DOI: 10.16725/j.1673-808X.202359
引用本文: 赵胤铎, 翟仲毅, 赵岭忠. 基于时序增强多模态注意力网络的行为识别模型[J]. 桂林电子科技大学学报, 2026, 46(3): 278-283.
Zhao Yinduo, Zhai Zhongyi, Zhao Lingzhong. Enhanced temporal multimodal attention network for action recognition[J]. Journal of Guilin University of Electronic Technology, 2026, 46(3): 278-283.

基于时序增强多模态注意力网络的行为识别模型

赵胤铎，翟仲毅，赵岭忠

（桂林电子科技大学 计算机与信息安全学院, 广西 桂林 541004）

摘 要： 行为识别是视觉智能的主流应用场景之一, 其比静态图像分类更关注时间特征。目前, 许多基于 2D 卷积神经网络的轻量级行为识别模型被提出, 然而, 这些模型未能很好地捕捉到运动特征的时间关系。为解决此问题, 提出了一种时序增强多模态注意力网络 E-TMAN, 通过两阶段结构充分构建时间特征。E-TMAN 包括局部运动增强(LME)和差分时间交互(DTI)2 个主要模块。LME 负责利用时间注意力机制将时间信息融合到图像特征中, 以增强短期运动特征的表达能力。DTI 采用差分范式学习长期的时间特征, 采用了 2 种注意力机制, 即通道注意力和空间注意力。这 2 种注意力机制通过不同的形式对语义特征进行补充。最后, 在 3 个公共数据集(SomethingV2、UCF101 和 HMDB51)上进行实验来评估 E-TMAN。实验结果表明, 该模型能够有效利用时间特征, 与现有方法相比, 能够以较低的开销获得较好的性能。

关键词： 行为识别; 注意力机制; 2D 卷积神经网络; 深度神经网络; 深度学习
中图分类号： TP391.4 **文献标志码：** A **文章编号：** 1673-808X(2026)03-0278-06

Enhanced temporal multimodal attention network for action recognition

ZHAO Yinduo，ZHAI Zhongyi，ZHAO Lingzhong

（School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China）

Abstract: Action recognition is one of the major application areas of computer vision and focuses more on temporal dynamics than static image classification. Recently, many lightweight temporal modeling methods based on 2D convolutional neural networks (CNN) have been proposed for action recognition. However, these models often fail to effectively capture long-range temporal dependencies in motion sequences. To address this limitation, an enhanced temporal multimodal attention network (E-TMAN) is proposed to model temporal features through a two-stage architecture. E-TMAN consists of two main modules: local motion enhancement (LME) and difference time interaction (DTI). The LME module employs a temporal attention mechanism to fuse temporal information into spatial features, thereby enhancing the representation of short-term motion patterns. The DTI module adopts a differential learning strategy to capture long-term temporal dependencies through two complementary attention mechanisms, namely channel attention and spatial attention. By integrating temporal and contextual information in different forms, the proposed network achieves more effective temporal feature modeling. Extensive experiments are conducted on three public benchmarks: SomethingV2, UCF101 and HMDB51. The results show that E-TMAN effectively exploits temporal features and achieves competitive performance with lower computational cost compared with state-of-the-art methods.

Key words: action recognition; attention mechanism; 2D CNN; deep neural network; deep learning

近年来, 视频数据爆发式增长。基于大数据与深度学习的视觉智能方法受到了诸多研究者的关注。随着深度学习的普及, 许多基于卷积神经网络 (convolutional neural networks, 简称 CNN) 的网络架构被应

收稿日期: 2023-08-20
基金项目: 国家自然科学基金 (62262009, 61902086)
通信作者: 翟仲毅(1986—), 男, 副研究员, 博士, 研究方向为物联网服务智能与安全、边缘计算与智能、行为识别。E-mail: zhaizhongyi@guet.edu.cn

