电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

专业学位硕士学位论文

MASTER THESIS FOR PROFESSIONAL DEGREE

论文题目 多模态视频语义分析理解研究

专业学位类别	电子信息
学　　号	202022080215
作者姓名	谭乾超
指导教师	姬艳丽　教授
学　　院	计算机科学与工程学院 (网络空间安全学院)

分类号 TP391 密级 公开

UDC 注 1 004.8

学 位 论 文

多模态视频语义分析理解研究

（题名和副题名）

谭乾超

（作者姓名）

指导教师 **姬艳丽 教授**

电子科技大学 成都

（姓名、职称、单位名称）

申请学位级别 **硕士** 专业学位类别 **电子信息**

专业学位领域 **计算机技术**

提交论文日期 **2023 年 3 月 27 日** 论文答辩日期 **2023 年 5 月 16 日**

学位授予单位和日期 **电子科技大学 2023 年 6 月**

答辩委员会主席

评阅人

注 1：注明《国际十进分类法 UDC》的类号。

Research on Semantic Analysis and Understanding of Multimodal Video

A Master Thesis Submitted to
University of Electronic Science and Technology of China

Discipline:	**Electronic Information**
Student ID:	**202022080215**
Author:	**Tan Qianchao**
Supervisor:	**Prof. Ji Yanli**
School:	**School of Computer Science and Engineering(School of Cyberspace Security)**

摘 要

随着存储技术的不断升级以及深度学习的不断发展，多模态数据的存储与处理也更加容易与高效，这也进一步为基于多模态视频学习的研究提供了支撑。目前音视频事件定位与音视频事件解析任务是多模态视频学习领域中的主要研究任务之一，该任务在人机交互、视频监控等领域具有广泛的应用，但是目前的一些研究方法还存在三个主要问题:（1）在对模态进行处理时，仅单方面地对视觉信息进行处理，而没有考虑到音视两种模态间的交叉关注以及对音频的特殊处理；（2）对音视频两种模态的模态内和模态间关系的探索上，还存在不足，现有大多数基于简单注意力机制的算法不能有效地利用不同片段上的关键信息，（3）在弱监督情况下，仅通过平均池化的方法对片段级别上的预测结果进行聚合并不有效，目前方法不能针对性地对不同模态不同片段上的结果进行聚合以形成更鲁棒的视频级别上的预测结果。因此，针对上述挑战，本文进行了如下工作：

（1）在音视频事件定位中针对音视模态间的交叉关注上，本文以预处理后的音视融合特征作为指导，在消除视觉背景区域干扰的同时聚焦整个视觉区域中与对应音频模态相关的关键区域，并对音频模态信息进行关注。此外，还通过残差连接的方式，对与之对应的音频特征进行了信息增强。

（2）在音视频事件定位中针对音视两种模态在片段内与片段间的关系探索上，本文一方面，基于自注意力机制对音视两种模态内在片段间的关系进行探索，另一方面，通过类似于异构图处理算法来对模态间的不同片段进行建模，并通过对比片段间的相似性值与阈值大小来更新片段间的关系，最后通过注意力加权的方式来聚合其它片段为当前片段所带来的信息。

（3）在弱监督情况下的预测上，针对音视频事件定位任务，本文基于多实例学习的方法，通过对所预测的事件相关性得分以及事件类别得分进行池化聚合以得到视频级别上的预测结果；针对音视频事件解析的任务，在多模态多实例池化方法的基础上，利用具有注意力的池化损失和对比学习损失对网络进行约束，并以一种三分支学习方式来更好的针对每种模态完成预测。

（4）本文对每个研究场景所提出的算法进行了充分的实验验证，同时在与该领域的最新方法进行了对比分析和消融实验，进一步证明了算法的有效性。

关键词：多模态学习，交叉注意力，音视联合学习，音视事件定位，音视事件解析

ABSTRACT

With the continuous upgrading of storage technology and the continuous development of deep learning, the storage and processing of multimodal data is also easier and more efficient, which further provides support for research based on multimodal video learning. At present, the task of audio-visual event location and audio-visual video parsing is one of the main research tasks in the field of multimodal video learning. This task also has a wide range of applications in the fields of human-computer interaction, video surveillance and other fields. However, there are still three main problems in some current research methods: (1) When paying attention to different modalities, they only unilaterally processes the visual information without considering the cross-focus between the two modalities of audio and video and the special processing of audio. (2) The exploration of the relationship between the two modalities of audio and video is not fully explored, and the processing methods are only based on the simple attention mechanism, which cannot effectively use the key information on different segments. (3) In the case of weak supervision, it is not effective to aggregate the prediction results at the segment level only by the average pooling methods. At present, the results of different segments for different modalities can not be aggregated specifically to form more robust prediction results at the video level. Therefore, in response to the above challenges, this thesis further does the following work on the audio-visual event location and audio-visual video parsing tasks:

(1) In terms of cross-focus between audio and visual modalities for audio-visual event location, this thesis uses the pre-processed audio-visual fusion features as the guidance to focus on the key areas related to the corresponding audio modality in the whole visual area while eliminating the interference from the visual background area, and pay attention to the audio modality information. In addition, the information of the corresponding audio features is enhanced by means of residual connection.

(2) On the exploration of the relationship between audio and visual modalities within and between segments for audio-visual video parsing, on the one hand, this thesis explores the relationship between the internal segments of audio and visual modalities based on the self-attention mechanism; on the other hand, this thesis model the relationship between segments of different modalities use the algorithm which is similar with the processing algorithm for different types of nodes in the heterogeneous graph, and the relationship

between segments is updated by comparing the similarity value between segments with the threshold value. Finally, the information brought by other segments for the current segment is aggregated by attention weighting.

(3) On the prediction under weak supervision, for the task of audio and video event location, this thesis based on the method of multi-instance learning, through pooling aggregation with the predicted event correlation scores and event category scores to get the prediction results at the video level, while for the task of audio-visual video parsing, based on the multimodal multiple-Instance learning method, the network is constrained by the loss of pooling with attention and the contrastive learning loss, and a three-branch learning method is used to get better prediction for each modality.

(4) In this thesis, the proposed algorithms in each research scenario are fully verified by experiments, and the effectiveness of the algorithms are further proved by comparative analysis and ablation experiments with the latest methods in the related fields.

Keywords: Multimodal, Cross-attention, Audio-visual learning, Audio-visual event location, Audio-visual video parsing

目　录

第一章 绪 论

1.1 研究工作的背景与意义

在人类的感知过程中，通常会涉及到多个感官通道，如视觉、听觉、触觉等，人们通过处理和融合源于不同感官通道的信息来理解周围的世界，并对其作出反应[1]。在现实生活中，很多应用场景都需要对声音和视觉信号进行联合处理，以获得更为准确的信息。然而在传统针对音频或视频的分析理解研究中，通常考虑的都只是单一模态的数据信息。例如，在语音识别中[2,3]，现有的一些研究一般只是单独分析音频信号；在视频监控[4]中，也只是对图像视觉信息进行分析，但人类的感知和认知过程是复杂的，它涉及到大脑的多个区域和不同的神经系统。这些神经系统的协同工作，可以使人们更加快速、准确地理解周围世界并做出相应的行为。例如，当人们在听到小狗叫的声音时，会下意识地去寻找发声源在哪里，当人们找到并看见这只正在叫的小狗时，会进一步地认识到目前正在发生的事情。目前，基于多模态的算法研究已经成为人工智能和机器学习领域的热门研究方向，同时随着计算机硬件处理性能的不断提高、传感器技术的不断发展以及存储技术的不断升级，多模态数据的处理、获取和存储都变得更加容易，这也为多模态视频语义分析理解研究的发展提供了坚实的基础。为此，本文不在局限与对单模态信息的处理，而是研究如何结合能够获得的多模态信息来理解所正在发生的事件。此外，对多模态视频语义分析理解的研究可以有助于一些计算机应用的发展，更有现实意义，如在智能交互系统[5,6]、智能家居[7]等产品中通过结合相应的语音、手势以及面部表情等多种感知源可以为人类带来更加智能与自然的人机交互体验。这些多模态技术的应用离不开相关研究的理论支持。

随着人工智能和深度学习的发展，机器感知模型也开始从利用单模态变得越来越趋向于使用多模态[8,9]。在机器学习和数据科学领域[8]中，多模态数据也被广泛应用于模型的训练和数据的分析。目前，在多模态视频语义分析理解研究中，主要存在两种研究方向，一种是通过对视频中多模态信息的处理来完成整个视频级别上的动作识别[10–12]，另一种是对视频中进行分段处理后，基于不同片段上的多模态信息以及不同片段间的联系，来完成对每个片段上所发生事件的理解。后者主要包括音视频事件定位[13]以及音视频事件解析[14]这两个任务。相比于前者，后者的事件理解范围所涉及到的区域更加广泛，不在局限与仅对人体动作的理解还包括对生活场景中的一些事情的识别，如狗叫、赛车、研磨等事件，会更加的符合现实情况，同时，后者在处理上也更加细粒度，具体处理到视频中的某个片

段，处理要求上相比前者也更加复杂。考虑到后者具有更广泛的现实意义与理论价值以及在对多模态信息利用上的挑战性，所以本文探究的是在多模态视频理解领域中针对片段级别上的识别，即针对音视频事件定位任务以及音视频事件解析任务进行研究。本文希望能够训练网络模型利用视频中至少两种模态的信息如音频与视频信息，在音视频事件定位任务以及音视频事件解析任务上取得比单模态要好的性能以完成多模态信息融合下的视频理解，以及为这两个任务的现有研究提供更有参考价值的处理方式。

音视频事件定位以及音视频事件解析这两个任务互相联系，后者为前者的进一步扩展。他们两者都旨在通过从视频中提取相应的音频特征信息与视觉特征信息，然后同时进行分析以准确地理解视频中某片段所正在发生的事件，两者的区别在于识别上的精度以及实现难度上的不同。音视频事件定位任务处理的对象是音视联合事件，所要识别的事件必须即是在视觉上可见的又是在音频上可听的，并不考虑单独的音频或者视觉事件，更侧重与对音视两种模态的联合处理，为此本文在第三章对该任务进行了探究。而音视频事件解析任务处理的对象不在是音视联合事件，而是视频中每种模态在不同片段上所发生的事件，同时，每个视频上不同模态所发生的事件也会存在多个，处理方式更具挑战性。此外该任务需要以一种弱监督的方式进行实现，在对具体实现上也更加复杂，为此，本文在第四章对该任务进行了探究进一步研究。虽然目前在这两个领域已经有了许多研究成果，但是这些方法主要是通过简单的中期或后期融合来处理各种模态的信息，仍然还是存在一些未解决问题。因此在结合视频中多模态的信息尤其是视觉上的信息与听觉上的音频信息来进一步对视频的语义进行分析与理解上，还是有许多可探索的空间。此外，基于多模态的音视事件定位与音视频事件解析技术在虚拟现实[15]、增强现实[16]、视频监控和语音识别等多个领域都具有重要的应用前景，通过将多种模态的数据结合起来，人们可以更全面地描述和模拟复杂的现实世界问题，从而提高网络模型预测和决策的准确性与可靠性。因此，对跨模态信息的提取和对应进行研究的极具实际应用价值的任务的，相关研究不仅可以推动学术领域的发展，也能为实际应用带来巨大的社会和经济效益。本文相信，随着相关技术的不断发展和突破，其应用范围也将越来越广泛。

1.2 国内外研究历史和现状

音视频事件定位与音视频事件解析任务是一个涵盖多个知识领域的研究任务，涉及计算机视觉、机器学习、音视对应、多模态学习等多个领域。目前在国内外已经有很多学者和机构从不同的角度对相关领域进行了研究，并提出了不同的算

法。在本节中，本文将主要阐述与音视频对应、音视频事件定位以及音视频事件解析相关工作的研究历史和现状。

1.2.1 音视频对应

音视频对应（Audio-Visual Correspondence, AVC）任务首先是由 DeepMind 团队的 Arandjelovic 等[17] 提出。该任务的目标是训练网络来判断同时输入的视觉信息和声音信息是否来自于同一个视频。该任务的一个重要假设是：同一个视频中的视觉信息和音频信息是天然对应的，而不同视频间的视觉信息与音频信息是不匹配的，如果一个网络能够很好地对输入音视对进行判断，那么该网络能够正确提取到关键信息。论文 [17] 通过实验证明这种方式所学习到的视觉特征与音频特征可以在一些下游任务上取得很好的效果。此外，Andrew Owens 等[18] 采用和论文 [17] 类似的网络结构来学习音视特征，只不过将视觉特征提取网络换成了 3D 网络，以用于捕获动态信息。论文 [19] 基于 AVC 任务，设计了一个名为 AVE-Net 的网络架构，用于跨模态检索以及在图像中进行声源定位。与利用 AVC 任务来训练网络学习关键特征的思想类似，Andrew Zisserman 等[20] 在音唇同步任务上利用视频中嘴巴的动作和语音传递之间的同步关系进行了算法设计。Joon Son Chung 等[21] 则是利用视频中人脸和音频之间的自然同步关系来构建自监督任务以完成说话者识别任务。与 AVC 任务不同的是，Bruno Korbar 等[22] 提出了一个辨别一段音频和一段无声视频是否在时间上同步的方法 (Audio-Visual Temporal Synchronization，AVTS) 来学习一个用于音视频特征提取的通用网络模型。此外，相比于 AVC 任务中样本对的选取，在论文 [22] 中，作者使用了“更难”的负样本对来训练模型。

在 AVC 任务之前，麻省理工学院（MIT）的学者们已经对音视频的对应关系进行了研究[23,24]，只不过其目的是为了训练单模态的特征提取网络。同时，Andrew Owens 等[23] 利用环境音来为视觉特征模型的训练提供超视觉信号。Yusuf Aytar 等[24] 为了使网络学习到更好的声音特征表示，他们利用一个教师学生网络，以未标记的视频作为桥梁将视觉网络中的视觉信息迁移到声音模态中。

在这个任务之后，Ying Cheng 等[25] 指出像 AVC、AVST 这样的方法忽略了模态之间的信息交换并且这些自监督任务在复杂场景如多音频下的情况，比较受限制。为此，他们提出了一种具有共同注意力机制的自监督框架以提供音频和视觉流之间的信息交换。Nuno Vasconcelos 等[26] 使用对比学习来跨模态地区分视频和音频。此外，他们通过跨模态的一致性来探索对比学习中的正负样本对的构建。Di Hu[27] 提出了一种称为深度协同聚类的无监督视听学习模型，以捕获多模态的视听对应关系。Vandana Rajan[28] 等提出了一种通过跨模态转换和对齐的方法来获

得强大的潜在特征表示，可在训练期间使用多种模态数据来提高单模态系统的测试性能。

针对音视频对应任务的相关研究是视频多模态学习领域中的一个重要方向。在音视频对应任务中，普遍认为视频中同时出现的音频信息与视觉信息存在天然的对应关系。这种对应关系的假设思想值得借鉴，可以用在相关领域的研究中。所以，以 AVC 任务为代表，衍生出一系列的相关研究，如在声源分离领域中对同时出现的视觉信息的利用[29–33] 等。不同于音视频对应任务中要求天然对应的假设，本文所要研究的场景更加现实，并不需要假设视频中的音频信息与视觉信息是一致对应的。在音视频对应中对音视两种模态特征的处理方式对本文的算法给予了启发，尤其是通过对音视频特征间的相似性计算来判断两者之间的联系，这在一定程度上能够启发本文更好地处理所要研究场景中不同音视片段间的关系。本文相信随着计算机视觉和语音处理技术的发展，音视频对应的研究将会更加深入和广泛。

1.2.2 音视频事件定位

音视频事件定位（Audio-visual event localization，AVE）任务首先是由罗彻斯特大学的田亚鹏在论文 [13] 中提出的。该任务的目标是要同时结合音频与视觉模态信息，以识别出视频中的哪些片段所发生的事件即是可见的又是可听的，并要确定所发生音视事件的类别。为了探索视觉模态与音频模态之间的相关性。基于注意力机制，论文 [13] 提出了一个针对音视双模态的双向残差网络来对两种模态的信息进行融合，并利用一个远程视听学习网络来处理跨模态间的定位问题。之后，Yan-Bo Lin 等[34] 提出了一种以序列到序列的方式来处理视频段中各种模态所带来的局部和全局信息，从而促进网络对正在发生的音视事件的理解。与之不同的是，台湾国立大学的 Yan-Bo Lin 等[35] 提出了一种基于视听转换器（AV-transformer）的框架，该算法利用视频帧内与帧间的视觉信息，并结合共同观察到的音频信息来学习三者之间的关系，从而获取利于事件定位的有用信息。随着研究的深入，学者们也越来越注重如何有效处理模态信息的方法，华南理工大学的 Haoming Xu 等[36] 提出了一种关系感知网络，其中关系感知模块能够在视觉和音频模态之间建立良好的联系。特别地，为了减少背景带来的干扰，他们还提出了一个音频引导的空间通道注意力模块，分别在空间与通道上来引导模型更加关注与事件相关的视觉区域。

之后也有许多学者主要从注意力机制的角度进行研究，并提出了许多新颖方法 [37–41]。如论文 [37] 所提出的视听融合模块以及片段级上的注意力模块能在解

决音视频事件定位任务上起到助力作用。此外，Bin Duan 等[39] 利用可迭代的联合注意力机制，提出了一种可迭代学习的网络模型，希望能够以一种正反馈的机制来从多种模态中学习特征表示的方法，从而关注视觉特征和听觉特征。在建立片段间联系上，合肥工业大学的 Jinxing Zhou 等[40] 利用正样本传播模块来评估每个可能的视听对中的关系以发现和探索密切相关的视听对。受人类多模态感知机制的启发，Hanyu Xuan 等[41] 提出了一种由三个注意模块（where, when, which）组成的跨模态注意框架，以充分挖掘模态内和跨模态片段在时间与空间上的潜在隐藏关联。考虑到片段持续时间通常很短，其所表达的信息具有局部性，百度研究所的 Yu Wu 等[38] 提出了一个双重注意匹配模块来覆盖更长的视频持续时间，以获得更好的高级事件信息，同时通过全局交叉检查机制在获取局部时间信息的同时筛去背景信息。

要完成音视事件定位这个任务，需要模型能够对视频有着很好的语义理解，同时也需要探究音视模态中在不同片段之间的关系。虽然上述的已有方法已经从不同方面对音视频事件定位任务进行了探索，但是这些方法并没有针对性地对两种模态间的信息进行交互关注，而且在处理不同模态在模态内与模态间的关系上探索得也不是很深入。为此，本文基于这些问题进一步地进行了相关研究。

1.2.3 音视频事件解析

音视频事件解析（Audio-Visual Video Parsing，AVVP）任务首先是由罗彻斯特大学的田亚鹏在论文 [42] 中提出的。与音视频事件定位任务中，只需要识别视频片段中所发生的音视事件类别不同，音视频事件解析任务所要做的是识别一个视频中在每个片段上基于单独的音频或者视觉模态所发生的事件类别以及基于音视联合模态所发生的事件类别，这需要网络模型对每个模态都有一定的认知能力。该任务相比于音视频事件定位任务，在视频理解上进行了进一步的细化处理，为此田亚鹏等[42] 在对视频特征进行提取后，利用模态内与模态间的混合注意力机制对视频上不同片段间的关系进行处理，以学习片段在时序上、空间上以及跨模态上的关联，之后利用一种多模态多实例学习（Multimodal Multiple Instance Learning，MMIL）的池化方法来解决弱监督下的音视频事件解析任务。而论文 [43] 则是在此基础上结合自监督损失，进一步地通过一种对抗学习以及跳接级联的方式来进行训练。为了在不同时间尺度和粒度上进行特征探索，复旦大学的冯瑞等[44] 提出了一种名为多模态金字塔注意力的网络（MM-Pyramid）来关注视频模态在不同尺度下的语义信息，其中每个特征金字塔模块都是由几个堆叠的金字塔单元组合而成，而这些单元又由不同尺度的卷积块和注意力块组成。与方法 [42–44] 不

同，百度研究所的 Yu Wu 等[45] 认为在弱监督下将视频级别上的标签分别赋值给音频模态与视觉模态是不合理的，虽然利用注意力下的 MMIL 池化方法可以减轻噪声标签的影响，但是效果还不够，为此他们通过交换两个不相关的视频之间音频或视频模态信息，然后利用一种两阶段的训练方法来对每种片段上不同模态上所发生的事件标签进行去噪。与之类似，论文 [46] 利用标签去噪领域中的一些方法 [47–49] 设计了一种新的联合模态噪音消除的训练策略，在噪声评估器的指导下动态地减轻模态特定噪声对弱监督下音视频事件解析的影响。

