从语义类别到注视点：一种新颖的弱监督视听显著性检测方法

王国涛 [1] 陈程立诏 [2]*范登平 [4] 郝爱民 [1,3,6] 秦洪 [5]

[1] 北京航空航天大学虚拟现实技术与系统全国重点实验室 [2] 青岛大学计算机科学技术学院

[3] 中国医学科学院虚拟人与虚拟手术研究单元 [4] 阿联酋人工智能研究院 [5] 石溪大学 [6] 鹏城实验室

摘要

近年来，随着深度学习技术的发展以及大规模训练数据的广泛应用，视频显著性检测取得了显著进展。然而，面向真实视听场景的注视点预测研究仍处于起步阶段，相关模型训练高度依赖真实人眼注视标注，而此类数据采集成本高、规模有限，严重制约了模型性能的进一步提升。针对这一问题，本文提出一种面向视听注视点预测的弱监督学习方法。该方法仅利用视频类别标签，引入选择性类别激活映射（Selective Class Activation Mapping, SCAM），通过由粗到细的方式，从空间、时间和音频等多源信息中筛选最具判别性的区域，并将其构造为伪注视点监督信号。在此基础上，进一步训练空间—时间—音频（Spatial-Temporal-Audio, STA）注视点预测网络。实验结果表明，在不依赖真实注视点监督的条件下，所提出方法仍能取得与全监督方法相当的性能。该研究为弱监督视听注视点预测提供了一条可行途径，也为视听信息融合研究提供了新的思路。

关键词：弱监督；视听显著性；注视点预测；选择性类别激活映射；伪注视点

1 引言与研究动机

随着深度学习的发展，视频显著性检测研究取得了长足进步，其核心任务是在视频序列中定位最能吸引视觉注意的区域 [14, 29, 34, 53]。现有研究大体可分为两个方向：一类是视频显著目标检测（图1-A），旨在分割具有清晰边界的显著目标 [4, 5, 8, 13, 19, 32, 41, 49]；另一类是视频注视点预测，关注人眼在动态场景中的注视分布，其结果通常表现为分散的注视坐标，而非具有明确边界的目标区域 [12, 18, 35, 44, 54]。

与主要面向纯视觉场景的已有研究不同 [29, 39, 51]，本文关注视觉与音频共同作用下的注视点预测问题，即视听注视点预测。该方向近年来开始受到关注，但整体上仍处于探索阶段。现有代表性方法大多基于深度学习框架构建，通常采用编码器—解码器结构，并结合注意力机制实现多模态融合 [45, 47]。尽管这些方法在一定程度上提升了预测性能，但其发展仍受到一个关键瓶颈的限制，即缺乏大规模、真实且可用于训练的视听注视数据。

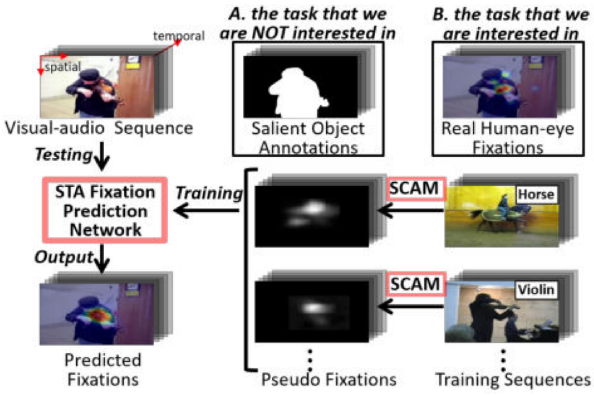

图 1: 本文提出一种弱监督的空间—时间—音频（STA）注视点预测框架，其核心思想是通过选择性类别激活映射（SCAM）将语义类别标签自动转换为伪注视点。

从数据条件来看，纯视觉场景下的显著性或注视数据集已相对丰富，而视听联合场景中的真实人眼注视数据却十分稀缺。其根本原因在于，视听注视数据需要在真实多模态环境下完成采集，过程耗时耗力，且采集成本较高。就目前而言，公开可用的视听注视视频序列数量仍然较少，真正可用于模型训练的数据规模更为有限。这使得现有基于深度学习的视听显著性预测方法 [45, 47] 普遍面临训练样本不足的问题，并容易出现过拟合现象。

针对上述问题，本文尝试从弱监督学习的角度出发，探索不依赖真实逐帧注视标注的视听注视点预测方法。具体而言，本文不再使用高成本采集得到的真实视听真值，而是仅利用视频类别标签生成伪注视点监督信号。事实上，现有视听分类任务中已经积累了大量带有语义类别标签的视频样本，例如 AVE 数据集 [46]，这为弱监督学习提供了可利用的数据基础。

本文的方法受到类别激活映射（Class Activation Mapping, CAM）[64] 的启发。CAM 已被广泛应用于图像目标定位 [43, 50, 57] 和视频目标定位 [2, 3, 33] 等任务中，其基本思想是：对分类任务贡献越大的区域，

*通讯作者。

往往越具有判别性，也更可能对应场景中的关键目标。换言之，分类模型在判别类别时最关注的区域，与视觉显著区域之间存在天然联系。

然而，若将传统 CAM 直接用于视听注视点预测，仍然面临两个突出问题。其一，传统 CAM 往往产生范围较大且较为分散的响应区域，而真实人眼注视通常更加集中，更偏向于落在最具判别性的局部区域上。其二，现有 CAM 大多仅依赖空间信息，而真实视听场景中的视觉注意不仅受到空间外观影响，还受到时间变化和音频线索的共同作用。因此，如何从多源信息中筛选出更接近真实注视行为的判别区域，成为本文需要解决的关键问题。

为此，本文提出选择性类别激活映射（Selective Class Activation Mapping, SCAM）方法。该方法采用由粗到细的处理策略，在空间、时间和音频等多种信息源之间进行选择与融合，以挖掘最具判别性的区域，并进一步生成更接近真实人眼注视分布的伪注视点。在获得伪注视点之后，本文进一步构建空间—时间—音频（Spatial-Temporal-Audio, STA）注视点预测网络，用于学习伪注视点之间的分布规律，从而实现对未知视听视频的注视点预测。

总体而言，本文尝试以弱监督方式研究视听注视点预测问题，旨在缓解真实注视标注稀缺所带来的训练困难。实验结果表明，基于语义标签构造的伪注视点能够有效替代部分真实监督信息，并为视听信息融合和多模态显著性建模提供新的研究思路。

