中图法分类号:TP37;TP18　　**文献标识码**:A　　**文章编号**:1006-8961(2026)01-0273-16

论文引用格式:Zhao X L, Zhao X, Zheng K, Yu H Y and Sun X. 2026. Pretrained hybrid architecture model for movie scene segmentation. Journal of Image and Graphics, 31(1):0273-0288(赵晓蕾, 赵鑫, 郑珂, 于浩洋, 孙旭. 2026. 预训练混合架构模型的电影场景分割方法. 中国图象图形学报, 31(1):0273-0288)[DOI:10.11834/jig.240532]

预训练混合架构模型的电影场景分割方法

赵晓蕾[1],赵鑫[1,2],郑珂[3*],于浩洋[4],孙旭[5]

1. 聊城大学东昌学院,聊城　252000; 2. 河南大学美术学院,郑州　450046; 3. 聊城大学地理与环境学院,聊城　252000; 4. 大连海事大学信息科学与技术学院,大连　116026; 5. 中国科学院空天信息创新研究院,北京　100094

摘　要:**目的**　随着电影内容的复杂化与多样化,电影场景分割成为理解影片结构和支持多媒体应用的重要任务。为提升镜头特征提取和特征关联的有效性,增强镜头序列的上下文感知能力,提出一种混合架构电影场景分割方法(hybrid architecture scene segmentation network, HASSNet)。**方法**　首先,采用预训练结合微调策略,在大量无场景标签的电影数据上进行无监督预训练,使模型学习有效的镜头特征表示和关联特性,然后在有场景标签的数据上进行微调训练,进一步提升模型性能;其次,模型架构上混合了状态空间模型和自注意力机制模型,分别设计Shot Mamba镜头特征提取模块和Scene Transformer特征关联模块,Shot Mamba通过对镜头图像分块建模提取有效特征表示,Scene Transformer则通过注意力机制对不同镜头特征进行关联建模;最后,采用3种无监督损失函数进行预训练,提升模型在镜头特征提取和关联上的性能,并使用Focal Loss损失函数进行微调,以改善由于类别不平衡导致的精度不足问题。**结果**　实验结果表明,HASSNet在3个数据集上显著提升了场景分割的精度,在典型电影场景分割数据集MovieNet中,与先进的场景分割方法相比,AP(average precision)、mIoU(mean intersection over union)、AUC-ROC(area under the receiver operating characteristic curve)和F1分别提升1.66%、10.54%、0.21%和16.83%,验证了本文提出的HASSNet方法可以有效提升场景边界定位的准确性。**结论**　本文提出的HASSNet方法有效结合了预训练与微调策略,借助混合状态空间模型和自注意力机制模型的特点,增强了镜头的上下文感知能力,使电影场景分割的结果更加准确。

关键词:电影场景分割;预训练模型;状态空间模型(SSM);自注意力机制;无监督相似性度量

Pretrained hybrid architecture model for movie scene segmentation

Zhao Xiaolei[1], Zhao Xin[1,2], Zheng Ke[3*], Yu Haoyang[4], Sun Xu[5]

1. Dongchang College, Liaocheng University, Liaocheng　252000, China; 2. College of Fine Arts, Henan University, Zhengzhou　450046, China; 3. College of Geography and Environment, Liaocheng University, Liaocheng　252000, China; 4. Information Science and Technology College, Dalian Maritime University, Dalian　116026, China; 5. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing　100094, China

Abstract: **Objective** With the rapid advancement of the film industry and the increasing complexity of movie content, automatic movie scene segmentation has emerged as a critical task for understanding film narrative structure and supporting various multimedia applications. Traditional scene segmentation methods often struggle with effectively extracting shot fea-

收稿日期:2024-09-29;**修回日期**:2025-08-22;**预印本日期**:2025-08-29

*通信作者:郑珂　zhengke@lcu.edu.cn

基金项目:山东省教育教学研究课题项目(2024JXQ260);聊城大学东昌学院博士科研启动基金项目(2024PHD009)

Supported by:Shandong Province Educational and Teaching Research Project (2024JXQ260); Doctoral Research Start-Up Fund of Dongchang College of Liaocheng University (2024PHD009)

tures and modeling the contextual relationships between shots in long video sequences. To address these challenges, this study proposes a novel hybrid architecture for movie scene segmentation called hybrid architecture scene segmentation network (HASSNet), which aims to enhance the effectiveness of shot feature extraction and intershot feature association while improving the contextual awareness capabilities of shot sequences. **Method** The proposed HASSNet consists of two coordinated modules and a two-stage training strategy. At the architectural level, we design 1) a Shot Mamba module for shot feature extraction and 2) a Scene Transformer module for intershot association. Shot Mamba adapts the Mamba state-space model to images by adopting a ViT style patching scheme on key frames, adding positional embeddings and a class token, and then modeling the patch sequence with stacked state-space blocks. Given a sequence composed of a center shot and its temporal neighbors, Scene Transformer takes the shot embeddings as input and applies multihead self-attention to explicitly model global dependencies among shots, yielding context-enriched "shot context embeddings" that capture short- and long-range relations needed for boundary reasoning. At the training level, we adopt a two-stage pipeline that leverages unlabeled and labeled movie data. In pretraining, we use all available movies without scene labels and drive learning with three unsupervised losses built upon a pseudoboundary generator. The pseudoboundary is obtained by scanning candidate split positions and selecting the one that maximizes the combined cosine similarity of the left-side shots to the left endpoint and the right-side shots to the right endpoint. On the basis of this proxy, we introduce 1) a shot-scene matching loss (SSMLoss) that uses an InfoNCE objective to maximize the similarity between the endpoint shots and the mean features of their respective sides while minimizing cross-side similarity; 2) a scene semantic consistency (SSC) loss that, for an anchor shot, treats a randomly chosen shot from the same side of the pseudoboundary as positive and one from the opposite side as negative, again optimized with InfoNCE; and 3) a boundary prediction (BP) loss that encourages the model to discriminate the pseudoboundary shot from nonboundary shots via binary cross entropy. Together, these losses guide the model to learn representations that increase intrascene coherence and interscene separability without requiring annotations. In fine-tuning, we transfer the pretrained parameters to the labeled subset of MovieNet, freeze the Shot Mamba parameters to avoid overfitting on limited labels, and train only the Scene Transformer with Focal Loss. This choice addresses the severe class imbalance in which true boundaries are sparse relative to nonboundary positions and empirically improves boundary sensitivity. **Result** We evaluate on MovieNet and test cross-dataset generalization on BBC and OVSD. On MovieNet, HASSNet outperforms recent approaches, including methods that also employ state-space models. In particular, relative to a state space-based baseline, HASSNet improves AP by 1.66%, mIoU by 10.54%, AUC-ROC by 0.21%, and F1 by 16.83%, demonstrating more accurate boundary localization and better scene grouping. Without any additional fine-tuning, the pretrained model also yields consistent gains on BBC and OVSD, indicating robust transferability. Among the unsupervised losses, tSSMLoss contributes most to boundary discrimination, while jointly optimizing SSMLoss, SSC, and BP yields the best performance. In fine-tuning, Focal Loss consistently outperforms the weighted cross entropy under severe class imbalance. **Conclusion** This study proposes HASSNet, a hybrid architecture for movie scene segmentation trained with a pretraining and fine-tuning strategy. Shot Mamba learns strong, position-aware shot embeddings efficiently from unlabeled data, while Scene Transformer explicitly captures global intershot dependencies for boundary decisions. The pseudoboundary mechanism and the three unsupervised losses enable effective representation learning without labels, and the fine-tuning regimen with Focal Loss mitigates class imbalance in supervised adaptation. Extensive experiments on MovieNet, together with cross dataset and ablations, verify that HASSNet enhances contextual awareness across shots and delivers accurate scene boundary detection. Results indicate that hybrid architectures with pretraining are promising for long-form video understanding.

Key words: movie scene segmentation; pre-trained model; state space model(SSM); self-attention mechanism; unsupervised similarity measurement

0 引　言

随着电影产业快速发展和数字技术的进步，电影的制作和发行变得更加便捷，电影作品数量也呈现爆发式增长。随着电影内容的日益复杂化，如何有效地理解和分割电影场景成为电影制作、编辑、分析及二次创作中的关键问题（Huang等，2020）。电影场景分割（scene segmentation）不仅可以帮助观众更好地理解影片结构，还能为电影推荐系统、视频摘要生成、内容检索以及其他多媒体应用提供技术支撑（Shou等，2021）。在电影制作中，故事线是电影的核心，通过一系列事件和情节的发展，推动故事的进展。故事线的发展需在时空维度上展开，形成多个场景（scene），每个场景包含一组特定事件与对话，服务于故事线的逻辑演进。场景由多个镜头（shot）组成，它是摄像机在某一位置所记录的连续画面，通过镜头的切换和组合，场景得以完整呈现。镜头则可以由一系列关键帧（key frame）描述连续画面中的重要的图像。图1展示了《碟中谍Ⅱ》电影中的开场视频，场景1展示了汤姆克鲁斯饰演的伊森与病毒科学家涅科维奇在飞机上相遇的多组镜头，场景2展示了涅科维奇遭到假冒伊森杀害的经过，场景3是真正的伊森在另一处攀岩并接受任务的过程。由此可见，电影场景是在一定的视频语义理解基础上，具有完整故事情节和连贯视觉效果的视频段落。

图1　电影结构示意图（以《碟中谍Ⅱ》为例）

Fig. 1　Film structure diagram (taking "Mission: Impossible Ⅱ" as an example)

在电影及视频分割领域，早期的任务集中于如何对电影和视频中的镜头进行分割（Baraldi等，2015b；Tang等，2019）。镜头分割（shot segmentation）通过检测镜头切换点实现，这些切换点可以是硬切换（即突然的场景变化）或是渐变过渡（如淡入淡出或叠化），通常是在较短的视频时序中定位连续图像变化的边界（Zhu等，2023）。与镜头分割任务不同，场景分割则需要在长视频中理解镜头之间的视觉和语义关联（Mun等，2022），以便将相关镜头聚合在一起，并定位场景边界。与镜头分割任务相比，场景分割需要在长时间序列和复杂语义信息中分析图像的上下文关系，因此，让计算机理解和分析如电影这类长视频内容通常更具有挑战性（Chen等，2021）。

