计算机科学
COMPUTER SCIENCE

基于两阶段时空对齐的小样本视频行为识别

王佳, 夏英, 丰江帆

引用本文

王佳, 夏英, 丰江帆. 基于两阶段时空对齐的小样本视频行为识别[J]. 计算机科学, 2025, 52(8): 251-258.

WANG Jia, XIA Ying, FENG Jiangfan. Few-shot Video Action Recognition Based on Two-stage Spatio-Temporal Alignment [J]. Computer Science, 2025, 52(8): 251-258.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240900127

基于两阶段时空对齐的小样本视频行为识别

王 佳 夏 英 丰江帆
重庆邮电大学计算机科学与技术学院 重庆 400065
旅游多源数据感知与决策技术文化和旅游部重点实验室 重庆 400065
(S220201092@stu.cqupt.edu.cn)

摘 要 小样本视频行为识别旨在利用有限的训练样本构建高效学习模型,从而减轻传统行为识别对大规模且精细标注数据集的依赖。目前,小样本学习模型大多依据视频之间的相似性进行分类,但不同的动作实例呈现出不同的时空分布,导致查询视频与支持视频之间出现时间错位和动作演化错位,从而影响模型的识别性能。针对此问题,提出两阶段时空对齐网络TSAN,以提高视频数据的对齐精度,进而提升小样本视频行为识别的准确率。该网络采用元学习的基本架构,第一阶段通过动作时间对齐模块ATAM,构建元组模式的视频帧对,将视频动作细分为子动作序列,并结合视频数据中的时序信息,提升小样本学习的效率;第二阶段通过动作演化对齐模块AEAM,及其中包含的时间同步子模块TSM和空间协调子模块SCM,对查询特征进行校准,以匹配支持集的时空动作演化,从而提高小样本视频行为识别的准确率。在HMDB51,UCF101,SSV2100和Kinetics100这4个数据集上的实验结果表明,TSAN网络相较于现有小样本视频行为识别方法,具有更高的识别准确率。

关键词:行为识别;视频分类;时空对齐;小样本学习;元学习
中图分类号 TP391

Few-shot Video Action Recognition Based on Two-stage Spatio-Temporal Alignment

WANG Jia,XIA Ying and FENG Jiangfan
College of Computer Science and Technology,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
Key Laboratory of Tourism Multisource Data Perception and Decision Technology,Ministry of Culture and Tourism,Chongqing 400065,China

Abstract Few-shot video action recognition aims to construct efficient learning models using limited training samples,thereby reducing the dependence of traditional action recognition on large-scale and finely annotated datasets. At present,most few-shot learning models classify videos based on their similarity. However,due to the different spatiotemporal distributions of action instances,there is a temporal and action evolution mismatch between the query video and the supporting video,which affects the recognition performance of the model. To address this issue,a two-stage spatiotemporal alignment network TSAN is proposed to improve the alignment accuracy of video data,thereby enhancing the accuracy of few-shot video action recognition. This network adopts the basic architecture of meta learning. In the first stage,the action time alignment module ATAM is used to construct video frame pairs in tuple mode,which subdivides video actions into sub action sequences and combines them with temporal information in video data to improve the efficiency of few-shot learning. In the second stage,the action evolution alignment module AEAM,along with its time synchronization submodule TSM and spatial coordination submodule SCM,are used to calibrate the query features to match the spatiotemporal action evolution of the support set,thereby improving the accuracy of few-shot video action recognition. The experimental results on the HMDB51,UCF101,SSV2100,and Kinetics100 datasets show that the TSAN network has higher recognition accuracy compared to existing few-shot video action recognition methods.

Keywords Action recognition,Video classification,Spatio-temporal alignment,Few-shot learning,Meta-learning

1 引言

视频行为识别[1]旨在从海量视频数据中提取不同行为的特征,从而实现对行为的自动检测、精确跟踪和高效分类,其被广泛应用于工业生产、城市管理、公共安全等领域。然而,视频行为识别方法大多依赖大规模标注数据的训练,这在现实中往往受到标注成本高昂、数据获取困难等因素的制约。如何在小样本视频数据中实现准确高效的行为识别,

到稿日期:2024-09-23 返修日期:2024-11-23
基金项目:国家自然科学基金(41971365);重庆市教委重点合作项目(HZ2021008);文化和旅游部重点实验室资助项目(E020H2023005)
This work was supported by the National Natural Science Foundation of China(41971365),Chongqing Municipal Education Commission Cooperation Projects(HZ2021008) and Key Laboratory Project from Ministry of Culture and Tourism,China(E020H2023005).
通信作者:夏英(xiaying@cqupt.edu.cn)

成为当前研究的热点。

视频中与运动相关的时间和空间特征在小样本行为识别中发挥着重要的作用。近年来,小样本视频行为识别算法主要是在元学习[2]的基本架构下,运用度量学习[3]进行视频间的相似性比较。研究者结合深度学习方法不断改进度量学习策略,关注视频特征的表示能力,提取更有效的时空特征并促进特征融合,使得小样本视频行为识别性能不断提升。Zhu等[4]提出CMN,利用复合记忆网络来存储矩阵表示,虽然减小了参数量与运算量,可以高效地检索和更新特征,但样本数量少会导致记忆网络中存储的类原型判别性较弱,从而影响分类精度。Bishay等[5]提出时间注意关系网络TARN,通过分段注意力模块进行特征层面的时间对齐。Zhang等[6]通过动作关系网络ARN学习查询视频与支持视频之间的相似性,同时引入增强引导的时空注意和辅助的自我监督训练损失。Cao等[7]提出有序时间对齐模块OTAM,利用动态时间规划算法的变体进行视频序列的显式对齐,利用查询视频中的时序信息,度量以查询集为中心的支持集样本的距离。这类方法都考虑了时间对齐,但对视频时空错位等问题没有开展深入分析。Fu等[8]提出深度导向自适应元融合网络AMeFuNet,将深度信息纳入分类过程,缓解了标注数据稀缺问题。Ni等[9]提出多模态原型增强网络MORN,利用标签文本的语义信息来增强原型,并通过多模态特征提取器实现特征提取与原型构建过程的深度整合。这类方法注重增加数据样本信息量,尽管单个类别的信息总量得到了提升,但也产生了冗余信息。Dwivedi等[10]提出生成对抗网络模型proto-GAN,使用带有类别原型的条件生成对抗网络生成额外的样本。Zhu等[11]提出以原型为中心的注意学习模型PAL,用所有的查询样本来匹配一个原型,解决了类间重叠和外围孤立样本的问题。Wang等[12]提出基于运动增强的长短对比学习网络MoLo,运用长短对比目标策略增强局部帧特征,并通过运动自解码器提取运动线索,实现对全局信息的感知。Wang等[13]提出了基于CLIP引导的原型调制网络CLIP-FSAR,充分利用CLIP模型的多模态知识,将大规模对比性语言图像预训练应用于小样本视频行为识别。这类方法虽然提升了模型的表征能力,但也增加了网络结构的复杂性,存在较大的计算负担和资源消耗。