目前，许多针对多模态的研究方法 [18,20–22,50,51] 通常都会假设一个视频中的音频信息与视觉信息都是紧密联系或者相互对齐的，然而在实际情况中，只能听见声音而具体发声物体却不在视觉场景中的例子无处不在，而且也不能否认这种情况下，单单借助音频或者视觉信息也是有助于理解当前所正在发生的事情[52]。音视频事件解析任务则是进一步地在此情况下进行探索研究。目前，该领域的主要研究方法可以分为两类，一类是基于视频中不同模态上的不同片段在内容以及时序关系上的探索与处理，而本文所提出的算法也是从这个角度出发的。另一类则是为了解决在弱监督下视频级别上的标签并不会和每种模态上所发生的标签一致的角度出发，通过设置不同的训练方法或者细化各种模态损失等方法来对每种模态上的冗余标签进行消除。本文一方面对视频中不同模态在模态内与模态间的关系上进行了探索，另一方面也在噪声标签消除的角度上，基于目前的一些方法进行了增强。

1.3 本文的主要研究内容与贡献

针对音视频事件定位任务，本文的主要贡献为：

（1）本文提出了一个音视间的共同注意力模块（Audio-Visual Co-attention Module，AVCM) 以预处理后的音视融合特征作为指导，在消除视觉区域中无关信息影响的同时聚焦整个视觉区域中与对应音频模态相关的关键事件区域，并对于与当前片段所对应的音频模态特征进行增强，从而使模型能够有效地聚焦不同模态中的关键信息来理解视频中事件。

（2）本文基于异构图算法中处理不同类型的邻接结点的思想，提出了一个多模态间的片段交互模块（Multimodal Segment Interaction Module，MSIM）来处理当前片段与其他同模态片段以及不同模态片段间的交互，并通过阈值的设定来更新片段间的联系，最后通过注意力加权的方式来聚合其它片段所带来的信息，一方面能够消除音视不一致的影响，另一方面，能够使当前片段学习到一些全局上的事件信息。

（3）本文所提出的基于模态间交叉注意力机制的多模态网络算法（Multimodal Network with Cross-Modal Attention，MCMA）通过结合所预测的事件相关性得分与事件类别得分在全监督与弱监督情况下完成音视频事件定位任务，并在 AVE 数据集上，通过大量的实验证明了本文所提算法的有效性。

针对音视频事件解析任务，本文的主要贡献为：

（1）本文提出了一种利用三分支网络来分别处理视频中的音视模态以及音视联合模态的方法。在对音视联合信息的处理上，一方面，利用多头的自注意力机制对同模态内在不同片段上的关系进行处理，另一方面，利用音视模态间的交互模块（Audio-Visual Interaction Module，AVIM）来处理不同模态在不同片段上的交互，从而更好的为网络学习到视频级别上整体信息。

（2）针对单独的音视特征学习分支，本文通过利用级联的 Transformer 编码器来对单模态内不同片段在上下文的关系进行建模，使其能够在保持自身信息的同时，通过均分误差损失对音视交互分支进行信息指导。

（3）本文以一种多模态多实例学习的方式[42]来解决弱监督下的音视频事件解析任务。利用一种注意力下的池化方法以及对比学习来缓解在训练过程中只有视频事件标签而没有片段级别标签的影响，并在 LLP 数据集上通过大量实验对比分析了本文所提出的算法相比于 baseline 的有效性。

1.4 本论文的结构安排

在本节中，将对本论文每章的节结构安排进行介绍，五个章节的具体安排如下：

第一章：首先介绍了多模态视频语义分析理解的研究背景与应用前景，然后阐述了本文所要研究的两个任务：音视事件定位与音视频事件解析，并对目前现有的一些主要研究方法，从算法的出发角度以及算法的优势与局限性进行了简要分类概述，最后对本文的主要研究内容与关键贡献进行总结。

第二章：主要介绍了本文算法在实现过程中所用到的相关技术与理论支撑。首先对两个关键的卷积神经网络进行介绍，然后对一些评价指标的计算方式进行解释说明，包括混淆矩阵、F1 分数等，之后进一步对与多实例学习以及异构图相关的知识点进行讲解，最后对常用的多模态特征融合方法进行阐述与总结。

第三章：对本文针对音视频事件定位任务所提出的基于模态间交叉注意力机制的多模态网络算法以及一些验证实验进行详细的说明。首先介绍了音视频事件定位任务的具体问题描述和数学符号定义，然后介绍了所提出的框架中每个模块的具体实现细节以及相关的计算公式，并对相关的损失函数以及最终任务的实现

流程进行阐述，最后介绍了算法所使用的数据集的情况、数据的预处理方式、实验中的具体参数设置以及在全监督与弱监督下与当下最新的研究方法的比较分析。此外，还通过消融实验来验证算法中每个模块的作用，并给出部分可视化结果以直观地进行对比分析。

第四章：对本文所提出的弱监督下的针对音视频事件解析的多模态学习算法以及实验流程和结果分析进行阐述。首先介绍了音视频事件解析任务的具体定义和实现方式，其次介绍了所提出的算法中每个模块的具体实现细节以及具体输入输出，并对弱监督下的多模态多实例学习池化损失的实现流程进行阐述，最后介绍了算法所使用的数据集的情况、实验中的具体参数设置以及与目前研究方法的比较分析。此外，还从不同角度对算法进行消融实验以验证每个模块的作用。

第五章：对全文工作进行总结并对后续工作进行展望。

第二章 相关理论与技术介绍

本章将对本文所提出的算法在实现过程中所涉及到的相关理论知识和技术进行解释说明，主要包括两种经典的卷积神经网络结构、相关评价指标的具体计算公式、与多实例学习以及异构图相关的知识点以及常用的多模态特征融合方法。

2.1 卷积神经网络

2.1.1 VGG

VGG[53]是由牛津大学的 Simonyan 等提出的，他们所提出的 VGG 网络结构在 2014 年的 ImageNet 挑战赛上针对相应的定位任务取得了当时最好的成绩。他们[53]通过实验结果证明了随着网络深度的增加，网络的性能也会表现得更好。VGG 常用的网络结构有 VGG16 和 VGG19，这两种网络结构本质上并没有什么区别，只是网络深度上的不一致，所以也有 VGG11、VGG13、VGG-like 等类似的网络结构。与 VGG 网络的相关 6 种配置如图2-1所示。

在图2-1中，可以把 VGG 网络看成是对多个 VGG 块的堆叠，其中每个 VGG 块中又是由多个卷积层，一个 ReLU 激活函数层，以及一个池化层组成。第三行中的数字所表示的是整个网络中包含的参数层的数量（不算池化层，只算卷积层与全连接层）。以 VGG19 为例，可以看出该网络是由 5 个 VGG 块堆叠共计 16 个卷积层，以及最后的 3 层全连接组成，一共 19 层，因此命名为 VGG19。

与其它一些网络结构[54,55]中所采用的大卷积核（5×5，7×7，11×11）不同，VGG 网络中采用重复堆叠的小卷积核（3×3）来替代大卷积核，从而在可以获取相同感受野的情况下，提升网络的深度，以进一步地提升网络的特征提取的能力，此外，ReLU 激活函数的使用也可以进一步提升网络非线性变化的能力。论文 [53] 中通过实验证明了 VGG19 的效果最佳，但由于 VGG19 参数量较大，实际中也会使用 VGG16，即图2-1中的 E 设置。

在音视频事件定位任务上，本文使用 VGG19 网络对音频进行特征提取并使用 VGG-like 网络对视觉图像进行特征提取。

2.1.2 ResNet

ResNet[56]是在 2015 年由何凯明所提出来的，该网络结构在当年的 ImageNet 挑战赛上针对图像任务取得了第一名的成绩。目前与 ResNet 网络相关的网络结构，如 ResNet18、ResNet50 以及 ResNet101 等在目标检测，图像分割，动作识别

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 LRN	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

图 2-1 VGG 网络结构示意图[53]。

等领域中得到广泛的应用。ResNet 的网络结构图如图2-2所示：

可以看到，不论是 ResNet18、ResNet50 还是 ResNet101 或者 ResNet152，他们结构类似。都是先针对输入进行一个 7×7 的卷积层，然后再经过一个 3×3 的最大池化下采样层，之后就是按照图中的 conv2_x 层到 conv5_x 层进行运算，最后再跟一个平均池化的下采样层和全连接层以及 Sofmax 函数的输出。其中每个层都由若干个卷积块堆叠而成同时还进行残差操作，这些层之间的区别就是通道数的变化以及输出尺寸上的变化。

VGG 网络的成功，证明了网络的深度是网络能够取得较好效果的一个重要因素，但是随着网络深度的提升，一方面会带来网络参数的激增，另一方面会带来梯度消失或者梯度爆炸。网络参数量的增加可以通过相应的硬件设备的升级进行解决，但是梯度消失或梯度爆炸却会导致网络无法收敛。这是因为损失的计算是通过反向传播函数实现的，在这个过程中若是每次的求导梯度值都大于 1，那么随

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

图 2-2 ResNet 网络结构示意图[56]。

着网络的深度增加，梯度堆叠值会趋于无穷大，即梯度爆炸。若是每次的求导梯度值都小于 1，那么梯度堆叠值会逐渐趋于 0，即梯度消失。虽然，通过初始化等操作可以在一定程度上进行缓解，但是此时会导致网络退化。而 ResNet 网络之所以应用如此广泛，就是因为何凯明针对在网络中出现的梯度消失或梯度爆炸设计了一种残差结构，这种残差结构的设置可以在很大程度上解决这类问题。残差块结构中的核心操作是使用短接（Shortcut Connection）的方法来构建，ResNet 中短接的结构示意图如图2-3所示：

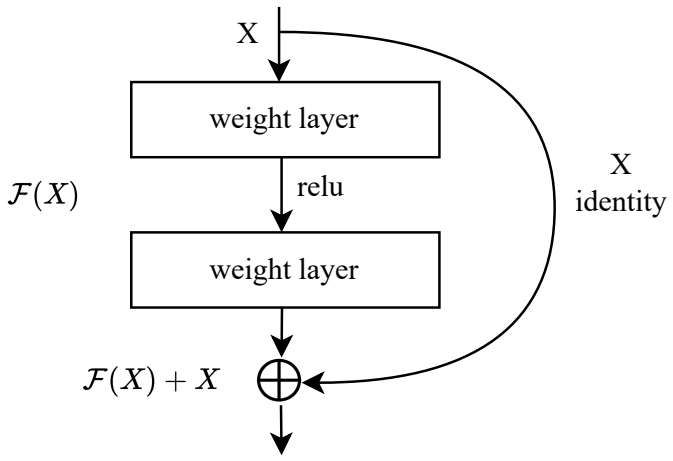

图 2-3 ResNet 短接结构示意图[56]。

图2-3中针对输入进行了两种映射，一种是针对输入的恒等映射（identity mapping），即右边的曲线部分以保证映射后的 X 与 $\mathcal{F}(X)$ 在维度上是一致的，另一种的是针对输入的残差映射（residual mapping），即左边的通过两个卷积层所获

得的 F(x) 部分。这个 F(x) 就是残差，短接指的是通过对残差 F(x) 和恒等映射后的 x 进行累加操作。事实上，短接也可以跨越多层来进行连接。如果网络在训练过程中已经达到最优，当再继续加深网络时，残差映射将不再起作用，那么就只剩下了恒等映射，在这样的情况下，网络在理论上就会一直处于最优的状态。

在音视频事件解析任务上，本文中所使用的是 ResNet152 网络来对视频特征进行提取，以获取较好的特征表达。

2.2 混淆矩阵与 F1 分数

2.2.1 混淆矩阵

混淆矩阵（Confusion Matrix）是针对分类算法模型的效果进行分析的一种常用方法[57]，有助于更加直观的对模型效果进行分析。混淆矩阵以矩阵的形式来对样本的真实属性以及分类模型所预测的结果进行关系分析，由此矩阵可得出一系列相关计算指标。一般来说，该矩阵横轴上的标签所代表的是样本的真实标签，而矩阵纵轴上的标签所代表的是网络模型针对样本的分类预测结果，那么由此便会产生四种样本情况。二分类下混淆矩阵的表现形式如图2-4所示：

		真实标签	
		1	0
预测标签	1	TP	FP
	0	FN	TN

图 2-4 二分类下混淆矩阵的表现形式[57]。

TP 代表的是真正 (True Positive, TP)，既样本的真实标签为正，模型所预测的结果也为正的正样本。

FP 代表的为假正 (False Positive, FP)，既样本的真实标签为负，但模型所预测的结果为正的负样本。

FN 代表的为假负 (False Negative, FN)，既样本的真实标签为正，但模型所预测的结果为负的正样本。

TN 代表的为真负 (True Negative, TN)，既样本的真实标签为负，模型所预测的结果也为负的负样本。

可以基于混淆矩阵来判断所训练模型将所测试的样本类别预测为错误的情况。它的每一行所代表的数据之和表示的是该类别所被预测的数目数，每一列所代表

的数据之表示的是该类别所对应的真实数目数，而矩阵在对角线上的数值所表示的是被网络模型所预测正确的样本数。进一步地，可以推理出下面这些常用的计算指标：

真正率 (True Positive Rate, TPR)，也被称为灵敏度 (Sensitivity)，表示的是被网络模型所预测为正的正样本数与数据集中实际的正样本数的比值，即：$TPR = \frac{TP}{TP+FN}$。

假负率 (False Negative Rate, FNR)，表示的是被网络模型所预测为负的正样本数与数据集中实际的正样本数的比值，即：$FNR = \frac{FN}{TP+FN}$。

假正率 (False Positive Rate, FPR)，表示的是被网络模型所预测为正的负样本数与数据集中实际的负样本的比值，即：$FPR = \frac{FP}{FP+TN}$。

真负率 (True Negative Rate, TNR)，也被称为特异度 (Specificity)：表示的是被网络模型所预测为负的负样本数与数据集中实际的负样本的比值，即：$TNR = \frac{FP}{FP+TN}$。进一步的，可以得到准确率与精确率以及召回率的计算公式。

准确率（Accuracy）表示的是被网络模型所预测正确的样本与总样本数的比值。它衡量的是样本分类正确的比例，具体公式为：

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2-1}$$

准确率计算方式比较简单而且经常在各类任务中充当对总体效果进行评价的指标，但是在数据中存在样本不均衡的情况下，准确率却不能很好地反应网络的效果，比如在某个数据集中，正样本占 95%，负样本占 5%，那么即使模型把全部的负样本都预测错误，网络模型的准确率也能达到 95%，但是这样的准确率是没有意义的。

精准率（Precision）也被称为查准率，表示的是模型预测正确的正样本数与所有预测为正样本数目的比值，具体公式为：

$$Precision = \frac{TP}{TP + FP} \tag{2-2}$$

召回率（Recall）也被称为查全率，表示的是模型预测正确的正样本数与全部实际为正样本数的比值，具体公式为：

$$Recall = \frac{TP}{TP + FN} \tag{2-3}$$

从公式上可以看出精准率和召回率的区别就是分母上的不同，一个分母表示的是所有预测为正样本的数目，另一个则是原始样本中所有的正样本数。对模型而言，精准率和召回率是一对矛盾的度量。若要让精准率比较高，则只需要挑选最有把握的正样本，但这样难免会放弃许多正样本，从而导致召回率较低；若要

召唤率比较高，则只需要挑选更多数量的样本来实现，此时，精准率可能会较低。

2.2.2 F1 分数

F1 分数（F1 Score）是对精确率和召回率的一种调和平均，常被作为一种评价指标来对分类模型中的精确度进行评价。他是对精确率与召回率的一种权衡，为的是让二者同时达到最高，从而能够更全面地评价一个分类器。具体公式为：

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2-4}$$

F1 的取值范围是从 0 到 1，一般来说 F1 值越大，就可以认为该模型的学习性能就越好。更一般的，有 F_β 分数，具体公式为：

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{2-5}$$

F_β 分数能够通过对 β 的设置来表达对精准率与召回率的不同偏好，当 β=1 时，F_β 就是 F1 分数，当 β>1 时，模型会更偏好与召回率，当 β<1 时，模型会更偏好与精准率。

2.3 多实例学习与异构图

2.3.1 多实例学习

为了探索大多数药物中一些小分子的构成与整体大型蛋白质之间的关系，作者首次在论文 [58] 中提出了多实例学习 (multiple-instance learning，MIL) 这个概念。与全监督中针对每个样本都会有利用其相应的标签进行训练的方式的不同，多实例学习在训练过程中是以多示例包 (bag) 为训练单元来进行学习的，其所利用的是其包级别上的标签。在多实例学习中，只能得到多实例包上的标签，但针对包中的每个示例（instance）却是没有标签的。如果一个包被标记为了正包，那么可以推出在这个多实例包中所包含的正示例样本数至少为 1。反之如果一个包被标记为了负包，那么可以推出在这个多实例包中所包含的样本全部都是负示例，也就是不包含正示例样本。多示例学习的目的是通过对具有分类标签的多示例包的学习来训练网络学习特征，然后基于每个示例上学习到的特征建立一个分类器以得到针对每个示例的预测结果，最后能够将分类器作用于一个新的多示例包中不同示例的预测 [59]。而多模态多实例学习则是进一步的基于多实例学习，在包中针对不同模态的实例进行划分，只要任何一个模态中包含了一个正实例，这个包就是正包。把视频级别上的标签当成一个包，视频中每个片段上的信息当成一个个示例，需要最终预测出每个片段的真实标签，本文在弱监督情况下针对音视频

事件定位以及音视频事件解析任务的实现就是把这种情况当成一种多实例问题进行解决。

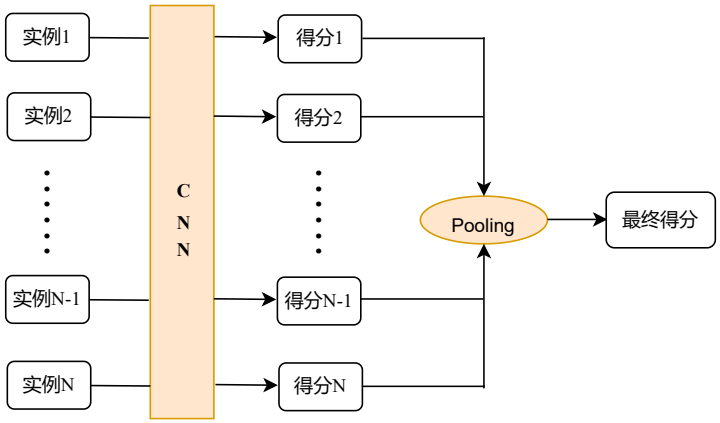

图 2-5 多实例预测流程示意图。

多实例学习常用于图像检索[60]、文本分类[61]等领域。多示例学习中的关键是要找到示例与包之间的逻辑关系，因为示例本身是无标签的而其所属的包是有标签的，所以这种学习方式也可以看成是一种弱监督学习框架下的特殊范式。图2-5给出了一个由多实例本身而预测出包级别标签的示意图。

图2-5中，一个包中的所有实例都会经过一个共享的卷积神经网络来提取特征，然后基于此特征，通过映射层来产生所预测的最终得分。如针对二分类情况, 每个得分值就代表针对当前样本所预测的为正的概率, 为了得到最终的包级别上的预测结果，会通过一个 Pooling 层，以聚合得到最终的预测分数, 这里的 Pooling 可以是最大池化或平均池化，也可以是其他的方法。本文在针对弱监督下的音视频事件定位任务的实现使用的是基于平均池化的 Pooling 方法，而在弱监督下的音视频事件解析任务使用的则是基于注意力下的 pooling 方法。

2.3.2 异构图

在同构图（Homogeneous Graph）中，结点的类型相同以及边的类型也相同，而在异构图中边和结点类型总和要大于 2。如图2-6所示，图左边显示的是同构图，图右边显示的是异构图，其边和结点类型总和为 7，是个典型的异构图。相比于同构图，异构图的使用范围更广，也能够更好地与真实生活相契合，如社交网络等通常所研究的图数据对象都是会包含多种类型的结点以及多种类型的边的。为了从不同类型的结点中进一步的聚合信息，目前的一些方法一方面是利用图中的元路径（meta-path）[62,63]来进行聚合，另一方面有的利用元关系（meta-relation）[64]进行计算聚合。两种类型的算法在聚合过程中都会考虑到结点级别上的注意力

