的区域被完全过滤，不同数据源之间的选择操作有助于揭示差别最大的区域，从而使伪视点更接近真实视点。

此外，由于人类视觉系统显然不独立于我们的大脑，因此真实的人眼视点通常受到多种生理活动的影响，例如短期记忆、长期记忆、联想记忆和语义推理，这使得我们的注意力机制本质上是一个“**全局**”过程。然而，提出的SCAM每

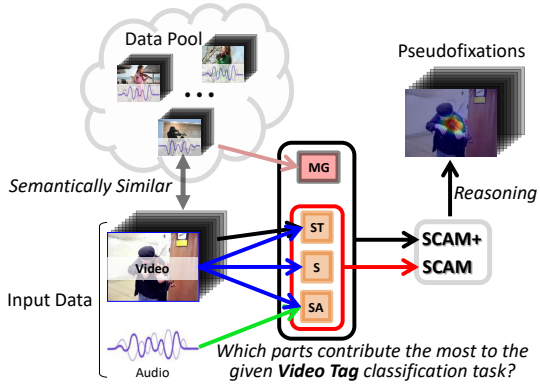


图 2. 作为第一次尝试，本文的方法旨在模拟多模态环境中的人类注意机制。关键的技术亮点是提出了一种可行的方法，通过多粒度（MG）感知，在不同模态（S、SA和ST）上定位最具辨别力的区域。 S:spatial、SA:Spatials+audio、ST:Spatiale+temporal。

次最多只考虑3帧和1秒音频信号，是一种典型的“**局部**”机制。因此，本文设计了SCAM的升级版本，名为SCAM+，它为SCAM配备了额外的多粒度感知机制，以进一步提高导出的伪视点和真实视点之间的一致性程度。通过同时使用SCAM和SCAM+，本文可以自动将视频标签转换为高质量的伪视点图，因此，理论上，本文可以轻松地将现有的视频视点训练集扩展到无限大小。为了更好地突出这种新颖性，本文在图 2中说明了本文方法的基本原理。此外，为了确保广泛的应用，本文将使用SCAM和SCAM+导出的伪视点来训练端到端视点预测网络，这个过程可以被视为知识蒸馏的一种变体，其中，端到端的视点预测网络显然是学生网络。具体而言，本文设计了2个spatial-temporal-audio（STA）视点预测网络（STA和STA+），相应地与基于SCAM和SCAM+的伪GTs相关。在测试阶段，这两个视点预测网络的输出以互补的方式融合为高质量的视点图。如图 1-D所示，整个测试过程不需要为给定的输入测试序列分配语义视频标记，该属性确保了所提出的方法在实际工作中的广泛应用。为了方便读者更好地阅读，本文将主要贡献总结如下：

1）从多源的角度来看，本文提出了一种将视频标签转换为视点图的新方法，即SCAM，它以一种从粗到细的方式，从多源方面定位图像判别性区域，这些判别性区域可能与真实的人眼视点高度一致。

2）受人类注意力生理机制的启发，本文设计了一个升级版的SCAM，名为SCAM+，它使SCAM具备多粒度感知能力。所提出的SCAM+能够压缩那些差别较小的图像区域，并将重点放在差别最大的区域，从而进一步提高产生的视点图的质量。

3）本文首次尝试以弱监督的方式预测基于深度学习的音视视点，这有望促进音视信息集成和计算机视觉中的相关应用的发展。本文的工作可以在面向机器的分类和真实人类视觉听觉系统之间架起桥梁，为在多模态环境中模拟人类注意力铺平了新的道路。

本文以会议版本 [1]为基础，在两个不同方面进行了扩展。首先，基于CVPR版本采用的多源和多尺度视角，本文深入了解了多粒度感知与视觉听觉环境中真实人类注意力行为之间的关系。第二，在不使用任何复杂网络的情况下，本文提供了一个优雅的框架来互补集成多源、多尺度和多粒度信息，以形成与真实信息高度一致的伪视点。除了获得显著的性能增益外，这项工作还为模拟多模态注意力机制提供了一个全面的解决方案。

2 相关工作

2.1 弱监督的方法

基于预给定的图像级标签 [32]， [38]， [39]， [40]，点 [41]、涂鸦 [42]和边界框 [43]，弱监督方法通常优于无监督方法。在这里，本文简单回顾了弱监督物体检测和分割方法。

一般而言，仅使用图像级标记作为监督的弱监督物体检测方法可以实现较好的性能；然而，它们有一个关键的局限性，即它们的检测往往覆盖最具辨别力的部分，而不是整个物体。为了更好地理解，本文将简要回顾几种最具代表性的方法。为了评估物体检测的置信度，Li等人 [44]提出通过掩模遮挡每个检测框，然后将遮挡的图像提供给图像级分类器。通过观察遮挡操作前后分类置信度的差异，仅保留差异较大的检测框，因为这些检测框对分类任务贡献最大。 Wan等人 [45]专注于物体检测任务，该任务试图将图像级物体标签转换为高质量物体检测框。基于选择性搜索生成的初始检测框，该工作提出了最小熵函数来弱监督额外分类器的学习过程，其中选择性搜索的那些不太可信的物体类别预测可能与新训练的分类器的预测存在很大冲突，仅保留那些高质量的物体检测框。在知识蒸馏框架下，Zeng等人 [46]将图像级分类置信度与超像素级跨接（一种测量属于物体的超像素置信度的度量）线性组合，以过滤一些不一致的初始物体检测框。然后，这些保留的检测框，作为老师网络，被用来训练学生网络物体检测器。该过程重复多次，以获得高质量的目标检测框。鉴于弱监督分割方法，关键原理是利用图像级分类任务生成伪掩码，这些掩码将用于促进分割任务。 Jiang等人 [47]提出了一种在线注意力累积策略，该策略利用训练阶段不同时期的注意力图来获得更完整的物体区域。然后，将累积的注意力图作为伪GT，用于训练单个分割模型。

针对分割任务， Choe等人 [48]提出了一种递归擦除策略，以获得完整的分割结果。关键原理是部分擦除当前最具辨别力的区域，然后将擦除后的图像反馈给图像级分类器以显示最具辨别性的区域。通过多次重复这个过程，通常表现出不同辨别度的所有物体部分最终都可以被显示出来。类似的方法可以在 [49]中找到，而主要差异可能是使用了额外的监督训练数据，因此，弱监督训练过程可以从这些额外的监督数据中受益。 Zhang等人 [50]提出了一种双分支结构，其中一个分支旨在图像级分类，另一个分支以像素级回归任务为目标，通过对预先从图像级标签转换的伪分割应用CRF生成像素级回归分割结果（有关伪任务生成的更多详细信息，请参见 [51]）。关键原理是在这两个分支之间执行交互交叉监督，其主要目标是确保对分类任务贡献相似的像素具有相似的回归值。为了将物体二进制掩码与语义标签相关联，Lu等人 [40]提出了一种联合L1优化，以确保具有相似外观的掩码应属于同一类别。通过将弱标签和噪声标签的学习问题作为标签降噪问题，将每个图像分割为一组超像素；然后提出了一种新的基于L1优化的稀疏学习模型。为了解决L1优化问题，他们引入中间标记变量开发了一种高效的学习算法。

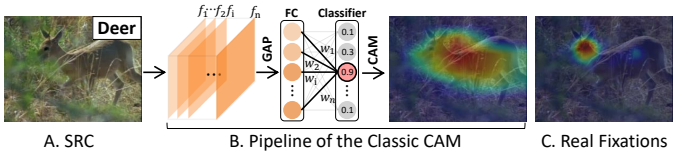


图 3. class activation mapping （CAM）的详细图示。 FC: 全连接层；GAP: 全局平均池化层分类器中的数字表示分类置信度。

2.2 类别激活映射

在典型的视频分类字段中，每个训练序列通常被分配一个语义标记，该语义标记将该序列与特定的视频类别相关联。通常，这些语义标记通过在多人之间执行多数投票来分配，旨在表示给定视频序列中最有意义的物体或事件。与标记分配过程类似，在观看视频序列时，真实的人类视觉视点倾向于关注最有意义和代表性的区域。因此，从视频类别标签生成伪视点在理论上是可行的。

Class activation mapping（CAM）的基本思想是使用最后一个卷积层中特征映射的加权求和来粗略定位与当前分类任务相关的最具代表性的图像区域。在实践中，如图 3-B所示，选择与最后一个全连接层中相对最高的分类置信度相关的那些权重（w_i）来加权特征图（f_i）。因此，一个二维矩阵可通过以下方式获得：

$$\mathrm{CAM} = \mathcal{Z}\Big(\sum_{i}^{d} w_i \times f_i\Big), \tag{1}$$

其中，d表示特征通道数，$\mathcal{Z}(\cdot)$是*min-max*归一化函数。

从定性的角度来看，在图 3-B的右部分可视化的CAM通常显示对分类任务贡献最大的帧区域（即“鹿”）的较大特征响应，这些区域通常与最显著的区域相关。

CAM可能与真实的人眼视点有很大不同，即图 3-B与图 3-C。实际上，在执行视频分类任务时，对于给定类别，挖掘对类别贡献最大的图像区域，能够突出相应物体（鹿）。根据这一原理，一些之前的工作 [52]–[56]使用CAM进行显著物体定位任务。然而，这些方法得出的CAM结果与真实人眼视点不同，主要包括以下三方面原因。

首先，由于局部和非局部深度特征都有助于分类任务，因此CAM往往是大的分散区域。例如，如图 3所示，鹿的主体可以帮助分类器将此图像与其他非动物情况区分开来，而

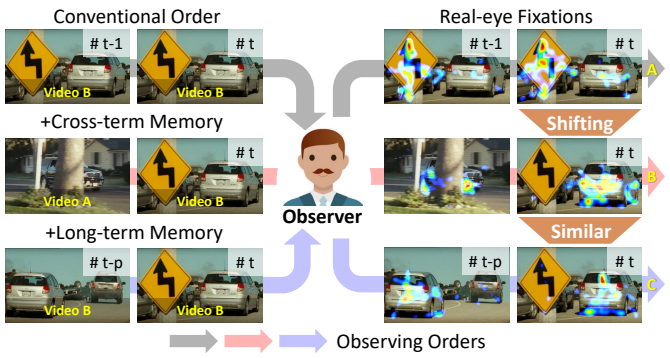


图 4. 多粒度感知的重要性说明($t-p \ll t$)。本文已经进行了额外的实验来支持本文的声明，可以在提交的“**补充材料**”中找到。

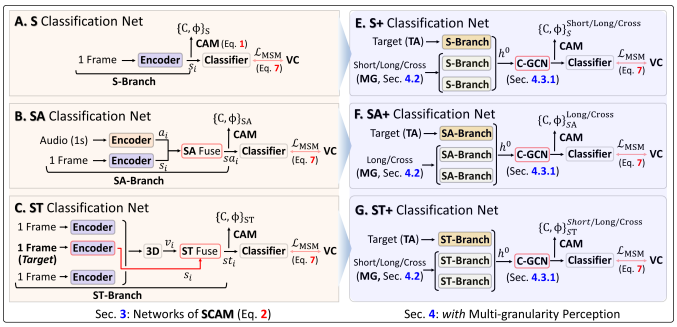


图 5. 本文使用的详细网络架构。子图A-C是本文的SCAM中使用的分类网；子图E-G是升级的分类网，将原始版本耦合在一起以服务于本文的SCAM+。 3D：3D卷积；VC：视频类别；SCAM和SCAM+都可以将视频标签转换为伪视点（pGTs），SCAM+的性能始终优于SCAM。通常，“SA/ST融合”的详细信息可在图 6中找到。 C-GCN是建议的m步推理，其详细信息可在图 11和 4.3.1章节中找到。

只有“鹿头”可以告诉分类器此场景中的动物是“鹿”。人类视觉系统倾向于将更多注意力集中在最具辨别力的图像区域（“鹿头”），图 3-C），而不是盯着“主体”。

其次，现有的大多数工作 [31]， [32]， [57]， [58]，[59]在计算CAM时只考虑了空间信息。然而，真实的人眼视点受到多个数据源的影响，包括空间、时间和音频源。事实上，本文的研究领域长期以来忽略了这种多源特性，因为与空间信息——这一稳定的数据源相比，其他两个数据源（时间和音频）到目前为止仍然被认为是不稳定的数据源，而这种不稳定的属性使得它们难以用于计算CAM。然而，在许多实际场景中，正是这两个数据源对分类任务最为有利。

第三，真实的人眼视点并不是完全通过生理反射产生的，但它也受到我们大脑的微妙影响，包括多种生理活动，这意味着在执行CAM时应考虑所有短期记忆、长期记忆、联想记忆和语义推理。例如，如图 4-A所示，如果仅显示了SRC图像，“道路标志”吸引了大量的眼球视点。然而，如果事先给出了长期信息图 4-B或关联信息图 4-C，其中关联信息可以通过将来自共享相同视频标签的其他视频序列的视频帧与目标序列对齐来获得。因此，当求助于CAM以产生高质量的视点时，需要多粒度感知。

3 选择性类别激活映射（SCAM）

与单个图像情况相比，本文的音视情况的问题域要复杂得多，需要同时考虑多个数据源，包括空间、时间和音频数据源。

如上所述，仅使用空间信息得到的常规CAM往往是大的散射区域，这与实际视点有很大不同；更糟糕的是，它完全忽略了不同数据源之间的互补地位。实际上，在时空音频环境中，编码器嵌入的特征图往往是多尺度、多层次和多源的，所有这些特征图都将共同参与分类任务。然而，如果本文在不考虑补充信息的排斥性的情况下，将所有这些数据源天然地结合起来，就会出现大量的虚警和冗余响应。

为了解决这个问题，本文建议将时空音频环境解耦为3个独立的数据源，即空间（S）、时间（T）和音频（A），并且这些数据源将被单独馈入3个分类网（即，S、ST和SA，参见图 5子图 A-C，下一小节将提供这种组合的原理）。通过在这些分类网络上执行CAM，最终可以获得多源CAMs。显然，这种分而治之的策略可以有效地缓解冗余问题；然而，为了在时空音频环境中模拟真实的人类注意机制，本文将“有选择地融合”它们，以定位最具辨别力的区域。因此，这里的关键问题是如何同时实现多源的互补状态并避免累积冗余特征响应。因此，本文提出了选择性类别激活映射（***selective* class activation mapping** （SCAM）），其技术细节可由等式 2表示。

$$\begin{aligned} \mathrm{SCAM} &= \mathcal{Z}\Big(\frac{\left\|\mathrm{UC}\odot\mathrm{UR}\right\|_1+\lambda}{\left\|\mathrm{UC}\right\|_1+\lambda}\Big), \\ \mathrm{UR} &: \Big[\Phi_{\mathrm{S}}\{i\},\Phi_{\mathrm{ST}}\{i\},\Phi_{\mathrm{SA}}\{i\}\Big], \\ \mathrm{UC} &: \Big[\oint\left(\mathrm{C_S}\{i\}\right),\oint\left(\mathrm{C_{ST}}\{i\}\right),\oint\left(\mathrm{C_{SA}}\{i\}\right)\Big], \end{aligned} \tag{2}$$

