中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

特征引导的多模态聚合低光环境行为识别方法

刘光辉, 王秦蒙, 孟月波, 陈廷廷, 张娅琳

引用本文:
刘光辉,王秦蒙,孟月波,陈廷廷,张娅琳. 特征引导的多模态聚合低光环境行为识别方法[J]. 控制与决策, 2024, 39(7): 2305–2314.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1753

特征引导的多模态聚合低光环境行为识别方法

刘光辉[1,2], 王秦蒙[1,2], 孟月波[1,2†], 陈廷廷[1,2], 张娅琳[1,2]

(1. 西安建筑科技大学 信息与控制工程学院，西安 710055；
2. 西安市建筑制造智动化技术重点实验室，西安 710055)

摘 要: 诸如夜间等低光场景下的行为识别对于安防、自动驾驶等领域具有重要意义,针对现有方法在低光环境下识别效果不佳、鲁棒性较差等问题,提出一种基于特征引导的多模态聚合低光环境行为识别方法(MALNFG). 首先,设计分层骨架特征融合网络(HSFIE),利用光照增强算法提升低光场景的骨架提取能力,采用层次化时空特征融合策略获取侧重于人体行为本身表达的动作特征,改善低光场景下因骨架缺失造成的精度下降问题;其次,设计高效表观特征提取模块(EAFEM),采用零参数时间位移模块在2D特征提取网络上高效捕捉包含丰富场景信息的时空特征;接着,设计特征引导多模态聚合网络(MNF),利用特征引导策略执行骨架特征与RGB表观特征的深层信息交互,实现行为特征的全面性表征;最后,采用全连接层进行特征分类,完成行为识别. 实验结果表明,所提出方法可以较好地适用于低光环境下的人体行为识别任务.

关键词: 行为识别；低光场景；多模态聚合；特征引导；光照增强

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1753

引用格式: 刘光辉, 王秦蒙, 孟月波, 等. 特征引导的多模态聚合低光环境行为识别方法 [J]. 控制与决策, 2024, 39(7): 2305-2314.

Night behavior recognition based on multi-mode feature fusion

LIU Guang-hui[1,2], *WANG Qin-meng*[1,2], *MENG Yue-bo*[1,2†], *CHEN Ting-ting*[1,2], *ZHANG Ya-lin*[1,2]

(1. College of Information and Control Engineering，Xi'an University of Architecture and Technology，Xi'an 710055，China；2. Xi'an Key Laboratory of Intelligent Technology for Building and Manufacturing，Xi'an 710055，China)

Abstract: Action recognition in low light scenes such as night is of great significance to the fields of security, automatic driving and so on. Aiming at the problems of poor recognition effect and poor robustness of existing methods in low-light environment, a multimodal aggregate low light environment action recognition method based on feature guidance is proposed. Firstly, the hierarchical skeleton fusion network for illumination enhancement is designed. The illumination enhancement algorithm and hierarchical spatiotemporal feature fusion strategy are used to obtain the action features that focus on the expression of human behavior itself, and improve the accuracy degradation caused by skeleton missing in low-light scenes. Secondly, the apparent feature extraction network based on a fusion spatiotemporal conversion module is designed. Spatiotemporal features containing rich scene information on a 2D feature extraction network are efficiently captured using a zero-parameter temporal displacement module. Then, a multi-modal aggregation network based on feature guidance is designed. The feature guidance strategy is used to perform the deep information interaction between skeleton features and RGB features, so as to realize the comprehensive characterization of behavior features. Finally, the full connection layer is used for feature classification to complete behavior recognition. The experimental results show that the proposed method can be well applied to human action recognition tasks in low light environment.

