DSconv-LSTM：面向边缘环境的轻量化视频行为识别模型

翟仲毅，赵胤铎

（桂林电子科技大学 广西可信软件重点实验室，广西 桂林 541004）

摘要：边缘智能是将 AI 技术应用于边缘嵌入式设备，提供一种智能化计算新范式，被广泛应用于物联网系统。智能摄像机是具有代表性的边缘产品，能够为智能家居、智能交通和智能监控提供低延迟的视频处理能力。由于摄像机计算资源的有限性，传统的行为识别模型难以在本地托管且及时地完成计算任务。本文设计了一种基于深度可分离卷积的长短时记忆学习模型-DSconv-LSTM，可以快速识别视频流中的目标行为。DSconv-LSTM 使用深度可分离卷积来处理卷积 LSTM 学习单元中四个门的时空数据，从而大大降低了模型的复杂性。最后，利用两个公共视频数据集对 DSconv-LSTM 进行了评估。实验结果表明，DSconv-LSTM 提升了模型的收敛性，大大减小了行为识别模型尺寸，加快了推理速度。

关键词：边缘智能；行为识别；资源受限；轻量化模型

中图分类号：TP18　　**文献标识码：A**

DSconv-LSTM：lightweight video action recognition model for edge embedded devices

ZHAI Zhongyi, ZHAO Yinduo

(Guangxi Key Laboratory of Trusted Software, Guilin University Of Electronic Technology, Guilin 541004,China)

Abstract: Edge computing provides an innovative construction paradigm for intelligent services on edge embedded devices with AI techniques, which has been used to improve the intelligence of IoT applications. Smart camera is one of representative intelligent edge products to be able to provide video-processing services for smart home, intelligent transportation and intelligent monitoring. Due to the resource constraint of cameras, some complex services, e.g., action recognition, are hard to be hosted locally to complete the computational tasks timely and precisely. In this paper, we design a lightweight model DSconv-LSTM based on depthwise separable convolution long short term memory learning unit for recognizing human behaviors locally through camera's video streaming. The DSconv-LSTM uses depthwise separable convolution operation to handle the spatio-temporal data of four gates in convolution LSTM learning unit whereby complexity of recognition model is reduced greatly. Finally, two public video datasets of human behavior are used to test the DSconv-LSTM. The experimental result shows that the DSconv-LSTM improves the learning property on convergence, reduces the model size of action recognition greatly, and shortens the interference time of action recognition.

Key words： Edge intelligence; Action recognition; Resource-constrained; Lightweight model;

1 引言

随着物联网和边缘计算技术的发展，许多嵌入式设备已经具备了较强的计算能力。边缘计算正在成为物联网数据处理的重要组成部分。此外，随着物联网设备的激增，网络边缘会产生大规模的感知数据。在大数据技术的推动下，边缘智能正在逐渐形成，即通过有效结合边缘数据和AI技术在本地完成计算并快速有效地提供智能服务[1]。在边缘智能

作者简介：翟仲毅（1986-），男（汉族），河南新安人，桂林电子科技大学副研究员，zhaizhongyi@guet.edu.cn

服务中，传感器数据由本地负责收集和处理，从而减少了对网络资源的需求。与云服务相比，边缘智能服务可以提高物联网环境下计算的实时性，避免浪费网络带宽资源。

智能摄像机是具有代表性的智能边缘产品，能够为智能家居、智能交通、智能监控等领域提供视频处理服务。智能摄像机服务通常需要从摄像机获取实时视频数据，并进行一系列视频帧处理操作，然后进行相应的行为识别。这意味着摄像机需要为这些服务提供相应的存储和计算资源。由于智能摄像机的资源限制，较多的行为识别模型[2]-[3]很难在本地托管并进行行为识别。这是由于常见的行为识别模型通常采用重量级的深度学习模型，计算复杂度较高且规模大。为了将行为识别服务引入边缘环境，通常需要降低学习模型的计算复杂度，从而减轻对本地设备的资源消耗。

本文提出了一种轻量级的学习模型，用于对边缘视频流中的目标行为进行识别。该动作识别模型主要基于DSconv-LSTM和自注意机制（Self-attention Mechanism）。DSconv-LSTM主要结合卷积LSTM（Conv-LSTM）和深度可分离卷积（Depthwise-Separable convolution，DSconv）进行设计。与Conv-LSTM[12]相比，DSconv-LSTM采用了一系列轻量级学习单元，通过深度可分离卷积运算[15]处理LSTM中四个门的时空数据流。