其中，$\odot$是元素乘法运算；$\|\cdot\|_1$表示L1范数；λ是一个非常小的常数，以避免被零除； Φ_{S}、Φ_{ST} 和 Φ_{SA} 代表分别来自于S、ST和SA 分类网络的CAM结果（公式 1）； $\mathcal{Z}(\cdot)$ 是*min-max*归一化操作。此外，假设S分类网的预激活类别标记是总计c（28）类中的第i个类，然后本文使用$\mathrm{C_S}\{i\}$来表示该置信度，其中，$\mathrm{C_S}\in(0,1)^{1\times c}$。 $\oint(\cdot)$是一种软滤波器（等式 3），旨在压缩在计算SCAM时要考虑的低分类置信度特征。

$$\oint\left(\mathrm{C_S}\{i\}\right)=\begin{cases}\mathrm{C_S}\{i\} & if\quad \mathrm{C_S}\{i\}>\mathrm{C_S}\{u\}|_{i\neq u,1\leq u\leq c}\\ 0 & otherwise\end{cases}. \tag{3}$$

在以下小节中，本文提供了该SCAM的基本原理。此外，为了使SCAM更符合真实的视点，本文将引入一种有效的阶段性（多尺度）方式，即阶段性SCAM。

3.1 SCAM 逻辑依据

一般来说，空间、时间或音频数据源都会影响人类的视觉注意力；然而，与后两者相比，空间源在实践中通常更为重要和稳定。例如，当视频物体长时间保持静止时，其空间外观

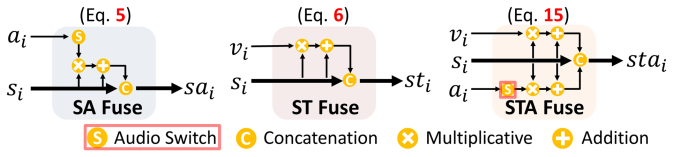


图 6. 在图 5 和图 12中使用的融合模块。

可能完全不变，而其时间信息则完全缺失。对于音频源也存在类似的情况，其中，音频信息可能与其空间对应完全无关。因此，在本文的分类网中，空间信息应被视为主力，而其他两个只能是其下属。这就是本文将S、T和A数据源分别重组为S（无变化）、ST和SA的原因。

考虑到所有的S、ST和SA分类网络都已经在仅标有视频类别标签的训练实例上进行了训练，如果将它们输入这些网络进行测试，则大多数训练实例通常表现良好。然而，由于这些网络的输入不同，因此从这些网络导出的CAM在本质上仍然存在很大差异，本文在图 8中展示了一些最具代表性的定性结果。

通常，多源CAM和真实视点之间的一致性水平通常与分类置信度正相关。通过使用分类置信度作为融合权重来选择性地压缩那些不太可信的CAM，通过选择性地融合所有这些

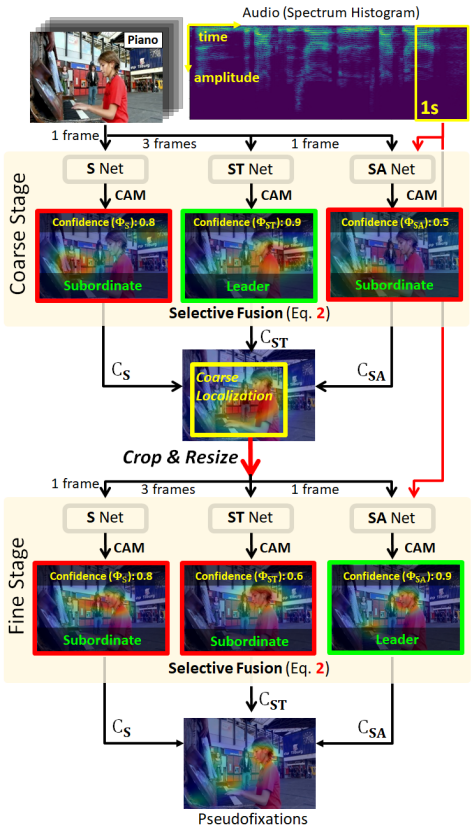


图 7. 本文的**s**elective **c**lass **a**ctivation **m**apping （SCAM）遵循从粗到精的策略，粗阶段定位感兴趣区域，然后细阶段显示具有最强局部响应的图像区域。 S：空间；ST：时空；SA：音频。S/ST/SA网络的结构可以在图 5的（A-C）中找到。

多源CAM（等式 2）获得的伪视点可以更接近真实视点。

3.2 多阶段SCAM

得益于多个数据源的选择性融合，提出的SCAM（等式 2）能够在揭示伪视点方面优于传统的CAM。然而，SCAM产生的伪视点有时可能与真实视点不同，特别是对于具有复杂背景的场景，其伪视点往往会引入错误信息。主要有两个原因：

- 复杂的视频场景通常包含更多的内容，但它只被分配了一个类别标记；因此，更多属于范围外类别的内容也可能有助于当前的分类任务。
- 上述SCAM遵循单尺度程序，而与此形成鲜明对比的是，真实人类视觉系统是一个多尺度（多阶段）系统，在将真实视点分配给局部区域之前，人们往往会无意识地快速定位感兴趣区域 [60]。

为了进一步改进，本文按照从粗到精的方法顺序执行两次SCAM。粗阶段减少了给定的问题域；因此，精细阶段（即第二阶段）中显示的伪视点更可能与最具辨别力的区域相关，从而显著提高整体性能。

如图 7所示，**粗**阶段使用矩形框（粗定位）来紧密变换已通过硬阈值二值化的伪视点（2倍平均值），并且给定的输入视频序列将通过这些矩形框裁剪成视频块。在**细**阶段，视频序列被这些视频块替换为分类网的输入，本文再次执行SCAM以获得多源伪视点。

3.3 SCAM中的融合模块

本文采用的所有网络均遵循最简单的编码器-解码器架构。遵循之前的工作 [29]，本文预先将音频信号转换为二维频谱直方图（U_i），随后将其输入到现成的VGGSound（$\mathcal{F}_{\mathrm{VggSound}}$）[61]获得相应的音频特征（$a_i$），类似地，可以通过将单个视频帧（$\mathrm{I}_i$）输入到现成的VggNet（$\mathcal{F}_{\mathrm{VggNet}}$） [62]来获得空间特征，这两个过程可以公式化如下：

$$a_i = \mathcal{F}_{\mathrm{VggSound}}\left(\mathrm{U}_i\right), \quad s_i = \mathcal{F}_{\mathrm{VggNet}}\left(\mathrm{I}_i\right). \tag{4}$$

所有这些实现都非常简单和直接，几乎所有的网络细节都已在图 5中清楚地表示出来。当然，增强的替代方案可能会带来额外的性能提升。接下来，本文将提供SA/ST分类网中采用的SA/ST融合模块的详细架构，即图 6的前2个。

SA 融合模块

SA融合模块的主要目标是将空间信息与音频信号集成。假设所采用的特征主干的输入大小为 256×256，空间特征s_i的确切大小应为$\{8 \times 8 \times 2048\}$，音频特征$\mathrm{a}_i$的大小为$\{1 \times 8192\}$。本文使用一系列反卷积将$\mathrm{a}_i$的大小转换为与$\mathrm{s}_i$相同的大小。详细的SA融合过程可详述如下：

$$sa_i = Relu\Big(\sigma\Big(DeConv\big(\phi(a_i)\big)\Big) \odot s_i + s_i\Big) \otimes s_i, \tag{5}$$

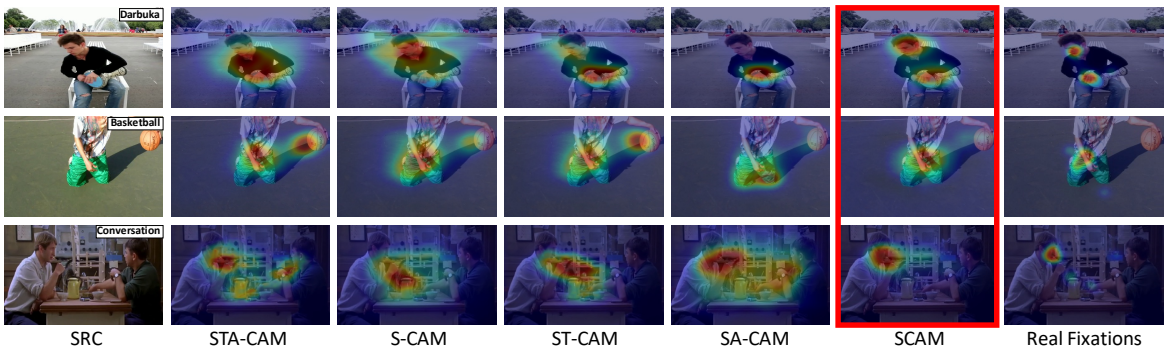


图 8. 来自不同数据源的CAMs的定性说明“STA(S/SA/ST)-CAM”：从时空音频（空间/空间音频/时空）环境中获得的CAM；“SCAM”代表由等式2获得的伪视点，可以很容易地观察到，本栏中的结果可以与GTs高度一致。

其中，$\otimes$是特征串联操作；$DeConv(\cdot)$是反卷积运算；$\odot$表示元素级乘法运算；$\sigma(\cdot)$是sigmoid函数；$Relu$是Relu函数。$\phi(\cdot)$是一个“二进制开关”（即本文的**audio switch**，将在后面详述），用于消除那些不相关的音频信号。提出的SA 融合模块的基本原理非常清楚，音频信号作为语义注意的空间对应来压缩那些不太容易区分的帧区域。

ST 融合模块

ST融合模块的方法与SA融合模块非常相似，主要区别在于ST融合组件省略了音频开关。主要原因是时间信息更有可能使空间信息受益，但音频信号可能与空间线索完全冲突，导致学习模糊。 ST融合的技术细节可以在图 6的中间列中看到，其数据流可以公式化为：

$$st_i = Relu\big(\sigma(v_i) \odot s_i + s_i\big) \otimes s_i, \tag{6}$$

其中，v_i表示使用3D卷积后的时间信息，其他符号和操作与等式 5的符号和操作完全相同。

音频开关

该模块的主要功能是在执行音视融合时减轻音频信号的潜在副作用。与时间源相比，音频源通常信息量较少，但与强语义信息相关，因此更容易影响其空间对应项。然而，音频源本身有一个严重的缺点，视频序列通常会与无意义的背景音乐或噪声耦合。在这种情况下，将音频源与空间源融合可能会使分类任务更加困难。

事实上，提出的“音频开关”的性质是一个插件工具，可以通过与“SA”分类网结构相同的单个网络实现。该插件不是针对视频分类任务，而是基于音视数据进行训练，以二进制标签作为学习目标，指示当前音频信号是否真正有利于空间源。

为了自动获得这些二进制标签，本文使用了一个现成的音频分类工具（VggSound [61]），该工具在一个包括近300个类别的大规模音频分类集上进行了训练。主要的理由是，只有当音频源与空间源同步，共享相同的语义信息时，音频源才能使空间源受益。因此，对于音视片段（1帧和1秒音频），如果音频分类工具预测的音频类别与预设置的视频类别相同，本文将其二进制标签指定为“1”；否则，本文将其二进制标签指定为“0”。

分类器

上述网络中采用的所有分类器都是简单分类器，仅由2个步骤组成。首先，本文使用1×1卷积将输入特征张量的通道大小（即：s_i、sa_i和st_i）细化为c。接下来，本文执行全局平均池化（GAP）操作，将张量转换为向量，其维度与视频类别数c相同。

分类器损失

SCAM中的所有分类网络都采用了相同的损失函数，即标准的**M**ultilabel **S**oft **M**argin (MSM) loss [63], [64]。为了更好地说明，本文以“ST网络”（图 5-C）为例，其分类损失如下所示：

$$\begin{aligned}\mathcal{L}_{\text{MSM}}(st_i, \text{VC}) = -\frac{1}{c}\sum_{i=1}^{c} \text{VC}_i \times \log\left(\frac{1}{1+e^{-st_i}}\right) \\ + (1-\text{VC}_i) \times \log\left(\frac{e^{-st_i}}{1+e^{-st_i}}\right),\end{aligned} \tag{7}$$

其中，“st_i”是“分类器”的输入，可通过等式 6获得，VC表示给定输入的确切类别，c是视频类别总数。对于本文采用的所有分类器，本文使用与等式 7相同的损失函数。

3.4 SCAM vs. Real Fixation

所提出的SCAM基于对多源分类网S、ST和SA网络显示的判别性区域执行选择性融合。尽管SCAM生成的伪视点图基本上与给定视听片段中最具辨别力的区域相关，但它们偶尔可能与真实人眼视点（GTs）不同，如图 9，即SCAM vs. GTs。

从数据角度来看，SCAM计算的总输入仅包括3个相邻视频帧和最多1秒的音频信号，这使得所提出的SCAM成为典型的***局部***方法。然而，真实的人类视觉系统遵循一种***全局***方式，真实的视点受到多种因素的共同影响（例如，短期/长期记忆）。因此，使用SCAM产生高质量伪视点的挑战依赖于将数据源选择过程（公式 2）与***全局***信息、多粒度信息结合起来。本文将在这里提供一个深入的解释。

图 9. 局部信息（SCAM）和全局信息（SCAM+）融合的视觉比较。可以观察到，得益于全局信息之间的关系约束，基于全局信息的方法能够处理各种挑战性因素，如位置偏差、颜色偏差和多个物体。