以及语义级别上的注意力，本文在后续算法中所提出的算法模型中便借鉴了这种基于异构图的注意力网络思想。具体计算方式会在算法章节中进行详尽的阐述。

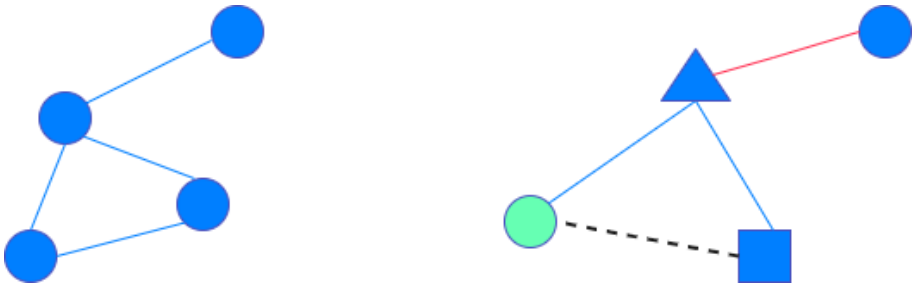

图 2-6 同构图与异构图示例。

2.4 多模态特征融合方法

如果说卷积神经网络的目的是为了获得一个较好的特征表达，那么多模态特征融合所要做的事情就是对两个或多个模态的信息进行整合，从而将不同模态特征融合为一个整体的特征，然后利用这个特征进行预测的过程[65]。在预测的过程中，相比于基于单个模态的预测，多模态特征融合方法能够结合多个模态的信息，实现信息互补，从而可以获取更高的准确率以及使网络模型更加鲁棒。目前主流的融合方式主要有两种：基于简单操作的融合和基于注意力操作的融合方法。

2.4.1 基于简单操作的融合

简单融合操作，顾名思义，就是将来自多个模态的特征向量使用简单融合的方式对各模态数据进行整体上的融合，主要包括早期融合、晚期融合和混合融合三种方式。图2-7中的三个子图，给出了这三种融合方式的示意图。

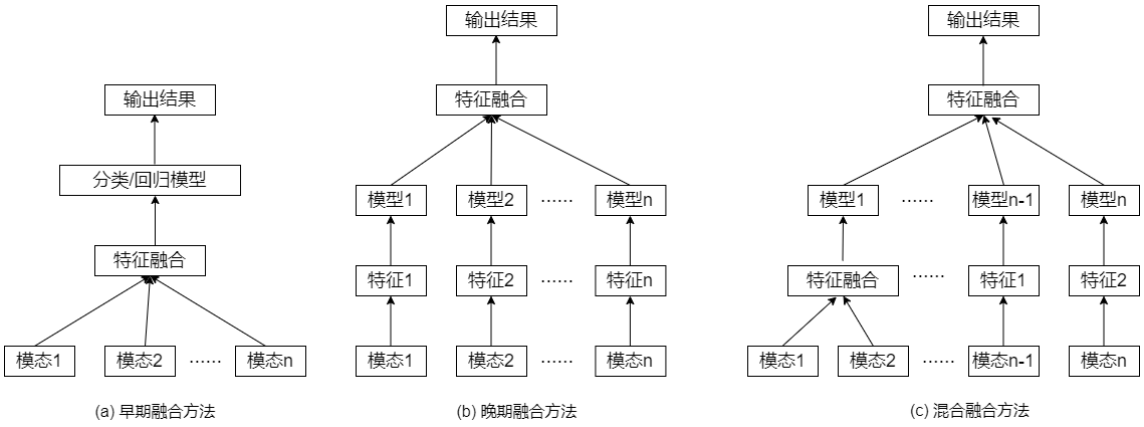

图 2-7 三种融合方式的示意图[66]。

在早期融合中，首先会针对性的对各输入模态进行提取特征，然后将所有提

取到的模态特征进行特征融合操作，从而将所有的特征融合到一个公共的融合特征中去，之后便以此融合特征作为模型的输入进行相应的模型训练，最后输出对应的预测结果。其中常见的早期融合操作可以是比较简单的对各模态的特征进行相同位置上元素间的相乘或相加。该方法的缺点是无法对多模态数据间的互补性进行合理性地利用，以及会存在信息冗余问题，所以可以进一步地通过构建编码器—解码器的方法或者用 LSTM 神经网络对时序信息进行整合以消除部分影响。

晚期融合方法也称决策级上的融合方法，该种方法指的是先针对每个模态用不同的网络模型进行训练，然后对每个模型的输出结果进行融合处理操作，以获得最终的结果输出。其中，常见的晚期融合处理操作可以是简单的取最大值、取平均值方法，可也利用贝叶斯规则进行计算以及借鉴集成学习中的集成方法等进行结合。与早期融合方法相比，晚期融合方法中不同分类器所产生的错误不会互相影响，因而可以在一定程度上处理数据间的异步性，模型的独立性以及鲁棒性也较强。然而该融合方法并未考虑到不同模态在特征层面上的相关性，此外实现难度也更难一些。

混合融合方法则是在增加了网络模型的训练难度和结构复杂度的同时，结合了早期融合方法与晚期融合方法中的优势，以更加灵活的方式选择要进行融合的位置[67]。实际上，论文 [66] 指出以上三种融合方式实际上并没有确定的优劣关系，只是在不同的相关任务以及实验条件下，不同融合方式的效果可能不一样，可以分别进行尝试，以取得最优结果。

2.4.2 基于注意力操作的融合

基于注意力操作的融合方法相比于基于简单操作的融合方法则更加复杂，前者需要更加细致的考虑不同模态间的交互处理。为了进一步地处理不同模态间的信息冗余以及学习模态间的互补性以得到更加丰富的特征信息，目前很多方法 [68–70] 都开始基于注意力来进一步融合不同模态的特征。该类方法的核心思想是针对不同的模态所预测的结果进行加权融合或者对不同模态的特征进行加权融合[69,70]，又或者在中间特征维度上进行权重融合处理[68]。具体方法不再阐述。

在本文所提出的算法中，不仅使用到了基于平均池化的后期融合方法，也对音视两种模态进行了注意力上的融合。

2.5 本章小结

本章主要介绍了针对音视频事件定位任务以及音视频事件解析任务所提出的相关算法在实现过程中所用到的一些相关理论与技术。包括所使用到的网络模型

的结构、评价指标的计算方式以及对相关的知识点如多实例学习，异构图算法以及多模态特征融合常用的方法进行总结。

第三章 基于模态间交叉注意力的音视频事件定位

针对音视频事件定位任务，本文提出了一个基于模态间交叉注意力机制的多模态网络算法（Multimodal Network with Cross-Modal Attention，MCMA）。算法整体流程示意图如图3-1所示。由于音频与视觉两种模态所侧重表达的关键信息体现在不同方面，所以不仅需要考虑到同一模态内部不同片段之间的关系，还需要考虑到不同模态之间众多片段相互的关系。为此，在对视频中的两种模态提取完基本特征后，本文首先通过自注意力机制来处理同一模态内在不同片段间的关系，然后通过音视间的共同注意力模块对视觉特征进行关键区域聚焦，并对音频特征进行增强，之后基于异构图算法当中的思想来建模单模态中的某一片段与其它所有片段间的关系，最后对所获得的最终信息进行融合以完成视频中片段上的音视事件识别。

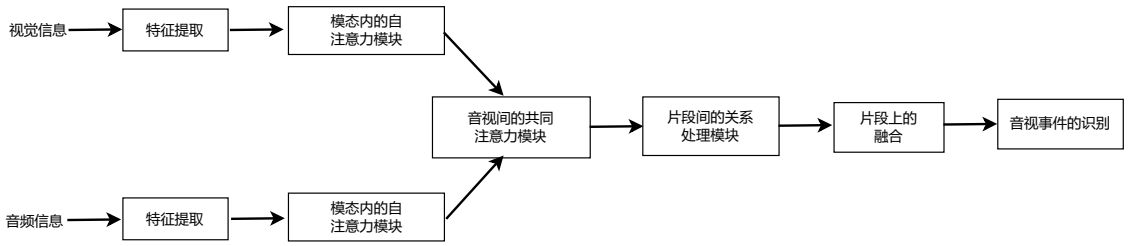

图 3-1 基于模态间交叉注意力机制的多模态网络算法流程示意图。

接下来本文将在 3.1 节中，给出音视频事件定位任务的具体问题描述与符合定义，在 3.2 节中针对音视频事件定位任务，就本文所提出的整体框架进行简要介绍，在 3.3 节中详细介绍算法框架中每个模块的具体实现细节，在 3.4 节中介绍实验中的参数设置并对实验结果进行对比分析，最后在 3.5 节中给出本章小结。

3.1 问题描述与定义

为了进一步结合音频与视觉信息来理解一个视频中所发生的内容，国内学者田亚鹏首先在论文 [13] 中提出了音视频事件定位这个任务。文中指出一个音视频事件是指这个事件在某个片段上既是可以看见相应视觉信息又是可以听见相应的音频信息。即若视频中的某些片段发生了某个音视事件，那么该事件的信息在这些片段中必须是既可以听见的又是可以看见的。如果在某个片段中，对应的只有事件的音频信息或者视觉信息又或者既没有音频信息也没有视觉信息，那么便认为该片段没有发生事件。该任务需要做的是预测出一个视频中的哪些片段是发生了什么类别的音视频事件，而那些未发生音视事件的片段应当被预测为背景。

图3-2给出了一个关于音视频事件的例子，可以看出在视频的第 2 个片段到第 4 个片段间发生了货车这个音视事件，因为在此持续范围内既可以听见货车发动机的声音又可以看见货车本身，所以这三个片段应当被预测为货车音视事件。虽然在第 5 个片段上能够听见货车发动机的声音，但是货车视觉信息并未出现，所以该片段应该被预测为背景。

图 3-2 音视频事件例子（货车音视事件）。

给定一个包含连续片段的长视频 S，可以将其分割成 T 个不重叠的音视频片段对 $S=\{S_t=(v_t,a_t)\}_{t=1}^{T}$。每个片段的时长都是 1 秒，$v_t$ 和 a_t 分别代表第 t 个片段所对应的视觉特征与音频特征。T 表示长视频 S 中所包含的片段总数。那么第 t 个片段上的标签为

$$y_t \in R^C=\left\{y_t^k | y_t^k \in \{0,1\}, k=1,...,C, \sum_{k=1}^{C} y_t^k = 1\right\}$$

其中 C 代表数据集中的类别总数（包括背景）。

针对音视频事件定位会分为全监督下的音视频事件定位与弱监督下的音视频事件定位两个子任务分别进行研究。在一个给定的时序视频片段中，这两个任务的目的都是为了预测视频中的哪些片段发生了什么类别的音视事件，同时需要将那些没有音视事件发生的片段需要预测为背景[13,34,36,38]。两者的区别是，在全监督的情况下，可以获得一个视频中所有视频片段上的标签，因此在训练时，模型可以通过细致的片段级标签进行有监督训练。而在弱监督的情况下，仅有视频级标签可用，缺失片段级类别标签，但这两个任务无论是全监督条件下还是弱监督条件下，在测试时都要求实现片段级别的事件预测。

具体而言：在全监督下，训练的时候有标签 $Y^{fully}=[y_1;y_2;...;y_T]\in R^{T\times C}$。通过 Y^{fully} 标签，网络可以知道一个视频中的哪些片段是发生了音视事件，哪些没有。而在弱监督下，训练的时候只有标签 $Y^{weakly}=\frac{1}{T}\sum_{k=1}^{T} Y_{kC}^{fully}$，其中 $Y^{weakly}\in R^{1\times C}$，它是对应的全监督标签沿着列求和的平均值，因此该标签只能表示一个视频中某个片段上所会发生事件的概率。由此也可以推测出弱监督情况下的实现会更难。本章主要探究的是在全监督以及弱监督情况下的音视频事件的定位。

3.2 整体框架结构

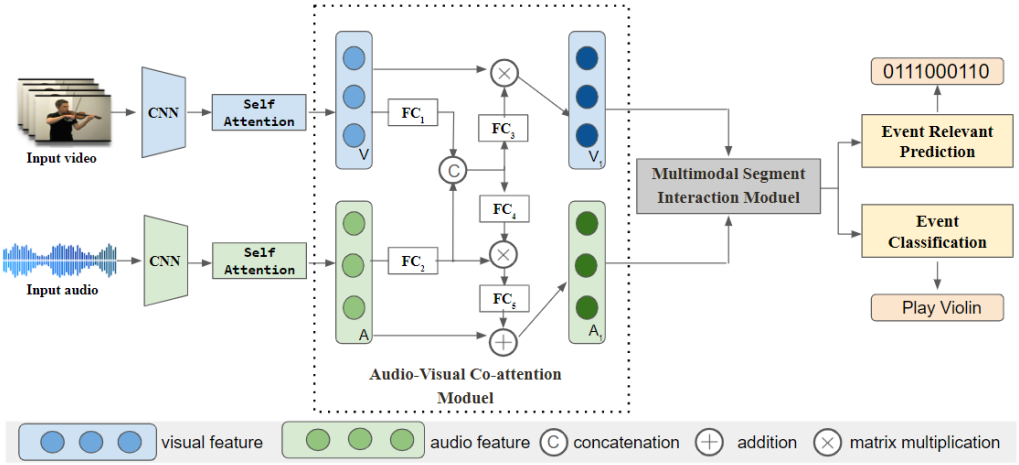

图 3-3 基于模态间交叉注意力机制的多模态网络算法模型。

针对音视频事件定位任务，本文所提出的整体框架如图3-3所示。该框架主要分为五个部分，分别是：针对视觉信息与音频信息的特征提取部分、同一模态内不同片段间关系的处理部分、音视两种模态间的共同注意力处理部分、多模态在不同片段间的交互部分以及最终的事件识别部分。首先针对输入的视觉信息与音频信息分别通过一个已经预训练好的卷积神经网络来提取对应的音视特征，然后基于注意力机制中的自注意力算法来建模同一模态内部不同片段之间的关系，之后通过音视间的共同注意力模块，以音视融合特征作为指导来聚焦视觉中与对应音频模态相关的事件区域，同时通过残差连接的方式增强对应的音频特征。处理后的特征会被送入多模态片段间的交互模块，以处理当前片段与其他同模态片段以及不同模态片段间的关系，从而学习其它片段所带来的信息。最后针对交互模块的特征输出进行融合，来完成最终的音视事件的识别。本文会在后续小节进行详尽的阐述。

3.3 算法模型

本章节将对所提出的基于模态间交叉注意力机制的多模态网络算法进行详细的阐述，其中主要包括初步的特征提取与编码部分、音视间的共同注意力模块、多模态片段间关系的建模模块以及最终的音视频事件的识别部分。

3.3.1 音视模态的特征提取与再编码

针对某个给定的视频 $S = \{S_t = (v_t, a_t)\}_{t=1}^{T}$，在对视频上的音频与视觉数据进行预处理后，本文首先利用已经预训练好的卷积神经网络来分别提取视频 T 个片

段上的视觉特征与音频特征。为了便于表示，本文把在第 t 个片段上所提取到的视觉特征与音频特征分别表示为 $v_t \in R^{H\times W\times d_v}$ 和 $a_t \in R^{1\times d_a}$，其中 $t \in \{1,2,...,T\}$，H 和 W 分别代表所提取的视觉特征在空间维度上的长和宽，d_v 代表的是视觉特征的通道维度，d_a 代表的是音频特征的通道维度。那么整个视频的视觉特征可以表示为 $F_V \in R^{T\times H\times W\times d_v}$，音频特征可以表示为 $F_A \in R^{T\times d_a}$。

考虑到在一个视频中所发生的事件通常是持续的，即使一个视频被分段了，这些片段之间依然有联系，而自注意力（Self-Attention，SA）机制则可以对视频片段中同一模态不同片段之间的相关性进行建模。为此本文首先使用自注意力机制来对视觉特征 F_V 与音频特征 F_A 进行片段上的时序处理，通过自注意力机制处理后的视觉特征与音频特征分别被表示为 V 和 A。为了避免重复描述，本文以对视频中音频特征 F_A 的处理为例进行阐述。具体计算公式为：

$$Q,K,V = Fun(F_A) = (W_qF_A, W_kF_A, W_vF_A) \tag{3-1}$$

$$A = SA(F_A) = Softmax(\frac{Q^TK}{\sqrt{d}})V \tag{3-2}$$

其中 Q，K，V 分别代表注意力机制中基于输入所产生的查询、键以及值，他们是通过不同的权重矩阵 W_q，W_k 和 W_v 与输入 F_A 进行相乘得到的，这些权重矩阵是通过映射函数 $Fun(\cdot)$ 所产生的。d 指的是查询值 Q 中的特征维度数，它与 K 和 V 的维度是一致的。注意力函数输出的结果是加权计算后的 A。它是通过采用点乘的方式计算矩阵 Q 和 K 之间的相关性，然后进行 softmax 归一化再与值 V 相乘得到的。同理可以对视觉输入 F_V 进行同样的处理。为此针对视觉输入 $F_V \in R^{T\times H\times W\times d_v}$ 和音频输入 $F_A \in R^{T\times d_a}$，可以得到通过模态内自注意力处理后的视觉特征 $V \in R^{T\times H\times W\times d_v}$ 和音频特征 $A \in R^{T\times d_a}$。

3.3.2 音视间的共同注意力模块

一般情况下，视觉模态中每张图片上所包含的信息比较多，会含有目标信息和背景信息，其中背景信息与当前音视事件无关，需要尽量消除这些无关区域对预测当前片段所发生音视事件的影响。为了使所提出的模型能够很好的聚焦于当前片段视觉区域中与对应音频信息紧密相关的局部区域，现有的一些方法 [13,34,38,41] 都是基于音频单方面指导视觉的方式，忽略了音视模态与视觉模态之间的相互对应关系。为了在筛选视觉区域的同时对音频特征进行增强，本文提出了音视间的共同注意力模块（Audio-Visual Co-attention Module，AVCM)。该模块以音视融合特征作为指导，一方面聚焦整个视觉区域中与对应音频模态相关的关键事件区域，另一方面增强与其对应的音频特征。

具体而言：基于上一个模块所获得的视觉特征 $V \in R^{T\times H\times W\times d_v}$ 和音频特征 $A \in R^{T\times d_a}$，本文首先通过两个单独的映射层对视觉特征 V 与音频特征 A 进行映射，使其特征分布到同一空间，然后通过级联二者获得音视融合特征 J，并以 J 为指导信息，在筛选视觉区域的同时对音频特征进行增强。之所以使用音视融合特征 J 做为指导，是因为相比于只使用音频信息作为指导，这个融合特征会带来更多的参考信息。即使当前的音频信息是无关的噪声，也不至于在指导聚焦的过程中很大程度上影响到原本的视觉信息，从而带来更少的误差。之后对指导信息 J 分别进行针对视觉特征和音频特征的映射以产生注意力权重矩阵。针对视觉特征的映射会产生一个空间注意力图 $att^v \in R^{T\times H\times W}$，以在空间上筛选掉与事件无关的区域，同时关注可能发生音视事件的区域。同理，为了获得更具有判别性的音频特征，也会针对音频特征的映射会产生一个通道注意力图 $att^a \in R^{T\times H\times W}$，同时为了防止信息的丢失，本文以残差的方式与原来的特征相加。本文把通过 AVCM 模块关注后的视觉特征与音频特征分别表示为 $V^{att} \in R^{T\times d_v}$ 和 $A^{att} \in R^{T\times d_a}$。具体计算公式为：

$$J = Concate(\sigma(fc_1(V)), \sigma(fc_2(A))) \tag{3-3}$$

$$att^v = Softmax(\sigma(fc_3(J))),\ \ att^a = Softmax(\sigma(fc_4(J))) \tag{3-4}$$

$$V^{att} = att^v \otimes V,\ \ A^{att} = fc_5(att^a \otimes \sigma(fc_2(A))) \oplus A \tag{3-5}$$

其中 $fc_i(\cdot)$ 表示的是图3-3中第 i 个由全连接层组成特征映射，$\sigma(\cdot)$ 代表的是 relu 激活函数，函数 $Concate(\cdot)$ 表示的是对输入的特征按照通道维度进行级联操作，函数 $Softmax(\cdot)$ 代表的是对输入进行 softmax 归一化操作，$\oplus$ 代表的是元素间的加和操作，$\otimes$ 代表的是元素间的乘积操作。

3.3.3 多模态片段间关系的建模模块

虽然一个视频被分割成了 T 个不重叠的片段，但是每个片段上的视觉信息或音频信息是会直接共享在整个长视频中所发生的事件信息，即两种模态的其他片段是能够为当前片段提供语义指导的。为了更加准确地理解当前片段无论是在音频模态还是在视觉模态中所发生的事件，为此本文基于异构图算法 [62–64] 中处理不同类型的邻接结点的思想，提出了一个多模态间的片段交互模块（Multimodal Segment Interaction Module，MSIM）来处理当前片段与其他同模态片段以及不同模态片段间的交互，从而学习其它片段所带来的自身所没有的信息。该方法能够忽略某一片段上两种模态在表达信息时的差异，并能够在探索跨模态关系的同时也不忽略模态间的关系。相比于其他方法，本文所提的方法能够只聚合与当前片