Keywords: action recognition；low light environment；multimodal aggregation；feature guidance；light enhancement

0 引 言

安防监控、视频检索、人机交互、智能家居等多领域的潜在需求极大促进了行为识别技术的发展[1],侧重不同应用场景,一系列表现优异的行为识别方法被相继提出. 当前,研究者对行为识别任务的探索大都集中在正常照明环境下,所提出的大部分数据集也

收稿日期: 2022-10-08；**录用日期:** 2023-04-15.
基金项目: 国家自然科学基金面上项目(52278125).
†通讯作者. E-mail: mengyuebo@163.com.

仅在白天拍摄. 随着夜间监控及夜间自动驾驶应用需求的进一步升级,实现低光环境下行为的精准识别日益迫切.

近年来,深度学习因其强大的特征挖掘与数据处理能力逐渐占据了行为识别领域的主导地位. 按照输入模态的不同,深度学习行为识别方法可以分为基于RGB模态方法和基于骨架模态方法两种[2]. 基于RGB模态方法以经典的Two-stream[3]、3DCNN[4]为基础,同时捕获RGB视频中体现行为变化的空间、时间特征,在行为的全面性表达方面优势明显[5]. 基于骨架模态方法主要是通过编码人体骨架上关键点的运动信息来表征人体动作[6],其更着重于动作本身的准确表达,突出行为的显著性表征. RGB模态方法与骨架模态方法分别从不同的角度描述人体行为,在不同的应用场景下各有优势.

随着场景复杂度的进一步升级,尤其是在目标丢失严重、噪声干扰严重的低光环境中,无论是侧重于场景表观的RGB模态还是侧重于人体行为本身表达的骨架模态,均无法取得理想的识别效果. 为解决此问题,研究者们尝试通过改善获取的数据质量提升识别精度. 如早期的一些工作中利用红外相机采集图像,从数据来源层面缓解低光环境下目标模糊的问题[7-8],但红外相机高昂的成本严重制约了其在现实生活中的普及,同时红外图像较难捕获运动上下文信息,在行为特征的整体表征上相对较弱[9]. 为寻求更加经济的方法,图像增强技术被应用在普通相机获取的低光视频数据预处理中. 文献[10]提出了首个以普通相机拍摄、专门用于低光环境下行为识别的数据集ARID (a comprehensive study on recognizing actions in the dark),并在该数据集上进行光照增强算法基准测试,探索其在低光行为识别任务中的应用. 随后文献[11]通过在3D-ResNext-18网络嵌入4种不同的光照增强策略,研究光照增强策略对低光环境行为识别的作用及局限性. 为进一步提升识别能力,文献[12]设计了光暗两个分支,通过自注意力机制同时提取光流特征与图像特征并进行融合,识别效果提升显著,验证了多特征融合在低光环境行为识别上的优势,但光流与图像特征在行为表达上侧重点相似,对局部细节变化及人体自身变化感知偏弱,且对光照变化较为敏感,识别效果不稳定. 相比之下,骨架数据以人类关节位置为基础,通过人体姿态表示行为动作[13],在行为的局部细节表征上具有优势且对光照变化鲁棒性强,与侧重整体表达的RGB特征有着明显的互补作用[14]. 研究者对RGB模态与骨架模态的融合识别进行了一定的尝试,并将其应用于正常光照条件下的行为识别任务中. 如: Zhao等[15]提出了一种双流结构,分别从RGB和骨架数据中提取特征,在网络末端进行特征融合后利用SVM完成动作分类,实现了相较于单一模态更好的识别效果; Liu等[16]通过融合骨架序列和RGB视频中心图像,设计一个骨架注意力模块实现对RGB表观信息的矫正,有效利用原始骨架序列对RGB特征的驱动能力,实现了两个模态的特征层面的交互,但该方法仅从单张图像中捕获RGB特征,且依赖于骨架数据的完整度,在复杂的场景中表现不佳;文献[17]提出一种词袋特征融合策略,通过多视图结构化稀疏学习优化了两类模态特征的互补能力,在行为的局部表征方面效果较好,但对全局信息的捕获能力较弱,导致识别效果不稳定.