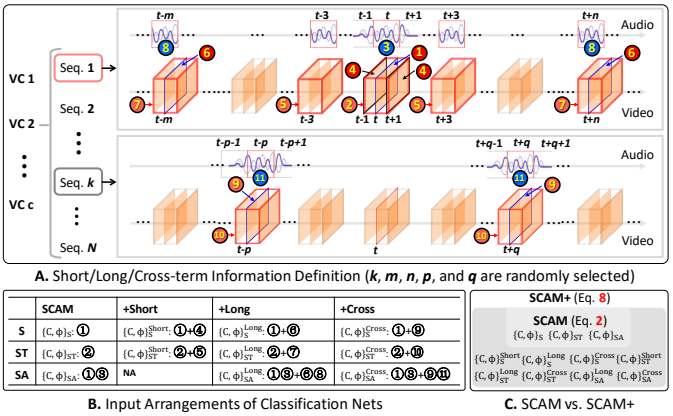


图 10. SCAM和SCAM+的输入数据公式/定义。 VC c：第c个视频类别类；Seq.k：第k个视频序列。

首先，人类视觉系统倾向于忽略在给定视频序列中重复出现的那些从数据源角度看不太突出的图像区域，即使是当前音视片段中最具辨别力的区域。例如，如图 9的第一行所示，真实的视点避免了“狗”，因为这种“狗”以前出现过多次，使其不如“鹰”突出。然而，该SCAM只考虑了当前的局部视听片段，因此其伪视点图仍然将“狗”视为显著物体，因为从局部角度来看，“狗”也可能对当前的分类任务有很大贡献。因此，本文将短期和长期信息纳入SCAM。

第二，真实的视点受嵌入我们大脑中的高级语义信息的影响 [63]–[65]。事实上，人类视觉系统从不单独工作，因为人脑在其中起着至关重要的作用，因此，不同年龄、职业等的受试者之间，真实的人类视点会有所不同。主要原因是我们在日常生活中学习到的高级语义知识可以直接决定我们真正看的地方。尽管存在这种现象，但认为区分性和显著性之间存在正关系的原则仍然成立，为了进一步完善SCAM衍生的伪视点，本文将在SCAM中引入高级语义信息。因此，本文建议考虑交叉信息，这可以通过引入从与给定序列共享相同标签的其他视频序列中剪切的多个音视片段获得。因此，在分类任务期间获取这些额外的交叉信息可以帮助压缩那些相对较少区分的区域，这最终可以改进伪视点。

总之，完美的伪视点生成框架不仅应考虑多源和多尺度方面，还应考虑多粒度感知，在执行SCAM时应考虑所有的短期（*short-term*），长期（*long-term*）和交叉（*cross-term*）信息。因此，本文提供了SCAM的升级版本，名为SCAM+，将在下一节中详细介绍。

4 选择性类别激活映射+ （SCAM+）

4.1 SCAM+ 概述和技术细节

如图 5所示，SCAM和SCAM+之间的主要区别在于所采用的分类网，SCAM只采用了3个分类网，即S、ST和SA（子图A-C），而在SCAM的基础上，SCAM+还采用了3种**类型**的分类网，分别为S+、SA+和ST+（子图E-G）。短期/长期/交叉信息的组合被用作这些新分类网的输入，产生了8个额外的分类网，具有8个分类置信度（C）和CAM映射（Φ），包括$\{C,\Phi\}^{\text{Short}}_{\text{S+}}$、$\{C,\Phi\}^{\text{Long}}_{\text{S+}}$、$\{C,\Phi\}^{\text{Cross}}_{\text{S+}}$、$\{C,\Phi\}^{\text{Short}}_{\text{ST+}}$、$\{C,\Phi\}^{\text{Long}}_{\text{ST+}}$、$\{C,\Phi\}^{\text{Cross}}_{\text{ST+}}$、$\{C,\Phi\}^{\text{Long}}_{\text{SA+}}$和 $\{C,\Phi\}^{\text{Cross}}_{\text{SA+}}$。其中，前3个可以从S+网导出，中间3个可以通过ST+网获得，最后2个可以通过SA+网得到。具体而言，本文省略了$\{C,\Phi\}^{\text{Short}}_{\text{SA+}}$，因为音频信号本身已经覆盖了短期信息。基于SCAM+的伪视点可通过以下等式获得。

$$
\begin{aligned}
&\text{SCAM+} = \\
&\frac{\left\|\text{UC}\odot\text{UR}+\text{SC}\odot\text{SR}+\text{LC}\odot\text{LR}+\text{CC}\odot\text{CR}\right\|_1+\lambda}{\left\|\text{UC}+\text{SC}+\text{LC}+\text{CC}\right\|_1+\lambda},\\
&\text{SR}:\left[\Phi^{\text{Short}}_{\text{S+}}\{i\},\Phi^{\text{Short}}_{\text{ST+}}\{i\}\right],\\
&\text{LR}:\left[\Phi^{\text{Long}}_{\text{S+}}\{i\},\Phi^{\text{Long}}_{\text{ST+}}\{i\},\Phi^{\text{Long}}_{\text{SA+}}\{i\}\right],\\
&\text{CR}:\left[\Phi^{\text{Cross}}_{\text{S+}}\{i\},\Phi^{\text{Cross}}_{\text{ST+}}\{i\},\Phi^{\text{Cross}}_{\text{SA+}}\{i\}\right],\\
&\text{SC}:\left[\oint\left(\text{C}^{\text{Short}}_{\text{S+}}\{i\}\right),\oint\left(\text{C}^{\text{Short}}_{\text{ST+}}\{i\}\right)\right],\\
&\text{LC}:\left[\oint\left(\text{C}^{\text{Long}}_{\text{S+}}\{i\}\right),\oint\left(\text{C}^{\text{Long}}_{\text{ST+}}\{i\}\right),\oint\left(\text{C}^{\text{Long}}_{\text{SA+}}\{i\}\right)\right],\\
&\text{CC}:\left[\oint\left(\text{C}^{\text{Cross}}_{\text{S+}}\{i\}\right),\oint\left(\text{C}^{\text{Cross}}_{\text{ST+}}\{i\}\right),\oint\left(\text{C}^{\text{Cross}}_{\text{SA+}}\{i\}\right)\right],
\end{aligned}
\tag{8}
$$

其中，UC和UR可以在等式 2中找到，而$\|\cdot\|_1$、λ、Φ、$\oint$和$\odot$的含义与等式 2的含义相同。

显然，提出的SCAM+在产生伪视点时包含了更多的粒度。为了便于更好地阅读，仍应提供2个技术细节：1）上述多粒度信息（Sec. 4.2）的准确定义和实现，以及2）所采用分类网的准确网络架构，特别是**融合部分**（Sec. 4.3）。以下小节将详细介绍这两个方面。

4.2 多粒度信息定义

与SCAM单独考虑的**局部**信息（即3个连续帧和1秒音频信号，本文称之为**目标**数据）不同，本文中提到的SCAM+的**多粒度信息**包括3种类型的数据公式，本文将给出它们的定义和详细的实现。

短期信息： 在本文的方法中，短期信息充当人类短期记忆，自动抑制那些不太重要的场景内容，并将本文的注意力集中在剩余的显著内容上。如图 10-A所示，假设第t个视频帧是由①标记的**目标（*target*）**帧；其短期时空信息可以详细描述为相邻的8个视频帧，标记为② + ⑤，并且所有9个帧一起可分为3组，以输入到三分支分类网络中，即图 5-G中详述的ST+网络。具体地说，从唯一的空间源角度来看（即，S+网络，图 5-E），短期信息是目标帧的两个相邻帧，即标记④，这可能与SCAM采用的ST网络（图 5-C）的输入完全相同，然而，基本原理在本质上是不同的。ST网络简单地利用3D卷积来获取时空信息，但与此形成鲜明对比的是，S+网络主要旨在学习帧级相似关系度量，以模仿真实人类注意机制的感知机制。因此，S+网和ST网本质上是互补的。

长期信息： 作为短期信息的补充部分，长期信息还可以帮助人类视觉系统将注意力集中在整个视频序列中最显眼的场景区域。与上面定义的短期信息类似，本文将在时间尺度上远离当前目标帧的视频帧和音频信号视为长期信息。假设第t帧是当前目标帧，长期信息的范围可以是同一视频序列中的任何帧，目标帧的相邻帧除外（8帧）。由于长期信息的准确表述可能因不同来源而异，本文将在此处提供其详细信息。从空间角度来看，第t帧的短期信息是时间上远离第t帧的另外两个帧，并且可以随机选择（例如，图 10-A中的标记⑥），这3个帧（1个目标帧和2个短期帧）将被输入到S+网络中，如图 5-E所示。类似地，从时空角度来看，短期信息变成2组3个连续帧（例如，图 10-A中的标记⑦），它们也是随机选择的。因此，总共有9帧（其中，3帧是目标连续帧，其他6帧是长期帧）将被输入到ST+网络，如图 5-G所示。关于视觉-音频感知，唯一的区别是额外考虑了音频信号（例如，图 10-A中的标记⑧）；因此，如图 5-F所示，总共有3帧和3秒的音频信号被输入到SA+网络。

交叉信息： 如前所述，真实的人类视觉系统受到联想记忆的影响，联想记忆是对不同物体类别累积的基本印象。例如，我们已经知道一辆车的外观，鉴于视频序列包含多辆车，我们可能会更加关注外观最独特的那辆车，此过程需要更多关于当前视频类别的信息。这种现象促使本文考虑交叉信息，可以将其他视频序列中与当前序列具有相同视频类别标签的帧视为交叉信息。如图 10-A所示，从唯一的空间域角度来看，序列1（视频类别2）中第t帧的交叉信息可以是从其他视频序列k中随机选择的任意2帧，这些视频序列也被分配了类别标签2，例如标记⑨。类似地，从时空或视听角度来看，交叉信息可以相应地表示为标记⑩或标记⑪。

4.3 多粒度感知

与SCAM相比，升级版SCAM+采用了几个新的分类网，即S+、SA+和ST+ 如图 5的子图E-G所示，所有这些分类网络都直接采用了SCAM的部分分类网（即：S、SA和ST）作为其主干，即S/SA/ST分支。由于这些特征主干几乎没有改变，本文将重点放在后续融合部分。

目前，有多种方法可以建模交叉数据关系，例如，循环神经网络（RNN） [66]、长短时序记忆网络（LSTM） [67]，[68]和门控循环单元（GRU） [69]。虽然这些现有工具可以感知随时间的细微变化，但它们显然不适合本文的情况，因为这些工具要求输入数据“空间”对齐（两个连续的视频帧在空间上对齐）。然而，在本文的例子中，确保属于不同视频序列的两个帧（例如，本文的交叉信息）的这种空间对齐几乎是不可行的。此外，考虑到人类的注意力机制，最有价值的短期/长期/交叉信息可能是高级语义部分，这促使本文对目标数据与其短期/长期或交叉之间的语义关系进行建模。为了实现这一点，本文设计了一种graph convolution network （GCN） [70]的变体，将多粒度信息与多数据源信息相结合，称为Conservative-GCN（C-GCN），其主要亮点包括：1）在多源和多粒度节点上执行m步的语义迭代（***semantic reasoning***）的全新方法，以及2）全新的视点精细化层（***fixation refine layer***，***FRL***），以确保整个过程足够精细化，消除冗余特征响应。

4.3.1 语义迭代

提出的C-GCN（图 11）由多个节点和边组成 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$，其中，$\mathcal{V}$表示节点集 $\{\mathrm{TA}, \mathrm{MG}_1, \ldots, \mathrm{MG}_n\}$，其中，TA表示**目标（*target*）**节点（即当前输入的视听片段），以及 MG包括与短/长/交叉信息相关的n个节点。本文使用$\mathcal{E}$表示边集 $\{e_{i,j}\}$，$i \neq j$；和 $e_{\mathrm{TA},\mathrm{MG}_i} \in \mathcal{E}$ 是在 TA 和 MG_i之间的边。

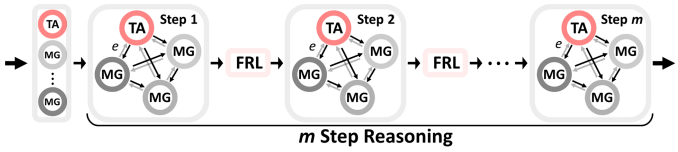


图 11. 提出的 C-GCN 的网络架构，主要由两部分组成，即：1）语义推理（**semantic reasoning**）（Sec. 4.3.1），2）视点精细化层（**fixation refine layer** ）（Sec. 4.3.2）。

如图 5（E-G）所示，C-GCN的输入由3部分组成，即S/SA/ST分别作为网络主干时，每个网络的三个分支的输出（$s_i/sa_i/st_i$)结果，其包含丰富的语义信息，为简单起见，本文统一将其表示为h_i^0，其中，下标i表示该特征属于第i个节点，上标0表示语义推理的开始。在第t推理阶段，本文使用*interattention*（F_{iatt}，公式 10）来建模每两个输入节点的语义关系，即边缘$e_{i,j}$，其公式如下：

$$\begin{aligned} e^t_{\mathrm{TA},\mathrm{MG}_i} &= Softmax\big(F_{iatt}(h^t_{\mathrm{TA}}, h^t_{\mathrm{MG}_i})^\top\big), \\ e^t_{\mathrm{MG}_i,\mathrm{TA}} &= Softmax\big(F_{iatt}(h^t_{\mathrm{MG}}, h^t_{\mathrm{TA}_i})^\top\big), \end{aligned} \tag{9}$$

其中，$e^t_{\mathrm{TA},\mathrm{MG}_i} \in \mathbb{R}^{1024\times 1024}$，$\top$是矩阵转置运算，$F_{iatt}(\cdot,\cdot)$测量其输入之间的一致性，具体如下：

$$F_{iatt}(h^t_{\mathrm{TA}}, h^t_{\mathrm{MG}_i}) = \left[\mathcal{R}_{28\times1024}\left(Conv_{1\times1}(h^t_{\mathrm{TA}})\right)\right]^\top \circledast \left[\mathcal{R}_{28\times1024}\left(Conv_{1\times1}(h^t_{\mathrm{MG}_i})\right)\right], \tag{10}$$

其中，$\circledast$表示矩阵乘法，$h \in \mathbb{R}^{28\times32\times32}$， 32是特征图的大小，28是视频类别总数（$c$），$\mathcal{R}_{28\times1024}(\cdot)$将其输入调整为大小为$28 \times 1024$的矩阵，而$Conv_{1\times1}$是典型的$1 \times 1$卷积。