2 相关工作

2.1 无监督视觉注视点预测

早期视觉注视点预测方法大多基于手工特征构建，通常可归入无监督方法范畴。代表性工作中，Fang 等 [15] 结合空间信息与时间信息中的不确定性度量进行视觉显著性检测；Hossein 等 [22] 提出了基于重建误差与粗粒度自信息度量的视觉显著性模型；Leboran 等 [30] 则通过对时空尺度分解特征进行显式短时视觉自适应来估计动态显著性。随着深度学习的发展，无监督显著性建模也逐步向数据驱动方式演进。Zhang 等 [62] 从多个带噪声的无监督显著性结果中学习显著图，并将该任务建模为潜在显著性预测与噪声建模的联合优化问题。Li 等 [31] 采用超像素级变分自编码器，以更好地保持目标边界并增强空间一致性。此外，也有研究从域适应与背景约束等相关角度，对模型的鲁棒性与适应性进行了探索 [27]。

2.2 弱监督视觉注视点预测

相较于无监督方法，弱监督方法通过引入低成本标注信息，通常能够获得更优的性能。现有弱监督研究主要利用图像级标签 [50]、点标注 [40]、涂鸦标注 [61] 以及边界框 [11] 等弱标注形式，为模型提供间接监督。Zeng 等 [58] 在弱监督目标检测任务中，将自底向上的目标证据与自顶向下的类别置信度结合起来，以提升目标定位效果。Zhang 等 [59] 利用图像级标签生成较为可靠的像素级标注，并设计端到端网络学习分割图。总体来看，弱监督学习为降低显著性建模对密集人工标注的依赖提供了有效途径，也为复杂场景下的视觉注意建模提供了新的思路。

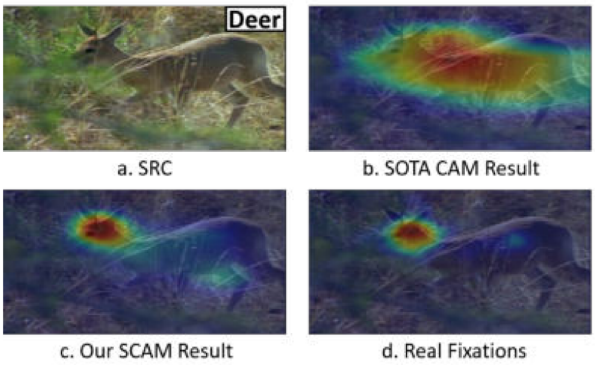

图 2: 现有最先进方法（如 Zeng 等 [57]）主要用于定位显著目标，而非模拟真实人眼注视，因此其结果通常表现为大范围分散响应区域（b），这与真实注视点分布（d）存在明显差异。

2.3 监督式视听注视点预测

近年来，随着多模态学习的发展，视听显著性检测逐渐受到关注，代表性工作包括 STAViS [47]、DAVE [45] 和 AVC [37]。这类方法通常基于这样一种假设：当视觉信息与音频信息在语义上保持一致时，音频线索会对人眼注视行为产生显著影响。因此，现有研究主要聚焦于如何设计更有效的视听融合机制，以提升注视点预测性能。然而，就目前而言，真实视听场景下带有人眼注视标注的视频数据仍然十分有限。现有公开可用的相关视频序列总数仅为 241 个，主要来自文献 [9, 10, 36]。训练数据的稀缺性在很大程度上制约了监督式视听注视点预测模型的进一步发展。基于这一现实背景，本文尝试以弱监督方式挖掘视听伪注视点，用于视听视频中的注视点预测任务。

3 本文提出的方法

3.1 视频标签与注视点之间的关系

在视频分类任务中，每个训练序列通常都被赋予一个语义标签，用于表征该序列所属的特定类别。一般来说，这些语义标签通过多位标注者的多数投票确定，其目的是概括视频中最具代表性、最有意义的目标或事件。与此类似，在观看视频序列时，人眼注视往往也

更容易集中在那些具有代表性和语义意义的区域。因此，从视频类别标签中挖掘伪注视点，在理论上可行。

3.2 预备知识：类别激活映射（CAM）

类别激活映射（Class Activation Mapping, CAM）的基本思想，是对最后一个卷积层输出的特征图进行加权求和，从而对当前分类任务中最具代表性的图像区域进行粗定位。具体而言，如图 3 所示，与最后一层全连接分类器中最高分类置信度相对应的权重 w_i 被选取出来，并用于加权对应的特征图 f_i。最终得到的 CAM 是一个二维响应矩阵，其形式可表示为：

$$\mathrm{CAM} = \xi\left[\sum_{i}^{d} w_i \times f_i\right], \tag{1}$$

其中，d 表示特征通道数，$\xi[\cdot]$ 表示归一化函数。

从定性结果来看，如图 3 右下角所示，CAM 往往会在对当前分类任务贡献最大的区域产生较强响应，例如图中的 *motorbike* 区域。这些高响应区域通常与场景中的显著目标相对应。

3.3 传统 CAM 的局限性

尽管 CAM 能够在一定程度上揭示对分类任务最重要的区域，但它与真实人眼注视之间仍存在明显差异。例如，如图 2(b) 所示，在视频分类任务中，那些对 *deer* 类别判别贡献最大的区域，通常会高亮整个显著目标，即鹿本身。基于这一现象，已有一些研究尝试利用 CAM 实现显著目标定位 [1,7,16,48,63]。然而，这类方法所得到的 CAM 与真实人眼注视分布仍然存在较大差距，其原因主要体现在以下两个方面。

首先，由于局部特征与非局部特征都可能对分类任务产生贡献，传统 CAM 往往呈现出范围较大且相对分散的响应区域。例如，如图 2 所示，鹿的主体部分（*main body*）能够帮助分类器将其与其他非动物图像区分开来，而鹿头区域（*deer head*）则进一步提供了更强的类别判别信息。相比之下，人类视觉系统通常不会平均关注整个目标，而是更倾向于聚焦在最具判别性的局部区域，例如图 2(d) 所示的鹿头区域。

其次，现有大多数 CAM 相关工作 [23,50,55–57] 在计算过程中仅考虑空间信息，而真实人眼注视往往受到多种信息源的共同影响，包括空间外观、时间变化以及音频线索。长期以来，这种多源驱动特性并未得到充分重视，主要原因在于，相比于相对稳定的空间信息，时间信息和音频信息通常被认为波动较大、难以直接利用。然而，在许多实际应用场景中，恰恰是这些动态线索对分类判断和注意分配起到了关键作用。