在实际场景中,由于视频中的动作实例常呈现出不同的时空分布特性,查询视频与支持视频之间常出现两种错位情形,影响识别精度。一是动作时间错位[14],由于动作起始时间和结束时间不同,一个动作的相对时间位置在视频之间通常是不一致的。二是动作演化错位[14],由于动作的非线性演化特性,即使属于同一语义类别,不同视频中的动作也可能在演化速度和状态上表现出差异。针对这两类情形,不少学者开展了深入研究。Perrett等[15]提出的TRX采用Transformer和交叉自注意力机制来捕捉视频序列中不同时间点之间的复杂关系,以元组匹配的方式减轻时间错位问题。Thatipelli等[16]在TRX基础上提出STRM,通过聚合局部以及全局的时空特征,来捕获视频中物体的外观和运动。Guo等[17]提出TSA-MLT,通过在帧序列进行时间维度上的仿射变换和多级Transformer融合来进行时空对齐。然而,由于动作实例间时空变化的多样性,用于缓解动作时间错位和动作演化错位等情形的视频语义对齐仍然是一个具有挑战性的问题。

本文受TRX的元组匹配思想启发,在优化元组匹配算法的同时对动作时间错位和动作演化错位进行特征对齐,提出基于两阶段时空对齐网络(Two-stage Spatio-temporal Alignment Network,TSAN)的小样本行为识别方法。TSAN网络包含两个阶段,在第一阶段采用动作时间对齐模块(Action Temporal Alignment Module,ATAM),通过引入元组匹配思想,构建子动作序列来校准动作的时间错位。在第二阶段构建动作演化对齐模块(Action Evolution Alignment Module,AEAM),该模块进一步分为时间同步模块(Temporal Synchronization Module,TSM)和空间协调模块(Spatial Coordination Module,SCM),分别从时间和空间两个维度对动作演化进行对齐。最终,通过两阶段对齐策略,有效提升分类性能。

2 元学习及训练单元

元学习[2]旨在通过少量的样本和较少的梯度迭代次数实现对目标模型的快速适应。其核心思想是通过训练,网络从大量模拟的小样本任务中学习和提取元知识,当面对一个全新的小样本任务时,模型就能利用这些习得的元知识,仅用极少的样本进行快速适应,从而指导模型在小样本任务中更加高效地学习。

在元学习模型的构建中,通常设定3个互不重叠的元集合,即元训练集、元验证集和元测试集。元训练集主要用于元学习过程的训练,通过最小化训练损失来优化网络性能。元测试集用于评估模型在新任务上的泛化效果。元验证集用于调整模型的超参数,以达到最佳性能。

小样本分类任务的基本单元被称为“情节”(Episode)。在每个训练批次中,将同时构建多个情节,每个情节由一个支持集和一个查询集组成。支持集包含带标签的训练视频,用于模型在情节内的学习,而查询集则包含未标记的视频样本,用于评估模型在当前情节上的泛化能力。

在具体的“n-way k-shot”小样本学习设定中,将在元集合的N个类别中随机选择n个类别(通常n不大于5)来构建每个情节。每个情节的视频和标签均从相应的元集合中采样。首先构建支持集S,其中包含n个不同类别,每个类别有k个支持视频,共$n\times k$个视频样本。然后从每个类中额外选取q个样本组成查询集Q,包含$n\times q$个样本。模型的目标是利用这有限的$n\times k$个支持样本对查询集中的$n\times q$个查询样本进行准确分类。

3 两阶段时空对齐网络模型设计

两阶段时空对齐网络TSAN采用元学习基本架构,整体结构如图1所示。该网络主要由5部分构成,分别为特征提取模块、特征预处理模块、ATAM模块、AEAM模块和距离度量模块。其中ATAM模块用于第一阶段动作时间对齐,AEAM模块用于第二阶段动作演化对齐。

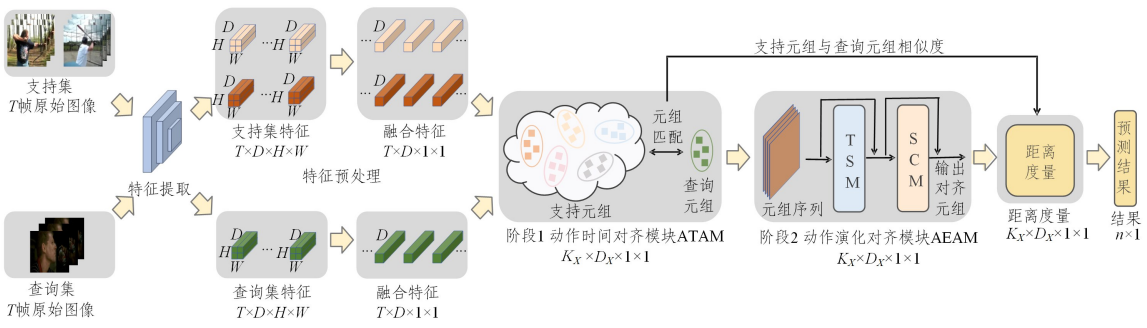


图 1　TSAN 网络框架

Fig. 1　Framework of TSAN network

首先，网络采用稀疏均匀采样策略处理支持集和查询集，获取 T 帧视频的原始图像。随后，在特征提取器的选择上，模型使用了当前主流的 ResNet50[18] 残差神经网络对视频帧进行特征提取，生成 $T\times D\times H\times W$ 的特征图，其中 D 代表输入特征的通道数，H 和 W 分别代表特征图的高度和宽度。为突出特征图中的显著信息，并缓解小样本行为识别中潜在的过拟合问题，网络采用全局最大池化操作替代原 ResNet50 网络中的全局平均池化，将特征图压缩为 $T\times D\times 1\times 1$ 的融合特征。