用于行为识别领域。学者们[1-5]都采用这种方法来处理视频任务, 在 CNN 出色的空间建模能力支持下, 这些方法侧重于提取视频片段的时间特征。

3D CNN[2,5,6]和双流网络[1,7,8]是 2 种经典的行为识别架构。其中, 3D CNN 通过提取显式空间关系和隐式时间关系来捕捉行为特征。基于 3D CNN 的方法在行为识别任务中表现出色, 然而其显著缺点是浮点运算量高、参数多, 进而导致计算开销很大, 且模型训练困难。双流网络利用堆叠光流提取时间特征, 利用 2D CNN 从 RGB 帧中提取空间特征。然而, 由于光流计算复杂, 双流网络的计算开销同样巨大。

为了以较小的计算代价提取时间特征, 一些基于 2D CNN 的模型被提出。例如, Wang 等[9]提出了一种分段采样方法来处理持续时间不同的运动; Lin 等[10]提出了 TSM, 以特征移位的方式来实现高效的行为识别, 在边缘设备上具有优异的性能; Wang 等[11]和 Jiang 等[12]通过独特的模块设计显式地提取时间信息, 从而以较低的开销获得良好的性能。在这些模型中, 通常采用固定滑动窗口机制来获得动态过程中的运动图像[10]。若采用稀疏采样的方式构建滑动窗口, 则会丢失一些局部信息。相反, 密集采样模型开销大, 且易丢失全局信息。此外, 由于视频剪辑中背景的时间冗余, RGB 帧的堆叠可能会导致模型过度关注背景信息而非运动特征。时序差分结构可以解决上述问题。

目前, 一些研究[11,13]利用残差帧来补充运动特征。具体来说, 利用残差帧来消除背景干扰, 突出运动物体的核心特征。相对于光流法, 残差帧能以较小的计算代价提取局部时间特征。然而, 这些方法通常以堆叠残差帧的方式来描述行为, 无法充分有序地提取时间特征。此外, 随着时间跨度的增加, 帧间差异会逐渐增大, 残差帧难以捕捉长期运动特征。

为了解决上述问题, 提出了一种时序增强多模态注意力网络 (enhanced temporal multimodal attention network, 简称 E-TMAN), 如图 1 所示。E-TMAN 旨在利用多重注意力机制显式捕获时间特征, 以构建完整的运动信息。E-TMAN 主要包括局部运动增强和差分时间交互 2 个核心模块。局部运动增强模块负责融合视频片段的局部运动信息, 它采用时间注意力机制来增强运动的时间特征。差分时间交互模块负责提取视频片段中的长期时间特征, 将段间信息引入模型深层, 通过注意力的方式增强运动的长期时间关系。