由上述分析可以看出,光照增强与多模特征融合是提升夜间行为识别能力的有效方法,模态的选择与融合的方式是影响识别精度的关键. 鉴于此,本文提出一种基于特征引导的多模态聚合低光环境行为识别网络(multimodal aggregate low light environment behavior recognition network based on feature guidance, MALNFG). 首先,分别从改善提取骨架质量、增强完备性有限骨架数据特征表达两个方面出发,采用伽马校正法与层次化结构挖掘思想,提高骨架模态的行为特征表达能力;然后,通过时间、空间分量捕获RGB模态时空特征,并利用时间位移模块降低时间分量的计算冗余;最后,设计特征引导策略,充分利用RGB与骨架模态的互补作用,进一步增强行为特征表达,并采用全连接层进行特征分类. 为验证算法性能,在低光与正常光照数据集上开展实验,结果表明MALNFG的总体性能优于对比方法.

1 特征引导下的多模态聚合行为识别网络

基于特征引导的多模态聚合低光环境行为识别网络MALNFG结构如图1所示. 具体的, MALNFG包括光照增强优化的分层骨架特征融合网络(hierarchical skeleton fusion network for illumination enhancement, HSFIE)、高效表观特征提取模块(efficient apparent feature extraction module, EAFEM)、特征引导多模聚合网络(multi-modal aggregation network based on feature guidance, MNF) 3个部分.

1.1 光照增强优化的分层骨架特征融合网络

骨架数据对光照、角度、距离等环境因素的干扰有着较强的鲁棒性,相较于对整个视频采样的图像特征,人体骨架更关注于动作本身的变化,对肢体的细微变化感知较为敏感,有利于人体动作的显著性表