随着人工智能的快速发展，视频理解（video understanding）受到广泛关注和研究（Huang等，2020），其旨在通过自动化的方法对视频内容进行分析和解读。视频理解任务涵盖了多个具体的研究方向，其中包括视频分类（Karpathy等，2014）、动作识别（Sun等，2023）、视频检索（曾润浩 等，2025）、视频摘要（Apostolidis等，2021）、视频异常检测（梁家菲等，2023）、视频对象检测和跟踪（Yao等，2020）以及场景分割（Wu等，2022）等。动作检测和视频对象检测及跟踪更专注于对特定动作与对象的识别，而场景分割更侧重于理解镜头之间的上下文关系，并准确定位和分割场景的边界。同时，场景边界的准确定位对于下游任务，如视频摘要、视频检索和视频编辑等，具有一定的推动作用。

目前，主流的电影场景分割方法是基于深度学习的方法，可分为两大类：基于监督学习的方法和基于预训练的方法。在监督学习方法的领域中，Baraldi等人（2015a）提出一种基于深度孪生网络的广播视频场景检测方法，该方法通过学习视频帧之间的相似性检测场景边界。进一步地，Rao等人（2020）提出一种局部到全局的多模态电影场景分割方法，通过融合视觉和文本特征，实现了多模态数据的场景分割。Huang等人（2020）不仅构建了MovieNet数据集，还开发了一种全局理解电影的框架，该框架通过监督学习来识别和分类电影中的场景和事件。然而，尽管MovieNet数据集涵盖了超过1 100部电影，但具有场景标签的电影样本数量仅为318部，这极大限制了模型学习到的特征表示能力和模型的

泛化性。

因此，基于预训练的自监督学习方法引起了研究学者的关注（Chen等，2020；Roh等，2021），其可以在无需过度依赖有标签样本情况下，提升模型的表示学习能力。相应地，在电影场景分割领域，Chen等人（2021）通过镜头对比自监督学习方法检测场景边界，利用视频片段之间的对比性信息提高检测效果。Mun等人（2022）提出边界感知自监督学习方法，通过学习视频中的场景边界特征优化场景分割。Wu等人（2022）研究了场景一致性表示学习，通过保持视频场景的一致性实现视频场景分割的自监督学习。

尽管这些方法在自监督损失函数的设计上进行了探索，但在镜头特征提取和不同镜头特征的上下文关联上仍存在一定的不足。例如，当前绝大部分方法仍采用卷积神经网络（convolutional neural network，CNN）作为特征提取模块，在捕捉全局上下文信息和长距离依赖关系方面具有局限性，从而限制了特征提取的有效性和场景分割的准确性。随着深度学习技术的不断进步，Transformer通过其自注意力机制，能够更加灵活地处理序列数据，捕捉长距离依赖关系。状态空间模型（state space model，SSM）则凭借其在长时序任务中优秀的时间和空间复杂度，逐渐在自然语言处理和计算机视觉等多种任务中引起了研究人员的重视。

在上述分析的基础上，为了提升镜头特征提取和镜头特征关联的有效性，本文提出一种混合架构的电影场景分割方法（hybrid architecture scene segmentation network，HASSNet）。与采用单一架构不同，本文构建Mamba + Transformer的混合架构模型，在镜头特征提取阶段，Mamba模型凭借其线性时间复杂度和优秀的位置感知能力，能够高效处理镜头图像的分块序列；在镜头关联阶段，Transformer的自注意力机制能够显式建模不同镜头间的全局依赖关系，更好地捕捉镜头序列的上下文语义信息。这种混合设计充分发挥了两种架构的互补优势，有效规避了纯Mamba架构在特征关联上的不足和纯Transformer架构在处理长序列时的高计算复杂度问题。具体而言，在模型构建方面，首先，提出Shot Mamba模块，使用状态空间模型进行空间建模和位置嵌入以提升镜头图像的特征提取；其次，提出Scene Transformer模块，对不同镜头提取特征进行全局关联性建模，获得具有镜头间上下文关系的镜头特征表示。在训练策略方面，首先，构建3个非监督损失函数，对大量无标签电影数据进行预训练，使模型的镜头特征提取和特征关联具备有效的初始特征表示能力和迁移学习基础；其次，在有标签数据上进行微调，获得具备有效电影场景分割能力的模型。

1 相关工作

电影场景分割是指在视频时序上进行镜头的聚类与分割，国内外研究学者针对场景分割提出多种处理方法，总体可分为3类：基于聚类的分割方法（Rasheed和Shah，2003，2005；Chasanis等，2009）、基于动态规划的分割方法（Han和Wu，2011；Tapaswi等，2014；Sidiropoulos等，2011）和基于深度学习的分割方法（Huang等，2020；Rao等，2020；Chen等，2021）。

基于聚类的视频场景分割方法旨在通过分析和聚合视频中的镜头，以识别和分割具有语义一致性的场景。Rasheed和Shah（2003）在好莱坞电影和电视节目数据集中进行场景分割时，使用了基于视觉内容和时间信息的镜头聚类方法，以捕捉镜头之间的连续性和场景转变。进一步地，Rasheed和Shah（2005）提出一个更为综合的方法，通过结合视觉特征和语义信息，利用聚类技术和图模型，将视频划分为语义上连贯的场景。Chasanis等人（2009）提出利用镜头聚类和序列对齐来进行场景检测的方法，通过提取视觉特征如颜色直方图和运动特征，并使用层次聚类算法将相似的镜头聚合成场景。

基于动态规划的视频场景分割方法利用动态规划算法，通过优化镜头分割点的选择，使视频被分割成语义连贯的场景。Han和Wu（2011）提出一种新颖的边界评价准则，并结合动态规划算法进行视频场景分割。Tapaswi等人（2014）采用动态规划方法优化角色互动的时间线，从而实现视频场景的自动分割和可视化。Sidiropoulos等人（2011）提出利用高层次视听特征进行视频时间分割的方法，结合视觉和音频特征，通过综合分析这些多模态信息，实现了视频场景分割。Rotman等人（2017）提出一种多模态融合的动态规划方法，通过整合视觉、音频和文本信息，利用动态规划进行最优序列分组，从而实现鲁棒的视频场景检测。

当前研究领域中，基于深度学习的方法在视频场景分割领域取得了显著进展。Huang等人(2020)通过建立MovieNet数据集，为电影理解提供了全面的基准，涵盖了镜头、场景和角色等多种注释，促进了电影分析的全面研究。Rao等人(2020)提出一种局部到全局的多模态电影场景分割方法，结合了局部视觉和音频特征与全局上下文信息，显著提高分割准确性。Chen等人(2021)引入了ShotCoL方法，一种基于对比学习的自监督学习方法，通过镜头的特征表示实现高效的场景边界检测。Mun等人(2022)提出一个边界感知的自监督学习框架，通过学习场景边界的显著特征，提升了视频场景分割性能。Islam等人(2023)提出使用状态空间变换器的高效电影场景检测方法，能够捕捉长程依赖和上下文信息，实现了准确且可扩展的场景分割。Tan等人(2023)通过角色注意力网络(character attention network，CAN)，利用角色互动链接镜头并识别场景边界，显著提高分割的准确性。Yang等人(2023)提出一个上下文感知的变换器模型，通过整合全视频的上下文信息，实现了连贯且准确的视频场景分割。Wei等人(2023)引入了多模态高阶关系Transformer，利用多模态特征之间的关联性，实现场景的边界检测。尽管上述部分方法通过引入额外的人物检测任务关联不同镜头，但其并没有充分考虑镜头特征间的相似性与差异性。另外，部分方法仅使用卷积神经网络作为镜头特征提取模块，使得预训练模型的特征表示能力不足，在镜头特征表示和镜头特征间的上下文关联上仍需进一步挖掘。

近年来，状态空间模型在序列建模任务中展现出优秀的性能，成为深度学习领域的重要研究方向。传统的状态空间模型起源于控制理论，通过描述系统状态的演化建模动态过程。Gu等人(2022)提出结构化状态空间序列模型(structured state space sequence model，S4)，首次将连续状态空间模型成功应用于深度学习中的长序列建模任务，通过零阶保持离散化、并行扫描和结构化参数化等技术实现了高效计算，在长程依赖建模上取得了突破性进展。在S4的基础上，研究者们进一步探索了状态空间模型的改进方向。Gu和Dao(2024)提出Mamba模型，通过引入选择性机制使得状态转移参数能够根据输入自适应调整，显著提升了模型的表达能力和计算效率。随后，研究者们提出多种Mamba的变体和改进，如Vision Mamba(Zhu等，2024)将Mamba成功扩展到计算机视觉任务，VMamba(Liu等，2024)在图像分类和目标检测等任务上展现出色性能。针对电影场景分割任务，镜头序列的时序建模和长程依赖关系捕捉是关键挑战。状态空间模型在处理长序列时的线性复杂度优势，以及其在视觉任务中展现的位置感知能力，为电影镜头特征提取提供了新的技术路径。同时，考虑到场景分割任务中镜头间复杂的语义关联性，如何将状态空间模型与注意力机制有效结合，构建适用于电影场景分割的混合架构模型，成为值得深入探索的研究方向。

针对上述问题，本文提出一个混合架构场景分割网络(HASSNet)，重点探讨了如何构建有效的基于预训练策略的电影镜头特征表示和关联模型，结合镜头特征的相似性与差异性，提升电影场景分割的性能。

2 研究方法

2.1 总体框架

本文提出方法的整体框架如图2所示，从训练策略上可以将其划分为预训练阶段和微调阶段。预训练阶段将大量无标签电影数据作为输入，以1个伪边界生成和3个自监督损失函数为牵引，通过构建及训练基于Shot Mamba的镜头特征提取模块和基于Scene Transformer的镜头关联模块，使得模型具备基本的镜头特征提取和上下文关联能力；微调阶段以有标签电影场景数据作为输入，迁移预训练阶段的模型参数，通过Focal Loss损失函数，进一步训练模型，以使模型适应场景分割任务。