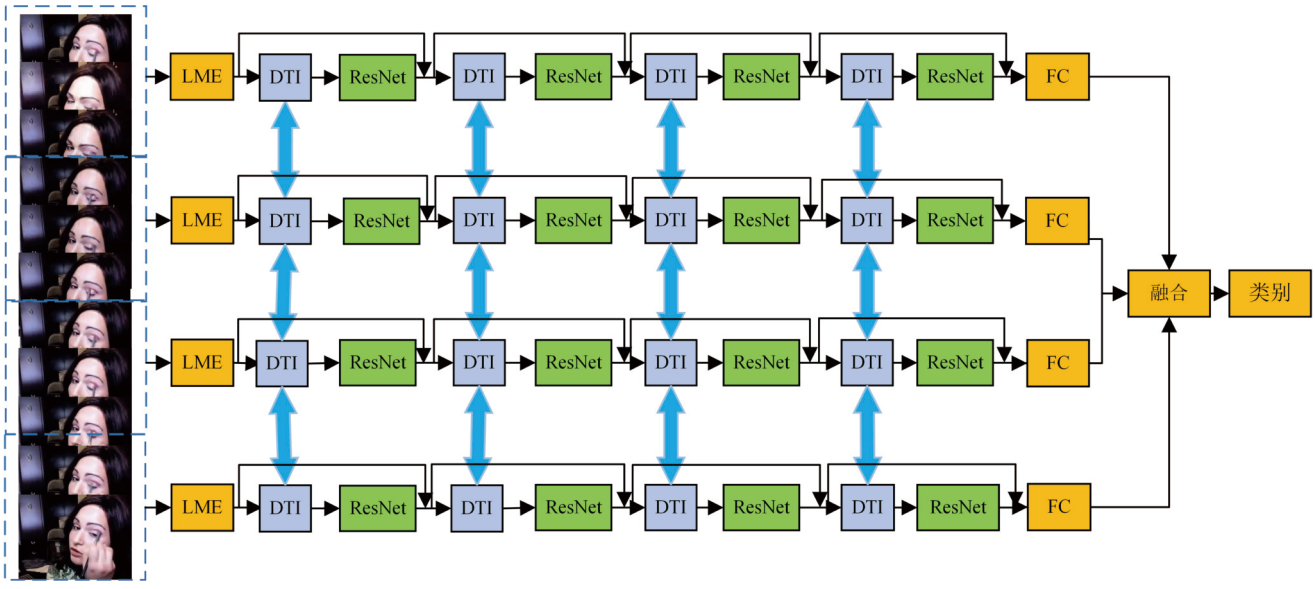

图 1　时序增强多模态注意力网络框架

具体来说, 局部运动增强模块以 RGB 帧和残差帧作为输入。RGB 帧主要提取运动完整的空间特征, 残差帧用来增强局部运动特征。差分时间交互通过特征差分注意力, 即通过通道注意力和空间注意力来融合全局运动信息。在计算空间注意力时, 提出一种轻量级空间注意力计算方法。实验结果表明, 该方法具有良好的性能和较低的开销。

1　基于卷积神经网络的视频行为识别相关研究

1.1　行为识别

2D 卷积近年来被广泛应用于行为识别[1,4,9-11]。与 3D 卷积结构相比, 2D 模型的参数更少, 计算效率更高。双流网络[7-8]利用堆叠光流或运动向量提取时

间特征，利用 2D CNN 从 RGB 帧中提取空间特征。由于光流计算复杂，双流网络的开销也很大。Wang 等[9] 提出时间分段网络 TSN，并设计了多种融合方法来聚合局部信息。在 TSN 基础上，Zhou 等[14] 提出了时间关系网络 TRN，通过融合采样帧之间的多尺度时间关系提高了模型性能。Shou 等[15] 利用压缩视频中的运动矢量信息而非光流进行特征提取，从而提高了计算效率。Lin 等[10] 提出了时间移位模块，通过将原始特征图上相邻时刻的特征图移位来实现高效的行为识别。Wang 等[11] 提出了时间差分网络，利用时序差分对局部信息进行堆叠，利用通道注意力提取长期信息。

3D CNN 提供了一种端到端的行为识别方法，它使用 3D 卷积内核来处理视频帧。Tran 等[2] 提出了一种 3D 卷积架构，可以直接从帧序列中提取时空特征，然而由于参数过多，模型训练困难。Carreira 等[16] 针对这个问题提出了 I3D，将二维卷积核扩展为三维卷积核，使 3D 模型可以使用 2D 模型的预训练参数，提高了训练效率。

1.2 注意力机制

注意力机制在计算机视觉[17–20] 中得到了充分运用。例如，Woo 等[17] 提出了 CBAM，采用通道注意力和空间注意力模块处理图像任务。Wang 等[18] 提出了时空激励、通道激励和运动激励 3 个注意力模块，显著提高了行为识别结果。Li 等[5] 提出了 STA 模块，增强了三维卷积的时空表达能力，在行为识别和定位方面表现良好。

2 时序增强多模态注意力网络

局部运动增强 (local motion enhancement，简称 LME) 模块主要关注运动本身的变化，而非背景的变化。差分时间交互 (difference time interaction，简称 DTI) 模块进一步集成长期的时序行为特征，以获取全局的上下文信息。