实际上，边$\boldsymbol{e}^t_{\mathrm{TA,MG}_i}$是两个相邻节点之间的特征相似度。由于推理过程的主要功能是在节点之间交换/共享信息，并针对给定的学习目标制定一系列隐式原则，因此本文在每个推理阶段后更新所有节点的状态。本文使用以下等式制定针对目标节点（TA）的更新过程：

$$h^{t+1}_{\mathrm{TA}} \leftarrow \mathcal{R}_{28\times32^2}\left(h^t_{\mathrm{TA}} \odot \sigma\Big(\sum_{i\in\mathcal{N}_{\mathrm{TA}}} \mathcal{R}_{28\times1024}(h^t_{\mathrm{MG}_i}) \circledast \boldsymbol{e}^t_{\mathrm{TA,MG}_i}\Big)\right), \tag{11}$$

其中，$\sigma(\cdot)$是典型的sigmoid函数； $\circledast$表示矩阵乘法；$\odot$表示元素乘法运算； $\times$是标准乘法； $\mathcal{N}_{\mathrm{TA}}$包括与目标节点（TA）相邻的所有节点，而$\mathcal{R}_{28\times32^2}(\cdot)$将其输入整理为大小为$28 \times 32^2$的张量。类似地，针对多粒度节点（例如，$\mathrm{MG}_i$）的更新过程可以表述如下：

$$h^{t+1}_{\mathrm{MG}_i} \leftarrow \mathcal{R}_{28\times32^2}\Big(h^t_{\mathrm{MG}_i} \odot \sigma\Big(\sum_{j\in\mathcal{N}_{\mathrm{MG}_i}} \mathcal{R}_{28\times1024}(h^t_{\mathrm{MG}_j}) \circledast \boldsymbol{e}^t_{\mathrm{MG}_i,\mathrm{MG}_j} + \mathcal{R}_{28\times1024}(h^t_{\mathrm{TA}}) \circledast \boldsymbol{e}^t_{\mathrm{MG}_i,\mathrm{TA}}\Big)\Big). \tag{12}$$

通过使用节点推理过程，可以获得目标输入与其多粒度信息之间的高级语义信息。目标节点的最终状态h^m_{TA}（m是总推理步骤）将被输入到分类器中。实际上，目标节点的相应特征响应图嵌入了多源和多粒度信息，它们将在生成伪视点时使用。

4.3.2 视点精细化层

在大多数情况下，基于C-GCN的推理过程可以适当地融合多源和多粒度信息。然而，关于要处理的推理过程仍然存在一个问题。尽管推理步骤可以隐式地包含多粒度信息，但基于加法运算的更新过程（例如，公式 11）可能导致目标节点的特征响应图冗余，导致伪视点与真实视点不同。因此，对于每个推理阶段，本文额外分配了一个**fixation refine layer**（FRL，图 11），以消除冗余特征响应。本文将提供拟定FRL的技术细节如下。首先，本文建立一个二进制矩阵$\mathrm{rMASK} \in \{0,1\}^{32\times32}$，以指示融合特征张量中的哪些元素具有相对较大的特征响应，并且这些限定元素更有可能属于最具区分性的帧区域。 rMASK的计算公式如下：

$$\mathrm{rMASK} = \left[max\Big(\mathrm{cMean}\big(h^t_i\big)\Big) \times \mathcal{T}_d - \mathrm{cMean}\big(h^t_i\big)\right]_+, \tag{13}$$

其中，$\times$是标准乘法，$\mathrm{cMean}(\cdot)$是信道平均运算，将张量转换为矩阵，$max(\cdot)$为典型的最大运算，返回其输入矩阵中的最大值。h^t_i 是张量特征，可以通过等式 11或等式 12获得$[\cdot]_+$将其所有负元素转换为0，并将剩余的正元素转换为1；而$\mathcal{T}_d$是预定义的硬阈值。

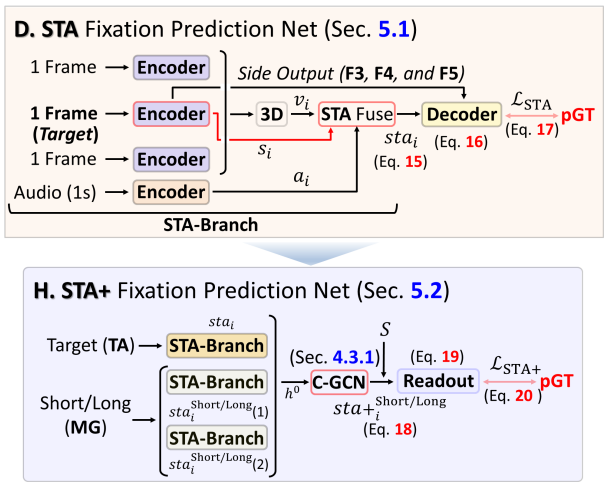


图 12. 全自动音视视点预测网络的体系结构，以伪视点为训练目标，在不使用视频标签的情况下实现通用视点预测。子图H为子图D配备了所提出的多粒度感知。STA融合模块的详细信息可参见图 6。

接下来，本文使用rMASK过滤h^t_i中的冗余信息，这个过程可以表述为：

$$h^{t'}_i = \frac{1}{2} \times \Big[h^t_i \odot \Big(\mathcal{T}_r \times \sigma\left(\mathrm{cMean}(h^t_i)\right) \odot \mathrm{rMASK} + \sigma\left(\mathrm{cMean}(h^t_i)\right) \odot (1 - \mathrm{rMASK})\Big) + h^t_i\Big], \tag{14}$$

其中，$\odot$是元素乘法运算，$\mathcal{T}_r$是由预定义值指定的精细化率。等式 14的基本原理可以解释如下：由于rMASK可以指示那些具有较大特征响应的空间区域，本文将其用作那些不太可信（即响应相对较低）区域的注意力过滤器（$\mathcal{T}_r$控制精细化率）——这些区域更可能是冗余的。

具体而言，提出的FRL可以进一步缓解特征冗余问题，使获得的伪视点与真实视点更加一致。 $\mathcal{T}_d$和$\mathcal{T}_r$ 的确切选择将在第 6.3.1节中讨论。

5 通用视点预测网络

通过上述SCAM+（公式 8），可以获得大量的伪音视视点图。虽然这些视点图通常与真实视点一致，但仍存在一个关键限制，即SCAM+的计算需要人工提供的视频标签，导致应用范围非常有限。因此，本文将基于SCAM+的伪音视视点图视为学习目标，以指导知识提取，因此，可以获得端到端视点预测模型，而无需使用任何人类提供的视频标签。这种策略可以使所提出的方法成为一种通用的视听视点预测工具。

5.1 基于多源的视点预测网络

如图 12-D所示，**spatial-temporal-audio**（STA）视点预测网络的实现非常直观，其中，空间特征（s_i）预先与时间特征（v_i）或音频特征（a_i）融合，然后通过最简单的特征串联操作进行组合。然后，使用具有3个反卷积层的典型解码器将从

“SA 融合”模块导出的特征图转换为视点图。“SA 融合”模块中的数据流可以表示为等式 15。

$$sta_i = Relu\Big[Cov\Big(\big(\sigma(\phi(a_i)) \odot s_i + s_i\big) \otimes \big(\sigma(v_i) \odot s_i + s_i\big)\Big)\Big], \tag{15}$$

其中，$\otimes$是典型的串联操作；$\phi(\cdot)$是建议的音频开关（第3.3章节），$Cov(\cdot)$表示1 × 1卷积；所有其他符号与等式 5中的符号相同。

sta_i将被输入到解码器中，解码器的输出（$\widehat{sta}_i$）可以公式化为：

$$\begin{aligned}\widehat{sta}_i = \uparrow(Ref(Ref(\mathrm{F3}) \otimes \uparrow(Ref(Ref(\mathrm{F4}) \otimes \\ \uparrow(Ref(Ref(\mathrm{F5}) \otimes sta_i)))))),\end{aligned} \tag{16}$$

其中，$\uparrow(\cdot)$是上采样操作，$Ref(\cdot)$将其所有输入细化为视点通道数（32），$\otimes$是特征级联操作，F3、F4和F5是与目标帧相关的编码器的侧输出。配备多尺度连接和信道关注的更强大的解码器可能会带来一些额外的性能增益，但充分证明这个问题超出了本文的主要主题，本文将把它留给未来的研究。对于训练过程，本文选择了两个典型的损失函数，即二元交叉熵损失（$\mathcal{L}_{\mathrm{BCE}}$） [71]和Kullback-Leibler散度损失（$\mathcal{L}_{\mathrm{KL}}$ [72]），且整体损失函数（$\mathcal{L}_{\mathrm{STA}}$）可以简单地表示为：

$$\mathcal{L}_{\mathrm{STA}} = \mathcal{L}_{\mathrm{BCE}}(\widehat{sta}_i, \mathrm{pGT}) + \mathcal{L}_{\mathrm{KL}}(\widehat{sta}_i, \mathrm{pGT}). \tag{17}$$

显然，所提出的STA视点预测网络的训练过程仅依赖于伪视点（pGT）；因此，它能够在没有任何类别标签的情况下对不可见的视听序列执行端到端视点预测。

5.2 基于多粒度的视点预测网络

如图 12-H所示，STA视点预测网络也可以通过将其与多粒度感知机制相结合来升级。本文将升级版命名为STA+，其输入也由3部分组成，这3部分是STA网络的输出，具有不同的输入数据。

在STA+网络中，本文将只额外考虑短期和长期信息，本文省略了交叉信息以确保其通用性。与前面提到的分类网类似，本文继续使用提出的C-GCN作为融合器。因此，本文独立训练2个STA+网络，其中，输入数据（3个部分）为{1目标+2短期}或{1目标+2长期}，这2个STA+网络中融合部分的数据流（见图 12-H）可以统一表示为以下等式：

$$\begin{aligned}sta{+}_i^{\mathrm{Short/Long}} &= \Big\{h^m_{\mathrm{TA}}, h^m_{\mathrm{MG}_1}, h^m_{\mathrm{MG}_2}\Big\} \\ &= \Omega\Big[sta_i,\ sta_i^{\mathrm{Short}}(1),\ sta_i^{\mathrm{Short}}(2)\Big] \\ &\quad or\ \ \Omega\Big[sta_i,\ sta_i^{\mathrm{Long}}(1),\ sta_i^{\mathrm{Long}}(2)\Big],\end{aligned} \tag{18}$$

其中，$sta{+}_i^{\mathrm{Short/Long}}$是$m$步进推理的输出，它包括大小为28 × 32 × 32的3个张量（即，$h^m_{\mathrm{TA}}, h^m_{\mathrm{MG}_1}, h^m_{\mathrm{MG}_2}$），这些张量与3个不同的节点相关（分别为1个TA节点和2个MG节点，见等式 11和等式 12）； $\Omega[\cdot]$表示第 4.3.1节中提到的C-GCN，$sta_i^{\mathrm{Long}}(1)$ 和 $sta_i^{\mathrm{Long}}(2)$ 是与2个不同的长期输入片段相关的2个输出。

“ReadOut”（读出）模块单独输入$sta{+}_i^{\mathrm{Short/Long}}$，然后输出大小为356 × 356的9张视点图。以目标节点为例，本文将$sta{+}_i^{\mathrm{Short/Long}}$缩写为$h^m_{\mathrm{TA}}$，其中，$m$是总推理步骤，“读出”的输出可表示为：

$$\hat{h}_{\mathrm{TA}} = RO(h^m_{\mathrm{TA}}) = Ref_1\Big(\uparrow\big(Ref_{32}(h^m_{\mathrm{TA}} \otimes S)\big)\Big), \tag{19}$$

其中，S表示与目标片段的中间帧相关的空间特征，例如，图 12-D中的s_i；RO表示图 12-H中的“读出”操作；$Ref(\cdot)_p$是一种细化操作，它使用1 × 1卷积将其输入的信道号减少到p； $\uparrow(\cdot)$将其输入向上采样为356 × 356。类似地，通过交替向STA+的“读出”模块输入$h^m_{\mathrm{MG}_1}$ 和 $h^m_{\mathrm{MG}_2}$，输出将变成$\hat{h}_{\mathrm{MG}_1}$ 和 $\hat{h}_{\mathrm{MG}_2}$

本文为每个视点预测网分配了一个单独的损失函数，这两个视点预测网采用的总损失函数（STA+$^{\mathrm{Short}}$和STA+$^{\mathrm{Long}}$，18）可以统一表示为：

$$\begin{aligned}\mathcal{L}_{\mathrm{STA+}} &= \mathcal{L}_{\mathrm{BCE}}(\hat{h}_{\mathrm{TA}}, \mathrm{pGT}_{\mathrm{TA}}) + \mathcal{L}_{\mathrm{KL}}(\hat{h}_{\mathrm{TA}}, \mathrm{pGT}_{\mathrm{TA}}) \\ &+ \mathcal{L}_{\mathrm{BCE}}(\hat{h}_{\mathrm{MG}_1}, \mathrm{pGT}_{\mathrm{MG}_1}) + \mathcal{L}_{\mathrm{KL}}(\hat{h}_{\mathrm{MG}_1}, \mathrm{pGT}_{\mathrm{MG}_1}) \\ &+ \mathcal{L}_{\mathrm{BCE}}(\hat{h}_{\mathrm{MG}_2}, \mathrm{pGT}_{\mathrm{MG}_2}) + \mathcal{L}_{\mathrm{KL}}(\hat{h}_{\mathrm{MG}_2}, \mathrm{pGT}_{\mathrm{MG}_2}),\end{aligned} \tag{20}$$

其中，$\mathrm{pGT}_{\mathrm{TA/MG_1/MG_2}}$表示不同节点的相应伪视点图；输入$\hat{h}_{\mathrm{TA}}$可以通过等式 19直接获得，并且$\hat{h}_{\mathrm{MG_1/MG_2}}$的计算应相应地将等式 19的输入替换为$h^m_{\mathrm{MG}_1}$ 和 $h^m_{\mathrm{MG}_2}$。

5.3 基于多粒度的音视视点预测

到目前为止，本文可以获得第 5.1和第 5.2节中提到的3个单独的视点预测网络，即，从多源角度来看，本文可以得到一个STA网络，从多粒度方面来看，还可以获得2个STA+网络，它们分别与短期或长期版本相关[1]（等式 18）。这3个视点预测网络在本质上相互补充，本文将结合它们的预测，得出最终的视点图，其性能可能优于它们中的每一个。

本文将3个预测网络（STA、STA+$^{\mathrm{Short}}$和 STA+$^{\mathrm{Long}}$）的预测视点图表示为PF_{sta}、$\mathrm{PF}^{\mathrm{Short}}_{sta+}$ 和 $\mathrm{PF}^{\mathrm{Long}}_{sta+}$。最终预测的视觉-音频视点图（$\mathrm{PF}_{final}$）可以公式化如下：