在阶段 1 中，为进一步提升特征的表达能力和区分度，网络利用动作时间对齐模块 ATAM 对支持集和查询集的视频特征分别构建支持元组和查询元组。通过构建 $K_X\times D_X\times 1\times 1$ 的元组，确保动作在时间维度上精确对齐，有效缓解动作时间错位问题。其中 K_X 表示在 X 元组设定下得到的元组总数，D_X 表示以 X 元组构建方案得到的特征维度。

在阶段 2 中，完成时间对齐后的特征进入动作演化对齐模块 AEAM，该模块由时间同步子模块 TSM 和空间协调子模块 SCM 组成。其中，TSM 利用自注意力机制对动作演化过程进行时间同步，以调节视频动作演化的速率。SCM 利用多层感知机对输入特征进行逐点细化，提升模型的空间泛化能力。

经两阶段对齐后，由距离度量模块计算特征之间的距离，并输出 $n\times 1$ 的预测结果，其中 n 表示“n-way k-shot”小样本学习任务中的类别数量。接下来将详细介绍 TSAN 网络的技术细节。

3.1　动作时间对齐模块 ATAM

动作时间对齐模块 ATAM 主要针对动作时间错位问题而设计，模块框架如图 2 所示。

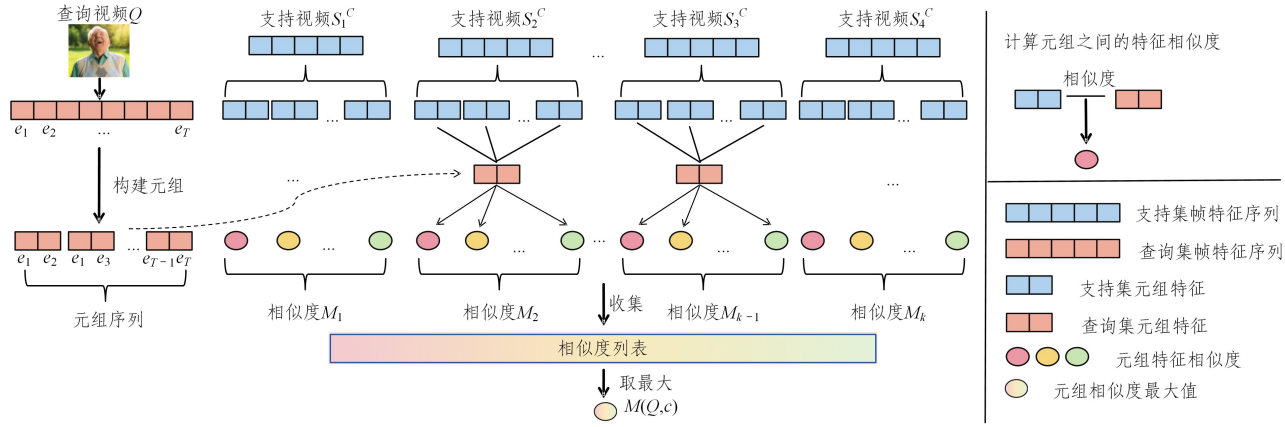


图 2　动作时间对齐模块 ATAM

Fig. 2　Action time alignment module ATAM

ATAM 模块采用元组匹配思想[15]，旨在捕捉高阶时序关系。相较于传统的逐帧匹配方法，元组匹配展现出更强的时空特征表示能力。在元组匹配中，用于编码子序列的帧是由不同基数的元组来表示的。

由特征预处理模块得到 T 帧视频特征序列，其中 $e_i\in R^D$ 表示视频第 i 帧的特征，D 表示特征的维度。由此，视频 Q 的 T 帧特征序列可以表示为：

$$(e_1,e_2,e_3,\cdots,e_T) \tag{1}$$

基于上述特征序列，构建二元组 $(e_i,e_j)\in R^{2\times D}$。其中，$(e_i,e_j)$ 表示特征 e_i 和 e_j 之间通过拼接得到的一个元组对，可以得到视频 Q 的全部二元组子序列表示：

$$[(e_1,e_2),(e_1,e_3),\cdots,(e_i,e_j),\cdots,(e_{T-1},e_T)],\quad 1\leqslant i<j\leqslant T \tag{2}$$

类似地，改变元组的基数，一个 X 元组可以表示为 $(e_i,e_j,e_k,\cdots,e_l)\in R^{X\times D}$，使得 $1\leqslant i<j<k<\cdots<l\leqslant T$。为了详细阐述 ATAM 模块的工作原理，本文选用二元组作为示例。

考虑到动作通常由视频帧的连续变化构成，用单个帧很难充分表示一个完整动作，采样较多帧虽能完整表示其动作，但会引入额外的计算成本。因此，ATAM 模块从视频中按时间顺序采样两帧来表示一个子动作。在查询视频 Q 中采样

一对有序帧构建二元组，索引为 $p=(p_1,p_2)$，其中 $1\leqslant p_1<p_2\leqslant T$，对应的视频帧表示为 $q=(q_{p1},q_{p2})$，定义该查询元组的特征表示为：

$$Q_p=[\Phi(q_{p1})+PE(P_1),\Phi(q_{p2})+PE(P_2)]\in R^{2\times D} \quad (3)$$

其中，$\Phi:R^{H\times W\times 3}\rightarrow R^D$ 表示通过一个卷积网络获得输入帧的 D 维特征，$PE(\cdot)$是给定帧索引的位置编码。

为了匹配支持集中不同速度和位置的动作，将查询表示 Q_p 与支持集视频的所有可能的元组表示进行相似度计算。通过二元组的方式定义所有可能元组的集合为：

$$\Pi=\{(n_1,n_2)\in N^2:1\leqslant n_1<n_2\leqslant T\} \quad (4)$$

因此，查询中的所有二元组表示的集合为：

$$Z^Q=\{Z_p^Q:(p\in\Pi)\} \quad (5)$$

对于支持集视频 $t(1\leqslant t\leqslant k)$在类别 $c(1\leqslant c\leqslant n)$中关于有序索引对 $m=(m_1,m_2)\in\Pi$ 的二元组特征表示为：

$$S_{tm}^c=[\Phi(s_{tm_1}^c)+PE(m_1),\Phi(s_{tm_2}^c)+PE(m_2)]\in R^{2\times D} \quad (6)$$

