激光与光电子学进展 先进成像

基于深度动态语义关联的短视频事件检测

井佩光*，宋晓艺，苏育挺
天津大学电气自动化与信息工程学院，天津 300072

摘要 现如今，短视频事件检测展现出广阔的应用前景。现有的事件检测研究普遍缺乏对关键帧重要性程度的考虑，且多是针对事件的显性语义进行学习，忽略了潜在语义及其相关性的学习在短视频事件检测中的作用。针对上述问题，提出了一种基于深度动态语义关联的短视频事件检测方法。首先，设计了帧重要性评估模块来获得具有区分度的帧重要性分数，其内嵌的变分自编码器和生成对抗网络联合结构可以最大程度地强化帧重要性信息；其次，设计了帧间自注意力增强模块，进一步协同帧间的重要性分数与其特征内在关联性的学习；最后，设计了动态图卷积下的隐藏属性关联学习模块来学习复杂事件的隐藏属性及事件之间的关联性，最终获得具有潜在语义信息感知的短视频检测系统并将其用于最终的短视频事件检测。在公开数据集和新构建数据集上进行了实验，实验结果表明了所提方法的有效性。

关键词 短视频；语义关联；特征表示；图卷积
中图分类号 TP183 **文献标志码** A **DOI**：10.3788/LOP230994

Micro-Video Event Detection Based on Deep Dynamic Semantic Correlation

Jing Peiguang*, Song Xiaoyi, Su Yuting
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract Nowadays, micro-video event detection exhibits great potential for various applications. As for event detection, previous studies usually ignore the importance of keyframes and mostly focus on the exploration of explicit attributes of events. They neglect the exploration of latent semantic representations and their relationships. Aiming at the above problems, a deep dynamic semantic correlation method is proposed for micro-video event detection. First, the frame importance evaluation module is designed to obtain more distinguishing scores of keyframes, in which the joint structure of variational autoencoder and generative adversarial network can strengthen the importance of information to the greatest extent. Then, the intrinsic correlations between keyframes and the corresponding features are cooperated through a keyframe-guided self-attention mechanism. Finally, the hidden event attribute correlation module based on dynamic graph convolution is designed to learn latent semantics and the corresponding correlation patterns of events. The obtained latent semantic-aware representations are used for final micro-video event detection. Experiments performed on the public datasets and the newly constructed micro-video event detection dataset demonstrate the effectiveness of the proposed method.

Key words micro-video; semantic correlation; feature representation; graph convolution

1 引 言

近年来，短视频作为一种新的多媒体信息传播载体出现在人们的视野之中。各种社交平台均嵌入了观看和分享短视频的功能，短视频也因此迅速成为了人们建立社交链接的重要途径。海量的短视频数据和更优质的用户体验需求使得面向短视频内容的理解与分析逐渐成为研究热点。

目前对短视频的研究方向主要有：短视频流行度预测[1-3]、短视频多标签分类[4-5]及个性化推荐[6-7]等。Su等[4]提出了一种低秩正则化深度协同矩阵分解模型来进行短视频分类；Chen等[7]利用用户的兴趣程度来进行短视频推荐。在短视频事件检测方向上：Redi等[8]提出了一种基于特征映射的自动检测短视频的方法；Zhang等[9]提出了一种基于低秩约束的自适应学习机制来检测短视频事件。事件检测中的“事件”是指发生在一个特定时间和地点的可被直接观察到的复杂活动[10]，如“婚礼”“生日会”“游行”等。不同于一般语义

收稿日期：2023-03-30；**修回日期**：2023-04-30；**录用日期**：2023-06-01；**网络首发日期**：2023-06-11
基金项目：国家自然科学基金(61802277)、天津市自然科学基金(20JCQNJC01210)
通信作者：*pgjing@tju.edu.cn

信息，事件通常涉及到更为复杂的目标间的交互，因而常伴随着高级抽象语义信息。相比于只关注特定目标特性的动作识别等视频分析任务，事件检测通过综合考量视频中的各个对象特性进而推断出视频所记录的事件，比如根据视频中的“人群”“鲜花”“蜡烛”“婚纱”等目标特性来推断所描述的事件是“婚礼”。现有的事件检测主要集中在面向传统视频的交通事件检测[11]和异常事件检测[12-13]等，很少有研究致力于解决短视频事件检测问题，其原因有以下几点：1）短视频事件相比于有明确特性的异常事件和交通事件来说具有更为复杂的特性，通常由多个对象、多种动作等共同构成[10]；2）由于短视频的拍摄环境复杂及剪辑多样等特点，其有效信息的学习往往更为复杂；3）主流数据集的缺乏限制了短视频事件检测的发展。

在短视频事件检测等任务中，短视频的视觉特征通常是一种具有稳定性和普适性的信息源并被广泛使用。当下短视频视觉特征的获取主要依赖所选取的关键帧，通常先对短视频帧序列进行等间隔采样得到关键帧，之后对关键帧进行特征提取。对于短视频关键帧序列来说，其包含的每一帧的重要程度都是不同的，信息相对丰富的帧重要程度更高。如何对这些相对更为重要的视频帧分配权重[14-15]逐渐成为研究的热门方向。Zhang 等[14]提出了一种将长短时记忆（LSTM）网络与确定点过程相结合的模型，该模型使用确定点过程增强的 LSTM 网络来对视频帧之间的时间依赖性进行建模，并利用多层感知机制来对视频片段的重要性分数进行预测。目前针对短视频事件检测的研究大多缺乏对帧重要性程度的考虑，为此本文设计了帧重要性评估（FIE）模块以获得具有区分度的帧重要性分数。同时，由于目前注意力机制在不同任务上均取得了出色的效果[16-18]，因此设计了帧间自注意力增强（ISAE）模块来协同帧间与特征空间的内在关联性，进而获取更具有注意力特性的特征表示。

短视频事件中往往包含着丰富的潜在语义信息，而图卷积网络（GCN）因其处理复杂关系的强大能力，目前在多标签图像分类、动作识别等方面被广泛应用。针对不同的任务通常会构建不同的图结构[19-23]，例如，Chen 等[20]将标签作为节点，标签间的相关性作为边，构建了用于多标签图像分类的 GCN。在短视频事件检测任务中，将复杂事件的隐藏属性视为节点，属性之间的关联度视为边，构建针对复杂事件的图结构。相比通常建立在数据集标签先验概率上的静态图而言，针对每个样本分别构建图更具有合理性。基于此，设计了使用动态 GCN 学习短视频事件的潜在语义及其关联性的嵌入模式。