2.1 局部运动增强

在行为识别中，时空特征是行为分类的基础。由于缺乏时间信息，2D 卷积不能直接应用于行为识别。为此，通过提出的 LME 模块来聚合局部运动信息。LME 着重于运动特征，进一步减少了背景信息的干扰。

LME 对原始帧和堆叠残差帧并行处理。利用残差通过卷积层提取空间特征。为了降低计算复杂度，对残差帧进行堆叠，并在时间通道上进行特征增强。这些信息以残差的形式叠加在主干网络上，可使用 RGB 帧进行运动识别。局部运动增强模块如图 2 所示。

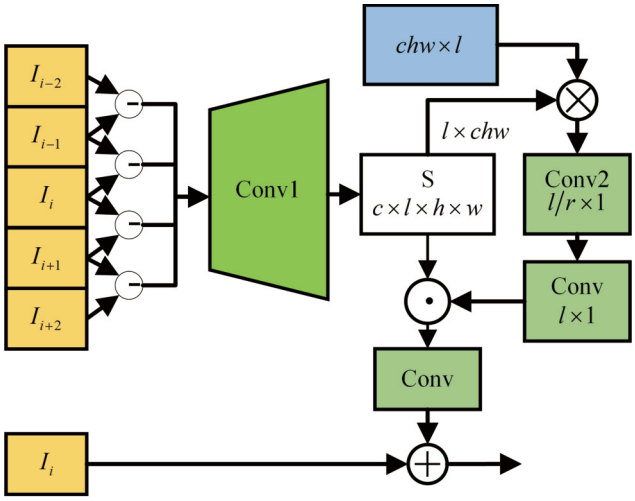

图 2 局部运动增强模块

给定一个视频序列$I_n=\{\boldsymbol{i}_k\in\mathbf{R}^{3\times h\times w},k=1,2,\cdots,l\}$，相邻帧之间存在较高的时间冗余。残差帧不仅有效解决了这一问题，也减少了背景因素对行为识别的干扰。残差帧可形式化表示为$D(I_n)=\{I_i-I_{i-1},i=1,2,\cdots,l\}$。通过堆叠残差帧的方式可以简单地表示运动的时间变化，但因为时间信息不够明确，需加入时间注意力模块。令S表示特征堆叠，则中间特征I'可表示为

$$I'=S(\text{Conv1}(D(I_i)))。\tag{1}$$

接着，I'的维度被调整为$I'\in\mathbf{R}^{n/l\times c\times l\times h\times w}$，$n$表示批大小，则时间注意力可表示为

$$l=\text{Conv2}(I'w)。\tag{2}$$

令⊙表示对应元素乘积，则残差信息可表示为

$$I=I'\odot\sigma(l),\tag{3}$$

其中，σ表示 sigmoid 激活函数。

从原始帧I_k中提取背景和行为对象的空间特征图，并将残差特征添加到网络中。LME 得到的特征F可表示为

$$F=\text{Conv}(I_k)+I。\tag{4}$$

通过这种方式，浅层视频特征包含了重要的局部运动信息，可以突出运动区域。

2.2 差分时间交互

局部运动增强模块加强了运动相关的局部窗口特征，即局部运动模式的时间演化。然而，LME 提取的浅层特征无法充分表达运动特征。因此，提出了差分时间交互模块，以低成本获取深层的时间上下文信息。

为了增强运动区域而非背景, 设计了一种利用上下文信息的空间注意力机制, 如图 3 所示。给出 2 个中间特征图 $\boldsymbol{F}_t$、$\boldsymbol{F}_{t-1}$, 其中 $\boldsymbol{F}_t \in \mathbf{R}^{c\times h\times w}$。DTI 利用 1×1 卷积, 将通道数减少到 c/r。该映射过程可表示为 $\boldsymbol{F}'_t = \text{Conv1}(\boldsymbol{F}_t)$, $\boldsymbol{F}'_{t-1} = \text{Conv1}(\boldsymbol{F}_{t-1})$, 其中 $\boldsymbol{F}'_t \in \mathbf{R}^{c/r\times h\times w}$。