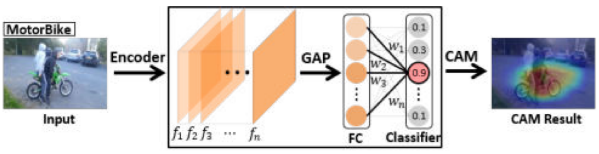

图 3: 类别激活映射（CAM）细节示意图。其中，FC 表示全连接层，GAP 表示全局平均池化，分类器中的数值表示对应的分类置信度。

3.4 多源条件下的 SCAM 计算

与单幅图像场景相比，本文所研究的视听注视点预测问题更为复杂，因为该任务需要同时考虑空间信息、时间信息和音频信息等多种信息源。如前所述，传统仅依赖空间信息构建的 CAM 往往会产生范围较大且较为分散的响应区域，这类区域与真实人眼注视分布之间通常存在较大差异。更重要的是，这种单一信息源建模方式无法充分利用空间、时间和音频之间的互补关系。其根本原因在于，视听联合场景下的特征通常具有多尺度、多层级和多来源等特点，不同来源的特征都可能对分类结果作出贡献，从而引入更多冗余响应和误激活区域，使最终得到的 CAM 偏离真正具有判别性的关键位置。

针对这一问题，本文将空间—时间—音频联合场景解耦为三个相对独立的信息源，即空间（S）、时间（T）和音频（A）。在此基础上，进一步构建三个不同的融合分类网络，分别对应 S、ST 和 SA 三种输入形式（见 Fig. 4 和 Fig. 6）。随后，对这些分类网络生成的 CAM 进行选择性融合，从而挖掘更接近真实人眼注视的判别性区域。本文将该过程称为**选择性类别激活映射**（Selective Class Activation Mapping, SCAM），其定义如下：

$$\mathrm{SCAM} = \xi\left[\frac{\phi(C_S^v)\,\Phi_S + \phi(C_{ST}^v)\,\Phi_{ST}}{\phi(C_S^v) + \phi(C_{ST}^v)} + \frac{\phi(C_{SA}^v)\,\Phi_{SA} + \lambda}{\phi(C_{SA}^v) + \lambda}\right], \tag{2}$$

其中，λ 为较小常数，用于避免分母为零；Φ_S、Φ_{ST} 和 Φ_{SA} 分别表示由 S、ST 和 SA 分类网络生成的 CAM；$C \in (0,1)^{1\times c}$ 表示关于 c 个类别的分类置信度向量。假设 S 分类网络预设的视频类别标签为 C 中的第 v 类，则 C_S^v 表示该类别对应的分类置信度；$\xi[\cdot]$ 表示归一化函数。$\phi(\cdot)$ 为软筛选函数，其作用是抑制分类置信度较低、因而不宜在 SCAM 计算中被重点考虑的特征，其定义为：

$$\phi(C_S^v) = \begin{cases} C_S^v, & \text{if } C_S^v > C_S^u \,|\, \forall u,\ 1 \le u \le c, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

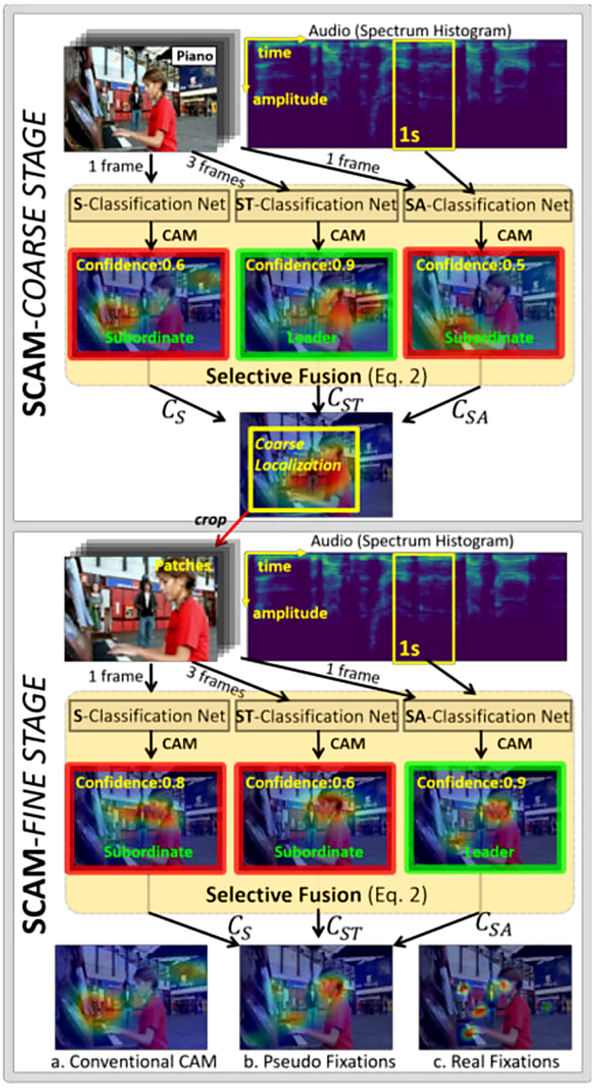

图 4: 本文提出的选择性类别激活映射（SCAM）遵循由粗到细的处理流程，其中粗阶段用于定位感兴趣区域，细阶段则用于揭示局部响应最强的图像区域。S 表示空间分支，ST 表示时空分支，SA 表示空间—音频分支。

3.5 SCAM 的设计动机

一般而言，空间源、时间源和音频源都可能对视觉注意分配产生影响。然而，与后两者相比，空间信息在大多数实际场景中更为稳定，也更具基础性。例如，在较长时间内保持近乎静止的视频片段中，时间信息能够提供的附加判别线索十分有限；类似地，当音频仅包含背景噪声或无关声音时，其对视觉注意的引导作用也会显著减弱。因此，在基于 CAM 进行分类建模时，空间信息应视为主要信息源，而时间信息和音频信息则更多起到补充和增强作用。这也正是本文将三类信息进一步组织为 S、ST 和 SA 三种分类网络，而不是简单地将所有信息一次性直接融合的主要原因。

进一步来看，S、ST 和 SA 分类网络均仅利用视频类别标签进行训练，因此当测试样本与训练样本在分布上较为接近时，这些网络通常都能够获得较好的分类效果。然而，由于输入信息源不同，不同网络生成的 CAM 在响应模式上仍然存在明显差异。本文在 Fig. 5 中给出了一些具有代表性的定性结果。总体上看，CAM 与真实人眼注视之间的一致性通常与分类置信度呈正相关关系。基于这一观察，本文采用分类置信度作为融合权重，对可信度较低的 CAM 进行抑制，并通过式（2）对多源 CAM 进行选择性融合，从而生成更接近真实人眼注视分布的伪注视点。