本文提出的方法从模型结构上可划分为镜头特征提取和镜头特征关联两部分。如图2所示，模型的输入为镜头S_i及其相邻的前后k个镜头，其中特征提取以Shot Mamba为骨干框架，对每个镜头单独进行特征提取，将镜头关键帧裁剪成切块(patches)，同时耦合位置嵌入(position embedding)，结合Shot Mamba的序列建模结构，获取该镜头关键帧的镜头编码(shot embedding)特征$\boldsymbol{e}$；对不同镜头获取的镜头编码$\boldsymbol{e}_{i-k}$至$\boldsymbol{e}_{i+k}$，采用基于Scene Transformer结构的镜头关联模块进行上下文关联性建模，生成镜头关联特征(shot context embedding)$\boldsymbol{v}_{i-k}$至$\boldsymbol{v}_{i+k}$，使得模型具备整合短时序镜头和潜在长序列镜头信息的能力。

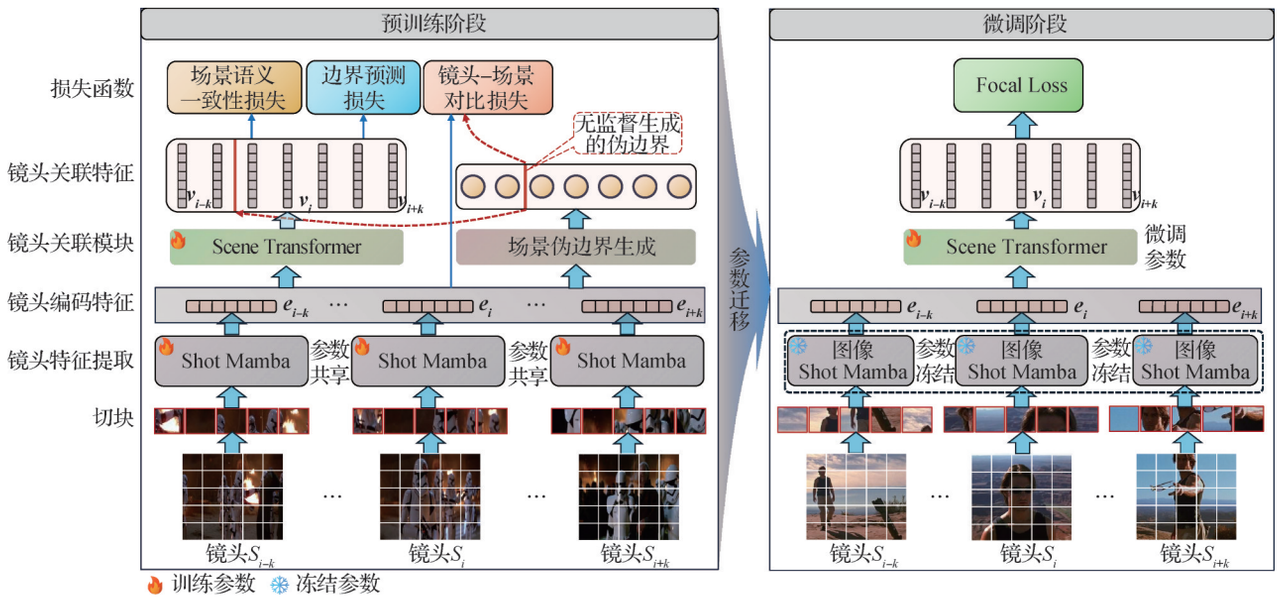

图2　本文方法的整体框架

Fig. 2　Overall framework of the proposed method

本节后续将从模型结构和损失函数两部分分别阐述本文方法的细节。

2.2　模型结构

2.2.1　基于Shot Mamba的镜头特征提取模块

1)状态空间模型。状态空间模型(SSM)是一个用于建模动态系统的数学框架,通常由输入$\boldsymbol{x}(t)$、输出$\boldsymbol{y}(t)$和隐藏状态$\boldsymbol{h}(t)$及一系列对应的系数A、B、C组成。该模型通过描述隐藏状态与输入之间的关系以及隐藏状态与输出之间的关系捕捉系统的动态特性。在状态空间模型中,隐藏状态的演变可表示为

$$\boldsymbol{h}'(t)=A\boldsymbol{h}(t)+B\boldsymbol{x}(t) \tag{1}$$

式中,$\boldsymbol{h}(t)$是在时刻t的当前隐藏状态,$\boldsymbol{h}'(t)$表示了在连续系统中$\boldsymbol{h}(t)$关于时间的微分,表示状态的变化率,A是状态转移系数,表示当前隐藏状态如何影响下一个隐藏状态,$\boldsymbol{x}(t)$是在时刻t的输入,B是输入影响系数,描述了输入对隐藏状态更新的贡献。同时,输出$\boldsymbol{y}(t)$与当前隐藏状态之间的关系可表示为

$$\boldsymbol{y}(t)=C\boldsymbol{h}(t) \tag{2}$$

式中,$\boldsymbol{y}(t)$表示在时刻t的输出,C是观测系数,描述了当前隐藏状态如何影响输出。

基于上述思想,结构化状态空间序列模型(S4)(Gu等,2022)和Mamba(Gu和Dao,2024)等连续系统的离散化模型相继被提出,其采用零阶保持(zero-order hold)方法,使用时间参数Δ将连续参数A和B转换为离散参数$\bar{A}$和$\bar{B}$,具体为

$$\begin{cases}\bar{A}=\exp(\Delta A)\\ \bar{B}=(\Delta A)^{-1}(\exp(\Delta A)-\boldsymbol{I})\cdot\Delta B\end{cases} \tag{3}$$

基于离散参数$\bar{A}$和$\bar{B}$,式(1)和式(2)的离散化版本可表示为

$$\begin{cases}\boldsymbol{h}(t)=\bar{A}\boldsymbol{h}(t-1)+\bar{B}\boldsymbol{x}(t)\\ \boldsymbol{y}(t)=C\boldsymbol{h}(t)\end{cases} \tag{4}$$

以输入序列为$\boldsymbol{x}(0)$、$\boldsymbol{x}(1)$和$\boldsymbol{x}(2)$为例,将$\boldsymbol{y}(2)$展开可表示为

$$\begin{aligned}&\boldsymbol{y}(2)=C\boldsymbol{h}(2)=\\&C(\bar{A}\boldsymbol{h}(1)+\bar{B}\boldsymbol{x}(2))=\\&C(\bar{A}(\bar{A}\boldsymbol{h}(0)+\bar{B}\boldsymbol{x}(1))+\bar{B}\boldsymbol{x}(2))=\\&C\cdot\bar{A}\cdot\bar{A}\cdot\bar{B}\boldsymbol{x}(0)+C\cdot\bar{A}\cdot\bar{B}\boldsymbol{x}(1)+C\cdot\bar{B}\boldsymbol{x}(2)\end{aligned} \tag{5}$$

因此,可通过构建一个全局卷积核$\bar{K}$,以卷积而非循环方式实现快速计算输出向量,具体为

$$\begin{cases}\bar{\boldsymbol{K}}=\left(C\bar{B},C\overline{AB},\cdots,C\bar{A}^{L-1}\bar{B}\right)\\ \boldsymbol{y}=\boldsymbol{x}*\bar{\boldsymbol{K}}\end{cases} \tag{6}$$

式中,L表示输入序列$\boldsymbol{x}$的长度,$\bar{\boldsymbol{K}}\in\mathbf{R}^{L}$表示结构化卷积核。采用卷积方式可以实现高速计算输出$\boldsymbol{y}$,以及并行训练。

算法1 Shot Mamba block

输入:特征向量序列$\boldsymbol{X}_{i-1}\in\mathbf{R}^{(L+1)\times D}$

输出:特征向量序列$\boldsymbol{X}_i\in\mathbf{R}^{(L+1)\times D}$

1) $\boldsymbol{X}'_{i-1}\in\mathbf{R}^{(L+1)\times D}\leftarrow\mathrm{Norm}(\boldsymbol{X}_{i-1})$

2) $\boldsymbol{z}\in\mathbf{R}^{(L+1)\times D_{\mathrm{inner}}}\leftarrow\mathrm{Linear}^{z}(\boldsymbol{X}'_{i-1})$

3) $\boldsymbol{x} \in \mathbf{R}^{(L+1)\times D_{\text{inner}}} \leftarrow$ SiLU(Conv1d(Linear$^{x}(\boldsymbol{X}'_{i-1})$))

4) $\boldsymbol{A} \in \mathbf{R}^{D_{\text{inner}}\times D_{\text{state}}} \leftarrow$Parameter

5) $\boldsymbol{B} \in \mathbf{R}^{(L+1)\times D_{\text{state}}} \leftarrow$Linear$^{B}(\boldsymbol{x})$

6) $\boldsymbol{C} \in \mathbf{R}^{(L+1)\times D_{\text{state}}} \leftarrow$Linear$^{C}(\boldsymbol{x})$

7) $\Delta \in \mathbf{R}^{(L+1)\times D_{\text{inner}}} \leftarrow$Linear$^{\Delta}(\boldsymbol{x})$

8) $\bar{\boldsymbol{A}} \in \mathbf{R}^{(L+1)\times D_{\text{inner}}\times D_{\text{state}}} \leftarrow \Delta \otimes \boldsymbol{A}$

9) $\bar{\boldsymbol{B}} \in \mathbf{R}^{(L+1)\times D_{\text{inner}}\times D_{\text{state}}} \leftarrow \Delta \otimes \boldsymbol{B}$

10) $h \in \mathbf{R}^{D_{\text{inner}}\times D_{\text{state}}} \leftarrow$Parameter

/* SSM */

11) for k in $(L+1)$ do

12) $\boldsymbol{h} \in \mathbf{R}^{D_{\text{inner}}\times D_{\text{state}}} \leftarrow \bar{A}_k \odot h + \bar{B}_k \odot x_k$

13) $\boldsymbol{y}_k \in \mathbf{R}^{D_{\text{inner}}} \leftarrow \boldsymbol{h}C_k$

14) end for

15) $\boldsymbol{y} \in \mathbf{R}^{(L+1)\times D_{\text{inner}}} \leftarrow$Concat$(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k, \cdots, \boldsymbol{y}_{L+1})$

16) $\boldsymbol{y}' \in \mathbf{R}^{(L+1)\times D_{\text{inner}}} \leftarrow y \odot$SiLU$(z)$

17) $\boldsymbol{X}_i \in \mathbf{R}^{(L+1)\times D} \leftarrow$Linear$^{X}(\boldsymbol{y}')+\boldsymbol{X}_{i-1}$