面向短视频事件检测任务，由于已有的工作大多缺乏对关键帧重要性程度的考虑，同时多是针对事件显性属性进行学习，缺乏对复杂事件下所隐藏的潜在语义的学习及语义间相关性的考虑，因此本文提出了一种基于深度动态语义关联的短视频事件检测方法，具体如下：1）设计了 FIE 模块，其内嵌的变分自编码器（VAE）和生成对抗网络（GAN）联合结构可以最大程度地强化重要性信息，获得具有区分度的帧重要性分数；2）设计了 ISAE 模块，通过协同帧间与特征空间的内在相关性来获取更具有注意力特性的特征表示；3）设计了隐藏属性关联学习（HACL）模块，通过引入隐藏属性激活单元及动态 GCN 来学习复杂事件的隐藏属性及其关联性，将最终得到的具有潜在语义信息感知的特征表示用于最终的短视频事件检测任务；4）构建了新的短视频事件检测数据集来解决该方向缺少主流数据集的问题。最后，在该数据集和公开数据集上进行了实验，证明了所提方法的有效性。

2　算法模型

所提方法整体框架如图 1 所示，根据作用分为 FIE、ISAE 和 HACL 3 大模块。其中：FIE 模块用来估算帧重要性分数；ISAE 模块则协同了帧间与特征空间

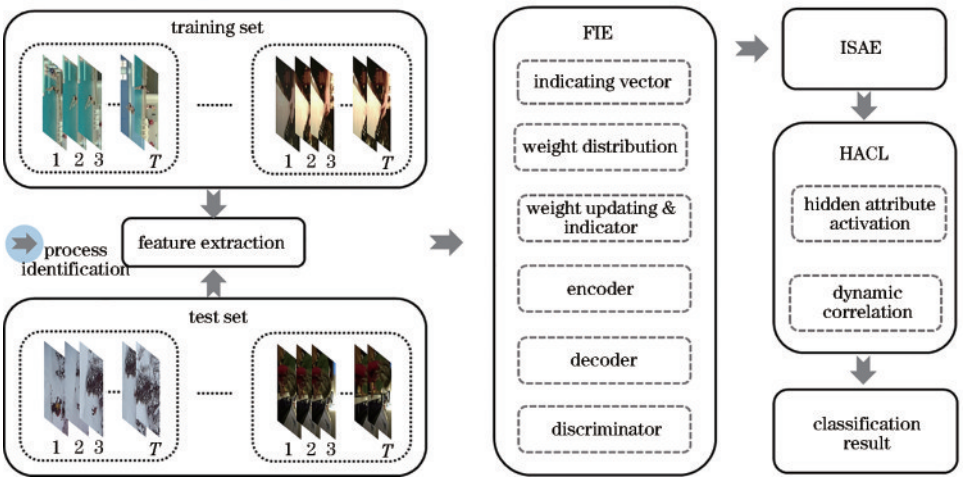

图 1　基于深度动态语义关联的事件检测方法

Fig. 1　Event detection method based on deep dynamic semantic correlation

的内在关联性，从而获取更具有注意力特性的特征表示；HACL模块由事件属性激活映射单元和动态关联性单元组成，主要进行复杂事件的隐藏属性学习及其关联性的计算；T为关键帧个数。

2.1　FIE

如图2所示，FIE模块主要由指示向量计算单元（indicating vector generator）、权重更新单元（weight updating）、指示器（indicator）、编码器、解码器、判别器以及权重分配单元（weight distribution）组成。在该模块中，指示向量计算单元用来产生初始短视频关键帧重要性权重；权重更新单元和指示器协同工作，用来更新重要性权重；编码器和解码器共同构成了一个VAE，来挖掘样本潜在的重要性信息。同时，解码器又充当了生成器的角色，和判别器共同构成了GAN，从而使得由VAE训练得到的潜在特征表示不会损失太多原始特征所含的信息。最后，经判别器学习得到的反馈值将作用于权重更新单元和指示器部分，用于指导重要性权重的更新。

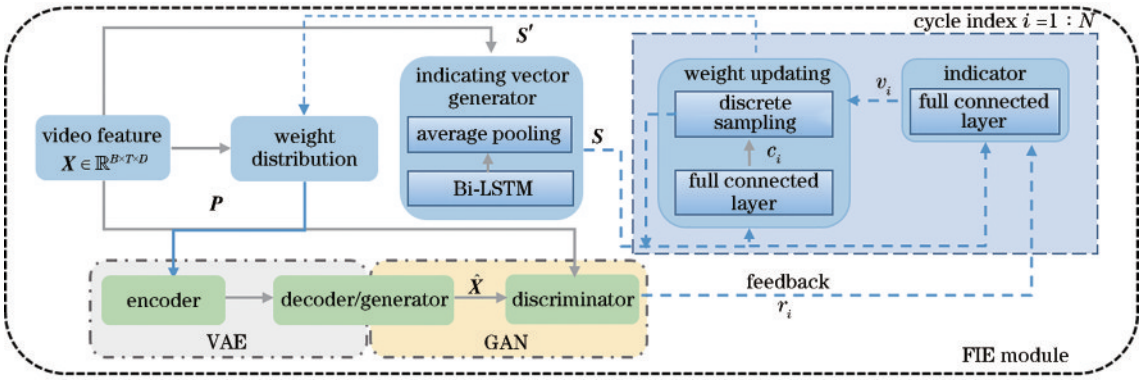

图2　FIE模块

Fig. 2　FIE module

不失一般性，假设经下采样后的短视频关键帧数量为T，提取到的原始视觉特征为$\boldsymbol{X}\in\mathbb{R}^{B\times T\times D}$，其中：$B$为短视频样本数；$D$为特征维度数。该特征经过由双向长短时记忆（Bi-LSTM）网络和平均池化（average pooling）操作组成的指示向量计算单元。Bi-LSTM用来捕获帧序列在向前和向后方向上的时间依赖性，从而获取短视频关键帧的重要性分数$\boldsymbol{S}=[\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_t,\cdots,\boldsymbol{s}_T]\in\mathbb{R}^{B\times T}$，其中，$\mathrm{s}_t$为$t$帧的分数。

权重更新单元和指示器共同负责帧重要性分数$\boldsymbol{S}$的更新，其中：权重更新单元包含一个全连接网络和归一化指数函数Softmax(•)；指示器为一个全连接网络。二者协同作用，交互过程重复次数为N，N为所要选取的重要关键帧的数量。当交互次数为i时，权重更新单元的Softmax(•)可获得当前状态下T帧的重要性分数$\boldsymbol{S}$的概率密度分布$\boldsymbol{c}_i\in\mathbb{R}^{1\times T}$，并对该分布进行随机采样。假设随机采样时选中了帧$k_1\in[1,T]$，此时对短视频关键帧的重要性分数$\boldsymbol{S}$施加权重因子$\boldsymbol{W}^i\in\mathbb{R}^{B\times T}$，$w_{ab}^i$为该权重因子中行数$a\in[1,B]$、列数$b\in[1,T]$的元素，定义为

$$w_{ab}^i=\begin{cases}\dfrac{N-(i-1)}{T-(i-1)}+1, & a\in[1,B],\ b=k_1\\ 1, & \text{otherwise}\end{cases}。\tag{1}$$