3.6 多阶段 SCAM

尽管通过多源信息的选择性融合，本文提出的 SCAM 在伪注视点生成方面已优于传统 CAM，但在某些复杂场景中，其生成结果仍可能与真实人眼注视存在差异，特别是在背景较为复杂时，伪注视点容易受到干扰而出现混杂现象。造成这一问题的原因主要有两方面。

一方面，复杂视频场景通常包含更丰富的语义内容，而训练样本往往只带有单一类别标签，因此那些不属于目标类别的语义内容也可能对分类结果产生影响。另一方面，前述 SCAM 本质上仍属于单阶段处理方式，而人类视觉系统则具有明显的多尺度处理特征：通常会先快速定位感兴趣区域，再在局部范围内完成更细粒度的注意分配。基于这一认识，本文进一步采用由粗到细的多阶段策略，对 SCAM 顺序执行两次。粗阶段首先缩小潜在关注区域，细阶段则在局部范围内进一步细化响应，从而使最终获得的伪注视点更接近真实的判别性区域，并提升整体预测性能。

具体而言，在粗阶段中，本文首先对伪注视点响应图进行二值化处理，并采用硬阈值（$2 \times \text{average}$）提取高响应区域；随后利用矩形框对这些区域进行紧致包围，并据此将原始视频序列裁剪为若干视频块。在细阶段中，分类网络的输入不再是原始视频序列，而是粗阶段得到的裁剪视频块；之后再次执行 SCAM，以生成最终的伪注视点。与传统 CAM（即由 S 分类网络直接生成的 CAM，对应 Fig. 4(a)）相比，多阶段 SCAM 所得到的伪注视点（Fig. 4(b)）与真实人眼注视（Fig. 4(c)）之间具有更高的一致性，其定量验证结果将在第 4 节中给出。

3.7 分类网络细节

本文所采用的各类网络均基于较为简洁的编码器—解码器结构构建。按照已有工作 [45] 的处理方式，本文首先将音频信号预处理为二维频谱直方图，并采用普通 3D 卷积来建模时间信息。这样的实现方式结构清晰、易于复现，相关网络细节已在 Fig. 6 中给出。需要说明的是，若引入更复杂的时空建模或特征增强机制，模型性能仍有进一步提升的可能，但这并非本文关注

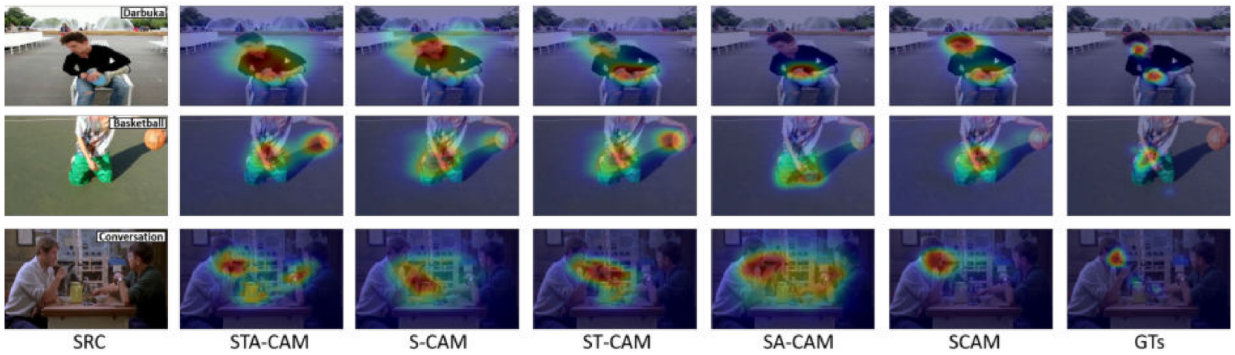

图 5: 不同信息源生成的 CAM 定性比较结果。其中，STA(S/SA/ST)-CAM 表示不同模态组合条件下得到的 CAM，SCAM 表示根据式（2）生成的伪注视点。可以观察到，SCAM 与真实真值之间具有较高一致性。

的重点。

音频开关（ϕ） 不同于传统视听融合方式，本文在 SA Fuse 和 STA Fuse 中均引入了音频开关模块（见 Fig. 6）。该模块的主要作用是在进行 SA 融合和 STA 融合时，抑制无效音频信息可能带来的干扰，从而减轻音频分支对整体分类过程的负面影响。

与时间信息相比，音频信息通常具有更强的语义指示性，因此也更容易对对应的空间分支产生影响。然而，音频信息并非始终有效。在实际视频中，往往伴随着背景音乐、环境噪声或其他与当前视觉语义无关的声音。若不加区分地将这些音频信息与空间特征直接融合，反而可能增加模型学习难度，并削弱分类效果。基于这一考虑，本文设计了音频开关模块，以判断当前音频信息是否值得参与融合。

从结构上看，音频开关本质上是一个插件式子网络，本文将其实现为与 SA 分类网络结构一致的独立网络。不同于视频分类网络，该插件网络利用带有二值标签的视听数据进行训练，用以判断当前音频信号是否真正有助于空间分支的判别。

为了自动构建这类二值标签，本文借助现成的音频分类工具 VggSound [6]。该工具是在包含近 300 个类别的大规模音频分类数据集上训练得到的。本文的基本假设是：只有当音频源与对应视觉内容在语义上保持一致时，音频信息才可能对空间分支产生正向帮助。因此，对于一个视听片段（1 帧图像与 1 秒音频），若音频分类器预测得到的音频类别与预设视频类别一致，则将其标签记为 1；否则记为 0。以 SA Fuse 为例，其融合过程可表示为

$$\mathrm{SA} \leftarrow \mathrm{ReLU}\big(\sigma(\phi(A)) \odot S + S\big), \tag{4}$$

其中，S 表示空间流，A 表示音频流，$\odot$ 表示逐元素乘法；$\mathrm{ReLU}(\cdot)$ 表示修正线性单元激活函数，$\sigma(\cdot)$ 表示 Sigmoid 函数。$\phi(\cdot)$ 表示本文提出的音频开关函数，当输入音频经 VggSound [6] 判定为与预设类别一致时，该函数输出为 1。实验结果表明，引入音频开关后，模型整体性能能够得到稳定提升，平均提升幅度约为 1.5%。