18) Return: $\boldsymbol{X}_i$

其中，$\otimes$表示外积运算，$\odot$表示逐元素乘。

2)镜头特征提取模块—Shot Mamba。由于原始Mamba模型(Gu和Dao，2024)针对一维序列数据设计，因此，本文提出的镜头特征提取模块示意图采用视觉Transformer(vision Transformer，ViT)(Dosovitskiy等，2021)的图像处理方式，如图3所示。

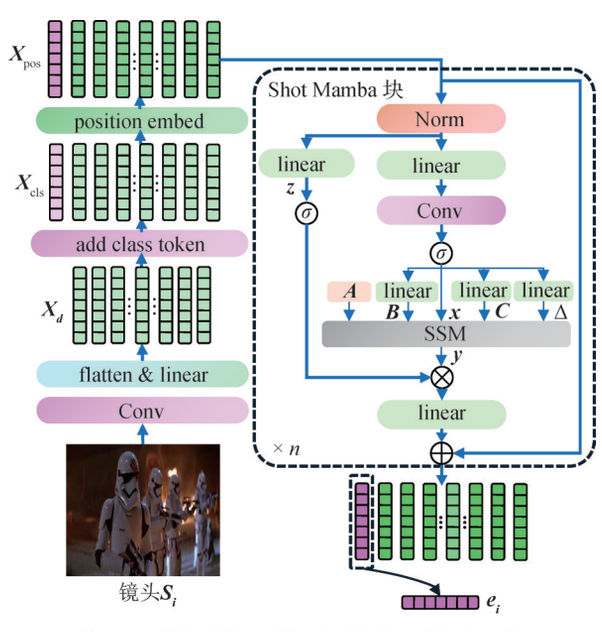

图3 Shot Mamba镜头特征提取模块

Fig. 3 Shot Mamba module for shot feature extraction

首先，将二维的镜头关键帧图像$\boldsymbol{S}_i \in \mathbf{R}^{h\times w\times c}$使用卷积(Conv)和拉伸(flatten)转换为L个切块向量$\boldsymbol{X}_p \in \mathbf{R}^{L\times(p\cdot p\cdot c)}$，其中$h$和$w$表示镜头图像的空间尺寸，$c$表示其波段数量，$p$表示切块大小。采用全连接层(linear)将切块向量转为特征维度为D的L个向量$\boldsymbol{X}_d \in \mathbf{R}^{L\times D}$。然后，将类别向量(class token)进行叠加，构建为新的向量序列$\boldsymbol{X}_{\text{cls}} \in \mathbf{R}^{(L+1)\times D}$，对其进行位置嵌入后可表示为特征向量序列$\boldsymbol{X}_{\text{pos}} \in \mathbf{R}^{(L+1)\times D}$。

随后，将$\boldsymbol{X}_{\text{pos}}$作为$n$个堆叠Shot Mamba模块的输入，设第$i-1$个Shot Mamba模块的输入为$\boldsymbol{X}_{i-1} \in \mathbf{R}^{(L+1)\times D}$，则Shot Mamba模块处理的伪代码可表示为算法1。通过n个Shot Mamba Block处理后，将类别向量位置的特征向量取出，作为该镜头的特征表示$\boldsymbol{e}$。

相比于S4模型(Gu等，2022)，Mamba(Gu和Dao，2024)将时不变系统中系数A、B和C转为随输入改变的参量，使得模型能够根据输入自适应调整对应的系数，具备输入感知能力。另外，Mamba采用并行扫描(parallel scan)方法将SSM的循环过程转化为并行操作，提高计算效率。

2.2.2 基于Scene Transformer的镜头关联模块

为了将每个独立的镜头编码特征进行关联，本文基于Transformer(Vaswani等，2017)构建具有自注意力机制的Scene Transformer模块，该模块能够有效地编码多组镜头特征，从而提升模型的上下文感知性能。

镜头$\boldsymbol{S}_{i-k}$至$\boldsymbol{S}_{i+k}$经过Shot Mamba的镜头特征提取后，可获得$2k+1$个镜头编码特征$\boldsymbol{e}_{i-k}$至$\boldsymbol{e}_{i+k}$，可表示为$\boldsymbol{E} \in \mathbf{R}^{(2k+1)\times D}$。

如图4所示，$\boldsymbol{E}$经过全连接层的特征变换改变特征维度后送入m个堆叠的多头自注意力模块(multi-head self-attention)，设第i个多头自注意力模块的输入为$\boldsymbol{E}_{i-1} \in \mathbf{R}^{(2k+1)\times D_i}$，有$h$个注意力头(head)，每个注意力的特征维度为$d_h$，则查询、键、值可以表示为$\boldsymbol{Q} \in \mathbf{R}^{(2k+1)\times h\times d_h}$，$\boldsymbol{K} \in \mathbf{R}^{(2k+1)\times h\times d_h}$和$\boldsymbol{V} \in \mathbf{R}^{(2k+1)\times h\times d_h}$，$Q$、$K$、$V$经过全连接层变换后分组进行缩放点积注意力(scaled dot-product attention)计算，具体为

$$Attention(Q_h, K_h, V_h) = softmax\left(\frac{\boldsymbol{Q}_h \boldsymbol{K}_h^{\mathrm{T}}}{\sqrt{d_h}}\right)\boldsymbol{V}_h \quad (7)$$

式中，$1/\sqrt{d_h}$为缩放因子(scale factor)，目的是让注意力得分(attention score)方差变为1，防止梯度消

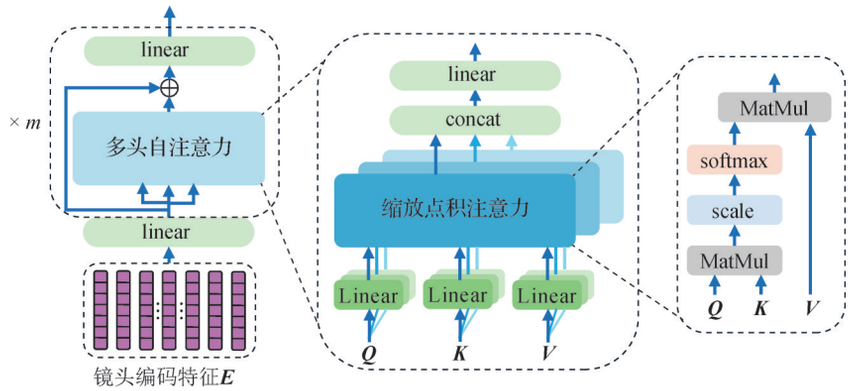

图4　Scene Transformer镜头关联模块示意图

Fig. 4　Scene Transformer module for shot feature association

失，Q_h、K_h和V_h为每个分组的缩放点积的输入。随后，将每个分组的输出叠加后使用全连接层进行整合，最后输出为镜头S_{i-k}至S_{i+k}的镜头关联特征$\boldsymbol{v}_{i-k}$至$\boldsymbol{v}_{i+k}$，可表示为$\boldsymbol{E}_i \in \mathbf{R}^{(2k+1)\times D_i}$。

通过使用自注意力模块，Scene Transformer模块可以对每个镜头之间的依赖性进行建模，通过后续的损失函数驱动，使得相同场景的镜头表征具有最大相似性，而不同场景的镜头表征具有最小相似性。

2.3　预训练及微调损失函数

2.3.1　场景伪边界生成

由于预训练阶段没有可参考的真实场景边界标签，无法直接使用交叉熵等有参考标签的损失函数驱动模型训练。而预训练阶段的目的是为了让模型学习场景的语义变化，使其具备初步的场景分割能力，因此，本文提出一种基于余弦相似度的场景伪边界生成方法，基于生成的伪边界构建后续的无监督损失函数，以此驱动预训练阶段的训练过程。

为了生成伪边界，核心思想是找到一个镜头S_k，使得其左侧的镜头编码特征序列与最左端的镜头编码特征具有最高的相似度，同时S_k右侧的镜头编码特征序列与最右端的镜头编码特征也具有最高的相似度。

针对$2k+1$个镜头序列编码特征，将序列两端的特征向量表示为$\boldsymbol{S} \in \mathbf{R}^{n_s \times d}$，其包含$\boldsymbol{e}_{i-k}$和$\boldsymbol{e}_{i+k}$两个镜头编码特征，$n_s$为2，$d$表示特征向量大小；序列中间的特征向量表示为$\boldsymbol{D} \in \mathbf{R}^{n_d \times d}$，其中$n_d$为$2k-1$，表示中间特征向量的数量。

首先，对特征向量$\boldsymbol{S}$和$\boldsymbol{D}$进行归一化处理，得到$\hat{\boldsymbol{S}}$和$\hat{\boldsymbol{D}}$，具体为

$$\hat{\boldsymbol{S}} = \frac{\boldsymbol{S}}{\|\boldsymbol{S}\|_2},\quad \hat{\boldsymbol{D}} = \frac{\boldsymbol{D}}{\|\boldsymbol{D}\|_2} \tag{8}$$

式中，$\|\boldsymbol{S}\|_2$和$\|\boldsymbol{D}\|_2$为对应特征向量的L2范数。

接着，计算序列中间每一个位于j的特征向量$\hat{\boldsymbol{D}}_j$与最左端特征向量$\hat{\boldsymbol{S}}_0$和最右端特征向量$\hat{\boldsymbol{S}}_1$之间的余弦相似度$sims_{l,j}$和$sims_{r,j}$，具体为

$$sims_{l,j} = \hat{\boldsymbol{D}}_j \cdot \hat{\boldsymbol{S}}_0^{\mathrm{T}} \tag{9}$$

$$sims_{r,j} = \hat{\boldsymbol{D}}_j \cdot \hat{\boldsymbol{S}}_1^{\mathrm{T}} \tag{10}$$