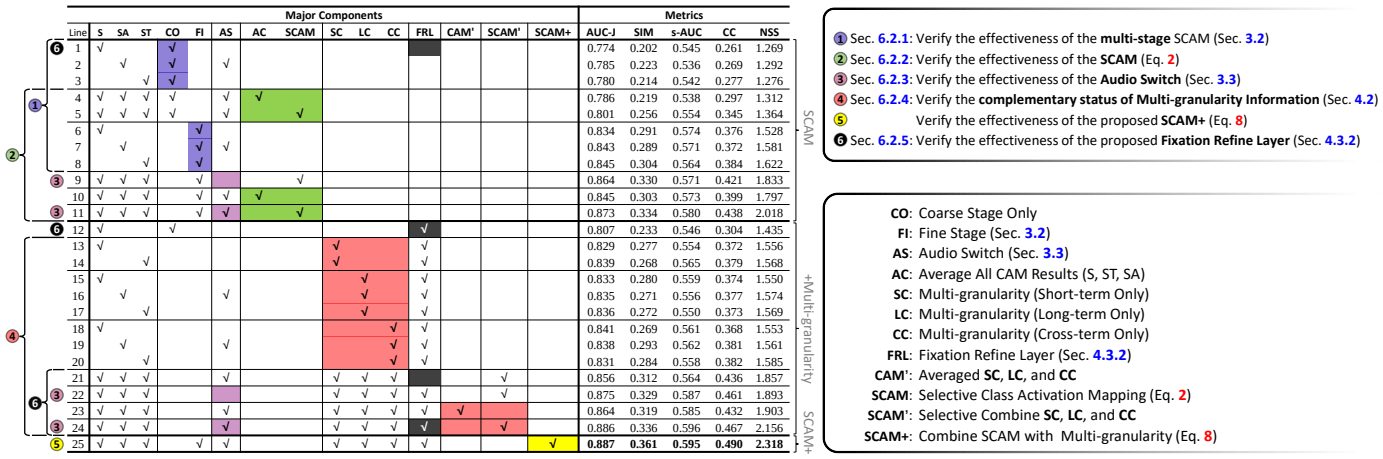
$$\begin{aligned}\mathrm{PF}_{final} &= \frac{1}{2} \times \mathcal{Z}\Big(\mathcal{Z}(\mathrm{PF}_{sta}) \odot \mathcal{Z}(\mathrm{PF}^{\mathrm{Short}}_{sta+}) \odot \mathcal{Z}(\mathrm{PF}^{\mathrm{Long}}_{sta+})\Big) \\ &+ \frac{1}{2} \times \mathcal{Z}\Big(\mathrm{PF}_{sta} + \mathrm{PF}^{\mathrm{Short}}_{sta+} + \mathrm{PF}^{\mathrm{Long}}_{sta+}\Big),\end{aligned} \tag{21}$$

其中，$\mathcal{Z}(\cdot)$是典型的*min-max*归一化函数，$\times$表示标准乘法运算。$\odot$表示元素乘法运算。等式 21主要由两部分组成，其中，左侧部分获得共同一致性，而右侧部分用作互补视点图，以确保良好的鲁棒性。

与传统的普通融合方案（基于相加/平均/最大值的融合）相比，所提出的融合方案可以使预测的视点图（即，PF_{final}）更符合真实的人类视点。

1. 本文省略了交叉版本，以确保在测试阶段具有良好的通用性。

表 1
STA和STA+的成分研究的定量证据，为了方便展示，本文用不同的颜色来区分不同的成分。本实验是在AVAD集上进行的 [73]。



	Major Components															Metrics				
Line	S	SA	ST	CO	FI	AS	AC	SCAM	SC	LC	CC	FRL	CAM'	SCAM'	SCAM+	AUC-J	SIM	s-AUC	CC	NSS
1	√			√												0.774	0.202	0.545	0.261	1.269
2		√		√		√										0.785	0.223	0.536	0.269	1.292
3			√	√												0.780	0.214	0.542	0.277	1.276
4	√	√	√	√		√	√									0.786	0.219	0.538	0.297	1.312
5	√	√	√	√		√		√								0.801	0.256	0.554	0.345	1.364
6	√				√											0.834	0.291	0.574	0.376	1.528
7		√			√	√										0.843	0.289	0.571	0.372	1.581
8			√		√											0.845	0.304	0.564	0.384	1.622
9	√	√	√		√			√								0.864	0.330	0.571	0.421	1.833
10	√	√	√		√	√	√									0.845	0.303	0.573	0.399	1.797
11	√	√	√		√	√		√								0.873	0.334	0.580	0.438	2.018
12	√			√								√				0.807	0.233	0.546	0.304	1.435
13	√								√			√				0.829	0.277	0.554	0.372	1.556
14			√						√			√				0.839	0.268	0.565	0.379	1.568
15	√									√		√				0.833	0.280	0.559	0.374	1.550
16		√				√				√		√				0.835	0.271	0.556	0.377	1.574
17			√							√		√				0.836	0.272	0.550	0.373	1.569
18	√										√	√				0.841	0.269	0.561	0.368	1.553
19		√				√					√	√				0.838	0.293	0.562	0.381	1.561
20			√								√	√				0.831	0.284	0.558	0.382	1.585
21	√	√	√			√			√	√	√			√		0.856	0.312	0.564	0.436	1.857
22	√	√	√						√	√	√	√		√		0.875	0.329	0.587	0.461	1.893
23	√	√	√			√			√	√	√	√	√			0.864	0.319	0.585	0.432	1.903
24	√	√	√			√			√	√	√	√		√		0.886	0.336	0.596	0.467	2.156
25	√	√	√		√	√			√	√	√	√			√	**0.887**	**0.361**	**0.595**	**0.490**	**2.318**

SCAM
+Multi-granularity
SCAM+

① Sec. 6.2.1: Verify the effectiveness of the **multi-stage** SCAM (Sec. 3.2)
② Sec. 6.2.2: Verify the effectiveness of the **SCAM** (Eq. 2)
③ Sec. 6.2.3: Verify the effectiveness of the **Audio Switch** (Sec. 3.3)
④ Sec. 6.2.4: Verify the **complementary status of Multi-granularity Information** (Sec. 4.2)
⑤ Verify the effectiveness of the proposed **SCAM+** (Eq. 8)
⑥ Sec. 6.2.5: Verify the effectiveness of the proposed **Fixation Refine Layer** (Sec. 4.3.2)

CO: Coarse Stage Only
FI: Fine Stage (Sec. 3.2)
AS: Audio Switch (Sec. 3.3)
AC: Average All CAM Results (S, ST, SA)
SC: Multi-granularity (Short-term Only)
LC: Multi-granularity (Long-term Only)
CC: Multi-granularity (Cross-term Only)
FRL: Fixation Refine Layer (Sec. 4.3.2)
CAM': Averaged **SC**, **LC**, and **CC**
SCAM: Selective Class Activation Mapping (Eq. 2)
SCAM': Selective Combine **SC**, **LC**, and **CC**
SCAM+: Combine SCAM with Multi-granularity (Eq. 8)

6 实验和验证

6.1 实施细节

训练集

最近，谷歌发布了Audioset [74]，这是迄今为止最大的音视集，本文使用其子集**a**udio **v**isual **e**vent（AVE） [30][2]定位数据集，数据集包含4143个序列，涵盖28个语义类别，作为S、ST、SA、S+、ST+和SA+分类网络的分类训练集（第3和第 4章节）。

训练过程

对于 S、ST、SA、S+、ST+ 和 SA+ 训练，由于计算机内存的限制，本文分别训练网络的这些分支，并使用 SCAM 和 SCAM+ 将它们融合成最终的伪 GT。对于SCAM训练，本文遵循广泛使用的多阶段训练方案。在粗略阶段，批次大小等于 20，所有视频帧都调整为 256×256。以裁剪后的视频块作为输入，将训练三个全新的精细阶段分类网络，其中，批次大小等于 3，所有视频块的大小调整为 356×356。对于SCAM+训练，本文将批量大小设置为 16，并将所有视频帧的大小调整为 256×256，并且 S+、ST+ 和 SA+多粒度推理网络使用相同的实验设置。所有分类网络都在 AVE数据集上进行训练。 STA 和 STA+ 视点点预测网络将伪视点作为 GT，其中，视频帧被调整为 356×356；因此，批次大小设置为 3。STA 和 STA+ 训练过程采用了 **s**tochastic **g**radient **d**escent（SGD） [75] 优化器，具有学习率0.00005。

测试集

为了测试所提出方法的性能，本文采用了 6 个测试数据集，包括 **AVAD** [73]、**Coutrot1** [76]、**Coutrot2** [77]、**DIEM** [78]、**SumMe** [79] 和 **ETMD** [80]。所有这些数据集（241 个序列）都配备了在视听环境中收集的像素级真实视点。

定量指标

2. https://sites.google.com/view/audiovisualresearch

继之前对显著性度量的研究 [81]， [82]， [72]， [83]，本文采用了5个常用的评估指标来衡量模型预测的显著性（PS）图与真实人眼运动（CF，FL），预测显著性图 PS $\in [0,1]$ 的范围，连续视点图 CF $\in [0,1]$ 和视点图 FL $\in \{0,1\}$。它主要包括基于位置的度量 **AUC-J**、**s-AUC**、**NSS**，以及基于连续分布的度量 **SIM**、**CC**。

6.2 不同组件的有效性评估

6.2.1 本文的多阶段的有效性

在 Sec. 3.2 中，本文采用从粗到细的原理来执行两次 SCAM，其主要目标是模仿真正的人类注意力机制——多阶段机制。为了验证所提出的从粗到细策略的有效性，本文在不同阶段测试了来自不同数据源（即，S、SA和ST）的所有 CAM 结果，即COARSE与FINE。对应的定量评价见表 1，用标记①表示。本文已经测量了使用不同数据源的CAM结果与真实的人眼视点数据之间的一致性程度，使用FINE阶段和不使用FINE阶段之间的差异是显著的。在使用单一数据源（S、SA 或 ST，参见第 1-3 行和第 6-8 行）的情况下，CAM 结果可以提高约 40%，例如，CC 度量值在仅使用空间源（S）的情况下，从 2.61 增加到 3.76。在使用多个数据源的情况下也可以观察到类似的趋势，例如，通过比较第 4 行和第 10 行，所有指标都可以获得显著的改进，表明简单融合的多源 CAM 结果也可以通过提出的多阶段策略得到改进。此外，从提出的 SCAM 的角度来看，还可以注意到多阶段策略可以带来可靠的性能增益，例如，第 5 行与第 11 行，这表明了多阶段策略的通用性。

6.2.2 本文的选择性融合的有效性

选择性融合多源 CAM 结果（即、Φ_S、Φ_{SA} 和 Φ_{ST}，公式 2），本文采用分类置信度作为融合权重（C_S、C_{SA} 和 C_{ST}）。实际上，这种实现的有效性是基于分类置信度确实与 CAM

表 2
对SCAM 有效性的定量证据。本实验是在 AVAD 数据集上进行的 [73]。

	Module	AUC-J ↑	SIM↑	s-AUC↑	CC↑	NSS↑
COARSE	CAM_{sta}	0.793	0.227	0.551	0.293	1.273
	SCAM	0.801	0.256	0.554	0.345	1.364
FINE	CAM_{sta}	0.856	0.296	0.579	0.415	1.803
	SCAM	**0.873**	**0.334**	**0.580**	**0.438**	**2.018**

结果和真实视点之间的一致性水平正相关的前提条件。为了验证这个问题，本文比较了提出的选择性融合（即，本文的 SCAM，公式 2）和“传统融合方案”——对所有数据源的CAM 结果进行平均，可以表示为以下等式。

$$\mathrm{AC} = \frac{1}{3}\left(\Phi_{\mathrm{S}} + \Phi_{\mathrm{SA}} + \Phi_{\mathrm{ST}}\right), \tag{22}$$

其中，所有符号与公式 2 的定义相同，本文在表 1 的“AC” 列中显示了该方案的相应定量结果。

表 1 中的标记 ② 突出了所提出的 SCAM 相对于传统方案（即，SCAM vs. AC）的优势，整体定量性能可以获得所有考虑的指标的持续改进。同时，与其他单数据源 CAM 结果（表 1中展示的S、ST 和 SA）相比，可以很容易地观察到，同时考虑所有数据源并不能充分利用它们之间的互补性；因此，AC 实现的性能改进是微不足道的，进一步显示了所提出的 SCAM 的性能优势和有效性。

此外，为了进一步验证所提出的选择性融合所采用的分治方案的有效性，本文将 SCAM 与直接从三分支时空音频分类网络获得的 CAM 结果进行了比较。本文使用‘CAM_{sta}’来表示从这个三分支网络获得的 CAM 结果，本文简单地采用一个三分支网络，采用与所提出的 STA 视点预测网络相同的融合单元（图 12-D），而解码器部分已替换为与其他分类网络相同的多类分类器。定量比较结果见表 2，其中，对于粗略和精细阶段，所提出的 SCAM 都可以优于 CAM_{sta}。

6.2.3 提出的音频开关的有效性

在表 1 中，本文报告了不使用建议的 **audio switch**（Sec. 3.3）的模型的性能，相应的定量结果已由标记 ③ 突出显示。

通过比较第 9 行和第 11 行，以及第 22 行和第 24 行，可以很容易地观察到，所提出的音频开关能够平均提高整体性能约 1.5%。主要原因是它可以过滤无意义的背景音频，减轻融合不同步的空间和音频信息时的学习歧义。

此外，由于本文的音频开关是本文提出的“SA融合”模块（Sec. 3.3）的主要组成部分，本文将另外验证SA融合模块是否可以优于传统的融合方案。因此，本文进行了额外的定量评估，在表 3 中，本文将提出的SA融合与现有简单融合的进行了比较，包括加法、乘法、连接和双线性。正如预期的那样，所提出的SA融合模块明显优于所有其他方法。

表 3
对所提出的“SA Fuse”模块有效性的定量证据，“Addition, Multiplication, Concatenation, Bilinear”是视觉和音频信息融合模式，“SA Fuse”是本文提出的音视融合模块。本实验是在 AVAD 数据集上进行的 [73]。

	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
Addition	0.857	0.325	0.581	0.438	1.880
Multiplicative	0.860	0.327	0.575	0.442	1.915
Concatenation	0.873	0.331	0.585	0.459	1.944
Bilinear	0.864	0.329	0.577	0.428	1.825
SA Fuse	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**