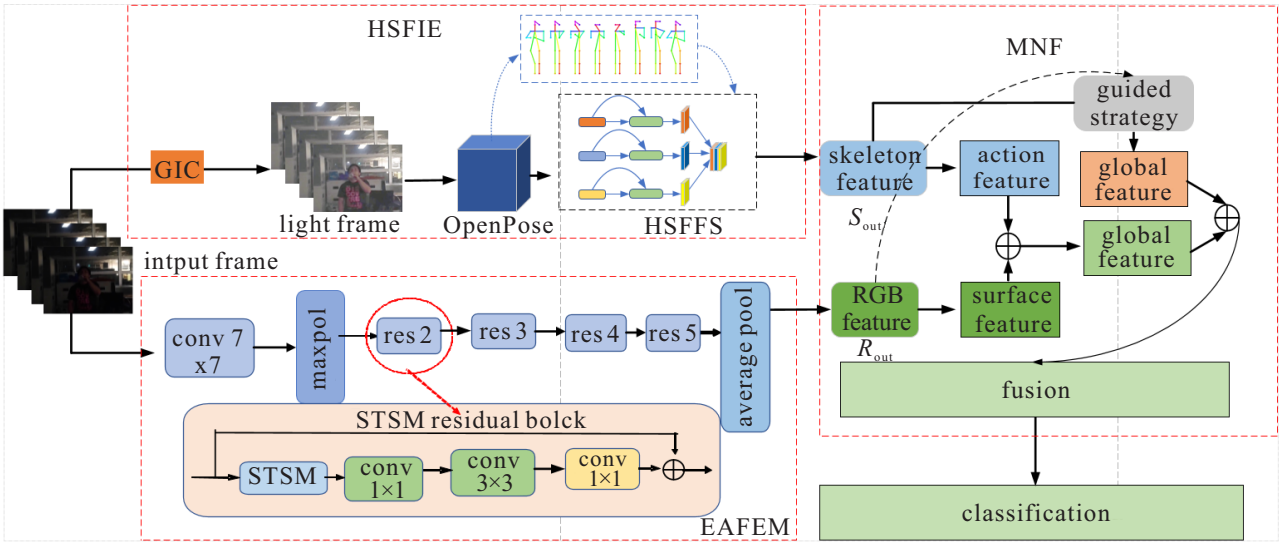


图 1　基于特征引导的多模态聚合低光环境行为识别方法

达. 然而,在低光场景中,由于夜间成像质量较差、噪声较多,导致人体目标不清晰,骨架提取完整度也随之受到影响,进而降低了行为识别的准确度. 本文采用光照增强策略与层次化结构挖掘思想,从改善提取骨架质量与增强时空特征表达两个方面出发,设计光照增强优化的分层骨架特征融合网络,提升骨架模态的行为特征表达能力.

1.1.1　骨架数据获取

光照增强是提升夜间图像质量、改善目标可见性的一种有效方法. 本文采用光照增强后的视频帧作为姿态估计网络的输入,获取相对较完整的骨架数据. 迄今为止,人们已经陆续提出了许多光照增强方法,包括传统的直方图均衡法、伽马校正法,基于光照的 Retinex 方法以及基于深度学习的方法. 虽然基于深度与基于光照的方法在提升夜间图像可见性上有着不错的表现,但通过实验证明,这两类方法在一定程度上破坏了原有的数据分布,增加了更多的噪声[10-12]. 因此,本文采用伽马校正法(Gamma intensity correction, GIC)对原始视频逐帧作光照增强处理. GIC 计算原理如下:

$$\mathrm{GIC}(p)=p_{\max}\left(\frac{p}{p_{\max}}\right)^{\frac{1}{\lambda}}. \tag{1}$$

其中: p 为像素值范围且 $p\in[0,255]$; $p_{\max}$ 为输入像素的最大值; γ 为光照增强的程度,当 $\gamma>1$ 时图像的整体灰度值开始增加,图像可见性逐渐提升.

OpenPose[18] 是目前应用最广泛的人体姿态估计算法之一,故本文选择其作为骨架数据的获取方式. 如图 2 所示, OpenPose 模型由 Branch1 和 Branch2 两个不同功能的分支组成, Branch1 用于提取人体姿态关键点的置信图(confidence maps),即获取人体关键点的位置信息; Branch 2 利用关联向量场(part affinity fields, PAFs)估计肢体区域的位置与方向,预测不同关键点之间的连接信息.

1.1.2　层次化时空特征融合策略 HSFFS

虽然光照增强在一定程度上提升了捕获骨架数据的质量,但依然可能存在的低光影响、视角、遮挡等因素造成的局部关节缺失仍是一个无法完全避免的问题. 本文认为,充分利用局部关节变化关系,是从完备性有限的骨架数据中充分挖掘人体行为时空特征的关键. 基于此,设计如图 3 所示的层次化时空特