类别 c 的支持集中所有二元组表示的集合为：

$$S^c=\{S_{tm}^c:(1\leqslant t\leqslant k)\wedge(m\in\Pi)\} \quad (7)$$

接下来，对于查询视频 Q 中的每个二元组特征 Z^Q，计算其在动作类别 c 的 k 个支持视频中所有元组的最高相似度。通过对 Q 中所有二元组的最高相似度得分进行加权平均，得到当前查询视频与类之间的相似度 $M(Q,c)$。

$$M(Q,c)=\psi\left[\frac{1}{|\Pi|}\sum_{r}^{|\Pi|}\sum_{j}^{k|\Pi|}\max\varphi(Z_r^Q,S_j^c)\right] \quad (8)$$

其中，$\varphi(\cdot,\cdot)$为欧氏距离函数，ψ 表示 softmax 归一化操作，$|\Pi|$表示元组的数量。

ATAM 模块包含两个输出，其一为类相似度 $M(Q,c)$，该值随后在距离度量模块中参与计算；其二是由支持集与查询集共同构建的元组序列，作为 AEAM 模块的输入进行后续的数据处理与分析。

对于查询视频 Q 中的每个二元组特征 Z^Q，需要计算其与支持集中 k 个支持视频的最高元组相似度，这涉及计算欧氏距离并找到最高值。对于查询中的每个二元组，需要与支持集中 $k\cdot|\Pi|$个二元组计算相似度。计算两个 D 维向量之间的欧氏距离的复杂度为 $O(D)$。因此，每个查询二元组的时间复杂度为 $O(k\times|\Pi|\times D)$。计算所有查询二元组的时间复杂度时，由于有 $|\Pi|$ 个查询二元组，因此总复杂度为 $O(k\times|\Pi|^2\times D)$。

3.2 动作演化对齐模块 AEAM

动作演化错位源于视频中动作的非线性演化特性，包括视频动作的演化速度（如用 3 秒完成击球/用 6 秒完成击球）和空间状态（如动作靠近镜头/动作远离镜头），这使得线性的 ATAM 模块在进行处理时存在局限性。为此，设计动作演化对齐模块 AEAM。该模块包含时间同步子模块 TSM 和空间协调子模块 SCM。TSM 子模块采用自注意力机制[19]，通过聚合元组上下文特征来生成全局融合特征，调节视频动作演化的速率。SCM 子模块通过构建多层感知机，实现对输入特征的逐点细化，从而生成空间信息更丰富的元组级特征。

TSM 子模块如图 3 所示。令 $x_i\in R^{2\times D}$ 表示经 ATAM 模块构建的支持集和查询集元组序列。

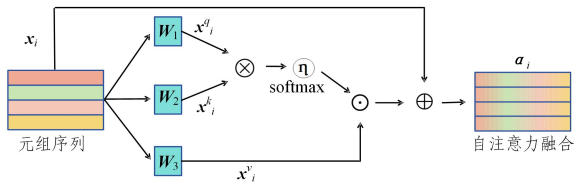


图 3 时间同步模块 TSM

Fig. 3 Temporal synchronization module TSM

首先通过权重 $W_1,W_2,W_3\in R^{D\times D}$ 对这些元组特征投影得到查询向量 x_i^q、键向量 x_i^k 和值向量 x_i^v。

$$x_i^q=x_iW_1,x_i^k=x_iW_2,x_i^v=x_iW_3 \quad (9)$$

当值向量保持当前元组的状态时，查询向量和键向量通过点积计算当前两个元组之间的相似度权重分数。每个值向量乘以它的权重分数并求和，得到当前元组对应的自注意力融合特征 α_i，给定：

$$\alpha_i=\psi\left(\frac{x_i^qx_i^{kT}}{\sqrt{D}}\right)x_i^v+x_i \quad (10)$$

其中，ψ 为 softmax 函数，D 为输入特征的维度。

TSM 子模块采用了自注意力机制，需要计算序列中每个元素与序列中所有其他元素的点积，假设输入序列的长度为 n，D 为输入特征的维度，则时间复杂度为 $O(n^2\times D)$。

SCM 子模块如图 4 所示。TSM 子模块得到了自注意力融合的元组特征，接下来通过 SCM 子模块来丰富元组的空间特征。

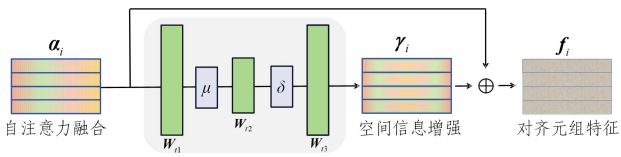


图 4 空间协调模块 SCM

Fig. 4 Spatial coordination module SCM

为了降低网络复杂度并提升泛化能力，SCM 子模块采用包含 3 个全连接层的瓶颈结构，首先将融合特征 $\alpha_i\in R^{2\times D}$ 输入 SCM 子模块，其中第一个全连接层起到降维的作用，降维系数 r 是超参数；然后采用 ReLU 函数激活，得到激活后的特征 $\beta_i\in R^{2\times D}$。

$$\beta_i=\mu(\alpha_i^TW_{t1}) \quad (11)$$

其中，μ 为 ReLU 激活函数，W_{t1} 为第一个全连接层的权重矩阵。

然后通过第二个全连接层，对这些激活后的特征 β_i 进行逐点细化，再使用 Sigmoid 激活函数激活。之所以选择使用两种不同的激活函数，是为了结合它们的优点，使网络在处理复杂特征关系时更加灵活和有效。最后的全连接层恢复特征原始的维度，并输出空间信息丰富的特征 $\gamma_i\in R^{2\times D}$。

$$\gamma_i=\{\delta[(\beta_iW_{t2})]\}W_{t3} \quad (12)$$

其中，δ 为 Sigmoid 激活函数，W_{t2} 和 W_{t3} 为第二个和第三个全连接层的权重矩阵。

将学习到的各个元组的激活特征 γ_i 加上原始特征 α_i，得到对齐后的特征 f_i，由此丰富了基于视频帧的元组空间特征的泛化能力。

$$f_i=\gamma_i+\alpha_i \tag{13}$$

SCM 子模块引入了多层感知机以增强模型的泛化性能。其中，每个全连接层的神经元均执行对输入信号的加权求和操作，该操作属于线性变换，其时间复杂度为 O(1)。由于多层感知器的每一层都包含多个这样的神经元，因此 SCM 子模块的时间复杂度主要由各层神经元的数量决定。假设第一、第二和第三个全连接层神经元数量分别为 n_1，n_2 和 n_3，则整体时间复杂度为 $O(n_1\times n_2)+O(n_2\times n_3)$。