最后，在UCF-11[12]和Olympic-sports[13]两个公共视频数据集上进行了一系列实验来评估DSconv-LSTM的性能和效果。结果表明：DSconv-LSTM能快速收敛到最优模型，并保持较高的识别精度。与Conv-LSTM相比，DSconv-LSTM的模型规模减小了约三倍，推理时间缩短了约50%。

2 相关工作

行为识别是计算机视觉领域一项常见的研究内容。随着深度学习的快速发展，许多学者都在关注视频行为的深层特征提取，以及行为分类模型和识别方法。双流融合模型（Two-Stream Fusion，TSF）[2]就是一种动作识别框架。TSF 结合时空网络，将RGB 图像和光流分别用卷积神经网络（Convolutional Neural Network，CNN）进行空间和时间的建模，然后进行融合得到最终结果。双流体系结构具有较好的分类效果，后续有许多工作在此基础上进行了相关的研究。Zheng 等人[3]提出了一种用于视频分类的混合深度学习模型，该模型首先使用卷积 LSTM 提取空间特征和短时记忆特征，接着通过注意力机制赋予不同权重来区分不同时刻序列的重要性，最后采用双向 LSTM 提取周期特征。Wang 等人[4]提出了一种基于随机抽样方案的分类模型，称为时间段网络（Temporal Segment Network，TSN），通过对视频端进行分段和随机窗口采样，降低了提取长程时间关系的计算量。为了进一步提高 TSN 的性能，Zhou 等人[5]中提出了一种多尺度采样和融合框架。

为了更直观的捕获视频中的时空特征，基于 3D 卷积的行为分类模型被较多人研究。文献[6]提出了一种基于 3DCNN（3D Convolution，C3D）的分类模型，可以通过卷积操作对空间和时间进行建模。由于额外的核维数，C3D 具有大量的参数和计算开销。文献[7] 提出了一种膨胀 3D 卷积（Inflated 3D ConvNet，I3D），通过在时域上进行额外的卷积运算，而不是直接使用 3D 内核，以减小网络的规模。Fan 等人[8]提出了 RubiksNet，通过一种 3D 时空移位操作，减少了卷积计算次数。Dong 等人[9]提出了 AR3D，通过构建一种注意力残差网络，减少了 3D 卷积的计算量，提升了模型的性能。

鉴于光流和 3D 卷积都需要较大的计算量，Lin 等人[10]提出了时间移位模型（Temporal shift module，TSM），通过将相邻帧的特征值移位，交换不同时刻视频帧之间的特征图，实现时间特征的提取，进而可以使用 2D 卷积来进行行为识别。相较于 3D 卷积的模型，TSM 在推理速度和准确率方面都有显著提升。Wang L 人[11]等提出了时间差分网络（Temporal Difference Networks，TDN）模型，采用 RGB 差分的方法融合时间特征，并通过 2D 卷积进行行为识别。

3 DSconv-LSTM

3.1 模型架构设计

本小节主要利用 DSconv-LSTM 和自注意力机制设计了一个行为识别模型。图 1 给出了分类模型的架构，主要包括：数据源模块、RGB 视频数据模块和 DSconv-LSTM 模块三部分。数据源模块负责获取边缘环境（如智慧家居、智慧交通等）下 RGB 视频信息，并进行视频帧的预处理。RGB 帧模块主

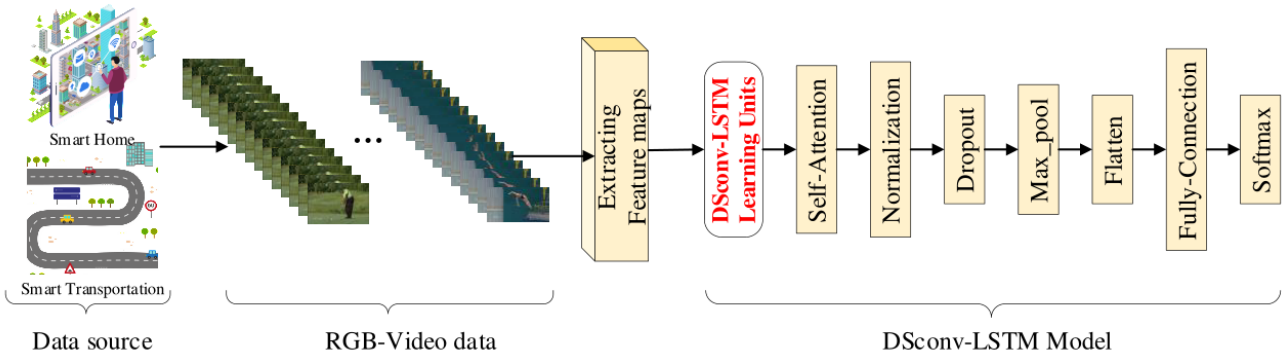

图 1　边缘行为识别系统

要通过 VGG-16[14]或 VGG-19[15]进行特征提取。由于视频是一种时空特征数据，需要通过 Dsconv-LSTM 模块对时间进行建模，接着通过自注意力模块提取特征映射，最后给出识别结果。