3.8 STA 注视点预测网络

STA 注视点预测网络的整体结构相对直观。具体而言，空间特征首先分别与时间特征和音频特征进行预融合，随后再通过特征拼接将二者结合起来。在此基础上，利用一个包含三层反卷积的解码器，将 STA 融合模块输出的特征图映射为最终的注视点预测结果。“STA Fuse” 模块中的数据流可表示为

$$\begin{aligned}\mathrm{STA} \leftarrow \mathrm{ReLU}\Big[\mathrm{Cov}\Big(\mathrm{Con}\big(&\sigma(\phi(A)) \odot S + S,\\ &\sigma(T) \odot S + S\big)\Big)\Big],\end{aligned} \tag{5}$$

其中，$\mathrm{Con}(\cdot)$ 表示特征拼接操作，$\mathrm{Cov}(\cdot)$ 表示 1×1 卷积，其余符号的含义与式（4）相同。

用于训练 STA 注视点预测网络的损失函数采用二值交叉熵形式，其定义为

$$\begin{aligned}L_B = -\frac{1}{N}\sum_{i}^{N}\Big[&\mathrm{PseudoGT}_i \log\Big(\mathrm{Dec}(\mathrm{STA}_i)\Big)\\ &+ \big(1 - \mathrm{PseudoGT}_i\big) \log\Big(1 - \mathrm{Dec}(\mathrm{STA}_i)\Big)\Big],\end{aligned} \tag{6}$$

其中，N 表示训练样本数量，PseudoGT 表示由双阶段 SCAM 生成的伪注视点真值，$\mathrm{Dec}(\cdot)$ 表示解码器映射过程。

需要指出的是，若进一步采用具有多尺度连接或通道注意机制的更强解码器，模型性能可能仍有提升空间。但这类结构增强并非本文的研究重点，因此相关讨论留待后续工作展开。由于本文提出的 STA 注视点预测网络仅依赖伪注视点进行监督训练，因此在测试阶段无需类别标签，也能够对未见过的视听视频序列进行注视点预测。

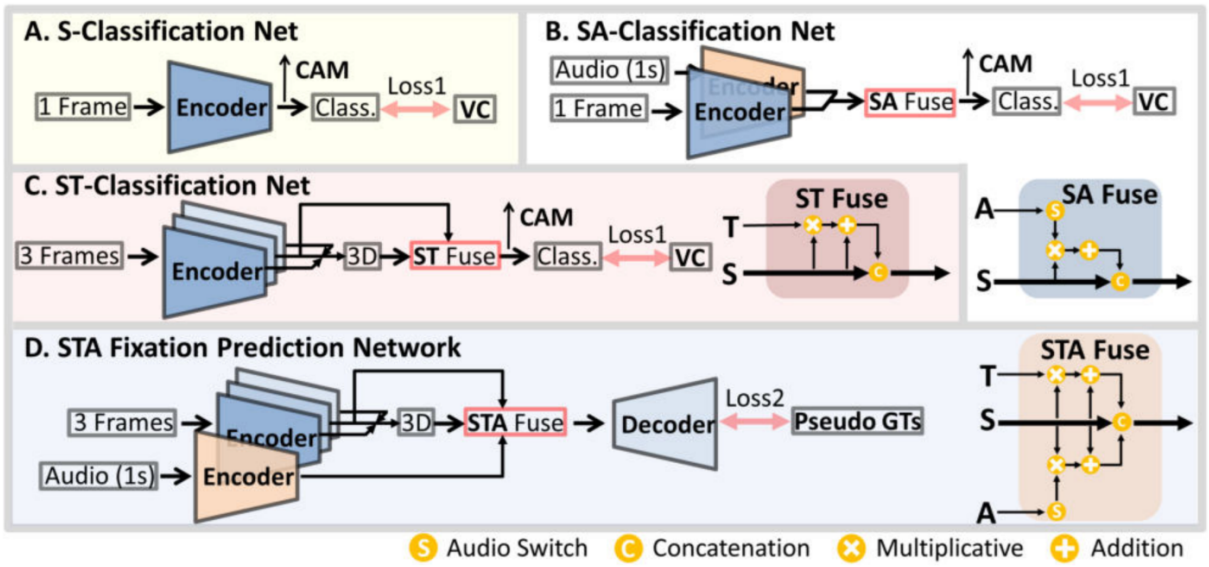

图 6: 网络结构细节示意图。其中，Loss1 为交叉熵损失，Loss2 为二值交叉熵损失，Class. 表示分类，VC 表示视频类别，Decoder 为 VGG16 解码器，3D 表示三维卷积。A–C 为分类网络，D 为 STA 注视点预测网络。

4 实验与验证

4.1 数据集与评价指标

测试数据集 本文在 6 个公开数据集上对所提出方法及其他对比方法进行了实验验证，分别为 AVAD [36]、Coutrot1 [9]、Coutrot2 [10]、DIEM [26]、SumMe [20] 和 ETMD [28]。上述数据集共包含 241 个视听视频序列，均提供了在真实视听场景下采集的人眼注视标注。

定量评价指标 参照已有工作 [52]，本文采用 5 个常用评价指标衡量模型预测结果与真实人眼注视之间的一致性，包括 AUC-Judd（AUC-J）、相似性指标（SIM）、shuffled AUC（s-AUC）、归一化扫描路径显著性（NSS）以及线性相关系数（CC）。上述指标数值越高，表明模型预测结果与真实注视分布越一致。

4.2 实现细节

训练集 近年来，Google 发布了目前规模较大的视听数据集 AudioSet [17]。本文选取其中的子集 Audio-Visual Event（AVE）定位数据集 [46]，作为 S、ST 和 SA 分类网络（见第 3.4 节）的训练数据。该数据集共包含 4,143 个视频序列，覆盖 28 个语义类别。

训练细节 本文采用多阶段训练策略。在粗阶段中，所有分类网络均在 AVE 数据集上进行训练，batch size 设为 20，所有视频帧统一缩放至 256×256。在细阶段中，以裁剪得到的视频块作为输入，重新训练三个新的分类网络，此时 batch size 设为 3，所有视频块统一缩放至 356×356。对于 STA 注视点预测网络，本文以伪注视点作为监督信号，输入视频帧同样缩放为 356×356，对应的 batch size 设为 3。所有训练过程均采用随机梯度下降（SGD）优化器，学习率设为 0.001。