段相关性比较大的片段的信息，而不是聚合一些无关的信息。

如何建立当前片段与其他所有片段之间的联系是一个需要考虑的问题。为此本文先假设当前片段与其他所有片段之间都是有联系的，然后通过计算当前片段特征与其他所有片段特征之间的相似性值，通过比较片段间的相似性值与所设置阈值 τ 的大小来更新当前片段的联系。两个片段间的相似性值越高就认为他们间的联系越紧密，只有当第 t 个片段与当前片段之间的相似性值大于阈值 τ 时，本文才认为他们两者是有联系，那么当前片段可以向第 t 个片段进行信息学习。以音频模态中的第 5 个片段 a_4 为例，图3-4展示了 a_4 与其他所有片段之间建立关系的示意图。

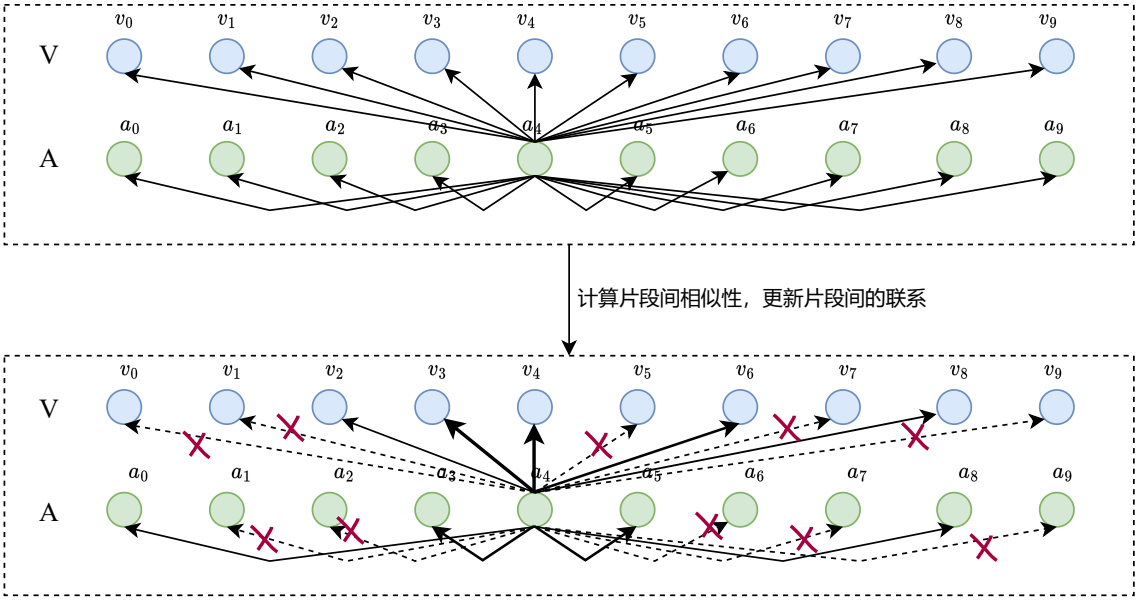

图 3-4 某视频中 a_4 片段与其他所有片段之间建立关系的示意图。图中上部分展示的是 a_4 与其他所有片段之间最初的联系，下部分展示的是通过计算 a_4 与其他所有片段特征之间的相似性值并与阈值 τ 比较后所更新的新联系。

图3-4中的每一个蓝色圆圈与绿色圆圈分别代表的是一个视频中每个片段上所对应的视觉特征与音频特征。一个视频中，每种模态上一共包括 10 个片段，每一个圆圈上都有对应的符号表示，a_i 表示的是第 i+1 片段上的音频特征，a_i 表示的是第 i+1 片段上的视觉特征。若两个片段之间是有联系的，就用连线箭头进行表示，如 $a_4 \leftrightarrow v_3$。虚箭头表示的意思是两个片段间的联系应当被切断，即在更新联系过程中两者之间的相似性值小于阈值 τ，如图3-4中下部分的 $a_4 \leftrightarrow v_0$ 连线。此外，若两个片段间的连线越粗，就表明他们两者之间的关系越紧密，从图中可以看出 a_4 与 v_3 之间的联系要强于 a_4 与 a_3 的联系，也即 $a_4 \leftrightarrow v_3$ 这两个片段之间的相似性值要更大。以音视片段间的相似性计算为例，他们之间的相似性可按照如下公式

进行计算：

$$\gamma^{va} = \frac{(V^{att}W_1^V)(A^{att}W_1^A)^T}{\sqrt{d}},\ \ \gamma^{av} = (\gamma^{va})^T \tag{3-6}$$

$$adj^{va} = \gamma^{va}\mathbb{I}(\gamma^{va} - \tau),\ \ adj^{av} = \gamma^{av}\mathbb{I}(\gamma^{av} - \tau) \tag{3-7}$$

其中 $W_1^V \in R^{d_v \times d}$ 和 $W_1^A \in R^{d_a \times d}$ 是在利用全连接层进行特征映射时所产生的可学习参数，d 是映射后视觉特征或者音频特征的维度数。$\gamma^{va} \in R^{T\times T}$ 和 $\gamma^{av} \in R^{T\times T}$ 分别代表视觉与音频片段间以及音频与视觉片段间的相似性矩阵，两者互为转置矩阵。$adj^{va} \in R^{T\times T}$ 与 $adj^{av} \in R^{T\times T}$ 是通过相似性矩阵与阈值进行运算所产生的分别代表视觉与音频间以及音频与视觉间的新联系矩阵，两者也互为转置矩阵。$\mathbb{I}(\cdot)$ 是一个指示函数，当输入大于等于 0 的时候会输出 1，当输入小于 0 的时候会输出 0。同理也可以获得音频与音频片段间的相似性矩阵 γ^{aa} 和更新后的联系矩阵 adj^{aa}，以及视觉与视觉片段间的相似性矩阵 γ^{vv} 和更新后的联系矩阵 adj^{vv}。

在给定了上述所得到的两种模态片段之间的联系矩阵后，本文把该视频中片段间的关系当成一个图来处理。那么片段间的联系矩阵就相当于图中的邻接矩阵，每个片段的特征（包括视觉和音频两种特征）可以当成图中的结点。为了聚合与当前结点相连的其他结点所带来的信息，本文以一种图学习的方式进行聚合。但由于音频特征与视觉特征是两种不同的类型的结点，把他们当成同一种结点进行特征聚合是不合适的。为此，参考异构图算法 [62–64] 中处理不同类型的邻接结点的思想，先分别聚合同种类型结点所带来的特征信息，之后在进一步聚合不同类型结点所带来的特征信息，同时为了防止自身信息的损失，本文以残差的方式进行级联。具体公式为：

$$V^{agg} = \beta_v \cdot adj^{vv}(V^{att}W_2^V) + (1-\beta_v)\cdot adj^{va}(A^{att}W_1^A) + V^{att} \tag{3-8}$$

$$A^{agg} = \beta_a \cdot adj^{aa}(V^{att}W_2^A) + (1-\beta_a)\cdot adj^{av}(V^{att}W_1^V) + A^{att} \tag{3-9}$$

其中 $V^{agg} \in R^{T\times d_v}$ 代表的是聚合后的视觉特征，$A^{agg} \in R^{T\times d_a}$ 代表的是聚合后的音频特征。$W_2^V \in R^{d_v\times d_v}$，$W_1^A \in R^{d_a\times d_v}$，$W_2^A \in R^{d_a\times d_a}$ 和 $W_1^V \in R^{d_v\times d_a}$ 分别是在利用全连接层进行特征映射时所产生的可学习的权重参数，$\beta_v \in R^{T\times d_v}$ 代表针对视觉模态中同类型结点聚合后所产生的权重，$\beta_a \in R^{T\times d_a}$ 是对音频模态中同类型结点聚合后所产生的权重。之所以有 β_v 与 β_a 这两个权重参数，是因为考虑到不同模态在表达不同信息时的重要性是不一样的。为此本文通过这种权重加权的方式来自适应的对不同模态提供关注。他们的计算方式如下：

$$C_v = Concate(adj^{vv}(V^{att}W_2^V), adj^{va}(A^{att}W_1^A)) \tag{3-10}$$

$$C_A = Concate(adj^{aa}(A^{att}W_2^A), adj^{av}(V^{att}W_1^V)) \tag{3-11}$$

$$\beta_v = Sigmoid(Linear(C_v)),\ \ \beta_a = Sigmoid(Linear(C_A)) \tag{3-12}$$

其中函数 $Concate(\cdot)$ 表示的是对输入的特征按照通道维度进行级联操作，函数 $Sigmoid(\cdot)$ 表示的是 sigmoid 激活函数，函数 $Linear(\cdot)$ 表示的是对输入进行一个线性映射层处理操作，C_v 与 C_a 分别代表的是针对视觉模态和音频模态级联后的多模态特征，之后分别针对 C_v 与 C_a 进行映射操作和 sigmoid 激活函数处理，来得到最终的针对不同模态的权重参数 β_v 与 β_a。

针对所有片段，在聚合完不同类型的结点的信息后，本文通过级联 V^{agg} 与 A^{agg} 以获得最终的视频级别的特征 F_{VA}。$F_{VA} \in R^{2\times d_l}$ 用于最终的音视事件的识别。公式如下：

$$F_{VA} = Concate(V^{agg}W_V^3, A^{agg}W_A^3) \tag{3-13}$$

其中 $W_3^V \in R^{d_v\times d_l}$, $W_3^A \in R^{d_a\times d_l}$ 分别代表由于映射而产生的权重参数。

3.3.4 音视频事件定位的实现

由于在全监督的情况下可以获得的训练标签与在弱监督情况下的标签是不一样的，所以这两种情况下音视频事件的识别处理也是不同的。在本小节中，将分别进行阐述。

3.3.4.1 全监督下的设置

在全监督的情况下，本文将针对某片段上所发生的音视事件的预测分解成综合两部分的得分。其中一部分得分为预测当前片段是否会发生音视事件的相关性得分 $p_t^r \in R^1$。p_t^r 这个得分指的是第 t 个片段发生音视事件的概率，其数值范围在 0 到 1 之间，数值越大则说明越有可能发生事件。为了预测某个视频所有片段上的事件相关性得分 $p^r = [p_1^r, p_2^r, ..., p_T^r] \in R^{1\times T}$，本文使用一个与事件相关的分类器对最终的输入 F_{VA} 进行预测，之后再通过一个 sigmod 激活函数，以得到最终结果。另一部分得分为预测当前片段所发生事件的类别得分 $p_t^c \in R^{1\times C}$。p_t^c 这个得分是对第 t 个片段所可能发生的音视事件的类别的一个预测，这个向量中的某个位置的值越大，就说明当前片段就越可能发生与该位置所对应的事件类别。为了预测某个视频在所有片段上的事件类别得分 $p^c = [p_1^c, p_2^c, ..., p_T^c] \in R^{T\times C}$，本文使用一个一层的线性映射层作为事件类别分类器来进行预测。事件类别分类器的输入是全局的视频特征 $O_{av} \in R^{1\times d_l}$。$O_{av}$ 是通过对视频特征 F_{VA} 进行全局最大池化得到的。具

体计算公式如下：

$$p^r = \sigma(Linear(F_{VA})) \tag{3-14}$$

$$O^{av} = Maxpooling(F_{VA}) \tag{3-15}$$

$$p^c = Softmax(Linear(O_{av})) \tag{3-16}$$

其中函数 $Linear(\cdot)$ 表示的是对输入进行一个线性映射层处理操作，函数 $Maxpooling(\cdot)$ 表示的是对输入进行最大池化操作，函数 $Softmax(\cdot)$ 代表的是对输入进行 softmax 归一化操作。

在训练阶段，是能够得到所有片段的音视事件类别标签的，因此可以由某片段是否发生了音视事件，来计算得到事件的相关性标签 Y^{re}。在计算损失时，针对事件相关性得分的预测，使用二元交叉熵损失 L_r 来进行计算。针对所发生的事件的类别的预测，使用交叉熵损失 L_c 来进行计算。具体公式为：

$$L_r = -\frac{1}{T}\sum_{t=1}^{T} Y^{re} log(p^r) \tag{3-17}$$

$$L_c = -\frac{1}{TC}\sum_{t=1}^{T}\sum_{c=1}^{C} Y^{fully} log(p^c) \tag{3-18}$$

综合上述损失，全监督下的总损失如式3-19所示：

$$L_{full} = L_c + L_r \tag{3-19}$$

在推理阶段，某个视频所发生的音视事件的预测结果将由 p^r 和 p^c 决定。针对第 t 个片段，如果 $p_t^r \geqslant 0.5$，那么就认为第 t 个片段发生了音视事件，同时它的类别为 p^c。如果 $p_t^r < 0.5$，就预测该片段为背景。

3.3.4.2 弱监督下的设置

在弱监督的情况下，和现有的一些方法 [13,34,38,41] 一样，本文把弱监督下的音视频事件定位当成一个多实例学习（multiple-instance learning，MIL）问题来进行处理。同样的，仍然会针对融合后的视频特征，通过分类器预测出对应的事件相关性得分 p^r 和事件的类别得分 p^c。但由于只能获得视频级别上的标签，所以本文把 p^c 复制 T 份，把 p^r 复制 C 份，然后通过元素间点积的方式将两者进行融合，以产生一个混合得分 $p^m \in R^{T\times C}$。之后，本文借鉴论文 [71] 中的方法，通过 MIL 最大池化的方法来聚合混合得分 p^m 在片段级别上的结果以形成视频级别上的预测结果，从而可以在训练过程中利用视频级别上的标签通过交叉熵损失进行训练。

在推理阶段，和上述全监督过程一样，在此不再叙述。

3.4 实验设置与结果分析

在本小节中，本文将主要介绍一些与音视事件定位任务相关的实验设置，如数据集的情况，评价指标的确定，一些实验参数细节，以及对相关的实验结果进行分析等。

3.4.1 数据集介绍

针对音视频事件定位任务，本章使用 Audio-Visual Event (AVE) 数据集[13] 进行算法的效果验证。AVE[13] 数据集是 Audioset[72] 的一个子集，一共包含了 4，143 个视频，其中每个视频中都包含了一个音视事件，且每个视频的时常为 10 秒，每个音视事件持续的时长从 2 秒到 10 秒不等，所持续的片段范围都被人工标记。此外，AVE 数据集包含了不同活动领域中共计 28 种类别的音视事件，如人类说话（human speeches）、煎食物（frying food）、狗吠（dog barking）、弹小提琴（playing violin）等音视事件，其中每个音视事件所包含的视频数量从 60 到 180 不等。图3-5中展示了一些音视事件类别的例子，并提供了部分该数据集的数据分析。在进行实验时，和大多数论文 [13, 34, 35, 37, 38, 73] 一样，本文随机将数据集分为三个部分，其中 80% 用于训练，10% 用于验证。10% 用于测试。

图 3-5 AVE 数据集示例图[13]。

3.4.2 评价指标

和大多数方法 [13, 34, 35, 37, 38, 73] 一样，针对视频中所有片段上所发生的音视事件类别的预测准确率是评价模型的主要指标。在全监督和弱监督下的两种设置下，本文都将全局上的 top-1 准确率作为评价指标。

3.4.3 实验细节

在本小节中，将分为对实验数据的预处理和实验中的参数设置两部分进行具体介绍。

3.4.3.1 实验数据的预处理

在对音频模态的处理上，针对给定的数据集，首先使用 ffmepg 音视处理技术提取视频中所对应的独立音频，所获取到的格式为.wav 格式。在处理过程中所使用的采样率为 11025Hz，那么针对一个 10 秒长的音频可以得到长度为 110250 的数据，之后进行短时傅里叶变换得到尺寸为 256×256 的频谱图。进一步使用在 Audio-set[72] 上预训练好的网络 VGG-like[74] 来提取对应的音频特征，每个片段上所获取到的音频特征维度为 1×128，最终一个视频上的音频特征维度为 10×128。

在对视觉模态的处理上，针对给定的数据集，由于视频中的图片在帧间是存在大量冗余的，并没有必要处理每秒中的所有图片帧。为此，本文同样使用 ffmepg 音视处理技术，对视频按照 fps=16 的采样率进行帧提取，一个 10 秒长的视频可以得到 10×16 共计 160 帧的图片。之后针对 1 秒中的 16 张图片，用在 ImageNet[75] 上预训练好的 VGG-19[76] 网络进行特征提取，再经过平均池化处理操作获得特征维度为 7×7×512 的视觉特征。最终一个视频上的视觉特征维度为 10×7×7×512。

3.4.3.2 实验中的具体参数

在训练过程中，本文选用 Adam 优化器来对训练中的参数进行优化更新。网络中参数的学习率为 0.0001，训练的批量大小为 64，共计训练 100 论。阈值 τ 的大小为 0.095，一个视频中片段的总长度 T 的值为 10，类别种类数 C 的值为 29（包含了背景），d_v 的数值大小为 512，d_a 的数值大小为 128，d_l 的数值大小为 128。在进行特征映射时所采用的都是一层的全连接层。

3.4.4 与其他最新方法的对比

在本小节中，本文将对所提出的基于模态间交叉注意力机制的多模态网络算法（MCMA）与音视频事件定位领域中的其他研究方法进行实验对比分析。为了公平地进行比较，本文采用和其他算法一样的经过预处理后的音频特征与视觉特征，分别在全监督设置下以及弱监督设置下进行实验对比。

全监督设置下的实验对比结果如表3-1所示，前两行的 Audio only 和 Visaul only 分别指的是单独使用音频模态或者单独使用视觉模态信息所获得的实验结果。从这两行结果中，可以看出针对于在 AVE 数据集上进行音视事件定位任务，单独使用音频信息相比于单独使用视觉信息能够获得更好的效果。这说明音频中包含了更多的关键信息，同时也能间接证明对音频模态进行关注以及对相应特征进行增强的必要性。所以在音视事件定位任务的设定下，更需要结合音频来理解视频中正在发生的内容，而不是向大多数如动作识别中的算法[77–80]，倾向于用视觉信

表 3-1 在 AVE 数据集上与其他方法在全监督下的对比。

Approaches	Feature	Accuracy(%)
Audio only [75]	VGG-like	59.5
Visual only [53]	VGG-19	55.3
AVEL [13](baseline)	VGG-like，VGG-19	71.4
AVSDN [34]	VGG-like，VGG-19	72.6
CMAN [41]	VGG-like，VGG-19	73.3
DAM [38]	VGG-like，VGG-19	74.5
AVRB [37]	VGG-like，VGG-19	74.8
AVIN [73]	VGG-like，VGG-19	75.2
AVT [35]	VGG-like，VGG-19	76.8
CMRAN [36]	VGG-like，VGG-19	77.4
CBS [81]	VGG-like，VGG-19	**79.3**
MCMA(Ours)	VGG-like，VGG-19	77.2

息来识别动作。

AVEL [13] 算法作为本文的 baseline，与之相比，本文的算法结果能够显著的超过其 6 个百分点，这也进一步论证了本算法的优越性。本文分析认为，一方面是因为 AVEL [13] 算法只注重对视觉模态的聚焦以及模态内的不同片段在时序上的联系，并未考虑到不同模态片段之间的联系，另一方面是因为本算法能够在关注视觉模态中关键信息的同时增强对应的音频模态信息，以及能够处理不同模态在片段内与片段间上的关系。这也能够佐证本文算法中的不同模块确实起到了作用。

相比于算法 CMRAN [36]，本文的实验结果与之相差 0.2%，本文认为这是由于 CMRAN 算法中在处理片段间的关系模块上采用了多个相对复杂的模块，而且该算法更倾向于全监督下的设置，在弱监督情况下，实验结果相对于本文的算法就差了一些。此外，与最新发表的算法 CBS [81] 相比，本文的结果还是有一定的差距。本文分析认为，这是因为 CBS 算法在处理背景信息以及在对网络的约束上，所设置的方法比较巧妙。该方法从时间级别和片段级别两个方面入手，结合门控机制以更加有效的方式剔除无关信息影响并关注关键事件信息。但与一些其他的方法 [34, 35, 37, 38, 41, 73] 相比，本文的实验结果相对较好。本文认为这是因为这些方法有的只是注意对模态内关系的建模，有些只是对不同片段间在时序上的联系进行处理，然而本文所提出的算法能够很好的联合处理不同模态在片段内与片段间上的关系，从而能够进一步的理解视频中所发生的事件。

弱监督设置下的实验对比结果如表3-2所示，同样的，前两行的实验结果指的是单独使用音频模态特征或者单独使用视觉模态特征所得到的实验结果。可以看出相比于全监督情况下的结果，由于训练时候的标签限制，弱监督情况下的准