3.2　Dsconv-LSTM 单元

DSconv-LSTM 单元将深度可分离卷积（DSconv）用于处理 Conv-LSTM 学习单元四个门的时空数据。

Dsconv 主要有 Dconv 和 Pconv 两个子操作组成，如图 2 所示。

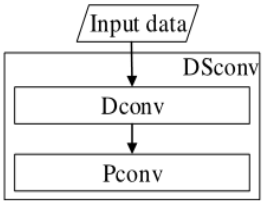

图 2　DSconv 结构

DSconv 首先使用 Dconv 进行空间建模，然后使用 Pconv 进行时间建模。Conv、Pconv、Dconv 和 DSconv 的数学公式分别如下：

$$Conv(W,x)_{(i,j)}=\sum_{k,l,\mathrm{m}}^{K,L,M} W_{(k,l,m)}\cdot x_{(i+k,j+l,m)} \quad (1)$$

$$Pconv(W^p,x)_{(i,j)}=\sum_{m}^{M} W_m^p\cdot x_{(i,j,m)} \quad (2)$$

$$Dconv(W^d,x)_{(i,j)}=\sum_{k,l}^{K,L} W_{(k,l)}^d\odot x_{(i+k,j+l)} \quad (3)$$

$$DSconv(W^p,W^d,x)=Pconv_{(i,j)}(W^p,Dconv_{(i,j)}(W^d,x)) \quad (4)$$

在公式(1)、(2)、(3)和(4)中，W、W^p和W^d分别是Conv、Pconv和Dconv的卷积核。K，L和M分别代表卷积核的宽度，高度和卷积核个数。(i,j)是每次卷积操作的起始位置，x是输入数据，$\odot$表示矩阵对应元素相乘。虽然DSconv需要两步来处理所有输入数据，但可以减少卷积的许多参数和计算。

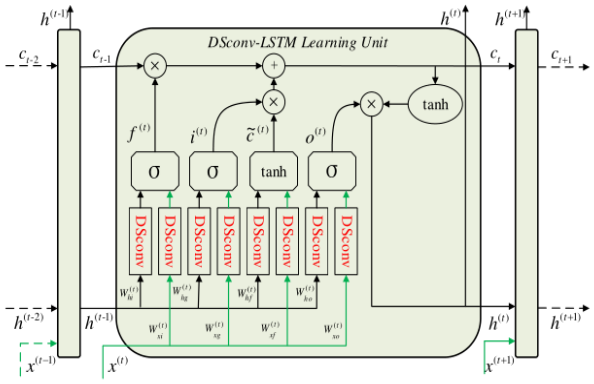

图 3　DSconv-LSTM 学习单元

DSconv-LSTM学习单元也有四个门来处理数据输入，这与Conv-LSTM学习单元类似，如图3所示。DSconv-LSTM学习单元的数学表示如下：

$$i^{(t)}=\sigma(DSconv(W_{x,i},x^{(t)})+DSconv(W_{h,i},h^{(t-1)})+b_i) \quad (5)$$

$$f^{(t)}=\sigma(DSconv(W_{x,f},x^{(t)})+DSconv(W_{h,f},h^{(t-1)}+b_f) \quad (6)$$

$$o^{(t)}=\sigma(DSconv(W_{x,o},x^{(t)})+DSconv(W_{h,\mathrm{o}},h^{(t-1)})+b_o) \quad (7)$$

$$\tilde{c}^{(t)}=\tanh(DSconv(W_{x,c},\mathrm{x}^{(t)})+DSconv(W_{h,c},h^{(t-1)})+b_{\tilde{c}}) \quad (8)$$