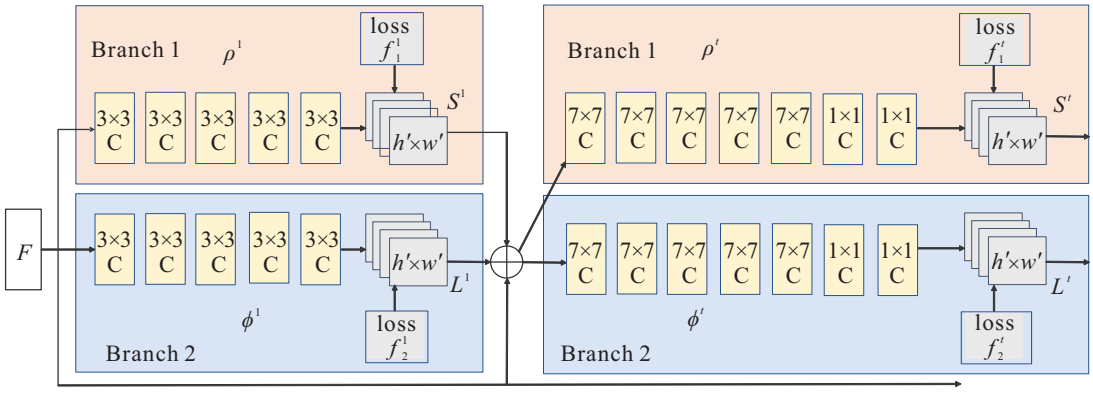


图 2　OpenPose 网络结构

征融合策略(hierarchical spatio-temporal feature fusion strategy, HSFFS),通过三级分层结构提取不同层次的相对速度、相对时间关系,提升局部肢体变化特性的挖掘力度,兼顾多层多尺度的丰富表达. 利用Bi-LSTM抽取时间序列的整体性双向时空关系,进一步缓解视角变换、骨架提取不全带来的干扰.

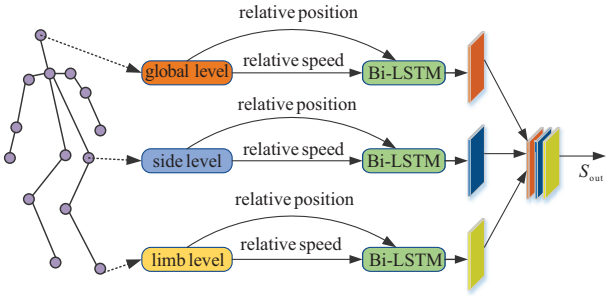


图 3 层次化时空特征融合网络

人体可被视作一个由铰接关节和刚性骨架组成的关节系统. 在人骨架数据中,可以读取身体关节的位置坐标,这些身体关节在时间空间上的变化关系能够直接表示人类的动作. 其中,同一帧各关节之间的距离关系可以提供人体行为中肢体间丰富的空间变化关系,本文将这种关系称为相对位置;不同帧之间关节点的动态变化关系蕴含着丰富的时间特征,可以较好地反映行为动作在时间序列上的差异与肢体运动的剧烈程度,本文将之称为相对速度.

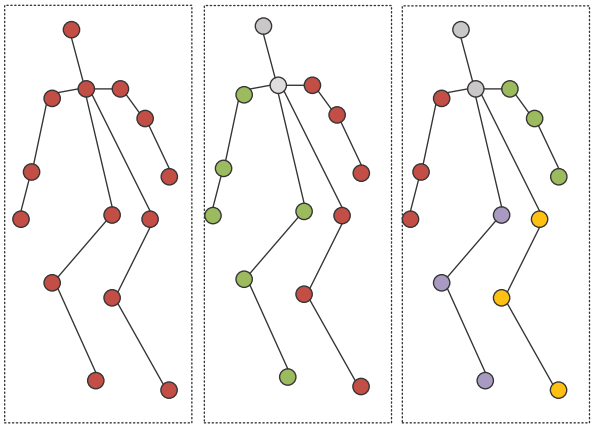
同时使用相对位置与相对速度对骨架序列进行特征提取,可兼顾行为发生时人体姿态的空间结构特征与动态变化特性. 若假设 P_i^t 是 t 时刻关节 i 的坐标,j 是 i 的相邻关节点,则相对位置 P_{ij}^t 和相对速度 S_{ij}^t 可定义为

$$P_{ij}^t = P_i^t - P_j^t, \tag{2}$$

$$S_{ij}^t = P_i^t - P_i^{t-1}. \tag{3}$$

当一些相似度较高的行为发生时,因视角变换,整体骨架表征能力较单薄,局部肢体变化更加明显. 为更好地区分相似动作,将提取的骨架序列划分成如图4所示的3个不同层次,整体骨架作为全局层次,两侧关节作为两侧层次,四肢关节特征作为四肢层次. 利用两侧及四肢局部细节骨架特征对全局骨架特征进行补充,放大肢体的微小变化,最后分别提取各层次的相对位置和相对速度特征对人体行为进行时空特征建模,提升行为特征的显著性表达能力.