表 1: 所提出选择性融合策略有效性的定量验证实验。该实验在 AVAD 数据集 [36] 上进行。'CO.' 表示粗阶段，'FI.' 表示细阶段。

阶段	模块	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
CO.	SCAM_{avg}	0.786	0.219	0.538	0.297	1.312
CO.	SCAM_{sel}	0.801	0.256	0.554	0.345	1.364
FI.	SCAM_{avg}	0.845	0.303	0.573	0.399	1.797
FI.	SCAM_{sel}	**0.873**	**0.334**	**0.580**	**0.438**	**2.018**

4.3 组件有效性分析

选择性融合策略的有效性 为实现多源 CAM 的选择性融合，本文采用分类置信度作为融合权重，即式（2）中的相关置信度项。该设计成立的前提在于：分类置信度与 CAM 和真实注视点之间的一致性程度呈正相关关系。为验证这一点，本文将所提出的选择性融合策略与传统融合策略进行了对比。传统策略直接对不同信息源生成的 CAM 进行平均，记为 SCAM_{avg}。

由表 1 可见，引入本文提出的选择性融合策略后，模型整体性能得到明显提升。具体而言，SCAM_{sel} 相较于 SCAM_{avg} 在多个评价指标上均表现更优，整体提升幅度约为 3%。这一结果表明，利用分类置信度对多源 CAM 进行加权筛选，能够更有效地突出与真实注视分布一致的判别性区域。

表 2: 所提出多阶段 SCAM 有效性的定量验证实验。该实验在 AVAD 数据集 [36] 上进行。

阶段	模块	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
粗阶段	CAM_S	0.774	0.202	0.545	0.261	1.269
	CAM_{ST}	0.785	0.223	0.536	0.269	1.292
	CAM_{SA}	0.780	0.214	0.542	0.277	1.276
	CAM_{STA}	0.793	0.227	0.551	0.293	1.273
	SCAM	0.801	0.256	0.554	0.345	1.364
细阶段	CAM_S	0.834	0.291	0.574	0.376	1.528
	CAM_{ST}	0.843	0.289	0.571	0.372	1.581
	CAM_{SA}	0.845	0.304	0.564	0.384	1.622
	CAM_{STA}	0.856	0.296	0.579	0.415	1.803
	SCAM	**0.873**	**0.334**	**0.580**	**0.438**	**2.018**

表 3: 所提出音频开关（AS，见式（5））有效性的定量验证实验。该实验在 AVAD 数据集 [36] 上进行。

设置	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
w/o Audio Switch	0.864	0.330	0.571	0.421	1.833
w/ Audio Switch	**0.873**	**0.334**	**0.580**	**0.438**	**2.018**

多阶段机制的有效性 为验证多阶段策略的作用，本文分别测试了不同信息源生成的 CAM 在粗阶段和细阶段下的性能，相关结果见表 2。其中，CAM_S 表示粗阶段中由空间分支生成的 CAM，该结果也可以视为经典 CAM 方法的一个代表。CAM_{STA} 则表示同时利用空间、时间和音频三类信息构建得到的 CAM，其中其融合方式与本文所提出的 STA 注视点预测网络（Fig. 6(D)）相近。

由表 2 可以看出，所有方法在细阶段下得到的结果均明显优于粗阶段，这表明由粗到细的处理机制能够有效提升伪注视点质量。同时也可以看出，若简单地同时利用所有信息源，并不能充分发挥多源信息之间的互补优势，因此 CAM_{STA} 的性能提升相对有限。相比之下，本文提出的 SCAM 在粗阶段和细阶段中均优于传统 CAM 方案。特别是从表 2 第一行与最后一行的对比可以发现，多阶段 SCAM 相较经典 CAM 具有显著优势，例如 CC 指标由 0.261 提升至 0.438，其余指标上也呈现出一致的提升趋势。

音频开关的有效性 表 3 给出了引入音频开关（AS，见式（4））前后的性能对比结果，其中 “w/o” 表示不使用音频开关，“w/” 表示使用音频开关。可以看出，引入音频开关后，模型整体性能提升约 2%。其主要原因在于，该模块能够有效抑制无关背景音频的干扰，从而缓解空间信息与音频信息在语义不同步条件下融合所带来的学习歧义。

4.4 定量比较

本文将所提出的模型，即仅利用伪注视点进行训练的 STANet，与另外 14 种最先进方法进行了对比，其中包括 5 种无监督方法、5 种弱监督方法以及 4 种全监督方法，实验覆盖全部 6 个测试数据集。由表 tab:table4 可见，本文方法显著优于所有无监督方法，同时也优于近年来具有代表性的弱监督方法，如 MWS [57] 和 WSSA [61]。进一步与全监督方法比较可以发现，本文方法同样表现出较强的竞争力。除 Coutrot2 数据集外，本文方法在其余测试集上的表现均优于全监督方法 DeepNet [39]。

造成上述结果差异的一个重要原因在于，Coutrot2 数据集的语义内容与 AVE 数据集之间存在较明显的分布差异，而本文方法的弱监督信号主要来源于 AVE 数据集中的类别标签。因此，当测试数据与训练数据在语义分布上存在较大偏移时，模型性能会受到一定影响。与此同时，这也说明，若后续能够引入更多带标签的视听视频序列，本文方法的性能仍有进一步提升的空间。

此外，本文方法与传统基于视频的 CAM 方法 [2, 3, 33] 在响应模式上也存在明显差异。传统方法往往倾向于持续高亮同一目标区域，而本文方法所强调的判别性区域会随着帧内容的变化而动态调整。这是因为在视听联合场景中，空间信息、时间信息和音频信息都可能在不同时刻对分类任务发挥主导作用。这样的动态响应特性与真实人眼注视行为更加一致，因为人类视觉注意通常不会长时间停留在同一位置，尤其是在视听共同作用的复杂场景中更是如此。

4.5 局限性

本文仅考虑了每个视听视频序列对应单一语义标签的情形，而在实际场景中，一个视频序列往往可能同时包含多个语义标签。因此，对于那些包含较多类别外语义内容的视频，本文方法的表现可能受到一定影响。未来若能够引入带有多标签标注的更多视听数据，这一问题有望得到缓解，这也是后续值得进一步研究的方向。

5 结论与未来工作

本文提出了一种由视听语义类别标签生成伪注视点的弱监督学习方法。与传统 CAM 方法相比，本文提出的 SCAM 能够生成与真实人眼注视分布更为一致的伪注视点。本文的关键改进主要体现在两个方面：一是基于多源信息的选择性融合机制，二是由粗到细的多阶段处理策略。相关组件实验表明，这两项设计均能有效提升伪注视点质量和模型整体性能。