$$c^{(t)}=f^{(t)}\odot c^{(t-1)}+i^{(t)}\odot\tilde{c}^{(t)} \quad (9)$$

$$h^{(t)}=o^{(t)}\odot\tanh(c^{(t)}) \quad (10)$$

Dsconv-LSTM 学习单元在t时刻的四个门分别表示输入门$i^{(t)}$，遗忘门$f^{(t)}$，输出门$o^{(t)}$以及输入调整门$\tilde{c}^{(t)}$。$x^{(t)}$，$c^{(t)}$和$h^{(t)}$分别表示t时刻的输入数据，细胞状态和隐藏层状态。$W_{x,i}$，$W_{x,f}$，$W_{x,o}$，$W_{x,c}$与$W_{h,i}$，$W_{h,\mathrm{o}}$，$W_{h,c}$，$W_{h,f}$分别代表 Dsconv 的$i^{(t)}$，$f^{(t)}$，$o^{(t)}$，$\tilde{c}^{(t)}$关于$x^{(t)}$和$h^{(t)}$的卷积核。b_i，b_f，b_o，$b_{\tilde{c}}$分别表示$i^{(t)}$，$f^{(t)}$，$o^{(t)}$，$\tilde{c}^{(t)}$的偏置。σ和tanh分别表示Sigmoid激活函数和双曲正切激活函数。最后，可以将多个 DSconv-LSTM 学习单元构成单层或多层结构，其中单层结构如图3所示。

表 1　UCF-11 数据集和 Olympic-sports 数据集

数据集	视频数量	行为类别
UCF11	1168	Basketball shooting, biking, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog.
Olympic-sports	770	Diving, golf swing, kicking, lifting, horse riding, running, skateboard, swing bench, swing side and walking.

3.3　自注意力机制

在传统结构中，LSTM 输出的最后一个特征映射会被用于数据处理的下一阶段。然而，这种方法可能会丢失许多重要的特征图，从而影响行为识别的准确性。这里采用自注意力机制来解决这个问题。自注意力机制可以提取 DSconv-LSTM 输出的特征映射中最重要的特征，数学表示如下：

$$u^{(t)} = \tanh(W_u o^{(t)} + b_u) \tag{11}$$

$$\alpha^{(t)} = \frac{\exp(W_l^T u^{(t)})}{\sum_{l=1}^{L} \exp(W_l^T u^{(t)})} \tag{12}$$

$$o' = \sum_{t=1}^{L} \alpha^{(t)} \cdot o^{(t)} \tag{13}$$

其中，$o^{(t)}$ 表示 Dsconv-LSTM 学习单元在 t 时刻的输出。$Tanh(\cdot)$ 表示双曲正切激活函数。自注意力机制首先通过(12)计算每个输入 $u^{(t)}$ 的权重 $\alpha^{(t)}$。$\exp(\cdot)$ 表示指数函数。o' 表示自注意力，即权重 $\alpha^{(t)}$ 和输入 $u^{(t)}$ 之间的线性组合。

4 实验

4.1　数据集和实验环境

本节主要介绍用到的数据集和实验环境。模型将通过 UCF11 和 Olympic-sports 两个公开数据集进行训练，如表 1 所示。实验中，两个数据集被随机分为 70%的训练集和 15%的测试集和 15%的验证集。基于 Tensorflow 构建了 DSconv-LSTM 模型和 Conv-LSTM 模型，GPU 采用了两个 16G 内存的 Tesla P100-PCIE。

4.2　模型参数设置

为了减少不同参数设置对结果的影响，实验将两个模型中除了卷积核之外的大多数参数都设置为相同的值，如表 2 所示。

表 2　模型参数设置

参数	值序列	参数	值
Dconv's 卷积核	(3×3×4096)	Dropout	0.5
Pconv's 卷积核	(1×1×4096)	学习率	0.0001
Conv's 卷积核	(3×3×4096)	批大小	6
Pool's 卷积核	(7×7×4096)	窗口大小	40
DSconv-LSTM 单元	1024	Conv-LSTM 单元	1024

在 DSconv-LSTM 模型中，Dconv 和 Pconv 卷积核分别设置为(3×3×4096)和(1×1×4096)。在 Conv-LSTM 模型中，Conv 的滤波器设置为(3×3×4096)。两个模型的 Dropout 和最大池的滤波器分别设置为 0.5 和(7×7×4096)。此外，模型的学习率、批大小和帧窗口大小分别设置为 0.0001, 6 和 40。DSconv-LSTM 单元和 Conv-LSTM 单元的个数都设置为 1024。本实验使用 Vgg-16 和 Vgg-19 从视频中提取特征图，以便观测不同预处理方法的影响。