该因子保证了模型先行选择的帧的重要性更高的特性。在N次交互结束后，对未被选择的帧施加权重因子$R_{\mathrm{RF}}\in\mathbb{R}^{B\times T}$。假设最终未被选择帧$k_2\in[1,T]$，则此时$R_{\mathrm{RF}}$中的$k_2$列的值全部为$(T-N)/T$，其余列的值全部为1。

在经过N次选择后最终产生新的短视频关键帧分数：$\boldsymbol{S}'=[\boldsymbol{s}_1',\boldsymbol{s}_2',\cdots,\boldsymbol{s}_t',\cdots,\boldsymbol{s}_T']\in\mathbb{R}^{B\times T}$，其中，$\boldsymbol{s}_t'$为$t$帧的当前分数。$\boldsymbol{S}'$扩展维度后和特征$\boldsymbol{X}$相乘得到施加重要性分数之后的特征$\boldsymbol{P}\in\mathbb{R}^{B\times T\times D}$：

$$\boldsymbol{P}=\hat{\boldsymbol{S}}\odot\boldsymbol{X},\tag{2}$$

式中：$\hat{\boldsymbol{S}}\in\mathbb{R}^{B\times T\times D}$由$\boldsymbol{S}'$扩展维度所得；⊙表示对应元素相乘。随后，$\boldsymbol{P}$经过由LSTM和线性层构成的编解码器，即VAE，产生一个维度相同但是含有短视频关键帧潜在重要性信息的重构特征结构$\hat{\boldsymbol{X}}\in\mathbb{R}^{B\times T\times D}$。重构特征$\hat{\boldsymbol{X}}$和原始特征$\boldsymbol{X}$共同经过由LSTM、线性层及Sigmoid激活函数构成的判别器进行判别，最终输出一个用于权重更新和指示器训练的反馈值r_i：

$$r_i=1-L_{\mathrm{recon}},\ r_i\in\mathbb{R},\ i\in[i,N],\tag{3}$$

式中，L_{recon}用来保证重构特征$\hat{\boldsymbol{X}}$没有损失过多原始特征$\boldsymbol{X}$所含的信息。L_{recon}使用均方误差来定义，具体表示为

$$L_{\mathrm{recon}}=\left\|\boldsymbol{X}-\hat{\boldsymbol{X}}\right\|^2。\tag{4}$$

上述权重更新的训练基于损失函数L_{act}，表示为

$$L_{\mathrm{act}}=-\frac{1}{N}\left[\sum_{i=1}^{N}\alpha_i\log\_\mathrm{prob}(\boldsymbol{c}_i)+\delta\sum_{i=1}^{N}\mathrm{entropy}(\boldsymbol{c}_i)\right],\tag{5}$$

式中：log_prob(•)为对概率密度分布取对数的函数，其所得值为标量值；$\boldsymbol{c}_i$为交互次数为i时重要性分数$\boldsymbol{S}$的概率密度分布；$\delta=0.1$是熵正则化系数；entropy(•)为熵

函数，其所得值为标量值，该值用来衡量概率分布的稀疏程度，熵越小概率分布越集中；$\alpha_i=d_i-v_i(i=1,2,\cdots,N)$表示交互次数为$i$时对随机采样选取的特定帧施加权重因子比直接对所有帧施加同样权重因子的优势所在，v_i是重要性分数$\boldsymbol{S}$经过指示器得到的对权重更新所进行的帧选择的评估值，该值是一个标量值，d_i为自交互次数$i\sim N$过程累积的判别器反馈值，定义为

$$d_i=\sum_{k=i}^{N}\gamma^{k-i}r_i, \tag{6}$$

式中：γ为该反馈值的影响程度($\gamma\in\mathbb{R}$, $0<\gamma<1$)。对γ施加的幂数$k-i(k\in[i,N])$保证交互次数i的反馈值的影响最大，后面交互产生的反馈值的影响依次减小。为使反馈值总体上有较大的影响程度，令$\gamma=0.99$。最终，指示器的训练基于损失函数L_{instru}，表示为

$$L_{\text{instru}}=\frac{1}{N}\alpha_i^2。 \tag{7}$$

综上，L_{act}和L_{instru}二者共同作用使c_i和α_i不断更新，最终达到最优值。

2.2　ISAE

本过程受多头注意力机制的启发，利用重要性分数加权的特征表示来引导注意力的学习，进而获得ISAE的特征表示。具体而言，假设头数为L，则头数l的$\boldsymbol{G}^l\in\mathbb{R}^{T\times d_v}$的定义为

$$\boldsymbol{G}^l=\text{Softmax}\left[\hat{\boldsymbol{P}}\boldsymbol{W}_{\text{Q}}^l\left(\hat{\boldsymbol{P}}\boldsymbol{W}_{\text{K}}^l\right)^{\text{T}}\Big/\sqrt{d_k}\right]\hat{\boldsymbol{P}}\boldsymbol{W}_{\text{V}}^l, \tag{8}$$

式中：$\hat{\boldsymbol{P}}\in\mathbb{R}^{T\times D}$为帧间重要性分数强化后的展开矩阵，即对上述重要性分数强化之后的特征$\boldsymbol{P}$沿第一阶进行切片后所得到的特征矩阵；$\boldsymbol{W}_{\text{Q}}^l\in\mathbb{R}^{D\times d_k}$、$\boldsymbol{W}_{\text{K}}^l\in\mathbb{R}^{D\times d_k}$和$\boldsymbol{W}_{\text{V}}^l\in\mathbb{R}^{D\times d_v}$分别为查询矩阵、键矩阵和值矩阵待学习的权重参数，$d_k=d_v=D/L$为比例缩放因子，$D$为特征维度数。式(8)的含义为$\hat{\boldsymbol{P}}$经Softmax(•)后得到每个键对应的特征空间权重，该权重再和$\hat{\boldsymbol{P}}$相乘得到特征空间权重分配后的特征表示。这一过程协同了帧间与特征空间的内在关联性，最终获得更加具有注意力特性的特征表示$\boldsymbol{F}\in\mathbb{R}^{T\times D}$：

$$\boldsymbol{F}=\boldsymbol{W}_{\text{o}}\,\text{Concat}\left(\boldsymbol{G}^1,\boldsymbol{G}^2,\cdots,\boldsymbol{G}^L\right), \tag{9}$$