在局部骨架缺失、视角变换剧烈的复杂场景中,骨架数据的空间信息此时对行为特征的表达能力较弱,增强时间维度的挖掘深度可以有效地利用整体的变化关系,提升行为的全局表达. LSTM作为一种



(a) 全局 (b) 两侧 (c) 四肢

图 4 骨架分层示意图

先进的循环神经网络架构,是处理时序信息常用的方法,可以学习远程依赖关系,在LSTM层中,存储器控制器用于确定哪些信息被遗忘和保留,并通过输入门、遗忘门和输出门这3种结构实现. 然而对于干扰严重、空间变化不明显的复杂行为,LSTM在对时间特征提取时无法兼顾上下文信息,只能进行单向的学习,对时间序列上下文关系感知较弱,不利于行为特征的整体表征. 相比之下,Bi-LSTM可以兼顾过去与未来的特征,在处理时序特征时具有明显优势,利用Bi-LSTM抽取各层时空特征的双向上下文关系,可增强行为特征的全局性表达,加强网络对时间特征的感知能力. Bi-LSTM结构如图5所示,其使用两个LSTM单元进行特征抽取,充分考虑了前后帧骨架序列的关系,将每个序列向前和向后表示为两个独立的隐藏状态,分别捕获前后帧信息.

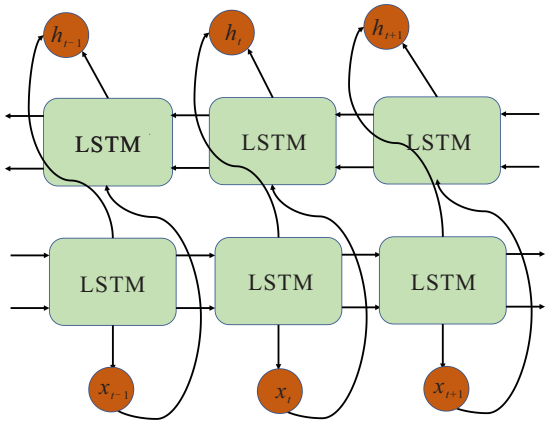


图 5 Bi-LSTM网络结构

分层特征被用于网络输入,Bi-LSTM对各层时空特征进行双向上下文关系的抽取,对于长度为 Q 的序列组合输出表示为

$$h_P^k = h_{1,P}^k \oplus h_{2,P}^k \oplus \ldots \oplus h_{Q,P}^k, \tag{4}$$

$$h_S^k = h_{1,S}^k \oplus h_{2,S}^k \oplus \ldots \oplus h_{Q,S}^k. \tag{5}$$

其中: h_P^k 为在第 k 级分层下相对位置的向量集合, h_S^k 为在第 k 级分层下相对速度的向量集合, $k = 1, 2, 3$; $h_{i,P}^k$ 为 i 点相对位置的向量, $h_{i,S}^k$ 为 i 点相对速度的向量, $i \in [1, Q]$. 随后, h_P^k 将与 h_S^k 进行融合, 组成蕴含丰富时空特性的 H_k, 即

$$H_k = h_P^k \oplus h_S^k. \tag{6}$$

其中: H_1 为全局时空特征, H_2 为两侧时空特征, H_3 为四肢时空特征.

为获取更加丰富的上下文尺度信息, 将全局层次、两侧层次以及四肢层次的输出进行融合, 有

$$X = H_1 \oplus H_2 \oplus H_3, \tag{7}$$

其中 X 为 Bi-LSTM 的最终级联输出. 最后使用全连接层对输出进行特征分类, 完成行为识别.

综上, 对于 SNLE, 若输入为 T 帧, V 表示骨架结构, 则输入骨架序列 $S \in \mathbb{R}^{T\times V}$, 骨架分支输出特征 $S_{\text{out}} = f_s(S, \theta_s) \in \mathbb{R}^{T\times C\times V}$. 其中: f_s 为时空特征提取操作, θ_s 为 HSFFS 在训练时的参数集合, C 为特征通道数.