6.2.4 多粒度选择性融合的有效性

为了研究提出的多粒度信息对本文的 SCAM 的影响，本文测试了来自不同数据源（即：S、SA和ST）的所有 CAM，以及多粒度信息 即，分别为短期（SC）、长期（LC）和交叉（CC）信息。

如表 1 所示，用 ④ 标记，本文发现使用每种类型的多粒度信息和多源CAM结果上的性能提升。同时，本文还同时使用所有类型的多粒度信息进行了测试，本文使用 CAM′ 和 SCAM′ 分别表示平均和选择性融合的 SC、LC 和 CC。

显然，与通过单一数据源获得的 CAM 结果（第 12 行）相比，多粒度信息（第 13 行）可以带来可靠的性能增益。但是，简单地将基于单数据源的 CAM 结果与一种类型的多粒度信息结合起来可能仍然不如 SCAM，这可以通过比较第 11 行和第 13 行来确认。主要原因是，与多数据源相比信息，多粒度信息可能对分类任务不太重要，激励本文使用它作为下属服务于 SCAM。

正如预期的那样，由于 CAM′（第 23 行）采用了多个多粒度信息，因此与基于单个多粒度信息的信息（例如，第 13-20 行）相比，整体性能可以得到提高。此外，得益于选择性融合，SCAM′（第 21、22 和 24 行）可以进一步提高整体性能。在结合完整的数据源信息和完整的多粒度信息后，所提出的以 ⑤ 标记的 SCAM+ 实现了最佳性能。

6.2.5 视点细化层的有效性

为了验证所提出的 **fixation refine layer** (FRL) 的有效性，本文在表 1 中进行了一系列评估，以及相应的结果由 ⑥ 标记。

首先，如第 12 行所示，本文将 FRL 应用于基于空间源的粗阶段特征，其中，FRL 直接应用于 S_i（图 5-A ），然后生成基于 FRL 的 CAM 结果。比较第 12 行和第 1 行，可以很容易地观察到明显的性能提升。

其次，本文还尝试从 SCAM′ 中删除 FRL，以验证性能差距。通过比较第 21 行和第 24 行，可以很容易地注意到性能明显下降，例如，AUC-J 指标从 0.886 下降到 0.856，这两个结果之间的唯一差异仅取决于是否使用FRL。

表 4
FRL（帧细化层）有效性的定量证据，$\mathcal{T}_d$ 是选择细化操作与否的阈值，$\mathcal{T}_r$ 是精细化抑制率。

Dataset		AVAD [73]			DIEM [78]			SumMe [79]		
$\mathcal{T}_d$	$\mathcal{T}_r$	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
w/o. FRL		0.856	0.436	1.857	0.835	0.482	1.799	0.844	0.330	1.598
0.8	0.4	0.860	0.426	1.671	0.859	0.504	1.832	0.849	0.335	1.602
0.8	0.8	0.843	0.393	1.802	0.844	0.511	1.815	0.853	0.341	1.609
0.9	0.6	0.868	0.421	1.790	0.865	0.498	1.856	0.848	0.332	1.611
0.7	0.6	0.849	0.425	1.699	0.853	0.487	1.862	0.841	0.340	1.600
0.8	**0.6**	**0.886**	**0.467**	**2.156**	**0.877**	**0.518**	**1.955**	**0.860**	**0.367**	**1.633**
Dataset		ETMD [80]			Coutrotl [76]			Coutrot2 [77]		
$\mathcal{T}_d$	$\mathcal{T}_r$	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
w/o. FRL		0.877	0.402	1.891	0.815	0.301	1.302	0.832	0.343	1.988
0.8	0.4	0.882	0.409	1.915	0.812	0.303	1.309	0.829	0.329	1.990
0.8	0.8	0.881	0.411	1.923	0.824	0.308	1.311	0.826	0.340	1.984
0.9	0.6	0.875	0.399	1.921	0.829	0.309	1.309	0.834	0.337	1.988
0.7	0.6	0.901	0.420	1.952	0.825	0.312	1.315	0.840	0.343	1.992
0.8	**0.6**	**0.903**	**0.423**	**2.056**	**0.833**	**0.330**	**1.326**	**0.858**	**0.371**	**2.105**

6.3 消融实验

6.3.1 *FRL*使用的阈值

在本文的 FRL 中存在 2 个预定义阈值，即，在公式 13 和公式 14中的$\mathcal{T}_d$ 和 $\mathcal{T}_r$，这会直接影响整体性能，本文将对确切的选择进行消融研究。本文使用不同的 $\{\mathcal{T}_d, \mathcal{T}_r\}$ 组合多次重新训练 STA+ 网络，对应的结果可以在表 4中找到，第一行表示不使用 FRL 的结果。

由于需要确定 2 个阈值，本文将采用一个固定住另一个求解的策略，其中根据经验分配 $\mathcal{T}_d = 0.8$ 和 $\mathcal{T}_r = 0.6$ 作为本文的初始选择，本文已经测试了其他几个选择。巧合的是，最初的选择变成了最好的，主要原因可能是其他主要组件的技术细节都是基于这个最初的选择来实现的，这使得它被动地优于所有其他选择。

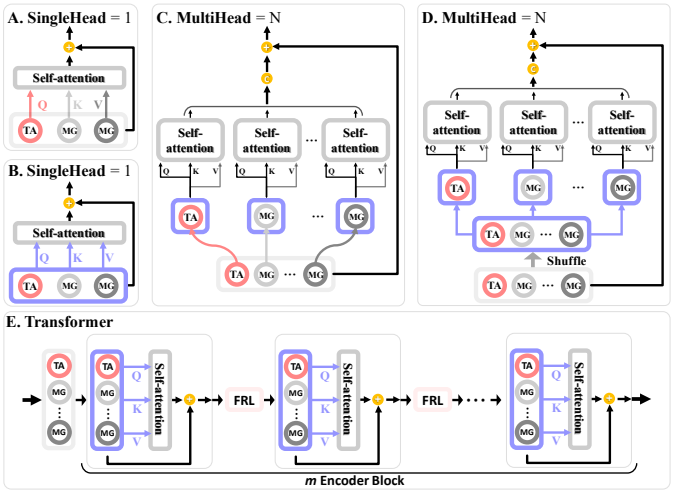


图 13. Transformer的详细数据流程采用SingleHead=1和MultiHead=N作为关系建模的输入。

表 5
AVAD数据集上语义推理阶段数（m, Sec. 4.3.1）和MG节点数（n, Sec. 4.3.1）的消融研究 [73]。 *w/o* Rea.：没有推理。本文测试了多个 GCN 已被 RNN 或 Transformer 替换的变体

			AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
GCN	*w/o* Rea.		0.825	0.253	0.556	0.338	1.519
	Reasoning Stage=m	m=2	0.873	0.316	0.575	0.442	1.980
		m=3	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**
		m=4	0.872	0.324	0.580	0.451	1.981
		m=5	0.870	0.309	0.576	0.439	1.967
	TA=1, MG=n	n=1	0.871	0.319	0.580	0.432	1.965
		n=2	**0.886**	**0.336**	**0.596**	**0.467**	**2.156**
		n=3	0.867	0.306	0.578	0.424	1.946
		n=4	0.852	0.289	0.565	0.415	1.863
RNNs			0.875	0.319	0.562	0.441	1.983
Transformer	**A**-SingleHead Realization		0.848	0.306	0.548	0.425	1.951
	B-SingleHead Realization		**0.882**	**0.339**	**0.599**	**0.464**	**2.157**
	C- MultiHead Realization		0.872	0.328	0.573	0.440	1.975
	D- MultiHead Realization		0.861	0.320	0.559	0.438	1.973
Version-**B**	*with* 2 Encoder Blocks		0.875	0.335	0.581	0.460	2.155
	with 3 Encoder Blocks		**0.882**	**0.339**	**0.599**	**0.464**	**2.157**
	with 4 Encoder Blocks		0.876	0.330	0.584	0.453	2.164
	with 5 Encoder Blocks		0.874	0.327	0.576	0.439	1.996

Modules	FLOPs↓	Params↓
RNNs	347,504,640.0	113,120.0
(**B**)SingleHead Realization 2	1,071,673,344.0	1,043,376.0
C-GCN (TA=1, MGNodes=2)	189,923,328.0	19,768.0

6.3.2 推理步数和节点数

为了将多粒度信息与提出的 SCAM 相结合，本文设计了推理单元（Sec. 4.3.1），其存在 2 个可能影响整体性能的关键参数，即，确切的推理步骤（m）和多粒度节点数（n）。

与之前的实验设置类似，本文测试了使用 SCAM+ 作为伪 GT 对数据进行训练的各种版本的 STA+，本文将 m 和 n 的范围分别设置为 $\{2, 3, 4, 5\}$和 $\{1, 2, 3, 4\}$。对应的定量结果见表 5。

从这张表中可以发现‘w/o Reasoning’表示简单地通过特征连接来组合目标节点和其他 MG 节点。由于这些连接的特征基本上没有很好地对齐，这个简单的实现已经证明了最糟糕的结果。然后，通过执行 3 次推理，本文的方法达到了最佳性能，如果本文继续推理步骤超过 3 次，整体性能可能会显著下降。主要原因是推理步数通常决定了网络的学习能力；给定的训练集大小有限，容量和数据大小之间应该有一个平衡点，$m = 3$ 是平衡点。

关于确切的 MG 节点编号，可以很容易地注意到，性能随着 MG 节点的增加而增加（1 →2），而如果本文使用太多的 MG 节点（3 →4），整体性能可能会下降。实际上，随着 MG 节点数量的增加，节点关系的特征空间会变得更加复杂，使得学习任务变得相当困难。虽然 MG 节点可以使 SCAM+ 的计算过程与真实的人类注意力机制更加一致，为了保持网络的精确性，因此本文选择$n = 2$ 作为最优选择。

本文测试了多个变体，他们的GCN 已被 RNN 或 Transformer 替换。如表 5 中所报告的，基于 Transformer 的变体

表 6
在 6 个测试数据集上生成通用视点图的不同融合方案的定量比较。

	AVAD [73]			DIEM [78]			SumMe [79]		
	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
PF_{sta}	0.873	0.438	2.018	0.861	0.469	1.716	0.854	0.368	1.647
$\mathrm{PF}_{sta+}^{\mathrm{Short}}$	0.881	0.465	2.151	0.872	0.486	1.843	0.855	0.331	1.466
$\mathrm{PF}_{sta+}^{\mathrm{Long}}$	0.886	0.469	2.159	0.874	0.501	1.896	0.860	0.383	1.650
PF_{agg}	0.886	**0.507**	**2.500**	0.862	0.500	1.920	0.844	0.359	1.593
PF_{final}	**0.887**	0.490	2.318	**0.884**	**0.544**	**2.042**	**0.866**	**0.402**	**1.761**
	ETMD [80]			Coutrotl [76]			Coutrot2 [77]		
	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
PF_{sta}	0.908	0.448	2.176	0.829	0.339	1.376	0.850	0.273	1.475
$\mathrm{PF}_{sta+}^{\mathrm{Short}}$	0.895	0.381	1.850	0.829	0.305	1.232	0.847	0.272	1.883
$\mathrm{PF}_{sta+}^{\mathrm{Long}}$	0.904	0.437	2.101	0.832	0.327	1.355	0.854	0.332	1.920
PF_{agg}	0.898	0.434	2.192	0.814	0.314	1.270	0.803	0.341	**1.952**
PF_{final}	**0.910**	**0.452**	**2.297**	**0.840**	**0.354**	**1.429**	**0.862**	**0.349**	**1.952**

中性能最好的版本与本文基于 C-GCN 的版本处于相同的性能水平，而基于 RNNs 的版本显然不如本文基于 C-GCN 的版本。由于本文已经测试了多个基于 Transformer 的变体，本文在图 13 中详细介绍了它们的实现/架构。在本文的实现中，本文分别测试了单头和多头 Transformer，每一个都有两个不同的实现，即，{**A**， **C**} *v.s.* {**B**， **D**}，其中 **C** 是 **A** 的 MultiHead 版本，**D** 是 **B** 的 MultiHead 版本。介绍了验证编码器块数量的消融研究（图 13-**E**）。如表 5 中所报告的，具有三个编码器块的基于单头 Transformer 的变体的性能与具有三步推理的基于 C-GCN 的模型相当。 Transformer 可能需要额外的学习参数 即，几乎是 C-GCN 的 52 倍模型大小和 5.6 倍 FLOPs。

6.3.3 最终视点预测网络的融合方案

正如本文在 Sec. 5.3 中所述，本文提出了一种新的融合方案来组合由 3 个不同网络预测的 3 个视点图，即，基于多源的 PF_{sta}以及基于多粒度感知的 $\mathrm{PF}_{sta+}^{\mathrm{Short}}$ 和 $\mathrm{PF}_{sta+}^{\mathrm{Long}}$。所提出的融合方案的基本原理仅仅是基于这三个视点图之间的互补事实，因此，本文从定量的角度验证了这种融合方案相对于传统简单方案的优势。

综合定量评价结果见表 6。与 3 个基本视点图（即，前 3 行）相比，简单的融合方案无法充分利用它们的互补属性，因此，性能增益往往会在不同方式之间转移，并且确实是微不足道的。

具体来说，由于提出的 PF_{final} 是基于对其两个部分（公式 21）执行加法运算，本文还测试了基于乘法的替代方案，记为PF_{agg}，可以表述为：

$$\mathrm{PF}_{agg} = \frac{1}{2} \times \mathcal{Z}\Big(\mathcal{Z}(\mathrm{PF}_{sta}) \odot \mathcal{Z}(\mathrm{PF}_{sta+}^{\mathrm{Short}}) \odot \mathcal{Z}(\mathrm{PF}_{sta+}^{\mathrm{Long}})\Big) \times \frac{1}{2} \times \mathcal{Z}\Big(\mathrm{PF}_{sta} + \mathrm{PF}_{sta+}^{\mathrm{Short}} + \mathrm{PF}_{sta+}^{\mathrm{Long}}\Big), \tag{23}$$

其中，所有符号和操作都与公式 21 的完全一样。

从表 6 中可以看出，本文发现 PF_{agg} 在某些情况下的表现优于 PF_{final}，数值差距是微不足道的。与此形成鲜明对比的是，PF_{final} 在许多情况下可以显著优于 PF_{agg}，例如，SunMe 数据集中的 CC 指标，$0.359 \rightarrow 0.402$。