式中：L是多头注意力机制的头数量；Concat(•)为矩阵的串联操作；$\boldsymbol{W}_{\text{o}}\in\mathbb{R}^{Ld_v\times D}$为待学习的权重参数。

2.3　HACL

为学习短视频事件的潜在语义及语义之间的关联性，将事件的隐藏属性视为节点，属性间的关联度视为边，为每个短视频样本构造特定图。首先通过隐藏属性激活映射单元(HAAU)来捕获隐藏属性响应矩阵：

$$\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_i,\cdots,\boldsymbol{e}_{\hat{C}}]^{\text{T}}\in\mathbb{R}^{\hat{C}\times D}, \tag{10}$$

式中：$\hat{C}$为隐藏属性数；D为特征维度数。每个隐藏属性响应向量$\boldsymbol{e}_i$的计算公式为

$$\boldsymbol{e}_i=\boldsymbol{m}_i^{T}\boldsymbol{F}=\sum_{q=1}^{T}\boldsymbol{m}_i^q\boldsymbol{F}^q, \tag{11}$$

式中：$\boldsymbol{F}\in\mathbb{R}^{T\times D}$为ISAE模块输出的包含更具有注意力特性的特征图；$\boldsymbol{F}^q$和$\boldsymbol{m}_i^q$分别为每个样本帧数为q的特征表示及隐藏属性激活图；$\boldsymbol{m}_i$为将卷积滤波器作为响应检测器时捕获的隐藏属性激活图。得到隐藏属性激活图矩阵：

$$\boldsymbol{M}=[\boldsymbol{m}_1,\cdots,\boldsymbol{m}_i,\cdots,\boldsymbol{m}_{\hat{C}}]\in\mathbb{R}^{T\times\hat{C}}。 \tag{12}$$

接着将$\boldsymbol{E}$输入到动态关联性单元(DCU)中。如图3所示，该单元通过构建静态图和动态图来获取隐

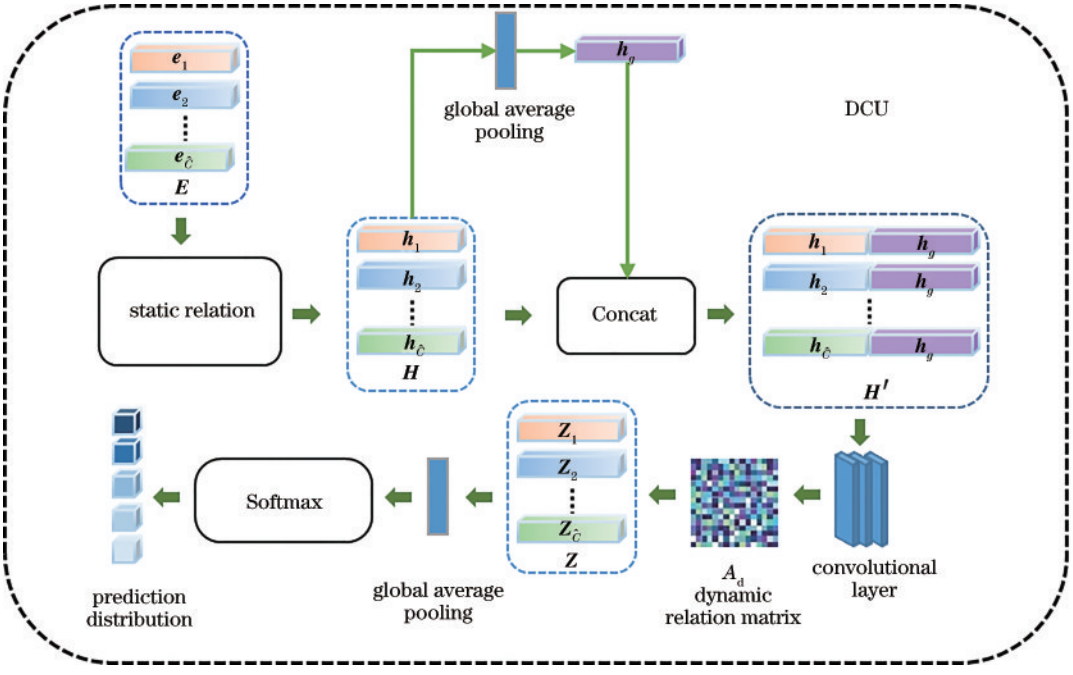

图3　动态关联性单元

Fig. 3　Dynamic correlation unit

藏属性之间的关联特性，最终得到具有潜在语义信息关联性的特征表示 $\boldsymbol{Z}$。静态图用于学习所有样本隐藏属性之间的总体关系，动态图用于学习每个样本所含有的特定属性关系。

具体而言，DCU 中静态部分定义为

$$\boldsymbol{H}=\mathrm{LeakyReLU}(\boldsymbol{A}_{\mathrm{s}}\boldsymbol{E}\boldsymbol{W}_{\mathrm{s}}),\tag{13}$$

式中：LeakyReLU(·)为激活函数；$\boldsymbol{E}\in\mathbb{R}^{\hat{C}\times D}$ 为隐藏属性响应矩阵，由式(10)获得；$\boldsymbol{A}_{\mathrm{s}}\in\mathbb{R}^{\hat{C}\times\hat{C}}$ 和 $\boldsymbol{W}_{\mathrm{s}}\in\mathbb{R}^{D\times D_{\mathrm{r}}}$ 分别为静态关联矩阵和静态权重更新矩阵，二者均为随机初始化并在训练过程中不断更新，D_{r} 为静态部分训练后得到的特征维度数。

动态部分定义为

$$\boldsymbol{Z}=\mathrm{LeakyReLU}(\boldsymbol{A}_{\mathrm{d}}\boldsymbol{H}\boldsymbol{W}_{\mathrm{d}}),\ \boldsymbol{A}_{\mathrm{d}}=\delta\left[f_{\mathrm{conv}}(\boldsymbol{H}')\right],\tag{14}$$