式中，“·”表示点积。

然后，对于每一个位于中间j位置的特征向量，计算左右相似度的平均值，具体为

$$SimLeft_j = \frac{1}{j}\sum_{i=1}^{j} sims_{l,i} \tag{11}$$

$$SimRight_j = \frac{1}{n_d - j}\sum_{i=j+1}^{n_d} sims_{r,i} \tag{12}$$

最后，计算每一个位置的总相似度，并找到可以最大化总相似度的分割点s，具体为

$$SplitIndex = \arg\max_s \left(SimLeft_s + SimRight_s\right) \tag{13}$$

2.3.2　镜头—场景匹配损失函数

基于场景伪边界生成的分割点s，构建镜头—场景对比损失函数（shot-scene matching loss，SSMLoss），用于最大化相同场景的相似性，最小化不同场景的相似性。对于分割点s的左右场景序列与左右两端的镜头特征分别做相似性约束，以左侧为例，最左端镜头特征向量$\boldsymbol{e}_{i-k}$应与分割点s左侧平均特征向量$\bar{\boldsymbol{e}}_l$有最大相似性，而与右侧的平均特征向量$\bar{\boldsymbol{e}}_r$有最小相似性。采用InfoNCE（information noise contrastive estimation）损失（van den Oord等，2019）约束整体相似性，具体为

$$\mathcal{L}_{\mathrm{SSM}}=\mathcal{L}_{\mathrm{NCE}}\left(f_{\mathrm{ssm}}\left(\boldsymbol{e}_{i-k}\right),f_{\mathrm{ssm}}\left(\bar{\boldsymbol{e}}_{l}\right)\right)+\mathcal{L}_{\mathrm{NCE}}\left(f_{\mathrm{ssm}}\left(\boldsymbol{e}_{i+k}\right),f_{\mathrm{ssm}}\left(\bar{\boldsymbol{e}}_{r}\right)\right) \tag{14}$$

$$\mathcal{L}_{\mathrm{NCE}}\left(f_{\mathrm{ssm}}\left(\boldsymbol{e}_{i-k}\right),f_{\mathrm{ssm}}\left(\bar{\boldsymbol{e}}_{l}\right)\right)=-\log\frac{\exp\left(f_{\mathrm{ssm}}\left(\boldsymbol{e}_{i-k}\right)\cdot f_{\mathrm{ssm}}\left(\bar{\boldsymbol{e}}_{l}\right)/\tau\right)}{\exp\left(f_{\mathrm{ssm}}\left(\boldsymbol{e}_{i-k}\right)\cdot f_{\mathrm{ssm}}\left(\bar{\boldsymbol{e}}_{r}\right)/\tau\right)} \tag{15}$$

$$\mathcal{L}_{\mathrm{NCE}}\left(f_{\mathrm{ssm}}\left(\boldsymbol{e}_{i+k}\right),f_{\mathrm{ssm}}\left(\bar{\boldsymbol{e}}_{r}\right)\right)=-\log\frac{\exp\left(f_{\mathrm{ssm}}\left(\boldsymbol{e}_{i+k}\right)\cdot f_{\mathrm{ssm}}\left(\bar{\boldsymbol{e}}_{r}\right)/\tau\right)}{\exp\left(f_{\mathrm{ssm}}\left(\boldsymbol{e}_{i+k}\right)\cdot f_{\mathrm{ssm}}\left(\bar{\boldsymbol{e}}_{l}\right)/\tau\right)} \tag{16}$$

式中，$f_{\mathrm{ssm}}()$表示针对SSMLoss损失的输入向量构建的投影函数，以全连接方式实现，exp()为自然指数函数，“·”表示点积，τ表示温度超参数，用于控制相似度的平滑度，$\bar{\boldsymbol{e}}_l$和$\bar{\boldsymbol{e}}_r$为分割点s左右序列的特征向量均值。

2.3.3　场景语义一致性损失函数

对于任意两个镜头，我们希望模型可以正确区分其是否在一个场景内，因此构建场景语义一致性损失（scene semantic consistency loss，SSC）。以位于中心位置的镜头关联特征$\boldsymbol{v}_i$为锚点，分别从分割点s的左侧和右侧各随机取出一个镜头关联特征$\boldsymbol{v}_l$和$\boldsymbol{v}_r$，如果分割点s位于中心位置左侧，则$\boldsymbol{v}_i$和$\boldsymbol{v}_r$同属一个场景，称$\boldsymbol{v}_r$为$\boldsymbol{v}_{\mathrm{pos}}$，$\boldsymbol{v}_l$为$\boldsymbol{v}_{\mathrm{neg}}$；反之，如果$s$位于中心位置右侧，则$\boldsymbol{v}_i$和$\boldsymbol{v}_l$同属一个场景，称$\boldsymbol{v}_l$为$\boldsymbol{v}_{\mathrm{pos}}$，$\boldsymbol{v}_r$为$\boldsymbol{v}_{\mathrm{neg}}$。使用InfoNCE损失则可以对场景语义一致性损失进行定义，具体为

$$\mathcal{L}_{\mathrm{SSC}}\left(\boldsymbol{v}_i,\boldsymbol{v}_{\mathrm{pos}},\boldsymbol{v}_{\mathrm{neg}}\right)=-\log\frac{\exp\left(f_{\mathrm{ssc}}\left(\boldsymbol{v}_i\right)\cdot f_{\mathrm{ssc}}\left(\boldsymbol{v}_{\mathrm{pos}}\right)/\tau\right)}{\exp\left(f_{\mathrm{ssc}}\left(\boldsymbol{v}_i\right)\cdot f_{\mathrm{ssc}}\left(\boldsymbol{v}_{\mathrm{neg}}\right)/\tau\right)} \tag{17}$$

式中，$f_{\mathrm{ssc}}()$为针对SSC损失的输入向量构建的投影函数，以全连接方式实现，exp()为自然指数函数，“·”表示点积，τ表示温度超参数。

2.3.4　边界损失函数

另外，分割点s的镜头特征与任意非分割点的镜头特征应该具有差异性，为了使模型可以更好地区分这种差异性，构建边界损失（boundary prediction loss，BP）。以$\boldsymbol{v}_s$表示分割点s处的镜头关联特征，$\boldsymbol{v}_{s*}$表示从分割点s以外位置随机采样的镜头关联特征，以二分类交叉熵可以构建如下损失，具体为

$$\mathcal{L}_{\mathrm{BP}}\left(\boldsymbol{v}_s,\boldsymbol{v}_{s*}\right)=-\log\left(f_{\mathrm{bp}}\left(\boldsymbol{v}_s\right)\right)-\log\left(1-f_{\mathrm{bp}}\left(\boldsymbol{v}_{s*}\right)\right) \tag{18}$$

式中，$f_{\mathrm{bp}}()$为针对BP损失的输入向量构建的投影函数，以全连接方式实现。

预训练阶段的整体损失函数可定义为

$$\mathcal{L}_{\mathrm{pretrain}}=\mathcal{L}_{\mathrm{SSM}}+\mathcal{L}_{\mathrm{SSC}}+\mathcal{L}_{\mathrm{BP}} \tag{19}$$

2.3.5　微调损失函数

在微调阶段，首先将预训练阶段的模型参数进行迁移，然后冻结Shot Mamba模块的参数，使其不再随微调阶段的迭代而更新，仅仅更新Scene Transformer模块的参数。根本原因在于微调阶段的训练样本相对较少，在预训练阶段，Shot Mamba模块已经可以较好地实现镜头特征编码，在有限的训练样本情况下，本文希望将模型聚焦在电影场景分割环节，因此整体的微调阶段模型结构如图2右侧所示。

对于微调阶段的有场景标签样本，场景边界（标签为1）仅占很少一部分，而大量的镜头都是非边界（标签为0），因此，存在极大的类别不平衡情况，如果仅仅使用交叉熵作为损失函数，模型则会偏向于非场景边界的类别，从而在边界预测上性能表现较差。本文采用Focal Loss（Lin等，2017）作为微调阶段的损失函数，以增强对场景边界这一类别的预测准确性，则微调阶段的损失函数可定义为

$$\begin{aligned}&\mathcal{L}_{\mathrm{finetune}}(z,y)=\begin{cases}-\alpha(1-p(y=1|z))^{\gamma}\log(p(y=1|z)) & y=1\\ -\alpha(1-p(y=0|z))^{\gamma}\log(p(y=0|z)) & y=0\end{cases}\\&z=f_{\mathrm{finetune}}(\boldsymbol{v})\\&p(y=1|z)=\frac{\mathrm{e}^{z_1}}{\mathrm{e}^{z_0}+\mathrm{e}^{z_1}}\\&p(y=0|z)=\frac{\mathrm{e}^{z_0}}{\mathrm{e}^{z_0}+\mathrm{e}^{z_1}}=1-p(y=1|z)\end{aligned} \tag{20}$$

式中，$f_{\mathrm{fine}}()$为针对微调损失的输入向量$\boldsymbol{v}$构建的投影函数，以全连接方式实现；z_0和z_1表示对应类别0和类别1的模型输出值；y为真实类别标签；$p(y=0|z)$和$p(y=1|z)$表示对应类别0和1的类别概率；α为平衡参数，本文设置为1，促使模型更加关注于边界标签；γ为聚焦参数，用于减少对非边界样本的关注度，增加对边界样本的注意力，本文设置为2。

3　实验结果与分析

本节通过消融实验以及与最新方法比较来验证所提方法在电影场景分割中的有效性。

3.1　实验环境及配置

本文所有实验均在Python3.10与PyTorch2.1.1环境中执行，操作系统为Ubuntu 18.04.6，CUDA环

境 11.8，硬件配置为 CPU：Intel（R） Xeon（R） Platinum 8352V CPU@2.1 GHz，GPU：4 × NVIDIA RTX A6000 48 GB。预训练阶段的数据集总迭代次数 epoch 为 10，batch size 为 120，优化器采用 Adamw（Loshchilov 和 Hutter，2019），使用线性预热（warmup）方式将学习率提升至 0.000 1，然后使用余弦退火（Loshchilov 和 Hutter，2017）方式进行学习率衰减。微调阶段的数据集总迭代次数 epoch 为 20，batchsize 为 240，优化器为 Adam（Kingma 和 Ba，2014），初始化学习率为 0.000 002 5，使用余弦退火方式进行学习率衰减。

由于不同电影的关键帧图像尺寸各不相同，因此，采用插值方法将输入图像尺寸缩放至 224 × 224 像素，使用卷积对输入图像进行切块的大小为 16 × 16 像素，邻域镜头的输入数量 k 为 8。镜头特征提取模块堆叠的 Shot Mamba 模块 n 为 24 层，镜头编码特征的特征维度 D 为 192。镜头特征关联模块的特征维度 D_l 为 768，堆叠的自注意力模块 m 为 3 层，多头注意力的头数 h 为 8，多头注意力的特征维度 d_h 为 96。