1.2　时空转换模块嵌入的表观特征提取网络

不同于关注动作本身且对环境适应性较强的姿态估计网络, RGB 特征提取网络关注场景表观特征, 且对背景干扰鲁棒能力较弱. 实验表明, 光照增强后的视频虽然增加了目标可见度, 但随之带来的噪声干扰会影响识别效果, 因此使用未经处理的原始低光视频帧作为 RGB 分支网络输入. RGB 表观时空特征提取, 可采用二维卷积、三维卷积提取算法. 二维卷积算法计算成本低, 但无法捕捉时间关系; 三维卷积算法可以较好地抽取时空特征, 但计算成本高, 参数量大. 文献 [19] 通过实验验证了将 3DCNN 分解为单独的时间和空间分量, 利用 2DCNN 结合 1D 时间卷积捕获时空特征, 可以显著提升精度并减少部分运算量; 但相较于 2DCNN, 时间卷积带来的参数量仍不可避免地带来了部分网络冗余. 为解决此问题, 设计了时空转换模块嵌入的表观特征提取网络 ANFSM 进行 RGB 模态分支的时空特征抽取, 结构如图 6 所示. 利用零参数量、零运算量的时间位移模块 (spatio-temporal shift module, STSM) 在 2DCNN 提取的空间分量特征上做移位运算, 高效捕获时空特征.

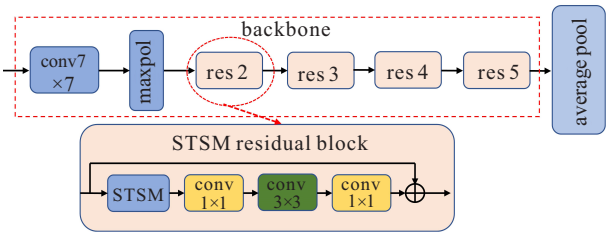


图 6　时空转换模块嵌入的表观特征提取网络结构

低光环境下 RGB 图像存在信息分布不均匀、特征提取难度大、噪声干扰严重等特点, 易造成模型训练效果差、学习能力退化、梯度消失等问题. Resnet50 是一种由残差结构堆叠而成的深层神经网络, 可以较好地实现图像深层特征的提取, 并且凭借其独特的残差连接方式, 能够缓解因网络深度增加造成的梯度消失等问题, 改善模型优化效果. 因此, 此处选择 Resnet50 作为 RGB 特征提取模态的骨干网络 (backbone).

STSM 是一种通过移动相邻帧之间部分通道信息实现序列数据上信息交换的时间位移模块[20], 结构如图 7 所示. 通过时间维度 (time Shift, T)、高度维度 (Height Shift, H)、宽度维度 (Width Shift, W) 三种一维移位操作, 同时进行时间特征与空间特征的学习, 实现时空特征的整体性表达. 将 STSM 嵌入至 Resnet50 网络的残差结构 (residual block) 中, 形成带有时序信息捕捉能力的时空残差结构 (STSM residual block), 以实现 RGB 模态时空特征高效提取.

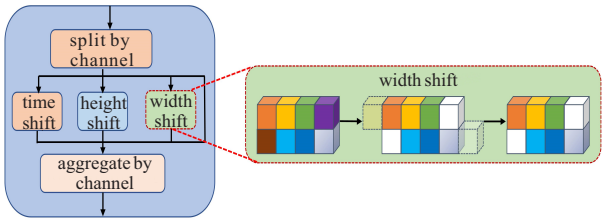


图 7　STSM 网络结构

综上, 对于 ANFSM, 若输入为 T 帧, 则输入图像序列 $R \in \mathbb{R}^{T\times 3\times H\times W}$. 其中: "3" 为 RGB 图像的 3 个通道, H 为输入图像高度, W 为图像宽度. RGB 分支输出特征 $S_{\text{out}} \in f_R(R, \theta_s) \in \mathbb{R}^{T\times C\times W\times H}$. 其中: f_R 为 RGB 分支时空特征提取操作, θ_R 为 ANSFM 在训练时的参数集合, C 为特征通道数, H' 为特征图高度, W' 为特征图宽度.