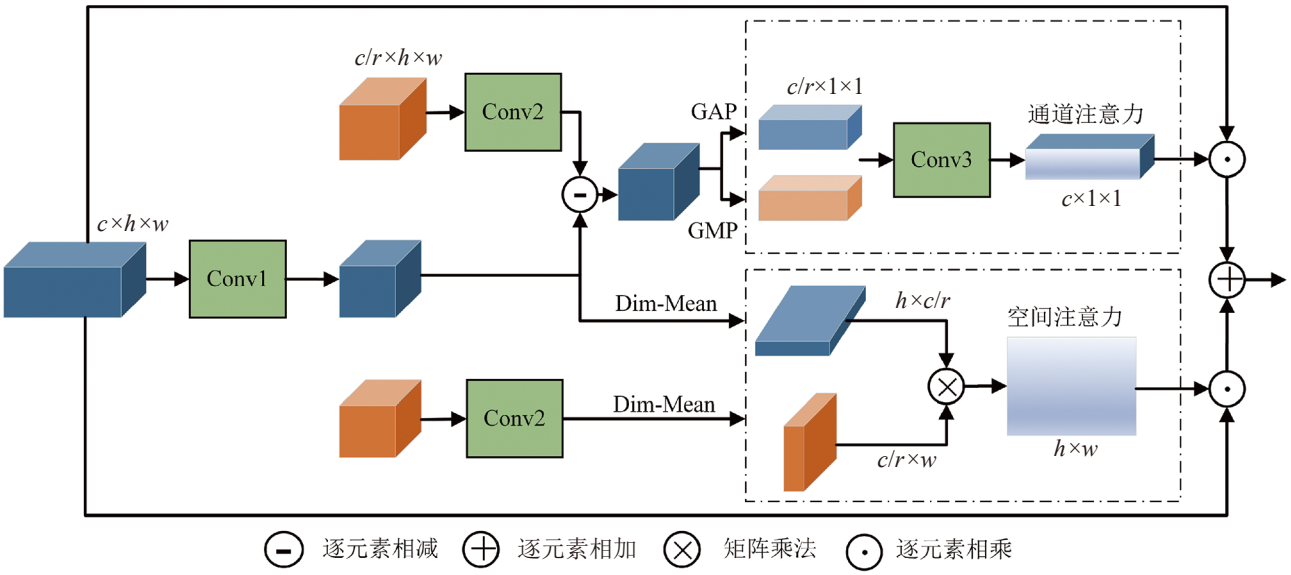

图 3　差分时间交互模块

在通道注意力子模块中, 采用差分卷积计算通道差分。该过程可表示为

$$f = \boldsymbol{F}'_t - \text{Conv2}(\boldsymbol{F}'_{t-1})。\tag{5}$$

通过全局池化和最大池化来细化通道特征, 分别表示为: 通道特征 $C_{\text{A}} = \{a_1, a_2, \cdots, a_c\}$, 特征向量 $C_{\text{M}} = \{m_1, m_2, \cdots, m_c\}$。由于目标运动会造成特征图的差异, 通道之间的差异代表了每个通道对运动过程的重要性。因此, 2 个多层感知器被建立, 分别处理 C_{A} 和 C_{M}, 并将返回的结果相加。该过程可形式化表示为

$$s'_{\text{A},t} = \text{Conv3}(C^t_{\text{A}} - C^{t-1}_{\text{A}}),\tag{6}$$

$$s'_{\text{M},t} = \text{Conv3}(C^t_{\text{M}} - C^{t-1}_{\text{M}})。\tag{7}$$

通过 1×1 的卷积将通道数恢复到 c, 通过 sigmoid 得到注意力权值:

$$s_f = \sigma(\text{Conv}(S(s'_{\text{A},t}, s'_{\text{M},t})))。\tag{8}$$

相邻矩阵之间的差异可清楚地表征物体的运动。矩阵乘法用于获得行和列之间的相关性。空间注意力计算式为

$$s_s = \sigma(\text{Conv}(\boldsymbol{F}'_t \otimes \boldsymbol{F}'_{t-1}))。\tag{9}$$