3.2 数据集与评价指标

3.2.1 数据集

MovieNet（Huang 等，2020）是一个用于电影理解的多模态数据集，其包含 1 100 部电影、160 万个镜头，其中 318 部电影具有场景边界标注，共 4.2 万个场景边界。在 318 部有场景边界的数据中，按 Chen 等人（2021）给出的训练集、验证集和测试集进行微调及验证，其中训练集电影数量 190 部，验证集电影数量 64 部，测试集 64 部。预训练时，使用 1 100 部电影的所有数据作为训练数据，其中 318 部电影数据同样作为预训练数据参与训练（无真值标签参与预训练过程）；微调阶段则在 318 部电影数据上进行训练、验证和测试。由于版权问题，数据集作者仅提供每个镜头的 3 个关键帧。

BBC（Baraldi，2015b）数据集收录了 BBC 制作的 11 集教育片《行星地球》，包含 4 900 个镜头和 670 个场景，由于该数据集可训练数据较少，且为非标准场景分割，采用 Islam 等人（2023）方法，仅使用在 MovieNet 数据集中的预训练模型对该数据的场景分割精度进行评估，而不在该数据中微调。

OVSD（Rotman 等，2016）数据集包含 10 000 个镜头和 300 个场景边界，与 BBC 数据集类似，该数据集的可训练数据较少，且为非标准场景分割，仅使用在 MovieNet 数据集中的预训练模型对该数据进行评估，而不在该数据集中微调。

3.2.2 评价指标

采用多种评价指标全面评估模型在电影场景分割任务中的性能。由于场景分割任务是一个二分类问题，主要关注边界与非边界的分割效果，使用的评价指标如下：

平均精度（average precision，AP）是衡量模型在不同阈值下对场景边界检测准确性的重要指标，它综合考虑了分割结果的精确率和召回率。

均值交并比（mean intersection over union，mIoU）用于评估模型分割结果的质量，通过计算预测场景边界与真实场景边界的交集和并集的比率衡量模型分割结果的效果。

接收者操作特征曲线下的面积（area under the receiver operating curve，AUC-ROC）用于衡量模型在不同阈值下区分边界和非边界区域的能力，反映了模型的整体分类性能。

F1-score 是精确率与召回率的调和平均值，能够有效评估模型在边界和非边界样本不均衡情况下的表现。

3.3 消融实验

为了验证不同损失函数、模型及超参数对场景分割精度的影响，本文以 MovieNet 数据集为例，分别对比了多种组合下的场景分割结果。

3.3.1 预训练损失函数对比

本文实验首先对比了采用不同预训练损失函数时的电影场景分割精度，如表 1 所示，SSMLoss 表示镜头—场景对比损失，SSC 表示场景语义一致性损失，BP 表示边界预测损失。单独对比 3 个预训练损失函数，SSMLoss 对场景边界的预测有着最大的影响，表明镜头—场景的对比损失对于引导模型无监督学习不同镜头之间的关系起到良好的作用，而 SSC 和 BP 损失的作用稍弱。另外，任意两个损失的组合对于模型的预训练都有正向反馈，相比而言，SSMLoss 和 SSC 的组合起到的作用最大，表明对比学习和一致性学习可以有效最大化相同场景的相似性和最小化不同场景的相似性。整体而言，当 3 个预训练损失组合在一起时可以达到最优效果。

3.3.2 微调阶段损失函数对比

本文采用 Focal Loss 解决场景分割中的类别不

表 1　不同预训练损失函数的电影场景分割精度对比

Table 1　Comparison of scene segmentation accuracy with different pre-training loss functions

SSMLoss	SSC	BP	AP	mIoU	AUC-ROC	F1-score
√	–	–	0.470 7	0.441 9	0.865 3	0.370 4
–	√	–	0.406 1	0.414 4	0.831 9	0.326 7
–	–	√	0.387 2	0.407 2	0.820 9	0.313 7
√	√	–	0.572 3	0.495 5	0.908 3	0.449 1
√	–	√	0.558 6	0.484 3	0.903 9	0.432 8
–	√	√	0.504 5	0.464 1	0.881 1	0.404 4
√	√	√	**0.617 9**	**0.573 8**	**0.920 8**	**0.565 0**

注：加粗字体表示各列最优结果。“√”和“–”表示使用和未使用对应损失函数。

平衡情况，而 BaSSL（boundary-aware self-supervised learning）（Mun 等，2022）方法采用的加权交叉熵同样有着平衡正负样本的功能，因此，本文测试了加权交叉熵与 Focal Loss 的分割精度，结果如表 2 所示。从表 2 中可以发现，相比于加权交叉熵，Focal Loss 的场景分割精度更高，可以有效提升模型在电影场景分割中类别不平衡场景下的性能。

表 2　微调阶段采用不同损失函数的精度对比

Table 2　Comparison of accuracy using different loss functions during the fine-tuning phase

损失函数	AP	mIoU	ACU-ROC	F1-score
加权交叉熵	0.607 9	0.507 4	0.920 4	0.464 6
Focal Loss	**0.617 9**	**0.573 8**	**0.920 8**	**0.565 0**

注：加粗字体表示各列最优结果。

3.3.3　模型结构对比

本文模型结构主要分为两部分：镜头特征提取模块和镜头特征关联模块，镜头特征提取模块的主要目的是提取每个镜头的特征编码，因此，对于该模型部分，对比了几个典型的图像特征提取模型：ViT-t、ViT-m（Dosovitskiy 等，2021）、SWIN-t（Liu 等，2021）及本文的 Shot Mamba。镜头特征关联模块的目的是关联镜头序列的特征编码，因此，对于该部分的序列建模，主要对比了 SSM（Gu 和 Dao，2024）和本文的 Scene Transformer。

表 3 展示了不同模型组合下的电影场景分割精度对比。可以看出，镜头关联模块采用 SSM 时，明显比使用 Scene Transformer 精度低，由于显存容量的限制，模型仅能处理镜头 i 及相邻的 $2k$ 个镜头，在镜头特征序列有限的情况下，相比于 Scene Transformer 的显式构建注意力模型，SSM 的特征关联能力弱于 Transformer。在镜头编码模块上，基于注意力机制的 ViT-t、ViT-m 和 SWIN Transformer 均弱于本文的 Shot Mamba，体现了 Shot Mamba 模型在图像编码方面的有效性。

3.3.4　超参数分析

本文模型的输入为镜头序列，其包含镜头 i 及前后相邻的 k 个镜头，共计 $2k+1$ 个镜头，实验对比了采用不同镜头数量时的场景分割精度，结果如表 4 所示。从表中可以发现，随着镜头数量的增加，模型的场景分割精度明显增加并趋于稳定，说明镜头序列的数量对模型确定场景边界有着积极的作用。

表 3　不同模型结构的场景分割精度对比

Table 3　Comparison of scene segmentation accuracy with different model structures

镜头编码	镜头关联	AP	mIoU	ACU-ROC
ViT-t	SSM	0.436 2	0.423 9	0.846 6
ViT-t	Transformer	0.572 3	0.495 5	0.908 3
ViT-m	SSM	0.436 6	0.424 5	0.848 6
ViT-m	Transformer	0.583 8	0.503 5	0.911 5
SWIN-t	SSM	0.490 6	0.451 0	0.874 5
SWIN-t	Transformer	0.597 4	0.507 4	0.914 8
Shot Mamba	SSM	0.504 5	0.464 1	0.881 1
Shot Mamba	Transformer	**0.617 9**	**0.573 8**	**0.920 8**

注：加粗字体表示各列最优结果。

表 4　镜头数量对分割精度的影响

Table 4　The impact of the number of shot on segmentation accuracy

k	AP	mIoU	ACU-ROC	F1-score
2	0.560 8	0.482 5	0.903 6	0.430 8
4	0.578 9	0.499 8	0.910 4	0.455 5
6	0.596 9	0.508 3	0.914 7	0.467 3
8	**0.617 9**	**0.573 8**	**0.920 8**	**0.565 0**
10	0.614 8	0.571 6	0.920 6	0.553 7

注：加粗字体表示各列最优结果。

为了探索式（15）和式（16）中温度系数 τ 对场景分割精度的影响，在预训练阶段分别设置 $\tau \in$

[0.05, 0.1, 0.2, 0.4, 1.0]，并保存数据集迭代4次的预训练模型，固定微调阶段的训练超参数，对比不同温度系数的敏感性，结果如表5所示。从表中可以发现，τ取值为0.1达到最优平衡，过小的温度系数会促使模型过度关注负样本，过大的温度系数导致相似度差异过度平滑。该结果与SimCLR（simple framework for contrastive learning of visual representations）（Chen等，2020）在ImageNet上的结论一致，验证了该参数的普适性。

表5　温度系数对分割精度的影响

Table 5　The impact of temperature coefficient on segmentation accuracy

τ	AP	mIoU	ACU-ROC	F1-score
0.05	0.572 5	0.554 9	0.907 4	0.536 6
0.1	**0.602 3**	**0.556 5**	**0.919 4**	**0.545 5**
0.2	0.561 1	0.545 5	0.900 5	0.516 7
0.4	0.555 8	0.543 7	0.897 8	0.511 6
1	0.536 4	0.530 1	0.890 1	0.492 1

注：加粗字体表示各列最优结果。

微调阶段采用Focal Loss作为损失函数，为了进一步探索Focal Loss中的平衡参数α和聚焦参数γ对场景分割精度的影响，分别设计了两组实验以覆盖更广泛的参数范围：

1）固定$\gamma = 1$，调整平衡参数α的取值（0.001，0.01，0.1，1，10，100），并评估其对分割精度的影响，实验结果如图5所示。从图5中可以发现，平衡参数α的极端情况（过大或过小）会导致分割精度大幅下降，较低的α取值会导致模型对分割边界的敏感性下降，而过大的α取值会导致模型过度关注分割边界，忽略非边界区域的语义一致性。本文采用$\alpha = 1$，在避免人为引入偏差的情况下，保持正负样本损失权重的自然平衡。