3.3　距离度量

经两阶段对齐后，需要计算支持集与查询集的元组特征间的距离，从而输出分类结果。

使用 f_i^S 表示属于类别 c 的支持视频 S 对齐后的元组特征，f_i^Q 表示查询视频 Q 对齐后的元组特征，首先对每个 f_i^S 与 f_i^Q 计算特征层面的平均距离，以量化两者之间的差异性。随后，再乘以动作时间对齐模块的类相似度 $M(Q,c)$，其反映了在动作时间对齐条件下，支持视频与查询视频的相似性程度。将两阶段时空对齐特性融合，计算分类概率 $P_{(Q\in c)}$：

$$P_{(Q\in c)}=\psi\left\{\left[\frac{1}{|\Pi|}\sum_i^{|\Pi|}\phi(f_i^S,f_i^Q)\right]\times M(Q,c)\right\} \tag{14}$$

其中，ψ 为 softmax 归一化函数，$\varphi(\cdot,\cdot)$ 为欧氏距离函数，$|\Pi|$ 表示元组的数量，$M(Q,c)$ 表示查询视频 Q 与类别 c 之间的相似度。

由此，与每个类别的支持视频计算分类概率后再取均值，得到当前查询视频的 n 个类概率，取类概率的最大值作为网络最终的预测结果。

对于任意给定的类别 c，需评估支持视频集合 S 与查询视频 Q 之间在特征空间上的平均距离。具体而言，每个视频均被表征为对齐后的元组特征集合，其中每个元组特征的维度均为 D。计算两个 D 维向量间欧氏距离的时间复杂度为 $O(D)$。在此基础上，为了获取平均距离，需针对每个类别 c，计算其内部所有元组特征对之间的距离，并对这些距离求均值，该步骤的时间复杂度为 $O(|\Pi|\times D)$，其中 $|\Pi|$ 代表需计算的元组特征对数量。鉴于总共有 n 个类别需进行此类计算，因此，整体的时间复杂度可表示为 $O(n\times|\Pi|\times D)$。

4　实验

4.1　数据集

为了检验 TSAN 网络的有效性，选取 HMDB51，UCF101，Something-Something-V2 和 Kinetics 这 4 个数据集进行性能评估。其中，HMDB51 主要收集了电影作品中的视频片段，涵盖了 51 个动作类别，每个动作类别至少包含 101 个视频，总计 6 849 个视频，分辨率为 320×240，完整的数据集大小约 2GB。该数据集的数据差异性主要体现在对象外观和人体姿态的变化上，动作类型涵盖了面部动作、身体动作以及人与对象的交互等多种类型。该数据集包含复杂画面背景、低光度等场景，可以检验 TSAN 网络在复杂环境中的有效性。

UCF101 包含 101 个动作类别，共计 13 320 个视频，平均视频长度为 7.21 秒。每个动作类别由 25 组不同的人员完成，同一组内的视频在背景、人物等方面具有一定的相似性。不同组之间的视频在各种不受约束的环境下录制，包括相机运动、不同光照条件、部分遮挡和低质量帧等。该数据集由用户拍摄并上传的视频组成，视频背景波动较小，可以检验 TSAN 网络在处理实际拍摄视频任务时的有效性。

Something-Something-V2 由大量的众包工作者创建，包含 174 个动作类别共 220 847 个视频。该数据集需要进行时间推理，其中大多数行为不能仅根据空间特征来推断（如开门/关门、抛起/坠落等）。从中随机抽取 100 个类别，每个类别随机选择 100 个视频，组成一个名为 SSV2100 的子数据集，用于小样本视频行为识别的训练及测试。该数据集的视频内容复杂，需要网络具有较强的时序建模能力，可以检验 TSAN 网络的时间推理能力。

Kinetics 由 400 个动作类别共 306 245 个视频组成，每个动作类别至少包含 400 个视频片段，每个片段的长度约为 10 秒。视频主要来源于 YouTube，视频背景相对固定，涵盖了人与物体、人与人等多种交互动作。在实验中，具体使用的是 Kinetics100 子数据集，即从 Kinetics400 中选取 100 个类别，每个类别取 100 个视频进行模型的训练和测试。该数据集的视频类间差异较大，视频时序性强，是比较综合的数据集，可以检验 TSAN 网络的泛化能力。

对于 HMDB51 和 UCF101，遵循 ARN[6] 中的划分标准，将数据集划分为训练集、验证集和测试集。而对于 Kinetics100 和 SSV2100，则按照 CMN[4] 中的划分标准进行数据集的划分。数据集的详细信息如表 1 所列。

表 1　实验数据集

Table 1　Experimental datasets

数据集	总类别(类)	训练集(类)	验证集(类)	测试集(类)
HMDB51	51	31	10	10
UCF101	101	70	10	21
SSV2100	100	64	12	24
Kinetics100	100	64	12	24

4.2　评价指标

使用“n-way k-shot”概念对小样本学习任务进行实验评估。其中，n 代表每轮次训练中使用的样本类别数，k 代表每个类别中选取的样本数量。实验方案包括“5-way 1-shot”和“5-way 5-shot”。这两种方案均从测试集中抽取 5 个类别的数据构成支持集，其中 1-shot 指的是每个类别仅选取 1 个样本，而 5-shot 则表示每个类别选取 5 个样本。除了支持集，还需要选取查询集，其类别与支持集一致，且二者包含的样本无交集，在每个类中选取 1 个样本作为查询视频。由支持集和查询集组成一个情节，用于对网络性能进行评估。

在最终评估查询集预测结果时，采用准确率作为判定标准。为增强实验结果的可靠性，进行 10 000 次重复实验并计算平均准确率，取 95%置信区间。

4.3　训练设置

在实验中，对于一个“n-way k-shot”训练任务，随机采样 15 000个情节来训练 TSAN 网络。视频预处理过程采用 TSN[26] 中介绍的方法。在训练过程中，首先将视频中的每一帧大小调整为 256 × 256，然后从视频片段中随机裁剪一个