在 DTI 中, 2 个并行子模块分别计算, DTI 的输出可表示为

$$F' = F \odot s_s + F \odot s_f。\tag{10}$$

根据上述方法, 将时间信息添加到中间特征图中。将这些特征以残差的形式添加到原始信息中。

3　实验与结果分析

通过实验来评估 E-TMAN。首先描述评估数据集和实现细节, 然后在 3 个公共视频数据集上进行实验, 将 E-TMAN 与现有方法进行比较。

3.1　数据集

Something-SometingV2[21] 是一个强时间关系数据集, 包含了大量发生于真实世界的基本行为。它包括了 220 847 个视频片段和 174 个类别。数据集中的视频集关注人类与日常物品的抽象交互行为, 在这些场景中, 时序关系的建模十分关键, 背景信息的干扰因素也很严重, 仅通过背景信息和目标特性无法区分这些行为。

UCF101[22] 和 HMDB51[23] 是 2 个公共数据集。UCF101 包含 101 个类别, 13 000 个视频, 而 HMDB51 包含 7 000 个视频, 涵盖 51 个类别。与 SomethingV2 不同, UCF101 和 HMDB51 是 2 个弱时间关系的数据集[24]。

3.2　实验设置

在实验中, E-TMAN 采用 ResNet 作为主干网络。采用在 ImageNet 数据集上预训练的参数。数据预处理过程: 首先将原始图像较短的一侧调整为 256, 然后使用中心裁剪和随机抖动。在将图像送入网络前, 将图像大小调整为 224×224, 初始学习率设为 0.02。在

SomethingV2 数据集中，训练轮数设为 60，在 UCF101 和 HMDB51 中设为 30。当验证集上的性能饱和时，学习率除以 10 的因子。使用动量为 0.9 的随机梯度下降算法和 5×10^{-4} 的权重衰减来训练模型。

3.3 实验结果分析

E-TMAN 分别在 SomethingV2、UCF101 和 HMDB51 数据集上进行对比实验。

表 1 为 E-TMAN 在 SomethingV2 数据集上的实验结果，其中采样数为 8。结果表明，E-TMAN 优于 TSN 和 TRN 两个经典的 2D CNN 框架，并且浮点运算量相差不大。与 TSM 相比，E-TMAN 无需在大数据上进行预训练，且在浮点运算量和准确率方面都有明显优势。与 CPNet 相比，E-TMAN 准确率高出 3.7 个百分点，且采样数更少。此外，在相同的浮点运算量下，准确率比 TEINet 高出 0.1 个百分点。

表 1 在 SomethingV2 上的实验结果

方法	主干	预训练	采样数	浮点运算量	Top1 准确率/%	Top5 准确率/%
TSN	ResNet50	Kinetics	16	33	30.0	60.5
TRN-Muliscale	BNInception	ImageNet	8	33	48.8	77.6
TRN-2stream	BNInception	ImageNet	8	—	55.5	83.0
TSM	ResNet50	Kinetics	16	65	61.2	86.9
CPNet	ResNet50	ImageNet	24	—	57.7	84.0
TEINet	ResNet50	ImageNet	8	33	61.3	—
E-TMAN	ResNet50	ImageNet	8	33.7	61.4	87.7

表 2 为 E-TMAN 在 UCF101 和 HMDB51 两个数据集上的实验结果。在这 2 个数据集上均采用了 10 个剪辑和 3 个裁剪，输入窗口为 8。在 UCF101 上，E-TMAN 表现出色，超越了 TSN、TSM 等基于 2D 卷积的行为识别方法，且优于 C3D、I3D 等 3D 行为识别模型。在 HMDB51 上 E-TMAN 略低于 TDN，超越了其他基线模型，展现出较强的竞争力。