2）固定$\alpha = 1$，调整聚焦参数γ的取值（0.01，0.1，1，10，100），并评估其对分割精度的影响，实验结果如图6所示。从图6中可以发现，聚焦参数γ在极端低值（$\gamma = 0.01$）的情况下仍然能保持良好的精度，而在γ大于10的情况下，分割精度出现断崖式下滑，较大的聚焦参数使得模型仅关注极难样本，极端压制简单样本的梯度，无法有效学习更具普适性的场景边界。超参数γ最优取值在1～10之间，由于在1～10之间的影响较为平缓，本文采用Focal Loss（Lin等，2017）结果中的$\gamma = 2$，温和调整样本权重，对模糊边界保持一定的敏感性，同时避免过度关注极难样本。

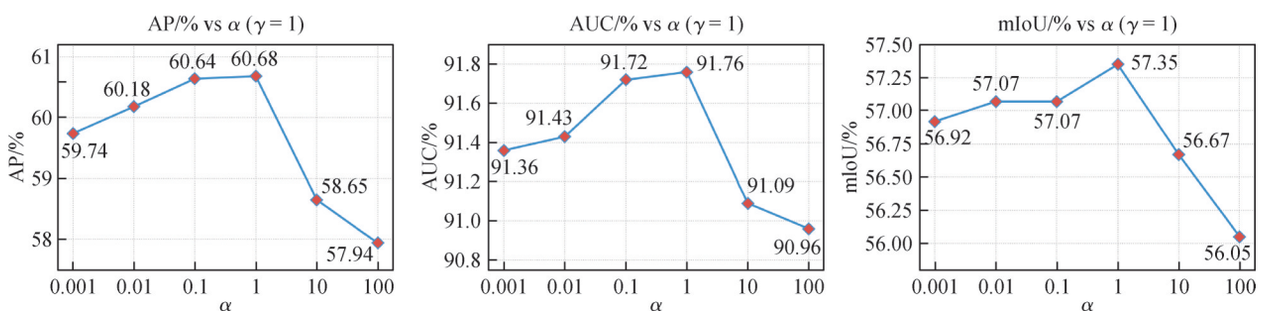

图5　固定$\gamma = 1$，平衡参数α对场景分割精度敏感性分析

Fig. 5　Sensitivity analysis of the parameter α on scene segmentation accuracy with fixed $\gamma = 1$

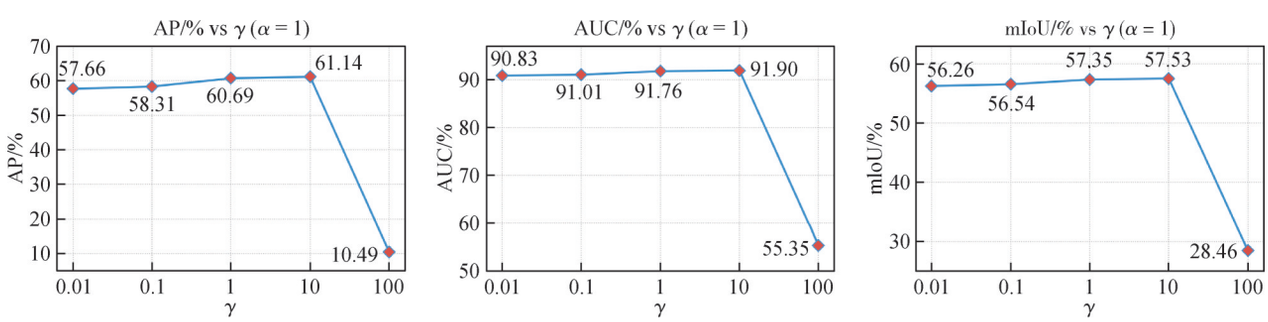

图6　固定$\alpha = 1$，聚焦参数γ对场景分割精度敏感性分析

Fig. 6　Sensitivity analysis of the parameter γ on scene segmentation accuracy with fixed $\alpha = 1$

3.4 与其他电影场景分割方法比较

为了验证本文算法在电影场景分割任务中的有效性，本节比较了不同场景分割方法在3个数据集中的表现。对比方法包括非监督类方法和预训练类方法，其中非监督类方法包括：GraphCut（Rasheed和Shah，2005）、SCSA（shot clustering and sequence alignment）（Chasanis等，2009）、DP（dynamic programming）（Han和Wu，2011）、Story Graph（Tapaswi等，2014）、Grouping（Rotman等，2016）和LGSS（local-to-global scene segmentation）（Rao等，2020）；预训练类方法包括：ShotCoL（shot contrastive learning）（Chen等，2021）、SCRL（scene consistency representation learning）（Wu等，2022）、BaSSL（Mun等，2022）、CAT（context-aware Transformer）（Yang等，2023）、TranS4mer（Islam等，2023）和HASSNet。

不同对比方法在MovieNet数据集上的分割结果如表6所示，由于所有对比方法的测试集一致，因此，对比方法的场景分割结果均来自于相关文献（Mun等，2022；Yang等，2023；Islam等，2023）。首先，相比于非监督方法，基于预训练及微调方法的场景分割精度明显高于非监督方法，非监督方法的AP普遍低于0.4，而采用预训练和微调的方法AP均高于0.5，因此，通过无标签的预训练及有标签的微调可以更好地使模型学习场景边界；其次，本文提出的HASSNet在评价指标上均得到了有效提升，相比于同样使用SSM结构的场景分割方法TranS4mer，AP、mIoU、AUC-ROC、F1-score分别提升1.66%、10.54%、0.21%、16.83%。实验结果表明，本文提出的混合架构电影场景分割方法在电影场景分割任务的性能上优于当前场景分割方法。

表6 不同方法在MoiveNet数据集上的场景分割精度对比
Table 6 Comparison of scene segmentation accuracy of different methods on the MoiveNet dataset

类别	方法	AP	mIoU	AUC-ROC	F1-score
非监督类	GraphCut	0.141 0	0.297 0	–	–
	SCSA	0.147 0	0.305 0	–	–
	DP	0.155 0	0.320 0	–	–
	Story Graph	0.251 0	0.357 0	–	–
	Grouping	0.336 0	0.372 0	–	–
	BaSSL (unsupervised)	0.315 5	0.393 6	0.716 7	0.325 5
	LGSS	0.449 0	0.465 0	0.877 3	0.385 2
预训练类	ShotCoL	0.528 9	–	0.491 7	–
	SCRL	0.548 2	–	0.514 3	–
	BaSSL	0.574 0	0.506 9	0.905 4	0.470 2
	CAT	0.595 5	<u>0.536 7</u>	0.918 1	<u>0.519 3</u>
	TranS4mer	<u>0.607 8</u>	0.519 1	<u>0.918 9</u>	0.483 6
	HASSNet	**0.617 9**	**0.573 8**	**0.920 8**	**0.565 0**

注：加粗、下划线分别表示各列最优、次优结果。“–”表示缺少相应实验结果。

基于Islam等人（2023）提供的BBC和OVSD数据集的场景分割精度，本文也在这两个数据集中使用预训练模型进行测试，表7的结果显示本文方法相较于其他方法均取得了一定程度的性能提升。

表7 BBC和OVSD数据集上的场景分割精度
Table 7 Scene segmentation accuracy on the BBC and OVSD datasets

方法	AP	
	BBC	OVSD
BaSSL	0.399 8	0.286 8
Transformer	0.418 6	0.331 2
TimeSformer（Bertasius等，2021）	0.422 3	0.338 7
SSM	0.425 6	0.342 1
TranS4mer	0.436 4	0.360 6
HASSNet	**0.442 3**	**0.368 2**

注：加粗字体表示各列最优结果。

另外，本文使用BaSSL（Mun等，2022）和TranS4mer（Islam等，2023）提供的预训练模型在MovieNet有场景标签数据集中进行微调测试，在《盗梦空间》电影中的场景分割部分可视化结果如图7所示。在该部分电影镜头中有连续的场景变化，受限于镜头特征表征和关联模块的有限性能，BaSSL和TranS4mer均无法准确预测场景边界，而本文提出方法则可以较好地对该场景边界进行预测。

4 结 论

为了提升电影场景分割任务中镜头特征提取和特征关联的有效性，增强镜头序列的上下文感知能

力，本文提出一种混合架构电影场景分割方法（HASSNet）。首先，采用预训练结合微调策略，通过在大量无场景标签的电影数据中无监督预训练，使得模型学习有效的镜头特征表示和关联特性，随后将预训练模型迁移至有场景标签数据中进行微调训练，进一步提升模型在电影场景分割任务中的性能；其次，分别设计了Shot Mamba镜头特征提取模块和Scene Transformer镜头特征关联模块，Shot Mamba模块将镜头图像进行分块建模，提取出最有效的镜头特征表示，Scene Transformer模块则将不同镜头特征进行注意力关联建模，生成具备镜头上下文感知能力的特征表达；最后，采用3种无监督损失函数对无场景标签数据进行预训练，并通过Focal Loss损失函数在有场景标签数据中进行微调，改善由于类别不平衡导致的场景分割精度不足问题。在3个数据集的实验以及消融实验表明，本文提出的混合架构电影场景分割方法可以增强模型对电影镜头的上下文感知能力，并可以有效提升场景分割精度。

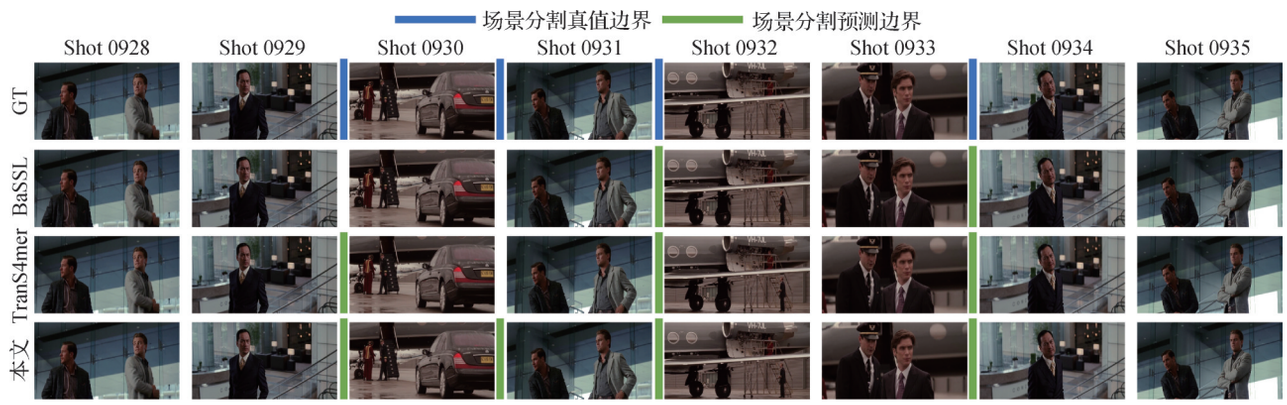

图7　场景边界分割可视化结果比较

Fig. 7　Comparison of scene boundary segmentation visualization results

参考文献（References）

Apostolidis E, Adamantidou E, Metsai A I, Mezaris V and Patras I. 2021. Video summarization using deep neural networks: a survey. Proceedings of the IEEE, 109(11): 1838-1863 [DOI: 10.1109/JPROC.2021.3117472]