表 3　模型性能的对比

模型	模型大小(MB)	参数	推理时间(s)
Conv-LSTM	660.213	57687051	0.2894
DSconv-LSTM	228.6875	19984912	0.1557

4.3　实验结果与分析

实验从性能和效果两方面对模型进行了评价。其中，表 3 展示了 DSconv-LSTM 和 Conv-LSTM 模型的大小，参数个数和推理时间。相比于 Conv-LSTM，DSconv-LSTM 的模型大小和参数个数减少了约 3 倍，推理时间减少了约 50%。

图 4 展示了不同预处理方法和数据集下 Conv-LSTM 模型和 DSconv-LSTM 模型对于测试精度和训练时间之间的关系。从图 4 可以看到，不同的预处理方法对模型训练有一定影响，但 DSconv-LSTM 模型在两种条件下都波动较小，并可以快速收敛到最优模型。此外，与 Vgg-16 相比，Vgg-19 的特征提取对模型性能有更高的提升。表 4 展示了两种模型在不同预处理模型和数据集下

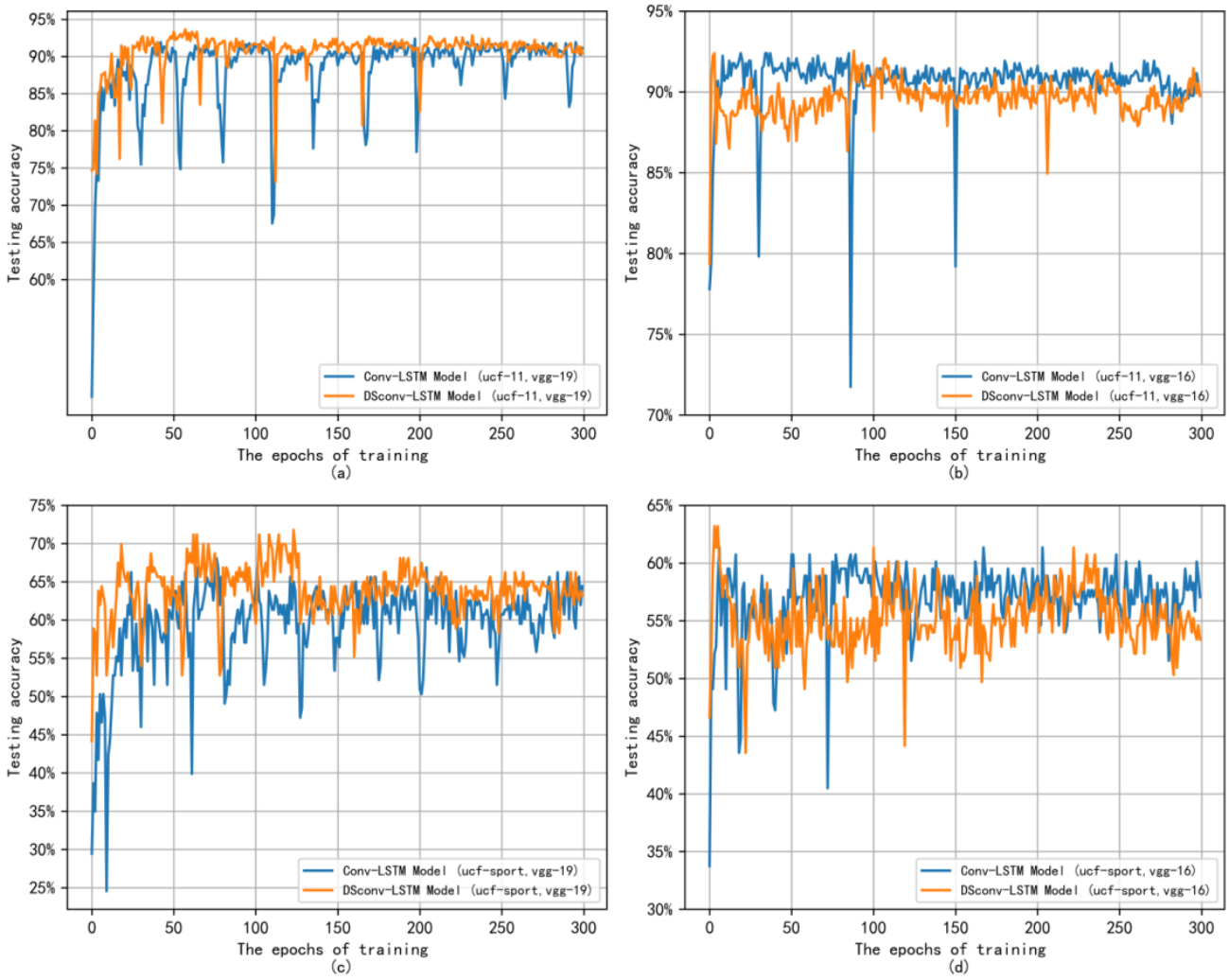

图 4　Conv-LSTM 和 DSconv-LSTM 模型的准确性。

表 4　识别效果评价

		Conv-LSTM	DSconv-LSTM
数据集	预处理模型	准确率	准确率
UCF11	Vgg-16	92.39130%	92.54658%
	Vgg-19	92.39137%	93.63354%
Olympic-sports	Vgg-16	61.96319%	63.19018%
	Vgg-19	68.09815%	71.77914%

行为识别的准确率。对于 UCF-11 数据集和 Vgg16 预处理模型，DSconv-LSTM 的识别准确率为 92.5466%，与 Conv-LSTM 模型相比，准确率高了 0.1553%。对于 UCF-11 数据集和 Vgg-19 预处理模型，DSconv-LSTM 模型的最高测试精度为 93.6335%，比 Conv-LSTM 提高了 1.2421%。对于 Olympic-sports 数据集和 Vgg16 预处理模型，DSconv-LSTM 模型的识别精度为 63.1902%，比 Conv-LSTM 提高 1.227%。对于 Olympic-sports 数据集和 Vgg-19 预处理模型，DSconv-LSTM 模型的最高测试精度为 71.7791%，比 Conv-LSTM 模型提高了 3.6809%。这表明 DSconv-LSTM 模型识别精度仍保持了较高水平。

综上可见，DSconv-LSTM 可以快速收敛到最优模型，并保持较高的识别准确率。与 Conv-LSTM 相比，DSconv-LSTM 的模型大小减少了约 3 倍，推理时间减少了约 50%。

5　结论

本文提出了一种基于 DSconv-LSTM 的轻量级视频行为识别模型，可以在边缘设备上进行应用。与 Conv-LSTM 模型相比，DSconv-LSTM 不仅可以保证行为识别的准确性，并减小了尺寸和参数数量，而且可以快速收敛，降低模型训练和推理的时间。该框架的不足之处是：预处理方法仍需要大量的时间和计算资源从视频中提取特征图。后续工作将关注预处理优化技术，进一步提高动作识别的实

时性，以及减少资源消耗。

参考文献

[1] Zhou Z, Chen X, Li E, et al. Edge Intelligence: Paving the last mile of artificial intelligence with edge computing[J]. Proceedings of The IEEE, 2019, 107(8): 1738-1762.

[2] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[A]. Advances in Neural Information Processing Systems[C]. Cambridge, MA: MIT Press:2014:568-576.