表 2 在 UCF101 和 HMDB51 上的测试准确率 %

方法	主干	UCF101	HMDB51
TSN	ResNet50	83.0	53.7
C3D	ResNet18	85.8	54.9
I3D	3D InceptionV1	84.5	49.8
TSM	ResNet50	82.4	—
TDN	ResNet50	85.0	57.1
E-TMAN	ResNet50	86.4	55.7

4 结束语

提出了一种时序增强多模态注意力网络 E-TMAN。E-TMAN 由局部运动增强和差分时间交互 2 个核心模块构成。局部运动增强模块负责融合视频片段的局部运动信息，差分时间交互模块负责提取视频片段间的长期时间特征。实验结果表明，该方法设计合理，与其他方法相比，具有较低的开销和良好的性能。

参考文献：

[1] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27: 568-576.

[2] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 4489-4497.

[3] Han Mingfei, Zhang Dejun, Wang Yali, et al. Dual-AI: dual-path actor interaction learning for group activity recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2022: 2980-2989.

[4] Sudhakaran S, Escalera S, Lanz O. Gate-shift networks for video action recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 1099-1108.

[5] Li Jun, Liu Xianglong, Zhang Wenxuan, et al. Spatio-temporal attention networks for action recognition and detection[J]. IEEE Transactions on Multimedia, 2020, 22(11): 2990-3001.

[6] Gavrilyuk K, Jain M, Karmanov I, et al. Motion-augmented self-training for video recognition at smaller scale[C]// IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 10409-10418.

[7] Feichtenhofer C, Pinz A, Zisserman A. Convolutional

two-stream network fusion for video action recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1933-1941.

[8] Li Bing, Cui Wei, Wang Wei, et al. Two-stream convolution augmented transformer for human activity recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(1): 286-293.

[9] Wang Liming, Xiong Yuanjun, Wang Zhe, et al. Temporal segment networks: towards good practices for deep action recognition[C]//European Conference on Computer Vision. Berlin: Springer Press, 2016: 20-36.

[10] Lin Ji, Gan Chuang, Han Song. TSM: temporal shift module for efficient video understanding[C]// IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 7082-7092.

[11] Wang Limin, Tong Zhan, Ji Bin, et al. TDN: temporal difference networks for efficient action recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 1895-1904.

[12] Jiang Boyuan, Wang Mengmeng, Gan Weihao, et al. STM: spatiotemporal and motion encoding for action recognition[C]// IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 2000-2009.

[13] Zhao Yue, Xiong Yuanjun, Lin Dahua. Recognize actions by disentangling components of dynamics[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 6566-6575.

[14] Zhou Bolei, Andonian A, Oliva A, et al. Temporal relational reasoning in videos[C]//Proceedings of the European Conference on Computer Visio. Berlin: Springer Press, 2018: 803-818.

[15] Shou Zhou, Lin Xudong, Kalantidis Y, et al. DMC-Net: generating discriminative motion cues for fast compressed video action recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 1268-1277.

[16] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 4724-4733.

[17] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer Press, 2018: 3-19.

[18] Wang Zhengwei, She Qi, Smolic A. Action-Net: multipath excitation for action recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 13209-13218.

[19] 张翼, 赵岭忠, 翟仲毅. 基于 DDQN 的多智能体冲突消解方法[J]. 桂林电子科技大学学报, 2022, 42(5): 366-370.

[20] 刘培培, 赵岭忠, 翟仲毅, 等. 基于空洞卷积的密集连接网络人流量预测模型[J]. 桂林电子科技大学学报, 2021, 41(5): 375-381.

[21] Goyal R, Ebrahimi Kahou S, Michalski V, et al. The "something something" video database for learning and evaluating visual common sense[C]//IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 5843-5851.

[22] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild[PP/OL]. arXiv(2012-12-03)[2026-03-01]. http://doi.org/10.48550/arXiv.1212.0402.

[23] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition[C]//International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2011: 2556-2563.

[24] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local neural networks[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 7794-7803.

编辑：张所滨