式中：$\delta(\cdot)$ 为激活函数；$f_{\mathrm{conv}}(\cdot):\mathbb{R}^{\hat{C}\times 2D_{\mathrm{r}}}\rightarrow\mathbb{R}^{\hat{C}\times\hat{C}}$ 为卷积层，用于维度转换；$\boldsymbol{H}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{\hat{C}}]\in\mathbb{R}^{\hat{C}\times D_{\mathrm{r}}}$ 为静态表示，由式(13)所得；$\boldsymbol{A}_{\mathrm{d}}\in\mathbb{R}^{\hat{C}\times\hat{C}}$ 和 $\boldsymbol{W}_{\mathrm{d}}\in\mathbb{R}^{D_{\mathrm{r}}\times D_{\mathrm{k}}}$ 分别为学习得到的动态图关联矩阵和动态权重更新矩阵，二者均为随机初始化并在学习过程中不断更新，二者协同作用将隐藏属性之间的动态相关信息不断向 $\boldsymbol{H}$ 传递，从而使 $\boldsymbol{H}$ 得到更新，D_{k} 为经过动态部分训练后得到的特征维度数。为获得动态特性，构建了矩阵 $\boldsymbol{H}'=\left[(\boldsymbol{h}_1;\boldsymbol{h}_g),(\boldsymbol{h}_2;\boldsymbol{h}_g),\cdots,(\boldsymbol{h}_{\hat{C}};\boldsymbol{h}_g)\right]\in\mathbb{R}^{2D_{\mathrm{r}}\times\hat{C}}$，该矩阵由 $\boldsymbol{H}$ 和它的全局表示 $h_g\in R^{D_{\mathrm{r}}}$ 组合而成。最终，获得包含潜在语义关联性的特征表示 $\boldsymbol{Z}\in\mathbb{R}^{\hat{C}\times D_{\mathrm{k}}}$，该表示经过一个全局平均池化层和归一化指数函数 Softmax(·)后得到事件类别得分 $\hat{\boldsymbol{y}}=\left[\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_R\right]$，其中，$R$ 为短视频事件类别数。该部分基于交叉熵构建的分类损失函数 L_{dcm} 表示为

$$L_{\mathrm{dcm}}=-\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{R}y_{i,j}\log\sigma\left(\hat{y}_{i,j}\right)-\left(1-y_{i,j}\right)\log\left[1-\sigma\left(\hat{y}_{i,j}\right)\right],\tag{15}$$

式中：$y_{i,j}\in\{0,1\}$，第 i 个样本若属于第 j 类则为 1，否则为 0；$\hat{y}_{i,j}$ 为第 i 个样本被预测为第 j 类的概率值。

综上，最终模型的损失函数为

$$L=\alpha L_{\mathrm{act}}+\beta L_{\mathrm{recon}}+\gamma L_{\mathrm{dcm}},\tag{16}$$

式中，α、β、γ 为平衡各损失间权重的超参数。

3　数据准备及实验分析

3.1　数据集准备、基本设置及评价方法

由于现有的事件检测数据集主要针对视频监控、体育事件等，在短视频方向上缺少可靠的主流数据集，因而在主要考虑以下 3 点的情况下构建了数据集：1)所选视频的标签要能充分体现复杂事件的特点；2)视频长度需要满足短视频的特点；3)每个标签下都要有足够多的样本数量。Flickr 是一个国外资源开放的短视频分享网站，且该网站上的大部分短视频都包含上传者添加的标签。针对以上特点，爬取 Flickr 网站的视频数据来构建短视频事件检测数据集。为确保短视频的实际内容与其标签相符，将批量下载的短视频进行人工清洗，最终，整个数据集共包含 20231 个短视频，每个短视频时长不超过 30 s，涉及棒球、音乐会及游泳等共 20 个类别。随机选择 80% 的短视频作为训练集，其余作为测试集，并将 ResNet(residual network)作为关键帧的特征提取器。

该实验所用服务器的中央处理器(CPU)为 Intel(R)Core(TM) i9-10920X CPU @3.5 GHz，图形处理器(GPU)为 RTX 3090，在 Python3.6 环境下使用 PyTorch 实现模型。训练过程中将迭代次数设置为 100，训练批次设置为 64，采用 Adam 作为训练优化器，并对模型中不同模块分设不同的学习率，其中：FIE 模块的初始学习率设置为 0.0001；其他部分的初始学习率设置为 0.05。在训练过程中使用 MultiStepLR 学习率调整机制不断调整训练阶段的学习率，防止模型由于初始学习率设置不当而无法达到全局最优的状况。该实验过程利用召回率 R(recall)、精度 P(precision)以及 mAP(mean average precision)3 种指标来进行模型评价。

3.2　结果分析

3.2.1　消融实验

在保证其他条件不变的情况下，通过混合叠加不同的模块对比其相应实验性能来验证本模型中各个模块的有效性。实际实验过程中，以 FIE 为基础模块，依次叠加其他模块来验证各个模块的性能。消融实验结果如表 1 所示，通过表 1 可以得出以下结论：1)不同模块的叠加均可使实验性能得到不同程度的提升，这表明每个模块添加的合理性，同时，由不同模块混合叠加的结果可见，4 个模块中 FIE 模块对结果影响最大，由此说明对帧进行重要性评估是十分必要的；2)加入 ISAE 的结果表明引入的自注意力机制可以有效获取帧间和特征空间的内在关联性，使模型效果得到进一步的提升；3)依次加入 HAAU 和 DCU 的结果表明对事件潜在语义及其关联性的学习有助于进一步提升事件检测效果，展现了这两个模块的有效性。

表 1　不同模块对实验结果的影响

Table 1　Influence of different modules on the experimental results

Module	R	P	mAP
FIE	0.723	0.708	0.754
FIE+ISAE	0.729	0.715	0.768
FIE+ISAE+DCU	0.731	0.719	0.772
FIE+ISAE+HAAU	0.742	0.736	0.783
FIE+HAAU+DCU	0.763	0.754	0.794
ISAE+HAAU+DCU	0.712	0.697	0.759
Ours	**0.774**	**0.769**	**0.817**

3.2.2 对比实验

为充分证明所提模型在事件检测任务上的优势，将其与应用于事件检测的典型3D卷积神经网络C3D（convolutional network with three-dimensional kernels）[24]、ResNet3D（residual network with three-dimensional kernels）[25]，基于子空间学习策略的SRRS（supervised regularization-based robust subspace）[26]、DTSL（discriminative transfer subspace learning）[28]及深度学习模型NI-SVM（nearly-isotonic support vector machines for event detection）[27]、RegNet（regulated residual network）[29]、KPGNN（knowledge-preserving incremental heterogeneous graph neural network）[30]、TSN（temporal segment network）[31]、Gated-VIGAT（gated bottom-up event recognition and explanation in video using factorized graph attention network）[32]、EventGraph（event extraction as semantic graph parsing）[33]、EITEST（testing for event impacts in time series）[34]、PP-GCN（pairwise popularity graph convolutional network）[35]、TSM（temporal shift module for efficient video understanding）[36]、EfficientNet（a new scaling method which achieves much better accuracy and efficiency than previous convolutional neural networks）[37]、X3D（expand three dimension）[38]进行对比。其中：NI-SVM、KPGNN、Gated-VIGAT、EventGraph、EITEST、PP-GCN为事件检测方法；其他为可用于事件检测任务的分类方法。实验结果如表2所示。