表 4: 本文方法与全监督、弱监督及无监督方法之间的定量比较。加粗表示该列最优结果。受篇幅限制，对应的定性比较结果已放入补充材料中。

方式	方法	AVAD [36]					DIEM [26]					SumMe [20]				
		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
无监督	ITTI [24]	0.688	0.170	0.533	0.131	0.611	0.663	0.217	0.583	0.137	0.555	0.666	0.151	0.559	0.097	0.436
	GBVS [21]	**0.854**	0.247	0.572	**0.337**	**1.556**	**0.830**	0.318	0.605	**0.356**	**1.277**	**0.808**	0.221	0.567	**0.272**	**1.134**
	SCLI [42]	0.747	0.210	0.535	0.170	0.792	0.739	0.267	0.590	0.207	0.779	0.746	0.209	0.577	0.184	0.796
	SBF [60]	0.833	**0.272**	0.576	0.308	1.489	0.759	**0.292**	0.608	**0.301**	1.081	0.783	**0.228**	0.590	0.230	1.023
	AWS-D [30]	0.825	0.221	**0.589**	0.304	1.378	0.733	0.250	**0.612**	**0.301**	1.128	0.747	0.192	**0.603**	0.186	0.853
弱监督	GradCAM++ [3]	0.777	0.273	0.559	0.255	1.217	0.732	0.216	0.583	0.271	0.778	0.774	0.217	0.593	0.225	0.924
	VUNP [33]	0.574	0.067	0.500	0.142	0.292	0.558	0.047	0.515	0.172	0.186	0.555	0.013	0.507	0.114	0.048
	WSS [50]	0.858	0.292	**0.592**	0.347	1.655	0.803	0.333	0.620	0.344	1.293	0.812	0.245	0.589	0.279	1.098
	MWS [57]	0.834	0.272	0.573	0.309	1.477	0.806	0.336	0.628	0.350	1.308	0.808	0.237	0.607	0.258	1.155
	WSSA [61]	0.807	0.261	0.574	0.285	1.339	0.767	0.305	0.608	0.311	1.178	0.755	0.225	0.585	0.231	1.058
	OUR(STANet)	**0.873**	**0.334**	0.580	**0.438**	**2.018**	**0.861**	**0.391**	**0.658**	**0.469**	**1.716**	**0.854**	**0.294**	**0.627**	**0.368**	**1.647**
全监督	DeepNet [39]	0.869	0.256	0.561	0.383	1.850	0.832	0.318	0.622	0.407	1.520	0.848	0.227	0.645	0.332	1.550
	SalGAN [38]	0.886	0.360	0.579	0.491	2.550	0.857	0.393	**0.660**	0.486	1.890	**0.875**	0.289	**0.688**	**0.397**	**1.970**
	DeepVS [25]	0.896	0.391	**0.585**	0.528	3.010	0.840	0.392	0.625	0.452	1.860	0.842	0.262	0.612	0.317	1.620
	ACLNet [52]	**0.905**	**0.446**	0.560	**0.580**	**3.170**	**0.869**	**0.427**	0.622	**0.522**	**2.020**	0.868	**0.296**	0.609	0.379	**1.790**

方式	方法	ETMD [28]					Coutrot1 [9]					Coutrot2 [10]				
		AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
无监督	ITTI [24]	0.661	0.127	0.582	0.083	0.425	0.616	0.178	0.529	0.082	0.319	0.694	0.142	0.530	0.040	0.331
	GBVS [21]	**0.856**	0.226	0.613	**0.299**	**1.398**	**0.798**	**0.253**	0.526	**0.272**	**1.055**	**0.819**	**0.189**	0.577	**0.183**	**1.071**
	SCLI [42]	0.761	0.165	0.570	0.129	0.617	0.754	0.216	0.536	0.239	0.883	0.669	0.137	0.510	0.014	0.093
	SBF [60]	0.805	**0.232**	0.641	0.262	1.298	0.726	0.187	0.530	0.215	0.789	0.827	0.152	0.583	0.131	1.101
	AWS-D [30]	0.754	0.161	**0.664**	0.181	0.907	0.729	0.214	**0.581**	0.207	0.872	0.783	0.170	**0.590**	0.146	0.842
弱监督	GradCAM++ [3]	0.575	0.124	0.157	0.576	0.736	0.704	0.137	0.537	0.210	0.511	0.733	0.114	0.567	0.168	0.625
	VUNP [33]	0.505	0.030	0.103	0.132	0.593	0.589	0.063	0.514	0.152	0.304	0.661	0.101	0.536	0.162	0.491
	WSS [50]	0.854	0.277	0.661	0.334	1.650	0.772	0.247	**0.547**	0.233	0.975	0.835	0.208	0.578	0.192	1.178
	MWS [57]	0.833	0.237	0.649	0.293	1.425	0.743	0.231	0.528	0.201	0.798	0.839	0.188	0.581	0.168	1.197
	WSSA [61]	0.793	0.201	0.622	0.222	1.075	0.701	0.180	0.535	0.169	0.780	0.797	0.185	0.571	0.180	1.263
	OUR(STANet)	**0.908**	**0.318**	**0.682**	**0.448**	**2.176**	**0.829**	**0.306**	0.542	**0.339**	**1.376**	**0.850**	**0.247**	**0.597**	**0.273**	**1.475**
全监督	DeepNet [39]	0.889	0.225	0.699	0.387	1.900	0.824	0.273	0.559	0.340	1.410	0.896	0.201	0.600	0.301	1.820
	SalGAN [38]	0.903	0.311	**0.746**	0.476	2.460	**0.853**	0.332	**0.579**	0.416	1.850	**0.933**	0.290	0.618	0.439	2.960
	DeepVS [25]	0.904	**0.349**	0.686	0.461	**2.480**	0.830	0.317	0.561	0.359	1.770	0.925	0.259	**0.646**	**0.449**	**3.790**
	ACLNet [52]	**0.915**	0.329	0.675	**0.477**	2.360	0.850	**0.361**	0.542	**0.425**	**1.920**	0.926	**0.322**	0.594	0.448	3.160

在此基础上，本文进一步利用伪注视点训练 STA 注视点预测网络，并与现有多种最先进方法进行了系统比较。实验结果表明，本文方法不仅优于现有无监督和弱监督方法，而且在若干测试集上还表现出可与部分全监督方法相竞争的性能。这说明，基于语义标签构造伪注视点并用于视听注视点预测，是一条切实可行的研究路径。

未来的研究将进一步面向更细粒度的语义监督展开。一方面，可以探索自动化的类别标签挖掘方法，从复杂视听序列中识别对分类任务贡献更大的关键标签子集；另一方面，也可结合多标签语义信息构建更高质量的伪注视点监督信号。通过引入更精细、更丰富的语义约束，所生成的伪注视点有望进一步逼近真实人眼注视分布，从而持续提升视听注视点预测模型的性能。

参考文献

[1] Aditya Arun, C. V. Jawahar, and M. Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9432–9441, 2019.