表 3-2 在 AVE 数据集上与其他方法在弱监督下的对比。

Approaches	Feature	Accuracy(%)
Audio only [13]	VGG-like	53.4
Visual only [13]	VGG-19	52.9
AVEL [13](baseline)	VGG-like，VGG-19	66.7
AVSDN [34]	VGG-like，VGG-19	67.3
CMAN [41]	VGG-like，VGG-19	70.4
AVRB [37]	VGG-like，VGG-19	68.9
AVIN [73]	VGG-like，VGG-19	69.4
AVT [35]	VGG-like，VGG-19	70.2
CMRAN [36]	VGG-like，VGG-19	72.9
CBS [81]	VGG-like，VGG-19	**74.2**
MCMA(Ours)	VGG-like，VGG-19	74.1

确率要差一些。这也印证了弱监督情况下音视频事件定位任务更具有挑战性。此外，与其它方法相比，从表中也可以看出本文所提出的算法的效果还是比较好的，甚至与最新发表的算法 CBS [81] 相比也只是差了 0.1%。而且，相比于 baseline 算法，本模型预测结果要高 7.4 个百分点。虽然在全监督情况下，本章的算法相比于 CMRAN 算法 [36] 差了 0.2%，但是在弱监督情况下，要比之要高 1.2%，这也间接印证了本文所提出的算法更加鲁棒，更能兼顾两种情况下的设定。本文分析主要是因为本算法在对不同模态不同片段间关系的建模上更加合理，相比于 CMRAN 算法 [36] 能够更加有效地聚合到有用的信息，而不是不加区分的接受。

3.4.5 消融实验

在本节中，本文将通过四个消融实验来从不同的角度对本文模型中的不同设计进行有效性验证。

3.4.5.1 算法中不同模块的作用

本文所提出的基于模态间交叉注意力机制的多模态网络算法主要包括三个模块：特征再编码模块（FAC）、音视间的共同注意力模块（AVCM)、多模态间的片段交互模块（MSIM）。为了对每个模块进行验证，本文提出了四个不同的变体方法：“w/o FAC”，“w/ LSTM”，“w/o AVCM”和“w/o MSIM”，其中“w/o”表示的是不使用（Without），“w/”表示的是使用（With）。“w/o FAC”代表的是只是去除整个算法模型中的特征再编码模块，即不使用 Self-attention 处理模态内的联系，而保持其他模块部分不变。同理，“w/o AVCM”代表的是只是去除整个算法模型中的音视间的共同注意力模块，然后用平均池化来进行代替。“w/o MSIM”代表

表 3-3 在 AVE 数据集上针对模型中不同模块的消融实验。其中“w/o”表示的是不使用（Without），“w/”表示的是使用（With）。

Approaches	Supervised	Weakly-supervised
w/o FAC	74.3	71.8
w/ LSTM	75.8	72.9
w/o AVCM	74.7	71.5
w/o MSIM	70.8	68.8
MCMA(Ours)	**77.2**	**74.1**

的是只是去除整个算法模型中的模态间的片段交互模块，然后用一个线性映射层来代替。而“w/ LSTM”表示用长短期记忆网络 (long short-term memory，LSTM) 来代替本文中所使用到的特征再编码模块。

不同变体算法在全监督与弱监督设置下，所产生的实验结果如表3-3所示。可以很明显的看出所有变体算法的效果相比于原模型都有很明显的下降，这也证明了本算法中每个模块的有效性。对比结果差异性，可以看出在不使用模态间的片段交互模块时，模型效果下降的最明显，在全监督情况下下降了 6.4 个百分点，在弱监督情况下下降了 4.3 个百分点。由此，本文认为基于异构图算法思想所提出的建模不同模态在不同片段上的关系模块是有很大贡献作用的，并认为该方法能够从所建立联系的不同片段上学习到有效信息。虽然在“w/o FAC”与“w/o AVCM”两种情况下，模型的效果下降效果相比“w/o MSIM”较小，但这也不能否认这两个模块对整体网络架构的贡献程度。

对比“w/o FAC”、“w/ LSTM”与 MCMA 三者，本文发现在通过预训练好的网络提取完特征后，进一步对特征进行处理，是对整个网络有益的。相比于用 LSTM 来建模模态内的片段在时间上的关系，通过 Self-attention 来建模单模态内的不同片段在特征上的关系会更加有效果。本文认为这是因为在识别音视事件这个任务上，网络最终是要落实到片段上的特征，而片段间的特征联系相比于片段上的时序处理会更加有益，此外使用 Self-attention 来多模态内关系进行建模也是能够在一定程度上学习到不同片段间的时序关系。所以与使用 LSTM 相比，使用 Self-attention 机制更加合理。

3.4.5.2 音视间的共同注意力模块的有效性

本文所提出的 AVCM 模块，利用的是音视间的共同注意力机制，在筛选视觉区域的同时对音频特征进行增强。为了单独验证这一模块的有效性，本文对比了两种常用的基于音频来单方面的指导视觉的方法以及其他论文中的共同注意力机制方法。对比结果如表3-4所示。

表 3-4 在 AVE 数据集上针对模型中音视间的共同注意力模块的消融实验。

Approaches	Supervised	Weakly-supervised
AGVA [13]	74.1	72.3
AGSCA [36]	76.4	73.3
co-att [82]	75.9	73.1
Ours	**77.2**	**74.1**

AVGA 是在论文 [13] 中通过对音视特征进行映射，然后计算音视特征间的相似性并针对视觉信息单方面产生注意力权重的方法，而 AGSCA [36] 虽然也是基于音频信息单方面地针对视觉信息产生注意力权重的方法，但是该方法针对视觉信息，不仅处理了音视两种模态在空间上的注意力关系还处理了音视两种模态在通道上的注意力。Co-att [82] 则是常用的互注意力的计算方式，通过权重矩阵同时对音视两种模态都进行处理。

从结果中可以看出，相比于这三种方法，本文的方法在同样的设定下取得了更好的结果，特别地，相对于 AVGA [13]，本文的算法在全监督下超其 3.1%，在弱监督下超其 1.8%。这也说明了对音视两种模态同时进行处理的必要性。对比 AVGA [13] 与 AGSCA [36] 两种方法可以说明在对视觉信息进行指导时，同时从空间与通道上进行计算处理，相比于单方面的在空间进行处理是更加有效的。因为视觉特征在不同的通道上对视觉信息所侧重的关注度也是不一样的。对比 Co-att [82] 与本文算法，本文的算法在全监督下超其 1.3%，在弱监督下超其 1%。虽然同样都考虑到了对音视两种模态同时进行处理，但是相比于前者单纯的计算音视两者间的相似性，然后进一步进行处理，本文利用的是音视融合特征进行的指导，这样即使音视两者间没有什么关系，也不至于学习到负面信息。此外，本算法在处理程度上也更加细致，除了针对性的维度映射还对音频进行了残差连接增强。

3.4.5.3 算法中不同阈值的设定对实验结果的影响

本文在建立当前片段与其他所有片段之间的联系时，是首先通过计算当前片段特征与其他所有片段特征之间的相似性值，然后通过比较片段间的相似性值与所设置阈值 τ 的大小来更新当前片段的联系的。只有第 t 个片段与当前片段之间的相似性值大于等于阈值 τ 时，当前片段才是可以向第 t 个片段学习信息的。那么这个阈值 τ 的选取就比较重要。如何选择一个相对较好的值，是算法中需要考虑的一件事情，为此本文针对 τ 的设置进行了一系列实验验证。

从表3-5中的结果可以看出，当阈值 τ 设定为 0.095 时，效果是最好的。阈值 τ 过大和过小都不合适。本文分析认为如果阈值 τ 的设置过小，就可能会在一些联

表 3-5 针对多模态片段交互模块中阈值 τ 选取的消融实验。

τ	0	0.025	0.050	0.075	0.095	0.115
Supervised	75.3	76.1	75.1	75.7	**77.2**	75.6
Weakly-supervised	73.1	73.2	73.8	73.1	**74.1**	73.0

系比较弱的片段间建立联系，那么这样就会导致当前片段学习到一些无效的信息，反而会干扰最终的判断。如果阈值 τ 的设置过大，当前片段就可能会失去一些本应该建立联系的片段，那么这样就会导致当前片段学不到足够的信息，反而不利于最终的判断。所以，根据此，本文的所有实验都是在 τ 值为 0.095 的设定下进行实验的。

3.4.5.4 算法中模态融合的方式对实验结果的影响

根据 3.3.3 节中的方案描述，本文在分别对与当前片段有联系的同类模态特征进行聚合后，会产生基于视觉模态的聚合特征以及基于音频模态的聚合特征。在进一步聚合形成当前片段的特征时，本文利用对级联后的特征进行映射所产生的这两个权重参数 β_v 与 β_a 来进行聚合。除此之外，常见的特征融合方式还有累加（Add）与级联（Concatenation）。为了验证这三种融合方式中，哪种方式是最佳的，本文在 AVE 数据集上进行了验证。对比结果如表3-6所示。

表 3-6 在 AVE 数据集上针对不同融合方式的消融实验。

Approaches	**Supervised**	**Weakly-supervised**
Add	76.2	72.5
Concatenation	75.7	73.0
Weight mapping	**77.2**	**74.1**

从表中可以看出与使用累加和级联这两种融合方式相比，利用映射所产生的注意力权重的方式更加有效果。本文分析认为这是因为不同模态在表达不同信息时对网络识别能力的提升的重要性是不一样的。如打开水龙头和关水龙头这两个动作，对网络而言，从视觉信息上判断的难度要高于从音频信息上的判断。因为从视觉信息上来看，这两个动作比较相似，而音频上的区别却很大。此外，不同模态的不同片段在传递同一事件信息时，也会是不一样的，也是有不同侧重的，所以通过学习所得到的权重来进行聚合更加合理。

3.4.6 实验可视化结果

在 3.4.4 节与 3.4.5 节中，本文对所提出的算法进行了定量的对比与分析。在本节中，将定性的角度对所提出的算法进行分析。

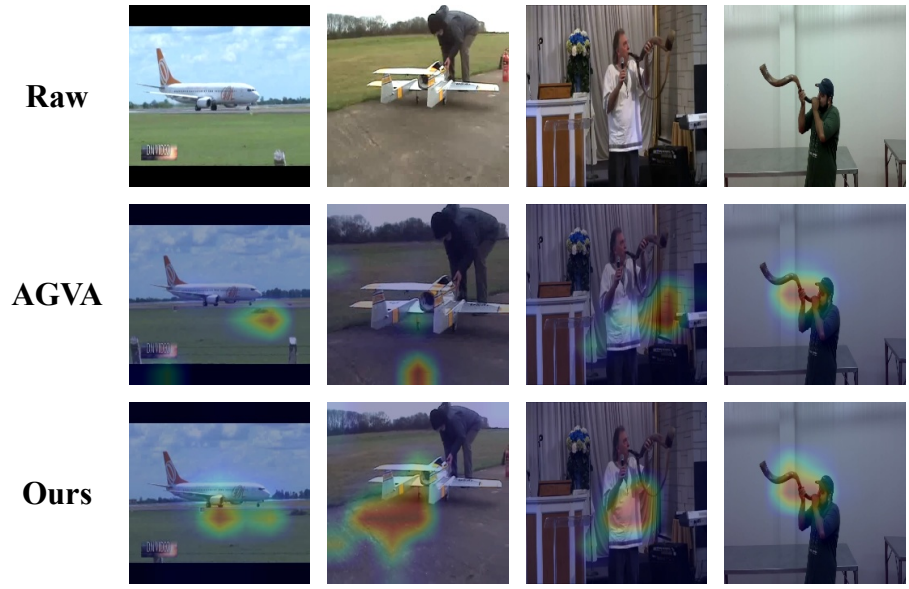

图 3-6 空间注意力图的可视化。第一行为原图，第二行为由 AVGA 方法所产生的注意力图，第三行为由本文 AVCM 模块所产生的空间注意力图。

图3-6中所展示的是由本文 AVCM 模块所产生的空间注意力图与原图以及由论文 [13] 中 AVGA 模块所产生的空间注意力图的对比可视化结果。从图3-6中，可以看出，相比于 AVGA，本文的算法能够更加聚焦于与当前音视事件有关的区域。如飞机起飞时，本文的模块能够更好到关注到引擎区域，而不是像 AVGA 算法一样关注到一些无关区域，在针对比较直接的乐器发声时，两种算法都能很好的进行区域聚焦，但可以看出，本文的算法能够聚焦的更加精确，这也从另一个角度证明了本文的音视共同注意力模块的有效性。

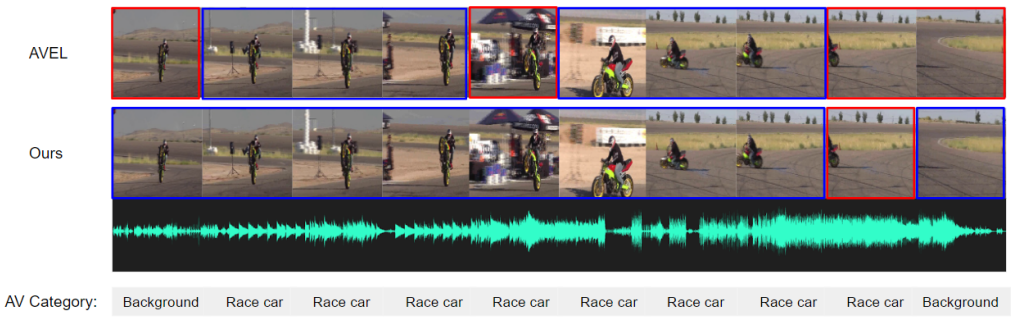

图 3-7 关于“Race car”音视事件定位结果的可视化。蓝色框代表预测正确的片段，红色框代表预测错误的片段。

图3-7所展示的是针对“Race car”这个视频上的音视事件定位结果的可视化，其中第一行为由本文的 baseline 方法 AVEL [13] 所预测产生的结果，第二行为由本文所提出的 MCMA 方法所预测产生的结果，第三行为与当前视频对应的音频信

号。从图3-7中，可以看出针对当前视频中所发生的“Race car”音视事件，在 10 个片段中本文能够预测对 9 个片段上所发生的音视事件，而 AVEL [13] 只预测对了 6 个。这也在一定程度上证明了本文所提出算法的有效性。此外，值得注意的是，两种算法在第九段视频上都预测错误。本文分析是因为在这个片段上，虽然能够在音频模态上听出摩托车引擎的声音，但是在视觉模态上，只能在片段最开始的时候看到部分摩托车，所以要识别出当前片段的音视事件还是有挑战性的。这也在一定程度上说明了同时结合音频与视觉模态识别音视事件的挑战性。

3.5 本章小结

针对音视频事件定位任务，本章提出了一个基于模态间交叉注意力机制的多模态网络算法，该算法框架中主要包含了特征再编码模块、音视间的共同注意力模块、多模态间的片段交互模块以及音视频事件定位的实现部分。本章先主要介绍了这个任务的具体问题描述与定义，然后对所提算法框架中模块的具体实现细节进行了介绍。之后本章对本文所使用的数据集信息、实验设置、实验的对比结果情况以及可视化效果进行了详尽的阐述。通过在 AVE 数据集上进行的对比实验验证以及消融实验和可视化效果的分析，从定量和定性的角度验证了本章所提出的模型在联合多模态进行学习并处理不同模态在片段内与片段间关系上的有效性。

第四章 基于弱监督学习下的音视频事件解析

在本章中，本文将进一步的研究音视频事件解析任务。这个任务相当于对音视频事件定位任务的扩展，该任务不再以一个音视对为目标对音视事件进行识别，而是需要分别在单模态上以及音视联合模态上对事件进行识别。为此，本文提出了一种在弱监督下针对音视频事件解析的多模态学习算法（Multi-Modal Learning Algorithms，MMLA）该算法利用三个分支分别对音频模态、视觉模态以及音视联合模态进行处理，在建模单模态内以及多模态内上下文关系的同时，结合弱监督下的学习损失，使网络产生更具有判别性的预测。

接下来本文将在 4.1 节中给出音视频事件解析任务的具体问题描述与相关符号定义，在 4.2 节中简要介绍所提出的多模态学习算法的整体框架，并在 4.3 节中详细地介绍所提出模型的具体实现细节，在 4.4 节中介绍实验的一些细节设置以及对相关的实验结果进行分析，最后在 4.5 节中给出本章小结。

4.1 问题描述与定义

为了使网络模型能够很好地理解一个视频中的每个单独模态以及音视联合模态在所有片段上所正在发生的事件，论文 [42] 首先提出了音视频事件解析任务。文中指出这个任务要做的就是预测出一个视频中的哪些片段，在哪些模态（音频模态、视觉模态或者音视联合模态）上发生了什么事件[42]。图4-1给出了一个音视频事件解析的例子，可以看出在视频中的不同片段上，针对音频模态与视觉模态发生了汽车（Car）和说话（Speech）这两个事件，而在音视模态上，则是在音频模态与视频模态的交集片段上发生了事件。

给定一个包含连续片段的长视频 S，将其分割成 T 个不重叠的片段对 $S=\{S_t=(V_t,A_t)\}_{t=1}^{T}$，其中，每个片段持续时间为 1 秒，$V_t$ 和 A_t 分别代表第 t 个片段上所对应的视觉内容与音频内容，T 表示的是长视频 S 中所包含的片段总数。针对第 t 个视频片段 (V_t,A_t)，对应的标签为

$$y_t=\left\{(y_t^a,y_t^b,y_t^{av}) \mid [y_t^a]_c,[y_t^v]_c,[y_t^{av}]_c\in\{0,1\},c=1...C\right\}$$

其中 C 代表的是数据集中所发生的事件类别总数，y_t^a，y_t^b 和 y_t^{av} 分别代表音频模态、视觉模态和音视模态在当前片段上所发生的事件标签。所以，y_t 是一个多类别的事件标签，即在第 t 个片段可能会发生 1 个到多个事件类别。此外，某个片段只有在音频模态与视频模态上都发生了同一事件，才认为该片段在音视模态上发

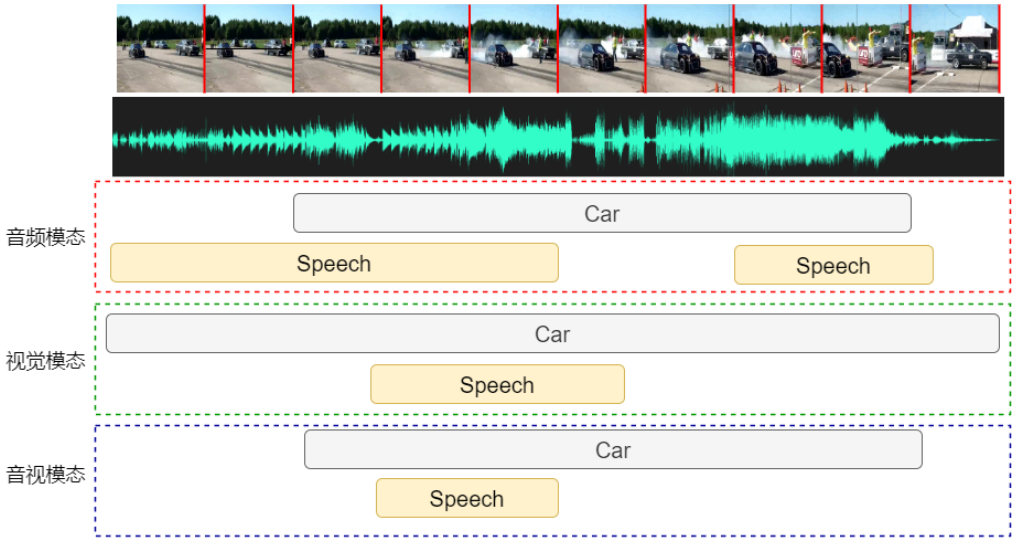

图 4-1 音视频事件解析例子示意图。

生了对应事件，所以 y_t^a，y_t^b 和 y_t^{av} 三者标签之间有这样一种关系：$y_t^{av} = y_t^a \cap y_t^b$。因此，只要能够准确预测视频中的音频模态与视觉模态在每个片段上所发生的事件，通过 y_t^a，y_t^b 和 y_t^{av} 之间的关系也能准确预测视频中的音视联合模态在每个片段上所发生的事件。

如果要以一种全监督的方式对模型进行训练，需要在数据集上针对每个视频中的所有片段都进行密集的标签标注工作，这将会是极其耗时和昂贵的，然而若是只需要使用视频级别上的弱标签则可以很容易的进行标注。为了避免繁琐地标记，和前人工作 [42,45,83] 一样，本文以一种弱监督的方式来进行训练。即在训练时，只知道在视频序列 S 中所发生的整体事件，但具体是哪种模态在哪个片段上发生了什么样的事件是不知道的，但是在预测时需要具体地预测出每种模态在每个片段上所发生的事件。图4-2给出了弱监督下的音视频事件解析任务示意图。在训练过程中，只能获取视频级别上的标签汽车（Car）和说话（Speech），而所要做的就是预测出音频模态，视觉模态，以及音视联合模态三者分别在这个视频上的哪些片段上，发生了什么事件类别。在这个例子上，针对音频模态，需要在第 3 到 9 个片段上预测出其发生了汽车事件，在第 1 到 5 个片段以及第 8 到第 9 个片段上预测出其发生了说话事件，而将其它片段预测为背景。此外，有些片段上可能会发生多个事件，例如视觉模态与音频模态在第 4 秒到第 5 秒这两个片段上，都发生了两个事件。这也在很大程度上增加该任务的实现难度。因此在弱监督情况下完成音视频事件解析是一件非常有挑战的工作。