表2 Flickr数据集下不同方法的性能对比

Table 2 Performance comparison of different methods on the Flickr dataset

Method	*R*	*P*	mAP	Training time /s
C3D	0.552	0.630	0.663	16560.490
ResNet3D	0.632	0.672	0.694	36720.080
SRRS	0.617	0.625	0.656	91440.120
DTSL	0.622	0.672	0.679	83160.370
NI-SVM	0.691	0.727	0.749	3048.720
RegNet	0.753	0.749	0.785	4505.810
KPGNN	0.705	0.731	0.756	2856.950
TSN	0.727	0.731	0.762	2568.560
Gated-VIGAT	0.733	0.746	0.778	2603.050
EventGraph	0.736	0.725	0.766	2516.970
EITEST	0,753	0.750	0.784	2478.230
PP-GCN	0.730	0.747	0.772	2756.280
TSM	0.732	0.726	0.757	2586.630
EfficientNet	0.745	0.752	0.781	2441.970
X3D	0.761	0.757	0.796	2391.710
Ours	**0.774**	**0.769**	**0.817**	**2370.560**

实验结果表明，所提模型在事件检测任务中表现最佳。由表2结果可见：1）所提模型相比3D卷积模型有较大的提升效果，3D卷积网络由于具有大量的参数，在训练过程中难以优化，最终检测效果不佳；2）相比SRRS和DTSL模型，所提模型的mAP分别提高了0.161和0.138，这是由于SRRS和DTSL模型在建模过程中仅使用了浅层特征，且所用子空间的维度较低，缺乏对高级语义表示及潜在信息的学习；3）相比NI-SVM模型，所提模型的mAP提高了0.058，这是因为NI-SVM模型利用不同帧的显著性动态来进行语义特征训练，仅对显示语义进行建模，缺乏对潜在语义及帧间相关信息的考虑；4）RegNet、TSM和EITEST模型均挖掘了帧的潜在关联特性，KPGNN、EventGraph和PP-GCN模型则利用图结构来完成检测任务，这几个模型均具有较好的实验效果，表明帧间特性的学习及图结构的引入确实有助于实验性能的提升；5）TSN和Gated-VIGAT模型的实验效果较好，这可能是由于二者均对视频帧进行了选择操作，该结果同时也表明帧选择对性能提升具有一定的促进作用；6）X3D通过对2D模型在空间、时间、深度和宽度上进行扩展来深度挖掘高级语义，进而提高实验性能，其实验效果较好，仅次于所提模型，这可能是因为X3D对事件中的高级语义进行了学习，但相比所提模型，仍缺乏对高级语义之间动态关系的考虑；7）EfficientNet模型通过平衡卷积网络维度来获得更好的分类性能，所提模型的mAP较之提升了0.036，其原因可能是相比传统卷积来说，所提模型使用的GCN能够更充分地挖掘短视频事件中包含的高级语义及语义之间的动态关系；8）通过对比表2中数据，所提模型所用训练时间最短，进一步表明其优势。

为进一步验证所提模型在公共数据集上的有效性，将模型在两个动作识别的公开数据集UCF101和HMDB51上进行训练并同表2所述方法对比最终实验效果。其中：UCF101数据集共101个类别，包含13320个短视频；HMDB51数据集共51个类别，包含6766个短视频。实验时均按照官方提供的第一种划分方式来获得训练集和测试集。实验结果如表3所示，实验结果表明，所提模型在UCF101和HMDB51数据集上的mAP可达0.885和0.622，实验效果最佳，这可能是因为所提模型在动作识别任务中充分挖掘了不同动作及其发生场景隐藏的潜在高级语义及其关联性，证明模型有效性的同时也表明模型具有一定的普适性。

3.2.3 模型收敛性

所提模型的收敛图如图4所示。从图4中可以看出，模型整体的损失随迭代次数的增加逐渐降低并最终趋于稳定，验证了模型的收敛性。

3.2.4 参数敏感性实验

为探求所提模型的最优性能，对隐藏属性数、所选

表 3　UCF101 和 HMDB51 数据集下不同方法的 mAP 对比

Table 3　Comparison of mAP of different methods on UCF101 and HMDB51 datasets

Method	UCF-101	HMDB51
C3D	0.785	0.533
ResNet3D	0.814	0.572
SRRS	0.763	0.517
DTSL	0.791	0.556
NI-SVM	0.827	0.573
RegNet	0.857	0.595
KPGNN	0.832	0.578
TSN	0.839	0.581
Gated-VIGAT	0.848	0.586
EventGraph	0.842	0.581
EITEST	0.846	0.575
PP-GCN	0.850	0.599
TSM	0.838	0.577
EfficientNet	0.856	0.595
X3D	0.867	0.608
Ours	**0.885**	**0.622**

重要帧数和 α、β、γ 共 5 个重要参数进行敏感性实验。事件中的隐藏属性数对模型性能的影响如图 5(a)所示，该结果表明当隐藏属性数设置为 200 时，模型性能达到最优。从图 5(a)中的变化趋势可见，隐藏属性的数量确实对模型性能有不同程度的影响，适当的隐藏属性学习可以使模型性能有所提升，但设置过高的隐藏属性数对于事件的学习具有反作用，推测可能是因为学习到的属性过于具象反而不利于复杂事件的表征。此外，对选取的重要帧的个数进行敏感性实验，实验结果如图 5(b)所示。从图 5(b)可以看出，mAP 先随重要帧数的增加上升，在重要帧数为 30 时取得最高值，随后逐渐下降且趋于平缓。由该趋势可见，重要帧的个数确实对模型效果有所影响，一定数量的重要帧的选取对模型效果有提升作用，但选取过少或过多的重要帧可能会忽略或者过于强调某些因素造成训练效果下降。