6.4 与SOTA工作的定量比较

本文在 6 个公开可用的测试集上比较了本文仅使用伪视点训练的模型STANet+，与 27 种其他 SOTA 方法，包括 5 种无监督方法、18 种弱监督方法和 4 种完全监督方法。

与无监督和弱监督方法的定量比较。

如表 7 所示，本文的方法（STANet+）明显优于所有无监督方法，并且也优于最近的弱监督竞争对手（例如，MWS [31] 和 WSSA [42]）。使得弱监督模型输出接近视点的小区域的最直接方法是使用阈值。结果可以在表 8 中看到。事实是，简单地通过阈值缩小显著区域是无法获得高质量的视点图的。主要原因是弱监督方法生成的显著性图与实际视点**理论上**是不同的。

不同于传统的基于视频的 CAM 方法 [35]， [36] 倾向于在物体级别持续突出单个显著物体，本文的方法突出显示的帧区域级别信息——最具区分性的信息可能会有所不同在视频中逐帧播放，因为在视听环境中，空间、时间或音频都可能对分类任务贡献最大。该属性与真实的人类注意力机制高度一致，因为人类在视听环境下不会长时间关视点位置。

与完全监督方法的定量比较。

如表 7 所示，本文的方法在除 Coutrot2 之外的所有测试集中取得了与完全监督方法相当的结果，甚至优于其中一些方法，例如，DeepNet [19]。主要原因是 Coutrot2测试集的语义内容与 AVE 集的语义内容有很大不同，而本文的模型受到 AVE 集的类别标签的弱监督。请注意，本文的方法的性能可以通过包含更多标记的视听序列来进一步提高。

为了公平比较，本文的方法和其他三种具有代表性的完全监督 SOTA 方法 [81]， [71]， [19] 都在相同的训练集上进行训练，即广泛使用的 168 个视频片段（70%从六个音视集获得的总 241 个序列的 [73]， [76]， [77]， [78]， [79]， [80]，所有这些片段都配备了真实的视点。在本文方法的训练中，本文省略了真实视点的使用，取而代之的是，本文为每个片段分配了适当的视频标签。表 9 中报告的结果表明本文的方法在如此小规模的数据集上表现不佳，并且其他 SOTA 方法也会退化。

如表 7 所示，本文可以通过使用 VggSound 数据集 [61]（STANet+‡）获得一些性能提升，但这种提升是以相对较大的数据量和计数据计算为代价。主要原因是 VggSound 中的大部分视频与采用的测试集完全无关。

局限性

在本文的实现中，本文只为每个视听序列考虑了一个语义标签，而在实践中，可以为一个视频序列分配多个标签。因此，对于具有大量超出范围的语义内容的视频，本文的方法可能无法很好地执行。这个问题可以通过包含更多带有多个标签的数据来缓解，这需要未来的研究。

表 7

本文的方法与其他完全/弱/无监督方法在 6 个数据集上的定量比较。 **Bold** 表示最佳结果，STANet+‡ 表示在 VggSound 数据集 [61]上训练的STANet+。此外，本文还提供了本文的方法与 SOTA 方法之间的一些代表性的'**定性比较**'，可以在提交的'**补充材料**'中找到。这里 max J、S、A、C 和 N 分别表示最大 AUC-J、SIM、s-AUC、CC 和 NSS。

Dataset	AVAD [73]					DIEM [78]					SumMe [79]					ETMD [80]					Coutrotl [76]					Coutrot2 [77]				
Methods	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N↑	J↑	S↑	A↑	C↑	N
ITTI [84]*	.688	.170	.533	.131	0.61	.663	.217	.583	.137	0.56	.666	.151	.559	.097	0.44	**.856**	.226	.613	**.299**	**1.40**	**.798**	**.253**	.526	**.272**	**1.06**	.819	**.189**	.577	**.183**	1.07
GBVS [85]*	**.854**	.247	.572	**.337**	**1.56**	**.830**	**.318**	.605	**.356**	**1.28**	**.808**	.221	.567	**.272**	**1.13**	**.856**	.226	.613	**.299**	**1.40**	**.798**	**.253**	.526	**.272**	**1.06**	.819	**.189**	.577	**.183**	1.07
SCLI [86]*	.747	.210	.535	.170	0.79	.739	.267	.590	.207	0.78	.746	.209	.577	.184	0.80	.761	.165	.570	.129	0.62	.754	.216	.536	.239	0.88	.669	.137	.510	.014	0.09
SBF [87]*	.833	**.272**	.576	.308	1.49	.759	.292	.608	.301	1.08	.783	**.228**	.590	.230	1.02	.805	**.232**	.641	.262	1.30	.726	.187	.530	.215	0.79	**.827**	.152	.583	.131	**1.10**
AWS-D [88]*	.825	.221	**.589**	.304	1.38	.733	.250	**.612**	.301	1.13	.747	.192	**.603**	.186	0.85	.754	.161	**.664**	.181	0.91	.729	.214	.581	.207	0.87	.783	.170	**.590**	.146	0.84
CAM [89]#	.743	.195	.542	.217	0.93	.730	.248	.589	.284	0.84	.744	.184	.570	.201	0.80	.715	.145	.556	.140	0.63	.680	.196	.508	.151	0.55	.482	.126	.425	.029	0.16
GCAM [90]#	.743	.196	.542	.217	0.93	.730	.249	.589	.236	0.84	.744	.185	.570	.201	0.80	.715	.146	.555	.141	0.63	.680	.197	.508	.151	0.55	.482	.126	.425	.030	0.16
GCAMpp [35]#	.777	.273	.559	.255	1.22	.732	.216	.583	.271	0.78	.774	.217	.593	.225	0.92	.575	.124	.157	.576	0.74	.704	.137	.537	.210	0.51	.733	.114	.567	.168	0.63
SGCAMpp [91]#	.809	.206	.550	.275	1.18	.802	.271	.620	.319	1.12	.786	.191	.591	.234	0.94	.791	.162	.599	.212	0.95	.750	.217	.523	.220	0.84	.618	.155	.449	.061	0.26
xGCAM [92]#	.743	.196	.543	.217	0.93	.730	.249	.589	.236	0.84	.743	.184	.570	.199	0.79	.715	.146	.555	.141	0.63	.680	.197	.508	.151	0.55	.482	.126	.425	.030	0.16
SSCAM [93]#	.777	.186	.531	.228	0.97	.750	.242	.604	.248	0.87	.763	.179	.591	.206	0.83	.730	.144	.595	.149	0.68	.686	.194	.521	.154	0.59	.502	.141	.417	.003	0.02
ScoCAM [94]#	.772	.196	.548	.237	1.02	.770	.257	.614	.279	0.98	.753	.182	.577	.202	0.80	.737	.148	.581	.157	0.70	.708	.203	.518	.176	0.67	.538	.143	.423	.018	0.07
LCAM [95]#	.776	.199	.542	.241	1.03	.773	.259	.616	.285	1.01	.778	.189	.593	.228	0.91	.749	.151	.581	.168	0.75	.699	.201	.516	.168	0.62	.511	.141	.400	.003	0.03
ISCAM [96]#	.774	.195	.545	.240	1.03	.774	.256	.619	.282	0.99	.761	.183	.582	.208	0.90	.738	.147	.585	.157	0.71	.704	.201	.520	.171	0.65	.480	.135	.409	.018	0.12
ACAM [97]#	.759	.198	.539	.231	0.98	.735	.247	.607	.245	0.88	.756	.183	.584	.209	0.84	.722	.145	.571	.147	0.66	.677	.198	.516	.154	0.55	.484	.130	.465	.030	0.17
EGCAM [98]#	.737	.222	.533	.212	0.91	.758	.310	.618	.308	1.10	.741	.220	.583	.215	0.87	.687	.156	.570	.124	0.58	.640	.193	.509	.114	0.41	.575	.107	.425	.031	0.18
ECAM [98]#	.725	.219	.526	.205	0.88	.740	.294	.605	.273	0.97	.727	.211	.570	.198	0.79	.683	.151	.555	.116	0.54	.647	.200	.508	.130	0.46	.610	.104	.415	.037	0.22
SPG [99]#	.662	.176	.506	.165	0.73	.713	.238	.579	.233	0.86	.714	.182	.561	.209	0.91	.695	.138	.550	.144	0.69	.650	.187	.505	.142	0.53	.511	.123	.464	.017	0.07
VUNP [36]#	.574	.067	.500	.142	0.29	.558	.047	.515	.172	0.19	.555	.013	.507	.114	0.05	.505	.030	.103	.132	0.59	.589	.063	.514	.152	0.30	.661	.101	.536	.162	0.49
WSS [32]#	.858	.292	.592	.347	1.66	.803	.333	.620	.344	1.29	.812	.245	.589	.279	1.10	.854	.277	.661	.334	1.65	.772	.247	.547	.233	0.98	.835	.208	.578	.192	1.18
MWS [31]#	.834	.272	.573	.309	1.48	.806	.336	.628	.350	1.31	.808	.237	.607	.258	1.16	.833	.237	.649	.293	1.43	.743	.231	.528	.201	0.80	.839	.188	.581	.168	1.20
WSSA [42]#	.807	.261	.574	.285	1.34	.767	.305	.608	.311	1.18	.755	.225	.585	.231	1.06	.793	.201	.622	.222	1.08	.701	.180	.535	.169	0.78	.797	.185	.571	.180	1.26
STANet [1]#	.873	.334	.580	.438	2.02	.861	.391	.658	.469	1.72	.854	.294	.627	.368	1.65	.908	.318	.682	.448	2.18	.829	.306	.542	.339	1.38	.850	.247	.597	.273	1.48
STANet+#	.887	.361	.595	.490	2.32	.884	**.436**	.679	.544	2.04	.866	**.323**	.634	.402	1.76	.910	.328	.683	.452	2.30	.840	.315	.552	.354	1.43	.862	.267	**.612**	**.349**	1.95
STANet+‡#	**.892**	**.366**	**.603**	**.508**	**2.43**	**.887**	.433	**.687**	**.558**	**2.13**	**.868**	.319	**.642**	**.417**	**1.86**	**.915**	**.345**	**.699**	**.493**	**2.48**	**.843**	**.315**	**.555**	**.370**	**1.54**	**.871**	**.269**	.609	.348	**1.97**
DeepNet [19]†	.869	.256	.561	.383	1.85	.832	.318	.622	.407	1.52	.848	.227	.645	.332	1.55	.889	.225	.699	.387	1.90	.824	.273	.559	.346	1.41	.896	.201	.600	.301	1.82
SalGAN [71]†	.886	.360	.579	.491	2.55	.857	.393	**.660**	.486	1.89	**.875**	.289	**.688**	**.397**	**1.97**	.903	.311	**.746**	.476	2.46	**.853**	.332	**.579**	.416	1.85	**.933**	.290	.618	.439	2.96
DeepVS [100]†	.896	.391	**.585**	.528	3.01	.840	.392	.625	.452	1.86	.842	.262	.612	.317	1.62	.904	**.349**	.686	.461	**2.48**	.830	.317	.561	.359	1.77	.925	.259	**.646**	**.449**	**3.79**
ACLNet [81]†	**.905**	**.446**	.560	**.580**	**3.17**	**.869**	**.427**	.622	**.522**	**2.02**	.868	**.296**	.609	.379	1.79	**.915**	.329	.675	**.477**	2.36	.850	**.361**	.542	**.425**	**1.92**	.926	**.322**	.594	.448	3.16

与其他全监督方法相比，本文的方法采用了不同的训练方式，SOTA全监督方法在广泛使用的带有人眼注视训练集（2,857个视频 [81], [101], [102]）上进行了训练，然而，本文的方法仅在 AVE 集上进行训练，AVE 集是 AudioSet [74] 的子集，包含 4,143 个配备语义标签的片段。

注意: */#/†代表无-/弱-/全-监督模型。

表 8

不同“阈值”对MWS [31]、WSSA [42]、WSS [32]、ScoCAM [94]、LCAM [95]和SGradCAMpp [91] 显著性检测结果的影响，“0.0’ 表示对显著性结果不应用阈值，“0.3’ 表示将小于 0.3 的显著性值设置为 0。

Methods	WSSA [42]			MWS [31]			WSS [32]		
Metrics	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
0.0	**0.807**	**0.285**	1.339	**0.834**	**0.309**	**1.477**	**0.858**	**0.347**	**1.655**
0.3	0.799	0.285	1.339	0.801	0.296	1.444	0.827	0.337	1.623
0.5	0.801	0.285	1.344	0.772	0.262	1.419	0.813	0.333	1.617
0.7	0.800	0.285	**1.347**	0.735	0.254	1.279	0.788	0.320	1.574
Methods	ScoCAM [94]			LCAM [95]			SGradCAMpp [91]		
Metrics	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑	AUC-J↑	CC↑	NSS↑
0.0	**0.772**	**0.237**	**1.016**	**0.776**	**0.241**	**1.033**	**0.809**	**0.275**	**1.181**
0.3	0.763	0.231	0.989	0.764	0.233	0.992	0.806	0.269	1.151
0.5	0.735	0.224	0.969	0.737	0.227	0.974	0.726	0.181	0.789
0.7	0.687	0.207	0.916	0.698	0.221	0.978	0.630	0.122	0.501

表 9

本文的方法与其他三个具有代表性的完全监督 SOTA 模型之间的定性比较。所有方法都在相同的数据集（168 个序列）上进行训练，在训练完全监督的 SOTA 时使用真正的视点，而在本文的训练中，仅使用分配的视频标签。数值结果是通过在 AVAD 数据集上测试这些模型获得的 [73]。

Methods	AUC-J↑	SIM↑	S-AUC↑	CC↑	NSS↑
STANet+	0.887	0.361	0.595	0.490	2.318
STANet+ (trained on 168)	0.856	0.264	0.609	0.378	1.711
ACLNet [81]	0.905	0.446	0.560	0.580	3.170
ACLNet (re-trained on 168)	0.838	0.329	0.579	0.393	1.843
SalGAN [71]	0.886	0.360	0.579	0.491	2.550
SalGAN (re-trained on 168)	0.845	0.327	0.556	0.382	1.854
DeepNet [19]	0.869	0.256	0.561	0.383	1.850
DeepNet (re-trained on 168)	0.852	0.247	0.582	0.410	1.750

7 结论

在本文中，本文详细介绍了一种将视频-音频语义类别标签转换为伪视点的新方案。与广泛使用的 CAM 机制相比，所提出的 SCAM 和 SCAM+ 能够产生更符合真实人类视点的伪视点。关键技术创新包括基于多源的选择性融合及其多阶段方法。此外，还考虑了多粒度信息，设计了一种将其与数据源特征相结合的新方法，最终实现了多粒度感知。在多粒度感知的帮助下，本文的方法可以更符合真实的人类注意机制，以在多模态环境中正确定位最具辨别力的空间区域。此外，

为了在没有任何人工提供的视频标签的情况下启用通用应用程序，本文设计了多个空间-时间-音频视点预测网络（STA 和 STA+），这些网络在由提出的 SCAM 和 SCAM+ 产生的伪视点上进行训练，从而实现最终端视点预测。本文还比较了提出的模型——使用本文的伪视点训练的 STA 和 STA+ 视点预测网络，以及其他 SOTA 方法。结果表明本文的新方法优于无监督和弱监督方法。此外，他们还表明提出的方法甚至比一些完全监督的方法更好。

参考文献

[1] G. Wang, C. Chen, D.-P. Fan, A. Hao, and H. Qin, "From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach," in *CVPR*, 2021.