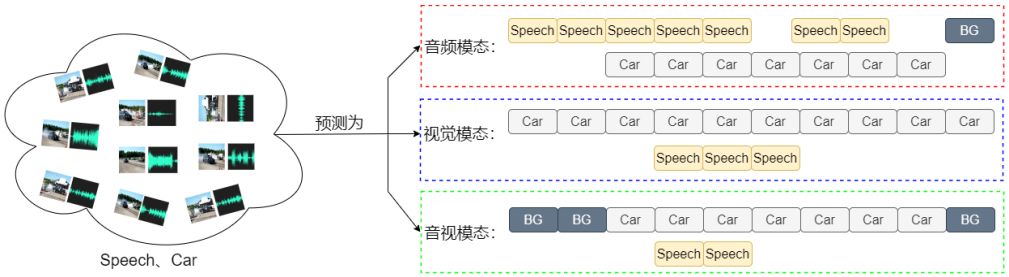

图 4-2 弱监督下的音视频事件解析任务示意图。

4.2 整体框架结构

针对弱监督下的音视频事件解析任务，本文所提出的整体框架如图4-3所示。首先针对输入的音视信息，分别通过已经预训练好的卷积网络来提取对应的音频特征与视觉特征，然后该框架通过三个分支来并行的处理音频模态、视觉模态以及音视联合模态上的信息。其中上下两个针对单模态的特征学习分支在结构上是对称的，主要是通过多个串联的 Transformer 编码器对模态内的上下文关系进行建模以及为了实现事件解析任务所进行的多实例池化处理。中间的针对音频联合模态进行处理的分支包括多头的自注意力处理部分、音视模态间的交互部分以及结合音视对比学习（Audio-Visual Contrastive Learning）和多模态多实例学习池化损失（Multimodal Multi-Instance Learning pooling，MMIL pooling）的约束部分。具体实现过程，本章将在后续小节进行详细阐述。

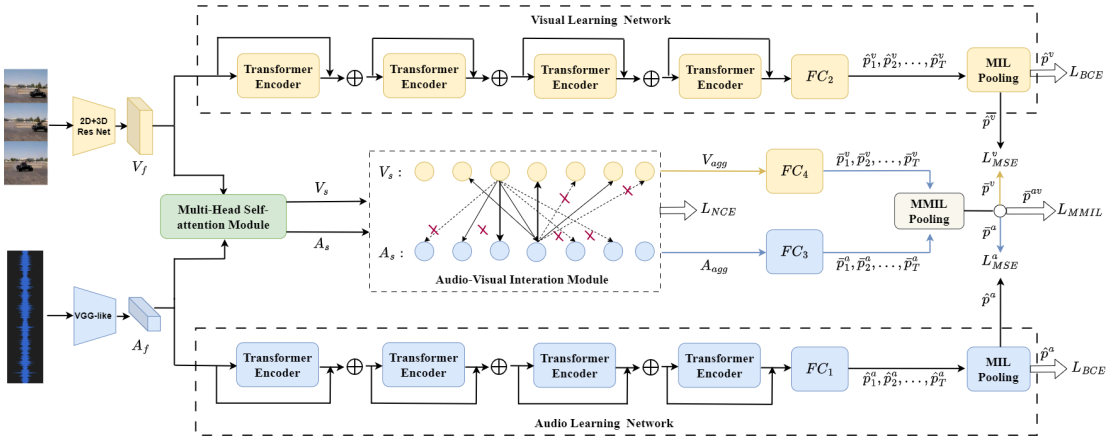

图 4-3 弱监督下的针对音视频事件解析的多模态学习算法整体框架。

4.3 算法模型

本章节将对所提出的在弱监督下的针对音视频事件解析的多模态学习算法进行详细的阐述，其中主要包括特征提取模块、多头的自注意力处理模块、音视片

段间的交互模块以及针对音频或者视觉的单模态学习网络。

4.3.1 特征提取模块

针对音视频事件解析任务，本文分别在特征层面上与概率空间上对不同模态的信息进行了处理。针对给定的视频信息输入 $S=\{S_t=(V_t,A_t)\}_{t=1}^{T}$。为了提取该视频上所对应的音视两种模态的特征，本文利用已经在其他数据集上预训练好的卷积网络来分别提取视频中 T 个片段上的视觉特征与音频特征，其中所提取到的视觉特征包括了片段上的 2D 特征与视频级别上的 3D 特征。本文将在第 t 个片段上所提取到的 2D 视觉特征表示为 $V_{2d}\in R^{1\times d_{v1}}$，提取到的音频特征表示为 $A_t\in R^{1\times d_a}$，其中 d_{v1} 代表的是 2D 视觉特征的维度，d_a 代表的是音频特征的维度，那么整个视频上的音频特征可以表示为 $A_f\in R^{T\times d_a}$。此外，本文将在整个视频上所提取到的 3D 视觉特征表示为 $V_{3d}\in R^{d\times d_{v2}}$，其中 d 和 d_{v2} 分别代表该特征在时间和空间上的维度。为了便于处理，本文首先对两种视觉特征进行融合以形成进一步的融合特征 V_f。首先针对 $V_{3d}\in R^{d\times d_{v2}}$，用大小为 (8, 1) 的池化核对其进行二维平均池化操作得到特征以在时间维度上进行缩减，由此可得到处理后的 3D 特征 $V_{3d}^{'}$，然后将 $V_{3d}^{'}$ 与映射后的 2D 视觉特征 V_{2d} 进行级联操作，最后再次通过一个映射层进行融合，由此形成整个视频级别上的视觉特征 $V_f\in R^{T\times d_v}$。具体计算公式如下：

$$V_{3d}^{'}=avgpool\_2d(FC(V_{3d}),(8,1)) \tag{4-1}$$

$$V_f=FC(Concate(V_{3d}^{'},FC(V_{2d}))) \tag{4-2}$$

其中函数 $avgpool\_2d(X,d)$ 表示的是对输入的特征 X 按照池化核大小为 d 的尺度进行二维平均池化操作，函数 $Concate(\cdot)$ 表示的是对输入的特征按照通道维度进行级联操作。之后本文以预处理后的音频特征 $A_f\in R^{T\times d_a}$ 与视觉特征 $V_f\in R^{T\times d_v}$ 为输入，在后续的三个分支中进行不同方面的处理。

4.3.2 多头的自注意力处理模块

在图4-3中对音视联合模态进行处理的分支中，考虑到同一模态内的不同片段是会共享一定语义信息的，所以在获得视频中所有片段级别上的音频以及视觉特征后，本文首先通过多头的自注意力机制来处理同一模态内部在不同片段间的关系。

多头自注意力机制的核心就是将输入的序列分别映射到多个“头”空间中，在每个“头”空间中都会进行一个自注意力操作，在将所有“头”空间的输出拼接在一起并通过一个线性映射层得到最终的输出。在这个过程中，每个“头”的关

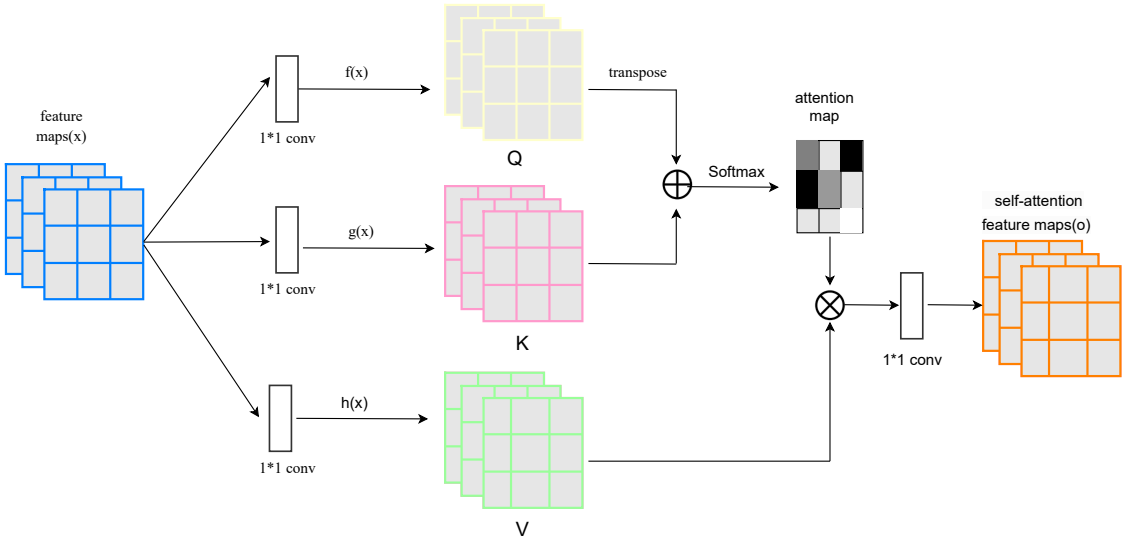

图 4-4 “头”空间中的自注意力处理流程示意图[84]。

注点都会不一样，所以通过对所有“头”的结果进行加权平均，可以在处理序列数据时获得更鲁棒的特征表示。每个“头”空间中的自注意力处理流程如图4-4所示。以对视觉特征 V_f 的处理为例，则进行一次 Self-attention 的具体计算公式为：

$$Q,K,V=(W^qV_f,W^kV_f,W^vV_f) \tag{4-3}$$

$$F=SA(V_f)=Softmax(\frac{Q^TK}{\sqrt{d}})V \tag{4-4}$$

其中 Q，K，V 分别代表注意力机制中基于输入 V_f 所产生的查询、键以及值，W^q、W^k 和 W^v 分别为对应 Q，K，V 进行映射时的权重矩阵，d 指的是查询值 Q 或者键值 K 中的特征维度数。

上述过程描述的是在一个“头”空间中进行的计算，在多头自注意力机制中，会将每个“头”空间的注意力权重与对应的值向量进行加权求和以得到最终的多头自注意力表示，公式如下：

$$head_i=Attention(QW_i^q,KW_i^k,VW_i^v) \tag{4-5}$$

$$MultiHead(Q,K,V)=Concate(head_1,...,head_h)W^o \tag{4-6}$$

其中 W_i^q，W_i^l 和 W_i^v 分别代表的是在第 i 个头空间中，进行映射时所产生的权重矩阵，函数 $Concate(\cdot)$ 表示的是对输入的特征按照通道维度进行级联操作。本文将通过多头自注意力机制处理后的视觉特征与音频特征分别被表示为 $V_s\in R^{T\times d_v}$ 和 $A_s\in R^{T\times d_a}$，然后在音视片段间的交互模块中基于此特征进行进一步的交互学习。

4.3.3 音视片段间的交互模块

在解决音视频事件解析任务时，仍然需要考虑不同模态在片段间的交互。目前，大多数方法 [42, 45] 采用的都是跨模态的自注意力机制来进行模态间的处理，但本文认为这种方法在对模态间片段上关系的探索并不充分，很可能引入不必要的信息。为了让音视两种模态能够充分的进行交互，从而可以学习到一些全局的语义信息，然后能够更加准确的完成视频级别上的事件预测。本文的音视模态间的交互模块（Audio-Visual Interaction Module，AVIM）通过一种有选择性的交互学习方式来对另一个模态在不同片段上的关系进行建模与交互，以使我们的模型能够感知到更多的事件级信息。此外，该模块基于对比学习和 MMIL pooling 操作来对模型在处理片段级别上的预测结果与视频级别上的预测结果的关系进行约束。

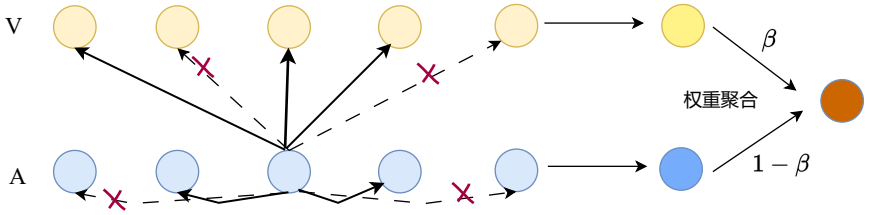

图 4-5 MSIM 模块中针对不同模态在片段上进行关系建模的简图。在对同模态与不同模态的片段信息进行有选择聚合后，再次通过加权的方式进行聚合。

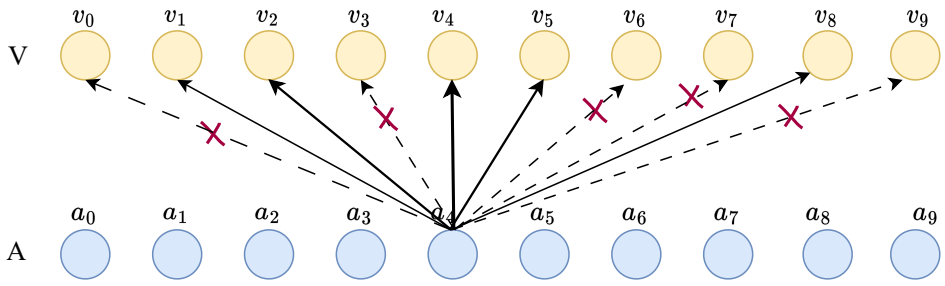

图 4-6 AVIM 模块中针对不同模态在片段上进行关系建模示意图。其中所展示的是针对视频中的第 4 个音频片段 a_4 与视觉模态中的其他片段建立关系的过程。

在第三章中本文基于异构图算法提出了一个 MSIM 模块来处理不同模态在不同片段上的交互，并通过实验进行了有效性验证，其简化示意图如图4-5所示。本章基于 MSIM 模块中的部分处理方式，提出了一个音视模态间的交互模块用于音视模态间的关系建模。两种处理方式的区别在于针对音视频事件解析任务，在此并没有对同模态上的片段进行建模，只是对不同模态的片段进行建模。之所以这

样处理，是因为该任务需要考虑每个单模态在不同片段上所发生的事件，在前面本文已经通过多头的自注意力机制对同模态内的不同片段进行了关系建模，此时再进行处理，反而会取得相反效果。图4-6所展示的是在音视频事件解析任务中，通过 AVIM 模块对视频中的 a_4 片段与另一个模态的其他片段建立联系的示意图。具体的处理过程如图4-7所示。图4-7中展示了以 V_s 与 A_s 为输入，通过音视模态间的交互模块处理后输出 V_{agg} 与 V_{agg} 的详细计算过程。

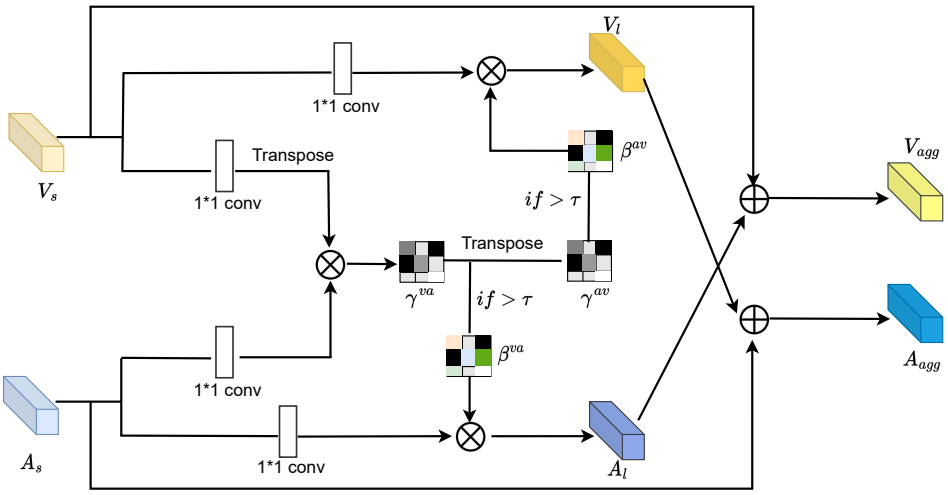

图 4-7 AVIM 模块的具体计算过程。

以针对音频模态中的片段计算为例，在计算得到音频片段与视觉片段之间的相似性矩阵 γ^{av} 后，通过指数函数 $\mathbb{I}(\cdot)$ 的运算，可以得到更新后的相似性矩阵 β^{av}，之后利用最终的相似性矩阵 β^{av} 来对视觉模态中的有联系的片段进行聚合，同时为了保持原有音频的信息，本文通过残差连接以获得最终聚合后的音频特征 $A_{agg} \in R^{T\times d_a}$。具体计算公式为：

$$\gamma^{av} = \frac{(A_s W_1^A)(V_s W_1^V)^T}{\sqrt{d}},\ \beta^{av} = \gamma^{av}\mathbb{I}(\gamma^{av} - \tau) \tag{4-7}$$

$$A^{agg} = \beta^{av}(V_s W_2^V) + A_s \tag{4-8}$$

其中 W_1^A、W_1^V 与 W_2^V 是在映射过程中产生的可学习的参数，d 是视觉特征或音频特征的维度数，$\mathbb{I}(\cdot)$ 是一个指示函数，当输入大于等于 0 的时候会输出 1，当输入小于 0 的时候会输出 0。同理，可以获得最终聚合后的视觉特征 $V_{agg} \in R^{T\times d_v}$。

由于在训练的过程中无法获取到不同模态在片段级别上所发生的事件标签，但是又要在测试的时候针对每种模态，预测出每个片段上所会发生的事件类别。所以，这就需要以某种方式去聚合片段上的标签以形成整体上的视频级别的标签来进行训练约束。本文把这种弱监督情况下的音视频事件解析任务当成多模态多

实例学习问题去解决。为此，本文利用一种注意力下的 MMIL 池化处理[42]去从每种模态的相关性得分上进行聚合以得到视频级别下的事件概率用于训练过程中的约束，同时会对每种模态进行约束。针对聚合后的音频特征 $A_{agg} \in R^{T\times d_a}$ 和视觉特征 $V_{agg} \in R^{T\times d_v}$，在通过一个分类器处理后，可以得到两种模态在每个片段上针对不同事件所发生的概率预测 $P^v = \{p_1^v, p_2^v, \cdots, p_T^v\}$ 与 $P^a = \{p_1^a, p_2^a, \cdots, p_T^a\}$。计算公式为：

$$P^v = Sigmoid(FC_4(V_{agg})) \tag{4-9}$$

$$P^a = Sigmoid(FC_3(A_{agg})) \tag{4-10}$$

其中 p_t^v 与 p_t^a 表示的是，在第 t 个片段上针对音频与视觉模态上所可能发生事件的预测结果，FC_3 与 FC_4 分别表示的是针对音视两种模态用于分类的的线性映射层。为了对片段上的预测结果 P^v 与 P^a 进行聚合以形成每种模态在视频级别上的结果，本文利用一种注意力下的 MMIL 池化方法[42]进行处理。首先基于聚合后的特征产生一个在片段维度上的权重参数 W_{tv} 与 W_{ta}，然后基于此权重系数来动态地对每个片段上的预测结果进行聚合，以形成最终的针对该模态的视频级别上的结果 $\bar{P}^v$ 和 $\bar{P}^a$。具体计算公式为：

$$W_{tv} = Softmax(FC_{tv}(V_{agg})), \quad W_{ta} = Softmax(FC_{ta}(A_{agg})) \tag{4-11}$$

$$\bar{P}^v = \sum_{t=1}^{T} W_{tv} \odot P^v, \quad \bar{P}^a = \sum_{t=1}^{T} W_{ta} \odot P^a \tag{4-12}$$

其中 FC_{tv} 和 FC_{ta} 表示的是针对视觉特征 V^{agg} 和音频特征 A^{agg} 的线性映射层。由于视频级别上的标签 Y 是音频与视觉两种模态上的标签交集，为了生成整个视频级别上的预测结果 $\bar{P}^{av}$，本文进一步的对音视两种模态在视频级别上的结果 $\bar{P}^v$ 和 $\bar{P}^a$ 进行处理。具体计算公式为：

$$W_{mv} = Softmax(FC_{mv}(V_{agg})), \quad W_{ma} = Softmax(FC_{ma}(A_{agg})) \tag{4-13}$$

$$\bar{P}^{av} = W_{mv} \odot \bar{P}^v + W_{ma} \odot \bar{P}^a \tag{4-14}$$

其中 FC_{mv} 和 FC_{ma} 表示的是在对不同模态进行关注时，针对视觉特征 V^{agg} 和音频特征 A^{agg} 的线性映射层，W_{mv} 与 W_{ma} 表示的是针对不同模态的权重参数。有了所预测的视频级别上的事件概率 $\bar{P}^{av}$ 和真实的视频级别上的训练标签 Y，可通过二元