Baraldi L, Grana C and Cucchiara R. 2015a. A deep siamese network for scene detection in broadcast videos//Proceedings of the 23rd ACM International Conference on Multimedia. New York, USA: Association for Computing Machinery: 1199-1202 [DOI: 10.1145/2733373.2806316]

Baraldi L, Grana C and Cucchiara R. 2015b. Shot and scene detection via hierarchical clustering for re-using broadcast video//Proceedings of the 16th International Conference on Computer Analysis of Images and Patterns. Valletta, Malta: Springer International Publishing: 801-811 [DOI: 10.1007/978-3-319-23192-1_67]

Bertasius G, Wang H and Torresani L. 2021. Is space-time attention all you need for video understanding?//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: ICML: 813-824

Chasanis V T, Likas A C and Galatsanos N P. 2009. Scene detection in videos using shot clustering and sequence alignment. IEEE Transactions on Multimedia, 11(1): 89-100 [DOI: 10.1109/TMM.2008.2008924]

Chen S X, Nie X H, Fan D, Zhang D Q, Bhat V and Hamid R. 2021. Shot contrastive self-supervised learning for scene boundary detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 9791-9800 [DOI: 10.1109/CVPR46437.2021.00967]

Chen T, Kornblith S, Norouzi M and Hinton G. 2020. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: PMLR: 1597-1607 [DOI: 10.5555/3524938.3525087]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Gu A and Dao T. 2024. Mamba: linear-time sequence modeling with selective state spaces [EB/OL]. [2024-09-29]. https://arxiv.org/pdf/2312.00752.pdf

Gu A, Goel K and Ré C. 2022. Efficiently modeling long sequences with structured state spaces//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: OpenReview.net

Han B and Wu W G. 2011. Video scene segmentation using a novel boundary evaluation criterion and dynamic programming//Proceedings of 2011 IEEE International Conference on Multimedia and

Expo. Barcelona, Spain: IEEE: 1-6 [DOI: 10.1109/ICME.2011.6012001]

Huang Q Q, Xiong Y, Rao A Y, Wang J Z and Lin D H. 2020. MovieNet: a holistic dataset for movie understanding//Proceedings of the 16th European Conference on Computer Vision 2020. Glasgow, UK: Springer International Publishing: 709-727 [DOI: 10.1007/978-3-030-58548-8_41]

Islam M M, Hasan M, Athrey K S, Braskich T and Bertasius G. 2023. Efficient movie scene detection using state-space transformers//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18749-18758 [DOI: 10.1109/CVPR52729.2023.01798]

Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R and Li F F. 2014. Large-scale video classification with convolutional neural networks//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1725-1732 [DOI: 10.1109/CVPR.2014.223]

Kingma D P and Ba J. 2014. Adam: a method for stochastic optimization [EB/OL]. [2017-01-30]. https://arxiv.org/pdf/1412.6980.pdf

Liang J F, Li T, Yang J Q, Li Y N, Fang Z W and Yang F. 2023. Video anomaly detection by fusing self-attention and autoencoder. Journal of Image and Graphics, 28(4): 1029-1040 (梁家菲, 李婷, 杨佳琪, 李亚楠, 方智文, 杨丰. 2023. 融合自注意力和自编码器的视频异常检测. 中国图象图形学报, 28(4): 1029-1040) [DOI: 10.11834/jig.211147]

Lin T Y, Goyal P, Girshick R, He K M and Dollar P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Liu Y, Tian Y J, Zhao Y Z, Yu H T, Xie L X, Wang Y W, et al. 2024. VMamba: visual state space model//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 103031-103063 [DOI: 10.5555/3737916.3741189]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Loshchilov I and Hutter F. 2017. SGDR: stochastic gradient descent with warm restarts//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization//Proceedings of the 7th International Conference on Learning Representations. New Orleans: OpenReview.net

Mun J, Shin M, Han G, Lee S, Ha S, Lee J, et al. 2022. BaSSL: boundary-aware self-supervised learning for video scene segmentation//Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer: 485-501 [DOI: 10.1007/978-3-031-26316-3_29]

Rao A Y, Xu L N, Xiong Y, Xu G D, Huang Q Q, Zhou B L, et al. 2020. A local-to-global approach to multi-modal movie scene segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE 10143-10152 [DOI: 10.1109/CVPR42600.2020.01016]

Rasheed Z and Shah M. 2003. Scene detection in Hollywood movies and TV shows//Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, USA: IEEE: II-343 [DOI: 10.1109/CVPR.2003.1211489]

Rasheed Z and Shah M. 2005. Detection and representation of scenes in videos. IEEE Transactions on Multimedia, 7(6): 1097-1105 [DOI: 10.1109/TMM.2005.858392]

Roh B, Shin W, Kim I and Kim S. 2021. Spatially consistent representation learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1144-1153 [DOI: 10.1109/CVPR46437.2021.00120]

Rotman D, Porat D and Ashour G. 2016. Robust and efficient video scene detection using optimal sequential grouping//Proceedings of 2016 IEEE International Symposium on Multimedia (ISM). San Jose, USA: IEEE: 275-280 [DOI: 10.1109/ISM.2016.0061]

Rotman D, Porat D and Ashour G. 2017. Optimal sequential grouping for robust video scene detection using multiple modalities. International Journal of Semantic Computing, 2017, 11(2): 193-208 [DOI: 10.1142/S1793351X17400086]

Shou M Z, Lei S W, Wang W Y, Ghadiyaram D and Feiszli M. 2021. Generic event boundary detection: a benchmark for event segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 8055-8064 [DOI: 10.1109/ICCV48922.2021.00797]

Sidiropoulos P, Mezaris V, Kompatsiaris I, Meinedo H, Bugalho M and Trancoso I. 2011. Temporal video segmentation to scenes using high-level audiovisual features. IEEE Transactions on Circuits and Systems for Video Technology, 21(8): 1163-1177 [DOI: 10.1109/TCSVT.2011.2138830]

Sun Z H, Ke Q H, Rahmani H, Bennamoun M, Wang G and Liu J. 2023. Human action recognition from various data modalities: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 3200-3225 [DOI: 10.1109/TPAMI.2022.3183112]

Tan J W, Wang H X and Yuan J S. 2023. Characters link shots: character attention network for movie scene segmentation. ACM Transactions on Multimedia Computing, Communications and Applications, 20(4): #94 [DOI: 10.1145/3630257]

Tang S T, Feng L T, Kuang Z H, Chen Y M and Zhang W. 2019. Fast video shot transition localization with deep structured models//Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer International Publishing: 577-592 [DOI: 10.1007/978-3-030-20887-5_36]

Tapaswi M, Bäuml M and Stiefelhagen R. 2014. StoryGraphs: visualiz-

ing character interactions as a timeline//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 827-834 [DOI: 10.1109/CVPR.2014.111]

van den Oord A, Li Y Z and Vinyals O. 2019. Representation learning with contrastive predictive coding [EB/OL]. [2024-09-29]. https://arxiv.org/pdf/1807.03748.pdf

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

Wei X, Shi Z X, Zhang T Z, Yu X Y and Xiao L. 2023. Multimodal high-order relation transformer for scene boundary detection//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22024-22033 [DOI: 10.1109/ICCV51070.2023.02018]

Wu H Q, Chen K Y, Luo Y N, Qiao R Z, Ren B, Liu H Z, et al. 2022. Scene consistency representation learning for video scene segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 14001-14010 [DOI: 10.1109/CVPR52688.2022.01363]

Yang Y, Huang Y R, Guo W L, Xu B H and Xia D Y. 2023. Towards global video scene segmentation with context-aware transformer//Proceedings of 2023 AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 3206-3213 [DOI: 10.1609/aaai.v37i3.25426]

Yao R, Lin G S, Xia S X, Zhao J Q and Zhou Y. 2020. Video object segmentation and tracking: a survey. ACM Transactions on Intelligent Systems and Technology (TIST), 11(4): #36 [DOI: 10.1145/3391743]

Zeng R H, Li J L, Zhuo Y S, Duan H H, Chen Q and Hu X P. 2025. Iterative optimization for video retrieval data using large language model guidance. Journal of Image and Graphics, 30(5): 1257-1271 (曾润浩, 李嘉梁, 卓奕深, 段海涵, 陈奇, 胡希平. 2025. 大语言模型引导的视频检索数据迭代优化. 中国图象图形学报, 30(5): 1257-1271) [DOI: 10.11834/jig.240545]

Zhu L H, Liao B C, Zhang Q, Wang X L, Liu W Y and Wang X G. 2024. Vision Mamba: efficient visual representation learning with bidirectional state space model//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: OpenReview.net: #2584 [DOI: 10.5555/3692070.3694654]

Zhu W T, Huang Y F, Xie X F, Liu W X, Deng J C, Zhang D B, et al. 2023. AutoShot: a short video dataset and state-of-the-art shot boundary detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada: IEEE: 2238-2247 [DOI: 10.1109/CVPRW59228.2023.00218]

作者简介

赵晓蕾,女,讲师,主要研究方向为电影场景理解。
E-mail:zhaoxiaolei@lcudcc.edu.cn
郑珂,通信作者,男,讲师,主要研究方向为计算机视觉和遥感图像处理。E-mail:zhengke@lcu.edu.cn
赵鑫,男,讲师,主要研究方向为电影场景理解。
E-mail:cmxzhaoxin@lcudcc.edu.cn
于浩洋,男,副教授,主要研究方向为计算机视觉和遥感图像处理。E-mail:yuhy@dlmu.edu.cn
孙旭,男,副研究员,主要研究方向为计算机视觉和遥感图像处理。E-mail:sunxu@aircas.ac.cn