[2] Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation backprop for RNNs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1440–1449, 2018.

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[4] Chenglizhao Chen, Shuai Li, Yongguang Wang, Aimin Hao, and Hong Qin. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. *IEEE Transactions on Image Processing*, 26(7):3156–3170, 2017.

[5] Chenglizhao Chen, Guotao Wang, Chong Peng, Xiaowei Zhang, and Hong Qin. A video saliency detection model in compressed domain. *IEEE Transactions on Image Processing*, 29:1090–1100, 2020.

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020.

[7] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xiangsheng Hua. SLV: Spatial likelihood voting for weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12995–13004, 2020.

[8] Runmin Cong, Jianjun Lei, Huazhu Fu, Fatih Porikli, Qingming Huang, and Chunping Hou. Video saliency detection via sparsity-based reconstruction and propagation. *IEEE Transactions on Image Processing*, 28(10):4819–4831, 2019.

[9] Antoine Coutrot and Nathalie Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, 14(8):5, 2014.

[10] Antoine Coutrot and Nathalie Guyader. Multimodal saliency models for videos. In *From Human Attention to Computational Attention: A Multidisciplinary Approach*, pages 291–304, 2016.

[11] Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643, 2015.

[12] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified image and video saliency modeling. *arXiv preprint arXiv:2003.05477*, 2020.

[13] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8554–8564, 2019.

[14] Yuming Fang, Guanqun Ding, Jia Li, and Zhijun Fang. Deep3DSaliency: Deep stereoscopic video saliency detection model by 3D convolutional networks. *IEEE Transactions on Image Processing*, 28(5):2305–2318, 2019.

[15] Yuming Fang, Zhou Wang, Weisi Lin, and Zhijun Fang. Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE Transactions on Image Processing*, 23(9):3910–3921, 2014.

[16] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-MIDN: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9834–9843, 2019.

[17] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[18] Siavash Gorji and James J. Clark. Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7501–7511, 2018.

[19] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Mingming Cheng, and Shaoping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 10869–10876, 2020.

[20] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 505–520, 2014.

[21] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 545–552, 2007.

[22] Sayed Hossein Khatoonabadi, Nuno Vasconcelos, Ivan V. Bajic, and Yufeng Shan. How many bits does it take for a stimulus to be salient? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, 2015.

[23] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5009, 2018.

[24] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[25] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. DeepVS: A deep learning based video saliency prediction approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 602–617, 2018.

[26] Parag K. Mital, Tim J. Smith, Robin L. Hill, and John M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011.

[27] Seunghyeon Kim, Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6092–6101, 2019.

[28] Petros Koutras and Petros Maragos. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 38:15–31, 2015.

[29] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE Transactions on Image Processing*, 29:1113–1126, 2020.

[30] Victor Leboran, Anton Garcia-Diaz, Xose R. Fdez-Vidal, and Xose M. Pardo. Dynamic whitening saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):893–907, 2017.

[31] Bo Li, Zhengxing Sun, and Yuqi Guo. SuperVAE: Superpixelwise variational autoencoder for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8569–8576, 2019.

[32] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7274–7283, 2019.

[33] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1120–1129, 2020.

[34] Panagiotis Linardos, Eva Mohedano, Juan Jose Nieto, Noel E. O'Connor, Xavier Giro-i-Nieto, and Kevin McGuinness. Simple vs. complex temporal recurrences for video saliency prediction. *arXiv preprint arXiv:1907.01869*, 2019.

[35] Kyle Min and Jason J. Corso. TASED-Net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2394–2403, 2019.

[36] Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang. Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia*

Computing, Communications, and Applications, 13(1):1–23, 2016.

[37] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020.

[38] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i-Nieto. SalGAN: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[39] Junting Pan, Elisa Sayrol, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–606, 2016.

[40] Rui Qian, Yunchao Wei, Honghui Shi, Jiachen Li, Jiaying Liu, and Thomas Huang. Weakly supervised scene parsing with point-based distance metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8843–8850, 2019.

[41] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. TENet: Triple excitation network for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 212–228, 2020.

[42] Dmitry Rudoy, Dan B. Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1147–1154, 2013.

[43] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6956–6965, 2019.

[44] Meijun Sun, Ziqi Zhou, Qinghua Hu, Zheng Wang, and Jianmin Jiang. SG-FCN: A motion and memory-based deep learning model for video saliency detection. *IEEE Transactions on Cybernetics*, 49(8):2900–2911, 2019.

[45] Hamed R. Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala. DAVE: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*, 2019.

[46] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[47] Antigoni Tsiami, Petros Koutras, and Petros Maragos. STAViS: Spatio-temporal audio-visual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4766–4776, 2020.

[48] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-MIL: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2199–2208, 2019.

[49] Bo Wang, Wenxi Liu, Guoqiang Han, and Shengfeng He. Learning long-term structural dependencies for video salient object detection. *IEEE Transactions on Image Processing*, 29:9017–9031, 2020.

[50] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 136–145, 2017.

[51] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2018.

[52] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video

saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4894–4903, 2018.

[53] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):220–237, 2021.

[54] Xinyi Wu, Zhenyao Wu, Jinglin Zhang, Lili Ju, and Song Wang. SalSAC: A video saliency prediction model with shuffled attentions and correlation-based ConvLSTM. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 12410–12417, 2020.

[55] Zhenheng Yang, Dhruv Mahajan, Deepti Ghadiyaram, Ram Nevatia, and Vignesh Ramanathan. Activity driven weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2917–2926, 2019.

[56] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2Det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9686–9695, 2019.

[57] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6074–6083, 2019.

[58] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8292–8300, 2019.

[59] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:12756–12772, 2020.

[60] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4048–4056, 2017.

[61] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12546–12555, 2020.

[62] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9029–9038, 2018.

[63] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1325–1334, 2018.

[64] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.