224×224 的区域，再对每段视频进行 $T=8$ 的稀疏均匀采样。特征预处理阶段的 $D\times H\times W$ 取值为 2 048×7×7。对于 HMD51，UCF101 及 Kinetics 数据集，在训练时随机施加水平翻转。由于 Something-Something V2 数据集中的标签包含了左右概念，如从左向右拉某物、从右向左拉某物等，因此对该数据集不使用水平翻转。

根据 CMN[4] 的实验设置，采用 ResNet50[18] 作为 TSAN 的骨干网络。使用在 ImageNet 上预训练后的权重初始化网络。采用 SGD 来优化 TSAN 网络，起始学习率为 0.001，每 30 个迭代点衰减 0.1。SCM 子模块中，降维系数 r 取 0.5。使用验证集对参数进行调优，当验证集的准确率即将下降时，停止训练。使用 PyTorch 实现 TSAN 网络训练框架。

为提升训练效率，使用两种硬件设备以适应不同实验的算力需求。“5-way 1-shot”实验在 Ubuntu18.04 服务器(GPU：NVIDIA RTX 2080 11 GB)上进行，全模型在 2 张 NVIDIA-SMI 2080 GPU 上训练 1 小时。“5-way 5-shot”实验在 Ubuntu18.04 服务器(GPU：NVIDIA RTX 3090 24 GB)上进行，全模型在 1 张 NVIDIA-SMI 3090 GPU 上训练 3 小时。

4.4 实验比较与分析

为了验证 TSAN 网络的识别性能，选用 HMDB51，UCF101，SSV2100 和 Kinetics100 数据集与以下模型进行比较。

TSN[20]：引入时序分段网络结构，通过均匀采样视频并分别对每个时间分段进行特征提取和分类。

TARN[5]：提出注意力关系网络，通过计算查询与支持视频在视频片段级别的关系分数进行分类。

ARN[6]：通过二次加权时空特征生成注意力掩码，并利用自监督学习的增强策略来增强其特征编码器和注意力机制。

OTAM[7]：设计了有序时间对齐模块，采用动态时间规划算法的变体来显式地对齐视频序列。

PAL[11]：包含以原型为中心的对比学习损失和混合注意力学习机制，解决了类间重叠和外围孤立样本的问题。

MoLo[12]：运用长短对比目标策略增强局部帧特征，并通过运动自解码器提取运动线索，实现对全局信息的感知。

TRX[15]：利用 Transformer 和交叉自注意力机制来捕捉视频序列中不同时间点之间的复杂关系。

STRM[16]：通过聚合空间上下文的局部块级特征和时间上下文的全局帧特征来捕捉视频中物体的外观和运动。

TSA-MLT[17]：通过在帧序列进行时间维度上的仿射变换，过滤关键信息较少的帧或可能误导视频语义的信息，并对不同基数的元组特征进行融合。

如表 2 所列，TSAN 网络在所有数据集上的识别准确率表现均优于对比模型。针对动作时间和动作演化两种错位，TSAN 网络以元组为单元同时处理时间和空间信息，基于元组注意力的时间同步模块配合增强非线性表达能力的空间协调模块，进一步增强时空特征融合效果，并成功捕获了数据中的动态时空特征。与具有良好识别准确率的 MoLo 和 TSA-MLT 相比，TSAN 网络的参数量更小，有利于实现更为轻量级的模型设计。与 TSA-MLT 相比，在“5-way 5-shot”设定下，TSAN 网络在 UCF101 和 Kinetics100 数据集上，识别准确率分别提高 2.4%和 3.1%；在视频场景更为复杂的 HMDB51 和 SSV2100 数据集上，识别率分别提高 1.7%和1.6%。

表 2 不同网络的识别准确率对比

Table 2 Comparison of recognition accuracy between different networks

方法	Backbone	Params	HMDB51		UCF101		SSV2100		Kinetics100	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
TSN[20]	ResNet50	**23.51×10^6**	48.6±0.4	57.2±0.4	69.2±0.3	78.6±0.3	30.6±0.4	38.7±0.4	58.9±0.4	67.0±0.4
TARN[5]	C3D	34.80×10^6	—	—	—	—	—	—	66.6±0.4	80.7±0.4
ARN[6]	C3D	34.80×10^6	44.6±0.4	59.1±0.4	62.1±0.4	84.8±0.4	—	—	63.7±0.4	82.4±0.4
OTAM[7]	ResNet50	**23.51×10^6**	50.7±0.4	64.8±0.4	76.3±0.4	86.4±0.3	36.1±0.4	47.8±0.4	68.7±0.4	76.9±0.3
PAL[11]	ResNet50	40.29×10^6	48.4±0.4	62.3±0.4	65.7±0.4	81.0±0.3	29.6±0.4	44.4±0.4	54.7±0.4	70.8±0.4
TRX[15]	ResNet50	47.12×10^6	49.5±0.4	70.1±0.4	79.3±0.3	93.3±0.2	34.4±0.4	52.3±0.4	69.4±0.4	84.8±0.4
STRM[16]	ResNet50	146.23×10^6	50.0±0.4	71.6±0.4	78.9±0.3	91.7±0.3	36.5±0.4	51.6±0.4	68.4±0.4	84.2±0.4
MoLo[12]	ResNet50	135.71×10^6	51.1±0.4	71.2±0.4	79.2±0.3	92.6±0.3	36.4±0.4	52.1±0.4	68.9±0.4	84.7±0.4
TSA-MLT[17]	ResNet50	122.40×10^6	52.4±0.4	72.3±0.4	80.6±0.4	93.3±0.2	35.6±0.4	52.6±0.4	67.7±0.4	85.6±0.4
TSAN	ResNet50	73.85×10^6	**54.1±0.4**	**74.0±0.4**	**81.8±0.3**	**95.7±0.3**	**38.4±0.4**	**54.2±0.4**	**70.3±0.4**	**88.7±0.4**

4.5 消融实验

实验 1 在上述实验基础上，为了进一步检验 TSAN 网络的泛化能力，以及 ATAM 和 AEAM 模块的有效性，设计了实验 1。考虑到 UCF101 和 Kinetics100 数据集均有较为丰富的视频类别，且其中的视频背景波动小，对动作捕获影响较小，能更好地验证 TSAN 网络模块对动作提取的有效性，因此选择在这两个数据集上开展实验。所有实验结果均基于“5-way 5-shot”设定。