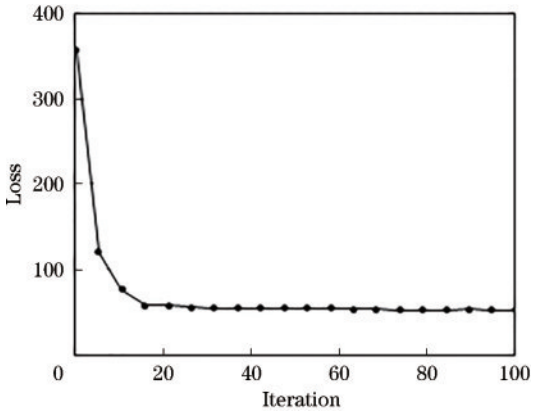

图 4　模型收敛性

Fig. 4　Convergence performance of the model

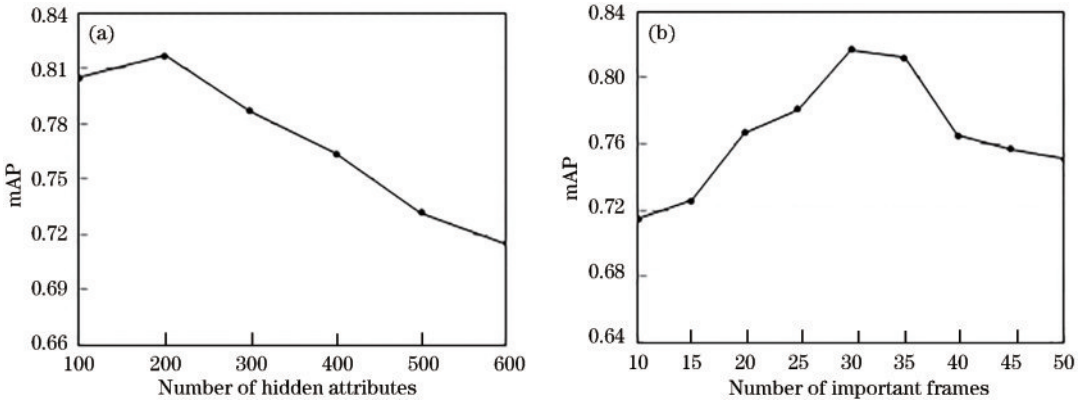

图 5　超参数对模型性能的影响。(a)隐藏属性数影响；(b)所选重要帧数影响

Fig. 5　Influence of parameters on model performance. (a) Influence of the number of hidden attributes; (b) influence of the number of important frames

根据式(16)，参数 α、β、γ 分别控制 FIE 策略损失、重构损失和分类损失对模型的影响。因此，通过评估不同参数值的影响程度可以获得模型的最佳参数组合。在本实验中，以启发式的网格搜索法来选择 3 个参数的最优化值，采用控制变量的思想，每次只改变其中的一个值，另外两个保持 1 不变，依次研究 3 个参数的影响程度。最终 3 个参数的影响结果如图 6 所示。该结果表明当参数 α 的值为 0.8、参数 β 的值为 0.6、参数 γ 的值为 1.2 时，所提模型的性能达到最优。由图 6 中的变化趋势可见，实验性能随 α 的增大先稳步上升后逐渐下降，考虑是 FIE 部分占比过多可能会使模型学习过程中忽略重要性分数较小帧中的特点，从而导致后续性能下降。β 的变化趋势表明重构特征和原特征过于贴近可能会产生过拟合的现象。由图 6(b)所示，相比于 α 和 β，γ 的变化对模型性能的影响更大，说明模型对分类损失最敏感，曲线最终逐渐趋于稳定也表明该模型具有较为稳定的分类效果。

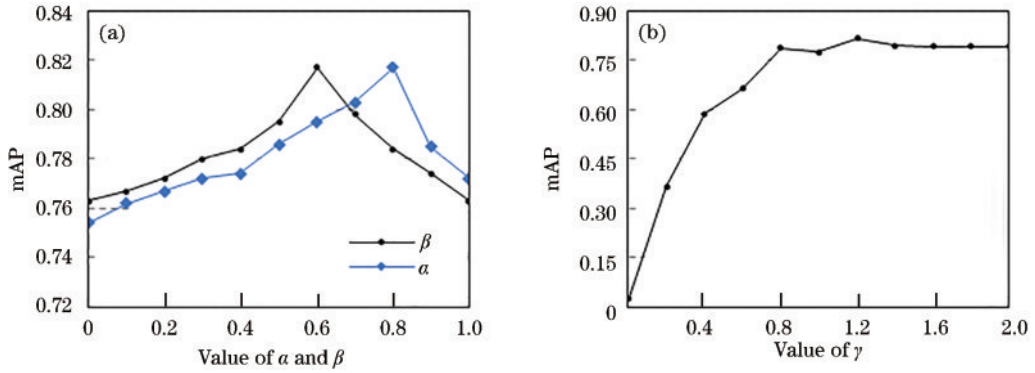

图 6　超参数对模型性能的影响。(a)α和β的影响;(b)γ的影响

Fig. 6　Influence of parameters on model performance. (a) Influence of α and β; (b) influence of γ

4　结　　论

针对目前的短视频事件检测任务,设计了一种基于深度动态语义关联的短视频事件检测模型。该模型依托深度神经网络,主要由 3 大模块构成,其中:FIE 模块最大程度地强化了重要性信息,获得了具有区分度的帧重要性分数;ISAE 模块协同帧间与特征空间的内在关联性获取了更具有帧间注意力的短视频内容表示;HACL 模块在挖掘事件潜在语义的同时还学习了语义间的动态关系。实验结果表明,与基准模型相比,所提方法具有更优的检测效果,能够进一步提升对短视频内容的分析和理解。

参　考　文　献

[1] Xie J Y, Zhu Y C, Zhang Z B, et al. A multimodal variational encoder-decoder framework for micro-video popularity prediction[C]//Proceedings of The Web Conference 2020, April 20-24, 2020, Taipei, Taiwan, China. New York: ACM Press, 2020: 2542-2548.

[2] Jing P G, Su Y T, Nie L Q, et al. Low-rank multi-view embedding learning for micro-video popularity prediction [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(8): 1519-1532.

[3] 井佩光, 叶徐清, 刘昱, 等. 基于双向深度编码网络的短视频流行度预测[J]. 激光与光电子学进展, 2022, 59(8): 0811009.
Jing P G, Ye X Q, Liu Y, et al. Micro-video popularity prediction with bidirectional deep encoding network[J]. Laser & Optoelectronics Progress, 2022, 59(8): 0811009.

[4] Su Y T, Hong D Z, Li Y, et al. Low-rank regularized deep collaborative matrix factorization for micro-video multi-label classification[J]. IEEE Signal Processing Letters, 2020, 27: 740-744.

[5] Liu M, Nie L Q, Wang X, et al. Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning[J]. IEEE Transactions on Image Processing, 2019, 28(3): 1235-1247.

[6] Wei Y W, Wang X, Nie L Q, et al. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video[C]//Proceedings of the 27th ACM International Conference on Multimedia, October 21-25, 2019, Nice, France. New York: ACM Press, 2019: 1437-1445.

[7] Chen J, Peng J J, Qi L Z, et al. Implicit rating methods based on interest preferences of categories for micro-video recommendation[M]//Douligeris C, Karagiannis D, Apostolou D. Knowledge science, engineering and management. Lecture notes in computer science. Cham: Springer, 2019, 11775: 371-381.

[8] Redi M, O'Hare N, Schifanella R, et al. 6 seconds of sound and vision: creativity in micro-videos[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 4272-4279.