的交叉熵损失对所提出的算法模型来进行约束，对应的损失计算方式为：

$$L_{ce} = BCE(\bar{P}^{av}, Y) = -\sum_{c=1}^{C} Y[c]\, log(\bar{P}^{av}[c]) \tag{4-15}$$

同时，由于视频级别上的标签包含了音视两种模态在所有片段上所发生的事件类别，如果只用 L_{ce} 损失进行约束，可能会在训练过程中更多的关注最具有判别性的模态而忽略了另一种模态信息，这样不利于对视频中不同模态在每个片段上所发生的事件的识别。所以，为了避免在训练过程中的模态偏差，本文针对音视两种模态也进行交叉熵损失约束。具体公式为：

$$L_g = BCE(\bar{P}^{a}, Y_a) + BCE(\bar{P}^{v}, Y_v) \tag{4-16}$$

其中，$Y_a = Y_v = Y$，此外，在实现过程中有些算法 [42, 43] 为了提升模型的泛化能力，会对 Y_a 与 Y_v 中的标签进行了标签平滑处理[85]。由此，有多模态多实例学习损失 L_{MMIL}，公式为：

$$L_{MMIL} = L_{ce} + L_g = BCE(\bar{P}^{a}, Y_a) + BCE(\bar{P}^{v}, Y_v) + BCE(\bar{P}^{av}, Y) \tag{4-17}$$

为了提升模型能够区别不同片段中所表达的信息的能力，本文利用对比学习损失来进一步进行约束。在对比学习中通常会构建正负样本对。在本文中，本文把同一个视频中的第 t 个片段上的视觉特征 f_t^v 与音频特征 f_t^a 作为一个正样本对，而把另一个模态的不在同一个时间片段上的所有片段特征（$f_{t'}^{a/v}, t' \neq t$）作为负样本集。本文利用 Noise Contrastive Estimation (NCE) 来拉近同一时刻上的正样本对 f_t^v 与 f_t^a，同时拉远当前片段上的视觉特征 f_t^v 与其它非对应片段上的音频特征 $f_{t'}{}^a$。因此，针对第 t 个片段，有如下对比学习损失公式：

$$L_c = -log\frac{exp(f_t^{v\,T} f_t^a/\tau)}{\sum_j exp(f_t^{v\,T} f_j^a/\tau)} \tag{4-18}$$

4.3.4 音频或者视觉的单模态学习网络

不同于音视频事件定位任务，在音视频事件解析任务中，是需要网络分别针对视频中的音频模态、视觉模态以及音视联合模态在不同片段上所发生的事件进行预测。在图4-3的中间分支上，本文对音视两种模态在不同片段上的特征交互进行了处理，更新后的特征同时融合了音视两种模态的特征，这种处理可以更好的形成视频级别上的结果，但对每种单模态内的独有信息获取上略有不足，所以本文提出了一个针对音频或者视觉的单模态学习网络（Modality-wise Learning Network，MLN）来提升模型区别不同模态的能力并可以针对性的对不同模态事件进行预

测。

如图4-3所示，针对音频模态和视觉模态，分别会有一个相应的学习网络。每个分支以预处理后的音频特征 A_f 或视觉特征 V_f 作为输入，分别通过两个不共享参数的 Transformer 编码器对模态内的上下文关系进行建模。参数的不共享可以限制了两种模态间的信息交互。在每个分支里面，每个 Transformer 编码器会迭代的处理输入的特征，但同一个分支中的编码器是参数共享的。特别地，给定音频特征 A_f 和视觉特征 V_f，对每种模态内的上下文关系进行建模的公式为：

$$E_a^{i+1} = T_a(E_a^i),\ \ E_v^{i+1} = T_v(E_v^i) \tag{4-19}$$

其中，$E_a^i \in R^{T\times d}$ 与 $E_v^i \in R^{T\times d}$ 分别代表的是在第 i 次迭代过程中，通过第 i 个 Transformer 编码器进行建模后的音频特征和视觉特征，经过学习网络编码后的音视特征分别表示为 $E_a \in R^{T\times d}$ 与 $E_v \in R^{T\times d}$。同样的，将编码后的单模态特征 E_a 与 E_v 送入一个分类器，会得到对应模态在不同片段上所预测的事件概率 $P^{v'} = \left\{p_1^{v'}, p_2^{v'}, \cdots, p_T^{v'}\right\}$ 和 $P^{a'} = \left\{p_1^{a'}, p_2^{a'}, \cdots, p_T^{a'}\right\}$，然后基于此，本文通过单模态的多实例学习来聚合生成视频级别上的标签 $\hat{P}^v$ 和 $\hat{P}^a$，以再次针对每种模态进行约束。具体计算公式为：

$$P^{a'} = Sigmoid(FC_1(E_a)),\ \ P^{v'} = Sigmoid(FC_2(E_v)) \tag{4-20}$$

$$W_{ta_1} = Softmax(FC_{ta_1}(E_a)),\ \ W_{tv_1} = Softmax(FC_{tv_1}(E_v)) \tag{4-21}$$

$$\hat{P}^a = \sum_{t=1}^{T} W_{ta_1} \odot P^{a'},\ \ \hat{P}^v = \sum_{t=1}^{T} W_{tv_1} \odot P^{a'} \tag{4-22}$$

其中，FC_1、FC_2、FC_{ta_1} 与 FC_{tv_1} 代表的都是针对特征进行映射的线性层，W_{ta_1} 与 W_{tv_1} 分别代表的是针对音视两种模态在时间片段维度上的权重参数，$\hat{P}^v$ 和 $\hat{P}^a$ 分别代表的是针对音视两种模态在视频级别上的事件预测结果。

由于每个分支中的 Transformer 编码器仅对单模态内的上下文关系进行建模，所以通过聚合所产生的预测结果 $\hat{P}^v$ 和 $\hat{P}^a$ 是高度依赖与单模态内的信息的。为了对这两个分支进行约束，本文一方面对每种分支所产生的预测结果与视频级别上的标签进行约束，故有损失 $L_{g'}$：

$$L_{g'} = BCE(\hat{P}^a, Y_a) + BCE(\hat{P}^v, Y_v) \tag{4-23}$$

另一方面，考虑到每种模态也是直接共享整个视频中的语义信息的，为了对音视交互分支进行模态指导，本文通过均方损失函数（Mean Square Error function，

MSE）来拉近与音视交互分支所产生的预测结果。所以有损失 L_m：

$$L_m = MSE(\bar{P}^a, \hat{P}^a) + MSE(\bar{P}^v, \hat{P}^v) \tag{4-24}$$

最终，弱监督下音视事件解析的总损失为：

$$L_{wsl} = L_{MMIL} + L_m + L_c + L_{g'} \tag{4-25}$$

其中，λ 是对应损失的权重系数。

4.4 实验设置与结果分析

在本小节中，本文将主要介绍一些与任务相关的实验设置，如数据集的情况，评价指标的确定等，以及对所得出的实验结果进行分析。

4.4.1 数据集介绍

针对音视频事件解析任务，本章使用 Look, Listen, and Parse（LLP）数据集 [42] 来进行算法效果的验证。LLP 数据集 [42] 中一共包含了 11,849 个来自于 YouTube 的时长总和为 32.9 小时的视频，这些视频都是源于 Audioset [72] 数据集。LLP 数据集包含了来自于不同活动领域中共计 25 种类别的事件，如人类说话（human speeches）、婴儿啼哭（baby crying）、狗吠（dog barking）、弹小提琴（playing violin）等事件。在 LLP 数据集中，每个视频的持续长度都是 10 秒，其中每个视频都至少包含了一个音频事件或者视觉事件。据统计，平均每个视频中有 1.64 个不同的事件，而且一共有 7,202 个视频发生了多于 1 个的事件。此外，为了对算法模型进行评估，作者从 LLP 数据集中随机选取了 1,849 个视频并人工在片段级别上进行标记。这 1,849 个视频一共产生了 6,626 个事件标注，其中有 4,131 个是音频上的事件标注，2,495 个是视觉上的事件标注，共形成了 2,488 个音视事件标注。

为了完成弱监督下的音视频事件解析任务，本文使用 10,000 个视频用于训练，649 个视频用于验证，1,200 个视频用于测试。用于训练的视频只有视频级别上的粗标签，而验证集和测试集用的都是片段级别上的细粒度标签。

4.4.2 评价指标

为了对所提出的算法进行全面的验证，本文在片段级别（Segment-level）和事件级别（Event-level）上对算法在音频事件、视觉事件以及音视事件的预测结果进行评估。其中，F-sores 被用来作为评估指标的计算。片段级别上的评价指标指的是在每个片段上的效果，而事件级别上的评价指标则是针对某个事件而言的。为了计算事件级别下的 F-sores，本文将同一事件下所有的连续片段作为一个整体，

然后以 mIou 为 0.5 的阈值进行 F-sores 计算。此外，本文还通过计算聚合后的指标 Type@AV 与 Event@AV 来评估全局下的音视事件解析效果。特别地，Type@AV 指的是平均计算所有音频、视觉以及音视事件后的结果，而 Event@AV 指的是针对每个事件, 平均计算音频事件和视觉事件的 F-sores 得分，而不是直接与 Type@AV 一样直接平均所有不同类别事件的结果。

4.4.3 实验细节

4.4.3.1 实验数据的预处理

和 3.4.3 节中对音频与视觉信息的处理类似。针对音频模态信息，在获得相应的频谱图后，进一步使用在 Audio-set[72] 上预训练好的 VGG-like 网络模型[74] 来提取对应的音频特征，最终针对每个片段所获取到的音频特征的维度为 1×128。那么一个视频上的音频特征维度为 10×128。针对给定数据集中的视觉信息，对视频按照 fps=8 的采样率进行帧提取，然后分别通过预训练好的卷积神经网络来提取 2D 和 3D 视觉特征。其中 2D 视觉特征的提取用的是在 ImageNet[75] 上预训练好的 ResNet-152 网络模型[86]，3D 视觉特征的提取用的是在 Kinetics-400[87] 上预训练好的深度为 18 的 R(2+1)D 网络模型[88]。片段上的 2D 视觉特征维度为 1×512，视频上的 3D 视觉特征维度为 80×2048。

4.4.3.2 实验中的具体参数

本文的网络模型在训练过程中采用的都是 Adam 优化器，其中参数的学习率设置为 0.0003，训练的批量大小为 16，共计训练 60 轮。用于片段间关系建模的阈值 τ 的大小为 0.09，用于对比学习损失中的参数 τ 的大小为 0.2。在音频或者视频的单模态学习网络中 Transformer 编码器所迭代的次数为 4，总片段数 T 的值为 10，类别种类数 C 的值为 25，d_v 的数值大小为 512，d_a 的数值大小为 128，d 的数值大小为 80，head 头的数量为 4。

4.4.4 与其他最新方法的对比

本小节将对所提出的弱监督下的多模态学习算法与其它相关算法进行分析比较。特别地，为了进行公平地比较，本文采用和其他算法一样的经过预处理后的音频特征与视觉特征。具体实验结果将会在下文进行分析。

在针对音频模态上的事件解析上，与音频事件检测算法 TALNet[89] 进行了对比，在针对视觉模态上的事件解析上，与弱监督下的动作定位算法 STPN[90] 和 CMCS[91] 进行对比，此外在针对整个视频上的音频模态与视觉模态以及音视模态上的事件解析上，一方面对比了本文的 baseline 算法 HAN[42] 以及最新发表的算法

MA [45]，另一方面，也与调整后的音视频事件定位中的两个算法 [13,34] 进行了对比。在弱监督下，本算法所得到的音视频事件解析结果与其它方法的对比结果如表4-1所示。

表 4-1 在 LLP 数据集上与其他方法的结果对比。

Event type	Methods	Segmet-level	Event-level
Audio	TALNet [89]	50.0	29.1
	AVE [13]	47.2	41.7
	AVSDN [34]	47.8	34.1
	HAN [42](baseline)	60.1	51.3
	MA [45]	60.3	53.6
	Ours	**63.6**	**55.3**
Visual	STPN [90]	46.5	41.5
	CMCS [91]	48.1	45.1
	AVE [13]	37.1	34.7
	AVSDN [34]	52.0	46.3
	HAN [42](baseline)	52.9	48.9
	MA [45]	**60.0**	**56.4**
	Ours	59.2	54.3
Audio-Visual	AVE [13]	35.4	31.6
	AVSDN [34]	37.1	26.5
	HAN [42](baseline)	48.9	43.0
	MA [45]	**55.1**	**49.0**
	Ours	54.7	48.4
Type@AV	AVE [13]	39.9	35.5
	AVSDN [34]	45.7	35.6
	HAN [42](baseline)	54.0	47.7
	MA [45]	**58.9**	**53.0**
	Ours	58.0	51.8
Event@AV	AVE [13]	41.6	36.5
	AVSDN [34]	50.8	37.7
	HAN [42](baseline)	55.4	48.0
	MA [45]	**57.9**	**50.6**
	Ours	57.1	50.4

从表中结果可以看出，除了与最新算法 MA [45] 相比，本文所提出的算法，在 10 种指标上的结果都高于其它算法的效果，这也能在很大程度上证明本算法的有效性。与最新算法 MA [45] 相比，本文算法的结果只能在音频模态取得更好的结果，但在其它几种指标下，结果相对会差一些，但差距不是很大，除了在视觉模态上的结果之差，基本都在 1% 以内。本文分析认为，这是因为 MA 算法中针对每种模态所进行的冗余标签消除操作起到了主要的贡献作用。该方法通过一种两阶段的训练方法能够为不同模态获得更加精准的约束标签，从而很好的训练网络以取

得较好效果。这也是本文算法的不足之处，后续需要进一步改进。

相比于 baseline 算法 HAN[42]，本文算法所取得的结果在每项评价指标上都能得到明显的提升，这也能间接反应出本算法的一个有效性。此外，本文发现无论是在哪种平均指标上，音频模态上的结果都要高于视觉模态上的结果，尤其是片段级别上的结果。这说明本文能够更好地分析与处理在音频模态上所发生的事件，但在对视觉模态上的处理与预测还有待改进，同时本文也认为这和数据集中音频事件的标注较多，在训练时会有更多的数据有关。从整体上的评价指标 Type@AV 与 Event@AV 来看，本文的算法也取得了较好的结果。

4.4.5 消融实验

4.4.5.1 算法中不同模块的作用

本文所提出的弱监督下的针对音视频事件解析的多模态学习算法主要包括三个分支，其中上下两分支是针对视觉模态的学习网络（VLN）和针对音频模态的学习网络（ALN）, 中间分支是针对音视模态的交互学习，主要包括一个多头的自注意力处理模块（MHSA）和一个音视间的交互模块（AVIM）。为了对这三个分支的有效性进行验证，本文提出了四个不同的变体方法："w/o MHSA"，"w/o AVIM"，"w/o VLN"，"w/o ALN"。其中"w/o MHSA"代表的是只是去除中间分支的多头自注意力处理模块，而保持其他部分不变，"w/o AVIM"代表的是只是去除中间分支的音视模态间的片段交互模块，"w/o VLN"代表的是只是去除模型中针对视觉模态的学习分支，而"w/o ALN"代表的是只是去除模型中针对音频模态的学习分支。

表 4-2 在 LLP 数据集上针对模型中不同模块在片段级别上的消融实验。其中"w/o"表示的是不使用（Without）。

Methods	Audio	Visual	Audio-Visual	Type@AV	Event@AV
w/o MHSA	62.3	52.1	51.8	55.4	55.4
w/o AVIM	59.7	47.0	45.9	50.9	51.8
w/o VLN	61.3	51.6	49.2	54.0	55.6
w/o ALN	61.9	52.5	50.7	55.0	56.0
Ours	**63.6**	**59.2**	**54.7**	**58.0**	**57.1**

四中变体方法在 LLP 数据集上，关于音视频事件解析任务在片段级别上以及在事件级别上的实验结果如表4-2和表4-3所示。可以看出在缺失任意一个模块时，算法效果都相比于原算法有不同下降，这在一定程度上可以论证算法中几个模块的有效性。对比"w/o VLN"与"w/o ALN"两种方法，可以看出相比于单独的音

频学习分支，单独的视觉学习分支对最终的效果会起到更多的贡献作用，本文认为，一方面这是由于网络确实需要基于单模态信息的指导，另一方面模型的学习过程中会更依赖视觉分支中的模态信息。同时相比于“w/o MHSA”方法，可以看到在缺失音视间的交互模块时，算法的下降的效果更为剧烈。本文分析认为，是因为 AVIM 模块能够很好的对音视两种模态在不同片段间的关系进行建模，能够使中间分支所学习到的特征更加有效。此外，可以看出，相比于片段级别上的结果，在事件级别上的预测准确率会相对低一些，这也说明进一步的预测连续片段上的结果会更加困难。

表 4-3 在 LLP 数据集上针对模型中不同模块在事件级别上的消融实验。其中“w/o”表示的是不使用（Without）。

Methods	Audio	Visual	Audio-Visual	Type@AV	Event@AV
w/o MHSA	53.4	47.2	45.7	48.8	47.3
w/o AVIM	49.7	34.8	32.7	39.1	39.1
w/o VLN	52.6	46.8	42.4	47.3	48.5
w/o ALN	52.8	47.9	44.2	48.3	48.7
Ours	**55.3**	**54.3**	**48.4**	**51.8**	**50.4**

4.4.5.2 多模态多实例损失计算方式的影响

本文把弱监督下的音视频事件解析任务当成一种多模态多实例学习问题来解决，利用一种注意力下的 MMIL 池化方法[42]去从每种模态所预测的事件类别得分上进行聚合以得到对应模态在视频级别上的事件概率，然后利用交叉熵损失用于训练过程中的约束。在池化方式的计算上，除了基于注意力的计算方法还有基于最大池化以及平均池化的方法，为此本文进行了消融验证以验证哪种方式能够取得最佳效果。在片段级别上以及事件级别上的实验结果如表4-4和表4-5所示，其中 Max pooling 表示的是针对每个训练视频，选择视频上所有片段中预测值最高的那个片段上的结果来代表整个视频级别上的预测结果，而 Mean pooling 则表示的是平均所有片段上的预测结果来当做最终的视频级别上的预测结果。

表 4-4 在 LLP 数据集上不同计算方式下的池化损失在片段级别上的结果。

MMIL Pooling	Audio	Visual	Audio-Visual	Type@AV	Event@AV
MAX	61.0	51.7	49.8	54.2	55.5
Mean	61.3	52.7	50.1	54.7	56.6
Attentive	**63.6**	**59.2**	**54.7**	**58.0**	**57.1**

无论是从在片段级别上的实验结果还是从事件级别上的实验结果，都可以从

表 4-5 在 LLP 数据集上不同计算方式下的池化损失在事件级别上的结果。

MMIL Pooling	Audio	Visual	Audio-Visual	Type@AV	Event@AV
MAX	51.8	45.8	43.2	46.9	46.8
Mean	51.6	48.5	43.3	47.8	47.9
Attentive	**55.3**	**54.3**	**48.4**	**51.8**	**50.4**

表4-4和表4-5中看出使用注意力下的 MMIL 池化方法会更加有效果，本文分析认为这是因为在时间维度上对不同片段上的预测结果以及针对不同模态进行加权融合会让网络能够动态的学习到更加的关键信息以完成预测。同时可以看出 Max pooling 与 Mean pooling 下的结果相差并不大，但 Mean pooling 的总体效果要更好，本文认为这是因为在一个视频中的不同模态上可能会发生多个事件，而只选取最大的那个作为整体结果可能会存在偏差，平均所有结果反而可能更好。所以，本文的实验都是基于注意力下的池化方式来进行计算的。

4.4.5.3 算法中不同损失的影响

在本文所提出的算法中，包括一个注意力下的多模态多实例池化损失 L_{MMIL}，单独针对音视两种模态进行约束的池化损失 $L_{g'}$，对中间分支进行指导学习的均方误差损失 L_m, 以及一个对比学习损失 L_c。为此，本文针对这四个损失，在其它条件不变的情况下，进一步的进行消融实验，以验证每个损失的有效性。在片段级别上以及事件级别上的实验结果如表4-6和表4-7所示。

表 4-6 在片段级别上不同损失函数组合下的结果。其中“w/o”表示的是不使用（Without）。

Loss	Audio	Visual	Audio-Visual	Type@AV	Event@AV
w/o L_c	62.1	53.5	50.6	55.4	56.5
w/o $L_{g'}$	61.2	51.1	49.7	54.0	54.9
w/o L_m	61.8	53.1	51.0	55.3	55.9
$L_{MMIL}+L_m+L_c+L_{g'}$	**63.6**	**59.2**	**54.7**	**58.0**	**57.1**