如表 3 所列，消融实验的结果展示了 ATAM 和 AEAM 在提升小样本动作识别准确率方面的积极作用。具体而言，在应用 ATAM 模块后，模型在 UCF101 和 Kinetics100 数据集上的识别准确率分别提高了 10.1%和 9.7%，表明了 ATAM 模块在增强模型时间对齐能力方面的有效性。

表 3 TSAN 网络各模块消融实验

Table 3 Ablation experiments of each module in TSAN model

ATAM	AEAM-TSM	AEAM-SCM	UCF101	Kinetics100
			78.6	67.0
√			88.7	76.7
√	√		93.6	86.4
√		√	90.2	77.3
	√		82.5	72.6
	√	√	85.2	76.4
√	√	√	**95.7**	**88.7**

在应用 ATAM 模块的基础上，进一步对比了单独添加 AEAM 中的时间同步模块 TSM（AEAM-TSM）和空间协调模块 SCM（AEAM-SCM）的性能差异。实验结果表明，单独添加 TSM 模块相较于单独添加 SCM 模块，在 UCF101 和 Kinetics100 数据集上的识别准确率提升更为明显，分别提高了 15.0%和 19.4%，说明了 TSM 在模型性能提升中的关键作用。

最后，当 ATAM 和 AEAM 以两阶段的方式共同应用于模型时，识别准确率得到了进一步提升，分别在 UCF101 和 Kinetics100 数据集上提高了 17.1%和 21.7%。这一结果支持了本文所提出的两阶段模型结构设计思想。

实验 2　元组数量的不同，对时序建模能力的要求也不同。为检验不同元组基数对 ATAM 性能的影响，设计了实验 2。考虑到 SSV2100 和 Kinetics100 数据集的视频类别较为丰富，视频内容复杂且具有更强的时序特点，能够更好地体现不同元组基数对性能的影响，因此选择在这两个数据集上开展实验。所有实验结果均基于“5-way 5-shot”设定。

如表 4 所列，在 SSV2100 数据集上，可以观察到从单帧匹配到二元组匹配的显著性能提升（4.3%）。随着元组基数的增加，三元组相较于二元组表现出进一步的性能提升（1.8%），而四元组相较于三元组则仅展现出微小的性能增长（0.4%）。这一趋势表明，随着元组基数的增加，性能提升逐渐趋于平缓。

表 4　ATAM 元组基数消融实验

Table 4　ATAM tuple cardinality ablation experiment

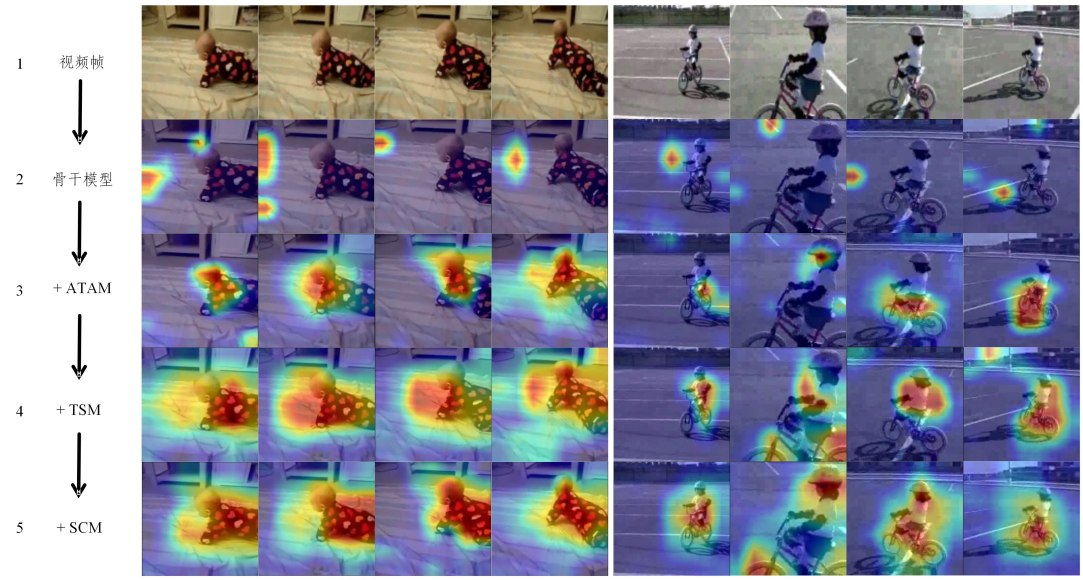
基数	元组数量	SSV2100	Kinetics100
$\Omega=\{1\}$	—	46.6	85.3
$\Omega=\{2\}$	28	50.9	87.7
$\Omega=\{3\}$	56	52.7	88.4
$\Omega=\{4\}$	70	53.1	88.3
$\Omega=\{2,3\}$	84	**54.2**	**88.7**
$\Omega=\{2,4\}$	98	53.3	88.4
$\Omega=\{3,4\}$	126	51.0	88.6
$\Omega=\{2,3,4\}$	154	51.1	88.0

在对比不同元组组合时，可以观察到，组合使用二元组和三元组时，即 $\Omega=\{2,3\}$，在 SSV2100 数据集上达到了最佳性能。这一结果揭示了二元组和三元组在 ATAM 模块中配合使用的积极作用，共同为模型提供了更丰富的时空信息，从而提高了动作识别的准确性。

然而，当组合使用二元组、三元组与四元组时，即 $\Omega=\{2,3,4\}$，性能却出现了下降（−3.1%）。这可能是因为使用四元组引入了过多信息导致了模型的过拟合或信息冗余，从而影响了模型的泛化能力。相较于 SSV2100 数据集，Kinetics100 数据集上的性能差异较小。与最佳元组组合 $\Omega=\{2,3\}$ 相比，$\Omega=\{2,3,4\}$ 在 Kinetics100 上的性能仅下降了 0.7%。这可能是由于 Kinetics100 数据集本身的复杂性较高，使得 ATAM 模块在不同元组基数下的性能差异相对较小。