[2] G. Wang, C. Chen, D.-P. Fan, A. Hao, and H. Qin, "Weakly supervised visual-auditory fixation prediction with multigranularity perception," *arXiv preprint arXiv:2112.13697*, 2021.

[3] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *TPAMI*, vol. 43, no. 1, pp. 220–237, 2021.

[4] P. Linardos, E. Mohedano, J. Nieto, N. OConnor, X. GiroiNieto, and K. McGuinness, "Simple vs complex temporal recurrences for video saliency prediction," *BMVC*, 2019.

[5] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *TIP*, vol. 29, pp. 1113–1126, 2019.

[6] Y. Fang, G. Ding, J. Li, and Z. Fang, "Deep3dsaliency: Deep stereoscopic video saliency detection model by 3d convolutional networks," *TIP*, vol. 28, no. 5, pp. 2305–2318, 2018.

[7] X. Wu, Z. Wu, J. Zhang, L. Ju, and S. Wang, "Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm," in *AAAI*, 2020.

[8] Y. Xu, S. Gao, J. Wu, N. Li, and J. Yu, "Personalized saliency and its prediction," *TPAMI*, vol. 41, no. 12, pp. 2975–2989, 2018.

[9] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, and S. Lu, "Pyramid constrained self-attention network for fast video salient object detection," in *AAAI*, 2020.

[10] B. Wang, W. Liu, G. Han, and S. He, "Learning long-term structural dependencies for video salient object detection," *TIP*, vol. 29, pp. 9017–9031, 2020.

[11] S. Ren, C. Han, X. Yang, G. Han, and S. He, "Tenet: Triple excitation network for video salient object detection," in *ECCV*, 2020.

[12] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in *ICCV*, 2019.

[13] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *CVPR*, 2019.

[14] C. Chen, S. Li, Y. Wang, A. Hao, and H. Qin, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *TIP*, vol. 26, no. 7, pp. 3156–3170, 2017.

[15] C. Chen, G. Wang, C. Peng, X. Zhang, and H. Qin, "Improved robust video saliency detection based on long-term spatial-temporal information," *TIP*, vol. 29, no. 1, pp. 1090–1100, 2020.

[16] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *TIP*, vol. 28, no. 10, pp. 4819–4831, 2019.

[17] B. Li, Z. Sun, and Y. Guo, "Supervae: Superpixelwise variational autoencoder for salient object detection," in *AAAI*, 2019.

[18] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *CVPR*, 2015.

[19] J. Pan, E. Sayrol, X. GiroiNieto, K. McGuinness, and N. OConnor, "Shallow and deep convolutional networks for saliency prediction," in *CVPR*, 2016.

[20] W. Wang and J. Shen, "Deep visual attention prediction," *TIP*, vol. 27, no. 5, pp. 2368–2378, 2017.

[21] ShewchenkoN, WithnallC, KeownM, GittensR, and DvorakJ, "Heading in football. part 1: development of biomechanical methods to investigate head response," *BJSM*, vol. 39, no. 1, pp. 10–25, 2005.

[22] M. Shigeoka, N. Urakawa, T. Nakamura, M. Nishio, T. Watajima, D. Kuroda, T. Komori, Y. Kakeji, S. Semba, and H. Yokozaki, "Tumor associated macrophage expressing cd 204 is associated with tumor aggressiveness of esophageal squamous cell carcinoma," *Cancer science*, vol. 104, no. 8, pp. 1112–1119, 2013.

[23] B. Mandal, L. Li, G. S. Wang, and J. Lin, "Towards detection of bus driver fatigue based on robust visual analysis of eye state," *TITS*, vol. 18, no. 3, pp. 545–557, 2016.

[24] J. Podlesny and D. Raskin, "Physiological measures and the detection of deception." *Psychological bulletin*, vol. 84, no. 4, p. 782, 1977.

[25] E. Kok and H. Jarodzka, "Before your very eyes: The value and limitations of eye tracking in medical education," *Medical education*, vol. 51, no. 1, pp. 114–122, 2017.

[26] Y. Zhu, G. Zhai, and X. Min, "The prediction of head and eye movement for 360 degree images," *SP:IC*, vol. 69, pp. 15–25, 2018.

[27] Y. Zhu, G. Zhai, Y. Yang, H. Duan, X. Min, and X. Yang, "Viewing behavior supported visual saliency predictor for 360 degree videos," *TCSVT*, 2021.

[28] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal audiovisual saliency network," in *CVPR*, 2020.

[29] H. Tavakoli, A. Borji, E. Rahtu, and J. Kannala, "Dave: A deep audio-visual embedding for dynamic saliency prediction," *arXiv preprint arXiv:1905.10693*, 2019.

[30] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018.

[31] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *CVPR*, 2019.

[32] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017.

[33] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *CVPR*, 2019.

[34] S. Bargal, A. Zunino, D. Kim, J. Zhang, V. Murino, and S. Sclaroff, "Excitation backprop for rnns," in *CVPR*, 2018.

[35] A. Chattopadhay, A. Sarkar, P. Howlader, and V. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *WACV*, 2018.

[36] Z. Li, W. Wang, Z. Li, Y. Huang, and Y. Sato, "Towards visually explaining video understanding networks with perturbation," in *WACV*, 2021.

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.

[38] Z. Shi, Y. Yang, T. M. Hospedales, and T. Xiang, "Weakly-supervised image annotation and segmentation with objects and attributes," *TPAMI*, vol. 39, no. 12, pp. 2525–2538, 2017.

[39] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, "Self paced deep learning for weakly supervised object detection," *TPAMI*, vol. 41, no. 3, pp. 712–725, 2019.

[40] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao, "Learning from weak and noisy labels for semantic segmentation," *TPAMI*, vol. 39, no. 3, pp. 486–500, 2016.

[41] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, and T. Huang, "Weakly supervised scene parsing with point-based distance metric learning," in *AAAI*, 2019.

[42] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *CVPR*, 2020.

[43] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *ICCV*, 2015.

[44] D. Li, J. Huang, Y. Li, S. Wang, and M. Yang, "Progressive representation adaptation for weakly supervised object localization," *TPAMI*, vol. 42, no. 6, pp. 1424–1438, 2019.

[45] F. Wan, P. Wei, Z. Han, J. Jiao, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," *TPAMI*, vol. 41, no. 10, pp. 2395–2409, 2019.

[46] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *ICCV*, 2019.

[47] P. Jiang, L. Han, Q. Hou, M.-M. Cheng, and Y. Wei, "Online attention accumulation for weakly supervised semantic segmentation," *TPAMI*, 2021.

[48] J. Choe, S. Lee, and H. Shim, "Attention-based dropout layer for weakly supervised single object localization and semantic segmentation," *TPAMI*, vol. 43, no. 12, pp. 4256–4271, 2021.

[49] K. Li, Z. Wu, K. Peng, J. Ernst, and Y. Fu, "Guided attention inference network," *TPAMI*, vol. 42, no. 12, pp. 2996–3010, 2019.

[50] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in *AAAI*, 2020.

[51] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, J. Alvarez, and S. Gould, "Incorporating network built-in priors in weakly-supervised semantic segmentation," *TPAMI*, vol. 40, no. 6, pp. 1382–1396, 2018.

[52] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *CVPR*, 2018.

[53] Y. Gao, B. Liu, N. Guo, X. Ye, F. Wan, H. You, and D. Fan, "C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection," in *ICCV*, 2019.

[54] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-mil: Continuation multiple instance learning for weakly supervised object detection," in *CVPR*, 2019.

[55] A. Arun, C. Jawahar, and M. Kumar, "Dissimilarity coefficient based weakly supervised object detection," in *CVPR*, 2019.

[56] Z. Chen, Z. Fu, R. Jiang, Y. Chen, and X. Hua, "Slv: Spatial likelihood voting for weakly supervised object detection," in *CVPR*, 2020.

[57] K. Ye, M. Zhang, A. Kovashka, W. Li, D. Qin, and J. Berent, "Cap2det: Learning to amplify weak caption supervision for object detection," in *ICCV*, 2019.

[58] Z. Yang, D. Mahajan, D. Ghadiyaram, R. Nevatia, and V. Ramanathan, "Activity driven weakly supervised object detection," in *CVPR*, 2019.

[59] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *CVPR*, 2018.

[60] Y. Lu, J. Yin, Z. Chen, H. Gong, Y. Liu, L. Qian, X. Li, R. Liu, I. Andolina, and W. Wang, "Revealing detail along the visual hierarchy: neural clustering preserves acuity from v1 to v4," *Neuron*, vol. 98, no. 2, pp. 417–428, 2018.

[61] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[63] W. Wang, X. Lu, J. Shen, D. J. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *ICCV*, 2019.

[64] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang, "Group-wise semantic mining for weakly supervised semantic segmentation," *AAAI*, 2021.

[65] G. Sun, W. Wang, J. Dai, and L. VanGool, "Mining cross-image semantics for weakly supervised semantic segmentation," in *ECCV*, 2020.

[66] C. Xia, J. Han, and D. Zhang, "Evaluation of saccadic scanpath prediction: Subjective assessment database and recurrent neural network based metric," *TPAMI*, vol. 43, no. 12, pp. 4378–4395, 2021.

[67] T. Yang and A. Chan, "Visual tracking via dynamic memory networks," *TPAMI*, vol. 43, no. 1, pp. 360–374, 2019.

[68] W. Sun, Z. Chen, and F. Wu, "Visual scanpath prediction using ior-roi recurrent mixture density network," *TPAMI*, vol. 43, no. 6, pp. 2101–2118, 2019.

[69] B. Zhang, D. Xiong, J. Xie, and J. Su, "Neural machine translation with gru-gated attention model," *TNNLS*, vol. 31, no. 11, pp. 4688–4698, 2020.

[70] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[71] J. Pan, C. Ferrer, K. McGuinness, N. OConnor, J. Torres, E. Sayrol, and X. GiroiNieto, "Salgan: Visual saliency prediction with generative adversarial networks," *CVPR SUNw*, 2017.

[72] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *TPAMI*, 2019.

[73] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM TOMM*, 2016.

[74] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[75] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *NeurIPS*, 2008.

[76] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *JoV*, vol. 14, no. 8, pp. 5–5, 2014.

[77] A. Coutrot and N. Guyader, "Multimodal saliency models for videos," in *From Human Attention to Computational Attention*, 2016, pp. 291–304.

[78] P. Mital, T. Smith, R. Hill, and J. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *COGN COMPUT*, vol. 3, no. 1, pp. 5–24, 2011.

[79] M. Gygli, H. Grabner, H. Riemenschneider, and L. VanGool, "Creating summaries from user videos," in *ECCV*, 2014.

[80] P. Koutras and P. Maragos, "A perceptually based spatio-temporal computational framework for visual saliency estimation," *SP:IC*, vol. 38, pp. 15–31, 2015.

[81] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *CVPR*, 2018.

[82] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *TPAMI*, vol. 35, no. 1, pp. 185–207, 2012.

[83] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *TPAMI*, vol. 41, no. 3, pp. 740–757, 2018.

[84] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[85] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NeurIPS*, 2007.

[86] D. Rudoy, D. Goldman, E. Shechtman, and L. ZelnikManor, "Learning video saliency from human gaze using candidate selection," in *CVPR*, 2013.

[87] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *ICCV*, 2017.

[88] V. Leboran, A. GarciaDiaz, X. FdezVidal, and X. Pardo, "Dynamic whitening saliency," *TPAMI*, vol. 39, no. 5, pp. 893–907, 2016.

[89] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.

[90] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.

[91] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models," *arXiv preprint arXiv:1908.01224*, 2019.

[92] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based grad-cam: Towards accurate visualization and explanation of cnns," *BMVC*, 2020.

[93] H. Wang, R. Naidu, J. Michael, and S. Kundu, "Ss-cam: Smoothed score-cam for sharper visual feature localization," *arXiv preprint arXiv:2006.14255*, 2020.

[94] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *CVPRW*, 2020.

[95] P. Jiang, C. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "Layercam: Exploring hierarchical class activation maps for localization," *TIP*, vol. 30, pp. 5875–5888, 2021.

[96] R. Naidu, A. Ghosh, Y. Maurya, S. Kundu *et al.*, "Is-cam: Integrated score-cam for axiomatic-based explanations," *arXiv preprint arXiv:2010.03023*, 2020.

[97] H. Ramaswamy *et al.*, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *WACV*, 2020.

[98] M. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *IJCNN*, 2020.

[99] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *ECCV*, 2018.

[100] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *ECCV*, 2018.

[101] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.

[102] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*, 2009.

Guotao Wang received the M.S. degree in computer science from Qingdao University in 2020. He is currently a Ph.D. student in Beihang University. His research interests include computer vision and deep learning.

Chenglizhao Chen is a Professor in College of Computer Science and Technology, China University of Petroleum (East China). His research interests include Virtual Reality, Computer Vision, Deep Learning, Data Mining, and Pattern Recognition.

Deng-Ping Fan received the Ph.D. degree from Nankai University in 2019. He is currently a Postdoctoral Researcher, working with Prof. Luc Van Gool in Computer Vision Lab at ETH Zurich. His current research interests include computer vision, image processing, and deep learning.

Aimin Hao is a professor in Computer Science School and the Associate Director of State Key Laboratory of Virtual Reality Technology and Systems at Beihang University. His research interests are on virtual reality, computer simulation, computer graphics, and computer vision.

Hong Qin is a professor of computer science in the Department of Computer Science, Stony Brook University. His research interests include geometric and solid modeling, graphics, physics-based modeling and simulation, computer-aided geometric design, visualization, and scientific computing.