表 4-7 在事件级别上不同损失函数组合下的结果。其中“w/o”表示的是不使用（Without）。

Loss	Audio	Visual	Audio-Visual	Type@AV	Event@AV
w/o L_c	53.2	49.6	44.6	49.2	49.8
w/o $L_{g'}$	52.3	46.8	43.6	47.6	47.9
w/o L_m	53.1	48.5	44.4	48.7	48.6
$L_{MMIL}+L_m+L_c+L_{g'}$	**55.3**	**54.3**	**48.4**	**51.8**	**50.4**

从表4-6和表4-7中的结果可以看出，无论哪一种损失的缺失都降低了模型的性能，这也间接证明了本文所使用到的每种损失的有效性。对比几种情况可以发现，在缺失针对音视两种模态进行约束的池化损失 $L_{g'}$ 时，网络模型的效果最差，本文分析认为，如果缺少了损失 $L_{g'}$，那么在算法框架中就会缺少对上下两个音视分支网络的约束，从而不能利用到两种单独模态的信息。在缺少损失 L_m 时，模型性能的下降也证明了通过均方误差损失的约束可以为中间音视处理分支提供一定的指导信息。同时可以发现，在加上对比学习损失后，效果得到了提升，本文认为，这是因为虽然通过 AVIM 模块已经对音视两种模态在片段间的关系进行了建模，但是在弱监督情况下需要进一步的通过对比学习损失对其进行约束，以提升模型能够区别不同片段所表达信息的能力。

4.4.5.4 算法中 Transformer 数量的影响

在图4-3所示的框架上，本文提出了一个针对音频模态或者视觉模态的单模态学习网络，其中每个网络分支是由多个串联的 Transformer 编码器所组成。为了确定对每个分支中最合适的编码器数量，本文在 LLP 数据集上针对音视学习网络中不同 Transformer 数量块进行了实验验证。实验结果如表4-8和表4-9所示。

表 4-8 在 LLP 数据集上针对音视学习网络中不同 Transformer 数量块在片段级别上的消融实验。

Number	Audio	Visual	Audio-Visual	Type@AV	Event@AV
1	63.3	55.5	52.7	57.2	**57.8**
2	63.3	57.4	53.4	57.0	57.1
3	**63.6**	58.3	53.7	57.3	57.6
4	**63.6**	**59.2**	54.7	**58.0**	57.1
5	63.1	58.2	**55.1**	57.9	57.6
6	63.2	54.8	51.2	55.5	57.5

表 4-9 在 LLP 数据集上针对音视学习网络中不同 Transformer 数量块在事件级别上的消融实验。

Number	Audio	Visual	Audio-Visual	Type@AV	Event@AV
1	**55.4**	51.5	47.1	51.3	**50.9**
2	54.9	52.8	47.3	50.8	50.4
3	55.3	54.1	48.3	51.4	50.7
4	55.3	**54.3**	48.4	**51.8**	50.4
5	55.0	53.8	**48.9**	51.6	50.6
6	55.1	49.5	44.5	49.3	50.8

从表4-6与表4-7中的结果可以看出，不同块数的 Transformer 编码器会对最终

的网络模态产生不同影响。如果串联的编码器数量较少会导致对单模态内的信息建模不充分，但若数量过多会导致网络学习到许多无用信息。此外，随着编码器数量的增加，预测效果在整体上会存在先增加后下降的趋势。从整体上的结果来看，在 Transformer 编码器数量为 4 的时候，模型效果较好。所以，本文在最终的实验过程中对音视两个分支所使用的 Transformer 编码器数量都是 4。

4.5 本章小结

针对音视频事件解析任务，本章提出了一种弱监督下的多模态学习算法，基于多模态多实例池化学习的方式以及对比学习，通过一种三分支的交互学习方式能够有效的完成该任务。在本章中，本文先主要介绍了这个任务的具体描述以及符号定义，然后对所提算法框架中不同模块的具体实现细节进行了介绍。之后通过在 LLP 数据集上进行的对比实验以及消融实验的验证分析，证明了本章所提出的模型的有效性。

第五章 全文总结与展望

5.1 全文工作总结

视频理解是人类与生俱来的能力，人类通过同时处理和融合来自视觉和听觉等多种模态的高维输入来感知这个世界。此外，随着深度学习的发展，机器感知模型也越来越趋向于使用多模态了，相关的任务也从早期的分别针对单独模态的处理发展到需要同时联合处理两种或者多种模态。音视频事件定位以及音视频事件解析这两个任务则是在视频领域，尤其是在多模态视频语义分析理解研究中的热门研究任务。这两个任务在实现难度上呈现递增趋势，后者在处理上也更加细化。虽然目前已经有了不少相关的研究成果，但目前仍然存在一些问题：（1）在结合多模态信息来理解视频中的事件时，需要更加精确的关注到相应模态的关键区域上，但目前的一些方法更多的是倾向于对视觉模态进行聚焦，并未考虑到音视两种模态间的交互处理。（2）在秒级别上对视频中的所有模态都进行事件标注是一个极其消耗人力与财力的事情，所以如何能够在弱监督情况下的训练来实现片段级别上的高准确率也是一件比较有挑战性的事情。（3）由于多模态数据对同一对象的描述会存在形式上的多样性、语义上的一致性的特点，但如何对这种冗余性进行消除以及对模态间的互补性进行利用却比较困难，目前一些方法常认为同一时间区域内的音频信息与视频信息应该是一致的，并未针对性的处理音视不一致问题以及在对多模态不同片段间的交互处理也较为简单。

为此，本文以解决上述问题为出发点，结合多模态数据的特点以及相关任务的设置，针对音视频事件定位任务提出了一种基于模态间交叉注意力机制的多模态网络算法，针对音视频事件解析任务提出了一种在弱监督下的针对音视频事件解析的多模态学习算法。本文的主要工作如下：

（1）本文从与音视频事件定位和音视频事件解析这两个任务相关的研究背景、应用前景和目前的研究难点等方面进行入手，首先系统的阐述了两个任务的研究背景以及研究目标，并对目前的主要研究方法进行了概括性总结，并对不同算法的优点与不足之处进行了分析。之后对与本文相关的网络结构以及相关的理论知识和技术手段进行了介绍，包括使用到的卷积网络结构、F1 分数等评价指标、多实例学习和异构图以及常用的多模态特征融合方法。

（2）针对音视频事件定位任务，为了保证对音视两种模态信息都能够关注到，本文以预处理后的音视融合特征作为指导信息，通过音视间的交叉关注，在消除视觉区域中干扰信息影响的同时聚焦整个视觉区域中与对应音频模态相关的关键

事件区域，与此同时对与之对应的音频特征进行信息增强。

（3）针对音视频事件解析任务，为了有效探究视频中音频模态、视觉模态以及音视联合模态中的信息，本文以一种三分支的处理方式针对性的进行处理，在保持模态自身信息的同时又进一步的交互学习，可有效地利用视频中不同模态的信息来完成针对模态上的片段事件识别。

（4）在对多模态数据利用上，本文基于异构图中处理并聚合不同类型结点的思想，在音视模态内与模态间的片段关系上进行探索，并通过对比片段间的相似性值与阈值的关系来更新片段间的关系，最后通过权重注意力的方式来聚合其它片段所带来的信息。这种处理方法能够在解决音视片段不一致问题的同时，获取更关键的信息从而显著提升预测结果。

（5）在弱监督情况下的监督上，针对音视频事件定位任务，本文基于多实例学习中的方法，通过结合所预测的事件相关性得分以及事件类别得分进行池化聚合以得到视频级别上的预测结果，而在针对音视频事件解析的任务时，则是在多实例聚合方法的基础上，同时利用具有注意力的池化损失以及对比学习损失进行约束，以更好的针对每种模态完成预测。

（6）设计了详细的实验验证方案。针对两种任务所提出的框架，本文首先对框架中的每个模块的具体实现细节以及相关的计算公式以及最终的实现流程进行了阐述。之后在相关的数据集上，对两个算法与当下最新的研究方法进行比较分析并从不同角度对算法进行消融实验以验证每个模块的作用。

5.2 后续工作展望

本文针对音视频事件定位任务与音视频事件解析任务进行了研究探索，并分别提出了一种基于模态间交叉注意力机制的多模态网络算法和一种在弱监督下的针对音视频事件解析的多模态学习算法。虽然本文的算法能够取得一定的结果，但还有许多方面需要进一步的研究：

（1）在音视事件定位的任务中，本文对多模态数据在模态内与模态间的关系进行了不同程度的探索与研究，但可能在最终获取到的特征上仍会存在冗余。需要更好的探究如何在有效地获取不同模态间的特征的同时又减小冗余。

（2）在音视频事件解析的任务中，本文所提出的算法无法很有效地将视频级别上的标签精确的分配到每种模态在片段上所发生的事件上去。这也是本文利于注意力下的多实例池化损失以及对每种模态进行约束和同时使用对比学习损失的原因。但本文认为理论上还会存在更加有效的方式进行噪声标签的消除，如何使在训练过程中针对每个模态以及模态中的片段获得相对更加正确的标签还需要进

一步研究。

（3）针对两种模态信息的匹配和关联，目前仅使用相似性计算进行处理，但这种对特征空间的距离的计算可能并不是最有效的方法，更多的还是没有从数据本质出发，此外在不同的数据集中存在模态的倾向偏差会不利于鲁棒性网络的训练。如何设计合理地计算以及有效的处理策略是一个极具挑战性的工作。

希望更多的学者能够从不同的角度，更深的深度，更加积极地进行相关探索。

参考文献

[1] Bulkin D A, Groh J M. Seeing sounds: visual and auditory interactions in the brain[J]. Current Opinion in Neurobiology, 2006, 16(4): 415-419.

[2] Rabiner L R. A tutorial on hidden markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[3] Graves A, Mohamed A r, Hinton G. Speech recognition with deep recurrent neural networks[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013: 6645-6649.

[4] Hameete P, Leysen S, Van Der Laan T, et al. Intelligent multi-camera video surveillance.[J]. International Journal on Information Technologies & Security, 2012, 4(4).

[5] Buehler E, Branham S, Ali A, et al. Sharing is caring: Assistive technology designs on thingiverse[C]. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015: 525-534.

[6] Kay M, Matuszek C, Munson S A. Unequal representation and gender stereotypes in image search results for occupations[C]. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015: 3819-3828.

[7] 刘勇, 谢若莹, 丰阳. 智能家居中的居民日常行为识别综述 [J]. 计算机工程与应用, 2021, 57(4): 35-42.

[8] Bowen J, Bradburne J, Burch A, et al. Digital technologies and the museum experience: Handheld guides and other media[M]. Rowman Altamira, 2008.

[9] Porcheron M, Fischer J E, Reeves S, et al. Voice interfaces in everyday life[C]. Proceedings of the 2018 CHI conference on Human Factors in Computing Systems, 2018: 1-12.

[10] Gao R, Oh T H, Grauman K, et al. Listen to look: Action recognition by previewing audio[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 10457-10467.

[11] Nagrani A, Yang S, Arnab A, et al. Attention bottlenecks for multimodal fusion[J]. Advances in Neural Information Processing Systems, 2021, 34: 14200-14213.

[12] Wu C Y, Feichtenhofer C, Fan H, et al. Long-term feature banks for detailed video understanding[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 284-293.

[13] Tian Y, Shi J, Li B, et al. Audio-visual event localization in unconstrained videos[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 247-263.

[14] Tian Y, Li D, Xu C. Unified multisensory perception: Weakly-supervised audio-visual video parsing[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, 2020: 436-454.

[15] Camporesi C, Kallmann M, Han J J. Vr solutions for improving physical therapy[C]. 2013 IEEE Virtual Reality (VR), 2013: 77-78.

[16] Sodhi R, Poupyrev I, Glisson M, et al. Aireal: interactive tactile experiences in free air[J]. ACM Transactions on Graphics (TOG), 2013, 32(4): 1-10.

[17] Arandjelovic R, Zisserman A. Look, listen and learn[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 609-617.

[18] Owens A, Efros A A. Audio-visual scene analysis with self-supervised multisensory features[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 631-648.

[19] Arandjelovic R, Zisserman A. Objects that sound[C]. Proceedings of the European Conference on C Vision (ECCV), 2018: 435-451.

[20] Chung J S, Zisserman A. Out of time: automated lip sync in the wild[C]. Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13, 2017: 251-263.

[21] Nagrani A, Chung J S, Albanie S, et al. Disentangled speech embeddings using cross-modal self-supervision[C]. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 6829-6833.

[22] Korbar B, Tran D, Torresani L. Cooperative learning of audio and video models from self-supervised synchronization[J]. Advances in Neural Information Processing Systems, 2018, 31.

[23] Owens A, Wu J, McDermott J H, et al. Ambient sound provides supervision for visual learning[C]. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, 2016: 801-816.

[24] Aytar Y, Vondrick C, Torralba A. Soundnet: Learning sound representations from unlabeled video[J]. Advances in Neural Information Processing Systems, 2016, 29.

[25] Cheng Y, Wang R, Pan Z, et al. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning[C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 3884-3892.

[26] Morgado P, Vasconcelos N, Misra I. Audio-visual instance discrimination with cross-modal agreement[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12475-12486.

[27] Hu D, Nie F, Li X. Deep multimodal clustering for unsupervised audiovisual learning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 9248-9257.

[28] Rajan V, Brutti A, Cavallaro A. Robust latent representations via cross-modal translation and alignment[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 4315-4319.

[29] Gao R, Feris R, Grauman K. Learning to separate object sounds by watching unlabeled video[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 35-53.

[30] Gao R, Grauman K. Co-separating sounds of visual objects[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 3879-3888.

[31] Gan C, Huang D, Zhao H, et al. Music gesture for visual sound separation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 10478-10487.

[32] Qian R, Hu D, Dinkel H, et al. Multiple sound sources localization from coarse to fine[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, 2020: 292-308.

[33] Oya T, Iwase S, Natsume R, et al. Do we need sound for sound source localization ?[C]. Proceedings of the Asian Conference on Computer Vision, 2020.

[34] Lin Y B, Li Y J, Wang Y C F. Dual-modality seq2seq network for audio-visual event localization[C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: 2002-2006.

[35] Lin Y B, Wang Y C F. Audiovisual transformer with instance attention for audio-visual event localization[C]. Proceedings of the Asian Conference on Computer Vision (ACCV), 2020.

[36] Xu H, Zeng R, Wu Q, et al. Cross-modal relation-aware networks for audio-visual event localization[C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 3893-3901.

[37] Ramaswamy J, Das S. See the sound, hear the pixels[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020: 2970-2979.

[38] Wu Y, Zhu L, Yan Y, et al. Dual attention matching for audio-visual event localization[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 6292-6300.

[39] Duan B, Tang H, Wang W, et al. Audio-visual event localization via recursive fusion by joint co-attention[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 4013-4022.

[40] Zhou J, Zheng L, Zhong Y, et al. Positive sample propagation along the audio-visual event line[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 8436-8444.

[41] Xuan H, Zhang Z, Chen S, et al. Cross-modal attention network for temporal inconsistent audio-visual event localization[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 279-286.

[42] Tian Y, Li D, Xu C. Unified multisensory perception: Weakly-supervised audio-visual video parsing[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[43] Lamba J, Akula J, Dabral R, et al. Cross-modal learning for audio-visual video parsing[J]. arXiv preprint arXiv:2104.04598, 2021.

[44] Yu J, Cheng Y, Zhao R W, et al. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing[C]. Proceedings of the 30th ACM International Conference on Multimedia, New York, NY, USA, 2022: 6241 6249.

[45] Wu Y, Yang Y. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[46] Cheng H, Liu Z, Zhou H, et al. Joint-modal label denoising for weakly-supervised audio-visual video parsing[C]. Proceedings of the European Conference on Computer Vision (ECCV), Cham, 2022: 431-448.

[47] Wei H, Feng L, Chen X, et al. Combating noisy labels by agreement: A joint training method with co-regularization[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[48] Yu X, Han B, Yao J, et al. How does disagreement help generalization against label corruption?[C]. Proceedings of the 36th International Conference on Machine Learning, 2019: 7164-7173.

[49] Nguyen D T, Mummadi C K, Ngo T P N, et al. Self: Learning to filter noisy labels with self-ensembling[J]. arXiv preprint arXiv:1910.01842, 2019.

[50] Owens A, Efros A A. Audio-visual scene analysis with self-supervised multisensory features[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 631-648.

[51] Zhou H, Xu X, Lin D, et al. Sep-stereo: Visually guided stereophonic audio generation by associating source separation[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, 2020: 52-69.

[52] Shams L, Seitz A R. Benefits of multisensory learning[J]. Trends in Cognitive Sciences, 2008, 12(11): 411-417.

[53] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[54] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[55] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[56] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[57] 周志华. 机器学习 [M]. 清华大学出版社, 2016.

[58] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles[J]. Artificial Intelligence, 1997, 89(1-2): 31-71.

[59] Zhou Z H. Multi-instance learning: A survey[J]. Department of Computer Science & Technology, Nanjing University, Tech. Rep, 2004, 1.

[60] Song L, Liu J, Qian B, et al. A deep multi-modal cnn for multi-instance multi-label image classification[J]. IEEE Transactions on Image Processing, 2018, 27(12): 6025-6038.

[61] Huang S J, Gao W, Zhou Z H. Fast multi-instance multi-label learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2614-2627.

[62] Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network[C]. The World Wide Web Conference, 2019: 2022-2032.

[63] Zhang C, Song D, Huang C, et al. Heterogeneous graph neural network[C]. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 793-803.

[64] Hu Z, Dong Y, Wang K, et al. Heterogeneous graph transformer[C]. Proceedings of The Web Conference 2020, New York, NY, USA, 2020: 2704 2710.

[65] Guo W, Wang J, Wang S. Deep multimodal representation learning: A survey[J]. IEEE Access, 2019, 7: 63373-63394.

[66] Ramachandram D, Taylor G W. Deep multimodal learning: A survey on recent advances and trends[J]. IEEE Signal Processing Magazine, 2017, 34(6): 96-108.

[67] 何俊, 张彩庆, 李小珍. 面向深度学习的多模态融合技术研究综述 [J]. 计算机工程, 2020, 46(5): 1-11.

[68] Liu Z, Shen Y, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[J]. arXiv preprint arXiv:1806.00064, 2018.

[69] Sahu G, Vechtomova O. Dynamic fusion for multimodal data[M]. , 2019.

[70] Li X, Wang C, Tan J, et al. Adversarial multimodal representation learning for click-through rate prediction[C]. Proceedings of The Web Conference 2020, New York, NY, USA, 2020: 827 836.

[71] Wu J, Yu Y, Huang C, et al. Deep multiple instance learning for image classification and auto-annotation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 3460-3469.

[72] Gemmeke J F, Ellis D P, Freedman D, et al. Audio set: An ontology and human-labeled dataset for audio events[C]. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017: 776-780.

[73] Ramaswamy J. What makes the sound?: A dual-modality interacting network for audio-visual event localization[C]. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 4372-4376.

[74] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25.

[75] Hershey S, Chaudhuri S, Ellis D P, et al. Cnn architectures for large-scale audio classification[C]. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017: 131-135.

[76] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[77] Kazakos E, Nagrani A, Zisserman A, et al. Epic-fusion: Audio-visual temporal binding for egocentric action recognition[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 5492-5501.

[78] Cartas A, Luque J, Radeva P, et al. Seeing and hearing egocentric actions: How much can we learn?[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019: 4470-4480.

[79] Alwassel H, Mahajan D, Korbar B, et al. Self-supervised learning by cross-modal audio-video clustering[J]. Advances in Neural Information Processing Systems, 2020, 33.

[80] Panda R, Chen C F, Fan Q, et al. Adamml: Adaptive multi-modal learning for efficient video recognition[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[81] Xia Y, Zhao Z. Cross-modal background suppression for audio-visual event localization[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 19989-19998.

[82] Feng G, Hu Z, Zhang L, et al. Encoder fusion network with co-attention embedding for referring image segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 15506-15515.

[83] Pasi P S, Nemani S, Jyothi P, et al. Investigating modality bias in audio visual video parsing[J]. arXiv preprint arXiv:2203.16860, 2022.

[84] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[85] Müller R, Kornblith S, Hinton G E. When does label smoothing help?[J]. Advances in neural information processing systems, 2019, 32.

[86] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.

[87] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 6299-6308.

[88] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 6450-6459.

[89] Wang Y, Li J, Metze F. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling[C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: 31-35.

[90] Nguyen P, Liu T, Prasad G, et al. Weakly supervised action localization by sparse temporal pooling network[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 6752-6761.

[91] Liu D, Jiang T, Wang Y. Completeness modeling and context separation for weakly supervised temporal action localization[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 1298-1307.