4.6　可视化分析

图 5 展示了 TSAN 网络在 UCF101 数据集上两个随机选择样例的注意力图，通过逐步集成所提出的模块，揭示了模型关注区域的变化过程。从原始视频帧（第 1 行）开始，逐步引入不同的模块来观察模型对关键区域的聚焦能力。第 2 行展示了仅包含 ResNet50 骨干模型的注意力图，它提供了基本的关注区域分布，但未能充分捕捉到与动作紧密相关的信息。第 3 行展示了集成元组匹配思想的 ATAM 模块的注意力图。通过引入 ATAM 模块，模型能够在骨干模型的基础上更加精确地关注到运动对象，如在样例（a）的第 2 帧和第 4 帧中，模型对运动对象的关注明显增强。第 4 行则展示了集成 TSM 子模块后的注意力图。TSM 子模块通过引入元组注意力机制，进一步编码了时间上下文信息。在样例（a）中，第 1 帧和第 3 帧的对象运动导致 ATAM 模块遗漏部分区域时，TSM 子模块能够有效地捕捉这些遗漏信息，从而提升模型对时间上下文的感知能力。第 5 行展示了完整 TSAN 网络的注意力图。SCM 子模块的引入，进一步增强了特征的可区分性，提高了模型对对象的关注度。例如在样例（b）中，第 3 帧和第 4 帧中对自行车对象的关注度得到了显著提升。



（a）BabyCrawling　　（b）Biking

图 5　样例可视化分析

Fig. 5　Sample visualization analysis

结束语 本文针对小样本视频行为识别中的时空错位问题，提出了两阶段时空对齐网络 TSAN。该网络能够充分利用有限的视频数据，有效应对视频中复杂的非线性时空变化。TSAN 网络主要由动作时间对齐模块 ATAM 和动作演化对齐模块 AEAM 组成。ATAM 引入元组匹配思想，在保持时间顺序的同时动态地对齐两个视频序列，解决了动作时间错位问题。AEAM 进一步分为时间同步模块 TSM 和空间协调模块 SCM 两个子模块，分别从时间和空间两个维度对动作演化进行对齐，能有效应对视频中复杂的时空演化错位。当 ATAM 和 AEAM 以两阶段的方式共同应用于模型时，显著提升了模型的学习效率和识别性能。实验结果表明，TSAN 网络的识别准确率优于对比模型。未来的工作中，将在此基础上进一步探索元组的组合策略，从而消除模型过拟合或信息冗余，提升模型的泛化能力。同时，结合城市管理、景区安全等实际场景，开展网络模型轻量化设计与应用。

参 考 文 献

[1] SHENG X X, LI K C, SHEN Z Q, et al. A Progressive Difference Method for Capturing Visual Tempos on Action Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(3): 977-987.

[2] COSKUN H, ZIA Z, TEKIN B, et al. Domain-Specific Priors and Meta Learning for Few-Shot First-Person Action Recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 6659-6673.

[3] WANG P, LI H B, ZHANG B X, et al. Metric-based few-shot learning method for driver distracted behaviors detection[C]// 2023 International Conference on Image Processing Computer Vision and Machine Learning(ICICML). IEEE, 2023: 959-963.

[4] ZHU L, YI Y. Compound Memory Networks for Few-Shot Video Classification[C]// Proceedings of the European Conference on Computer Vision. 2018: 751-766.

[5] BISHAY M, ZOUMPOURLIS G, PATRAS I. TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition[C]// British Machine Vision Conference. 2019: 154-168.

[6] ZHANG H, ZHANG L, QI X, et al. Few-shot Action Recognition with Permutation-invariant Attention[C]// Proceedings of the European Conference on Computer Vision. 2020: 525-542.

[7] CAO K, JI J, CAO Z, et al. Few-Shot Video Classification via Temporal Alignment[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2020: 10615-10624.

[8] FU Y, ZHANG L, WANG J, et al. Depth Guided Adaptive Meta-Fusion Network for Few-shot Video Recognition[C]// Proceedings of the 28th ACM International Conference on Multimedia. ACM, 2020: 1142-1151.

[9] NI X Z, WEN H, LIU Y, et al. Multimodal Prototype-Enhanced Network for Few-Shot Action Recognition[C]// Proceedings of the 2024 International Conference on Multimedia Retrieval. ACM, 2020: 1-10.

[10] DWIVEDI S K, GUPTA V, MITRA R, et al. ProtoGAN: Towards Few Shot Learning for Action Recognition[C]// International Conference on Computer Vision Workshop (ICCVW). IEEE, 2019: 1308-1316.

[11] ZHU X, TOISOUL A, PEREZ-RUA J M, et al. Few-shot Action Recognition with Prototype-centered Attentive Learning[C]// British Machine Vision Conference. 2021: 249-259.

[12] WANG X, ZHANG S, QING Z, et al. MoLo: Motion-augmented Long-short Contrastive Learning for Few-shot Action Recognition[C]// Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2023: 18011-18021.

[13] WANG X, ZHANG S W, CEN J, et al. CLIP-guided Prototype Modulatingfor Few-shot Action Recognition[J]. International Journal of Computer Vision, 2024, 132(6): 1899-1912.

[14] LI S, LIU H, QIAN R, et al. TA2N: Two-Stage Action Alignment Network for Few-shot Action Recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 36(2): 1404-1411.

[15] PERRETT T, MASULLO A, BURGHARDT T, et al. Temporal-RelationalCrossTransformers for Few-Shot Action Recognition[C]// Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2021: 475-484.

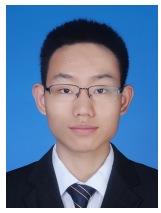
[16] THATIPELLI A, NARAYAN S, KHAN S, et al. Spatio-temporal Relation Modeling for Few-shot Action Recognition[C]// Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 19926-19935.

[17] GUO F, ZHU L, WANG Y K, et al. Task-Specific Alignment and Multiple Level Transformer for Few-Shot Action Recognition[J]. Neurocomputing, 2024, 32(5): 598-612.

[18] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// Conference on Computer Vision and Pattern Recognition(CVPR). IEEE, 2016: 770-778.

[19] HAKIM N, INSAF B, HASSAN S. Improving Human Action Recognition in Videos with Two-Stream and Self-Attention Module[C]// Colloquium in Information Science and Technology. IEEE, 2023: 215-220.

[20] WANG L, XIONG Y, WANG Z, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition [C]// Proceedings of the European Conference on Computer Vision. 2016: 20-36.



WANG Jia, born in 1996, postgraduate. His main research interests include deep learning and video action recognition.

XIA Ying, born in 1972, professor, Ph.D supervisor. Her main research interests include spatio-temporal big data and cross-media retrieval.

（责任编辑：喻藜）