[9] Zhang J, Wu Y T, Liu J H, et al. Low-rank regularized multimodal representation for micro-video event detection [J]. IEEE Access, 2020, 8: 87266-87274.

[10] Over P, Fiscus J, Sanders G, et al. TRECVID 2014-an overview of the goals, tasks, data, evaluation mechanisms and metrics[C]//Trecvid Workshop Participants Notebook Papers, November 10-12, 2014, Orlando, FL, USA. Orlando: NIST, 2014: 1-53.

[11] Alomari E, Katib I, Mehmood R. Iktishaf: a big data road-traffic event detection tool using twitter and spark machine learning[J]. Mobile Networks and Applications, 2020: 1-16.

[12] Wan S H, Xu X L, Wang T, et al. An intelligent video analysis method for abnormal event detection in intelligent transportation systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(7): 4487-4495.

[13] 杨先斌, 党建武, 王松, 等. 基于双流网络与多示例学习的异常事件检测[J]. 激光与光电子学进展, 2021, 58(20): 2015006.
Yang X B, Dang J W, Wang S, et al. Anomaly event detection based on two-stream network and multi-instance learning[J]. Laser & Optoelectronics Progress, 2021, 58(20): 2015006.

[14] Zhang K, Chao W L, Sha F, et al. Video summarization with long short-term memory[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9911:

766-782.

[15] Rochan M, Ye L W, Wang Y. Video summarization using fully convolutional sequence networks[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision – ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11216: 358-374.

[16] 王慧赢, 王春平, 付强, 等. 面向嵌入式平台的轻量级光学遥感图像舰船检测[J]. 光学学报, 2023, 43(12): 121-134.
Wang H Y, Wang C P, Fu Q, et al. Lightweight ship detection based on optical remote sensing images for embedded platform[J] . Acta Optica Sinica, 2023, 43(12): 121-134.

[17] 吕梦凌, 何玉青, 杨峻凯, 等. 基于循环注意力机制的隐形眼镜虹膜防伪检测方法[J]. 光学学报, 2022, 42(23): 2315001.
Lü M L, He Y Q, Yang J K, et al. Anti-counterfeiting detection method of contact lens iris based on cyclic attention mechanism[J]. Acta Optica Sinica, 2022, 42(23): 2315001.

[18] 刘昊鑫, 赵源萌, 张存林, 等. 基于改进U-net的牙齿锥形束CT图像重建研究[J]. 中国激光, 2022, 49(24): 2407207.
Liu H X, Zhao Y M, Zhang C L, et al. Study on reconstruction of tooth cone beam CT image based on improved U-net[J]. Chinese Journal of Lasers, 2022, 49(24): 2407207.

[19] 储光涵, 范大昭, 董杨, 等. 结合图论的异源影像点云配准方法[J]. 光学学报, 2023, 43(12): 264-272.
Chu G H, Fan D Z, Dong Y, et al. A cross-source image point cloud registration method combined with graph theory[J]. Acta Optica, 2023, 43(12): 264-272.

[20] Chen Z M, Wei X S, Wang P, et al. Multi-label image recognition with graph convolutional networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 5172-5181.

[21] 刘磊, 李元祥, 倪润生, 等. 基于卷积与图神经网络的合成孔径雷达与可见光图像配准[J]. 光学学报, 2022, 42(24): 2410002.
Liu L, Li Y X, Ni R S, et al. Synthetic aperture radar and optical images registration based on convolutional and graph neural networks[J]. Acta Optica Sinica, 2022, 42(24): 2410002.

[22] Bastings J, Titov I, Aziz W, et al. Graph convolutional encoders for syntax-aware neural machine translation[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, September 9-11, 2017, Copenhagen, Denmark. Stroudsburg: Association for Computational Linguistics, 2017: 1957-1967.

[23] 田晟, 龙安洋. 基于图卷积和多层特征融合的点云分类方法[J]. 激光与光电子学进展, 2023, 60(14): 281-288.
Tian S, Long A Y. Point cloud classification method based on graph convolution and multi-layer feature fusion [J]. Lasers & Optoelectronics Progress, 2023, 60(14): 281-288.

[24] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2016: 4489-4497.

[25] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 6546-6555.

[26] Li S, Fu Y. Learning robust and discriminative subspace with low-rank constraints[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(11): 2160-2173.

[27] Chang X J, Yu Y L, Yang Y, et al. Semantic pooling for complex event analysis in untrimmed videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1617-1632.

[28] Xu Y, Fang X Z, Wu J, et al. Discriminative transfer subspace learning via low-rank and sparse representation [J]. IEEE Transactions on Image Processing, 2016, 25(2): 850-863.

[29] Xu J, Pan Y, Pan X L, et al. RegNet: self-regulated network for image classification[J/OL]. IEEE Transactions on Neural Networks and Learning Systems: 1-6[2023-03-28]. https://ieeexplore.ieee.org/document/9743274.

[30] Cao Y W, Peng H, Wu J, et al. Knowledge-preserving incremental social event detection via heterogeneous GNNs[C]//Proceedings of the Web Conference 2021, April 19-23, 2021, Ljubljana, Slovenia. New York: ACM Press, 2021: 3383-3395.

[31] Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks for action recognition in videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2740-2755.

[32] Gkalelis N, Daskalakis D, Mezaris V. Gated-ViGAT: efficient bottom-up event recognition and explanation using a new frame selection policy and gating mechanism [C]//2022 IEEE International Symposium on Multimedia (ISM), December 5-7, 2022, Italy. New York: IEEE Press, 2023: 113-120.

[33] You H, Samuel D, Touileb S, et al. EventGraph: event extraction as semantic graph parsing[C]//Proceedings of the Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, December 7-8, 2022, Abu Dhabi, United Arab Emirates. Abu Dhabi: ACL, 2022: 7-15.

[34] Scharwächter E, Müller E. Two-sample testing for event impacts in time series[C]//Proceedings of the 2020 SIAM International Conference on Data Mining, May 7-9, 2020, Cincinnati, Ohio, USA. Philadelphia: Society for Industrial and Applied Mathematics, 2020: 10-18.

[35] Peng H, Li J X, Song Y Q, et al. Streaming social event detection and evolution discovery in heterogeneous information networks[J]. ACM Transactions on

Knowledge Discovery from Data, 2021, 15(5): 89.

[36] Lin J, Gan C, Han S. TSM: temporal shift module for efficient video understanding[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 7082-7092.

[37] Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]//International conference on machine learning, June 9-15, 2019, Long Beach, California, USA. Long Beach: PMLR, 2019: 6105-6114.

[38] Feichtenhofer C. X3D: expanding architectures for efficient video recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 200-210.