[3] Zheng H, Lin F, Feng X, et al. A hybrid deep learning model with attention-based conv-LSTM networks for short-term traffic flow prediction[J]. IEEE Transactions on Intelligent Transportation Systems(Early Access) ,2020.

[4] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]. European Conference on Computer Vision(ECCV), 2016:20-36.

[5] Zhou B, Andonian A, Oliva A, et al. Temporal relational reasoning in videos[C]. Proceedings of the European Conference on Computer Vision (ECCV) ,2018:803-818.

[6] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231.

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

[8] Fan L, Buch S, Wang G, et al. Rubiksnet: Learnable 3d-shift for efficient video action recognition[C]. European Conference on Computer Vision, 2020:505-521.

[9] Dong M, Fang Z, Li Y, et al. AR3D: Attention Residual 3D Network for Human Action Recognition[J]. Sensors, 2021, 21(5): 1656.

[10] Wang L, Tong Z, Ji B, et al. TDN: Temporal difference networks for efficient action recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021:1895-1904.

[11] Lin J, Gan C, Wang K, et al. TSM: Temporal shift module for efficient and scalable video understanding on edge devices[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access), 2020.

[12] Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]. Advances in Neural Information Processing Systems, 2015:802-810.

[13] Liu J, Luo J, Shah M. Recognizing realistic actions from videos “in the wild”[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2009:1996-2003.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[DB/OL]. arXiv:1409.1556v6,2014.

[15] Barsoum E, Zhang C, Ferrer C C, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution[C]. Proceedings of the 18th ACM International Conference on Multimodal Interaction,2016:279-283.