1.3　基于特征引导的多模聚合网络

HSFIE 与 EAFEM 分别从不同的角度提取时空特征, 前者侧重于行为本身的姿态变化, 后者更着重于整体场景特征的感知. 在低光环境下, 特征提取较难, 单一模态较难提供足够的信息, 制约了行为识别精度的进一步提升. 因此, 充分利用不同模态之间的特性, 实现高效的特征聚合是提升行为识别效果的重心. 为解决常用聚合方法存在的问题, 本文设计如图 8 所示的基于特征引导的多模态聚合网络 MNF, 利用特征引导策略实现两个模态时空特征的细密性联系, 加深特征融合力度, 并进行多层多级融合, 实现完整性表达.

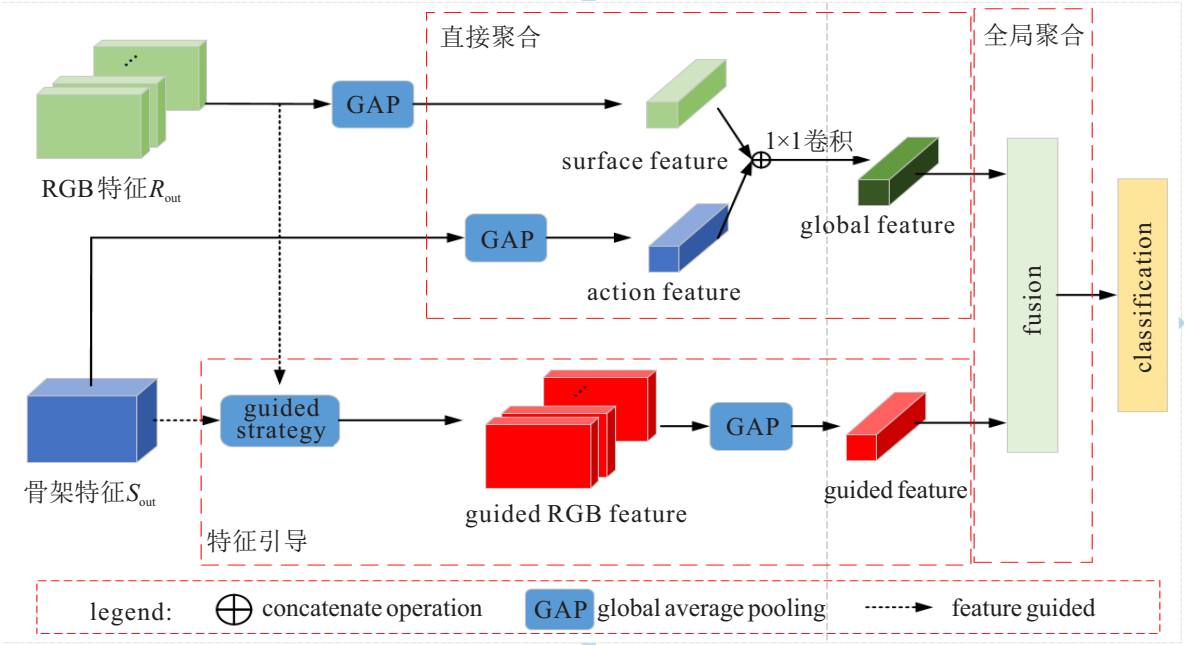


图 8 基于特征引导的多模态聚合网络

MNF 特征聚合分 3 个层次: 直接聚合、特征引导、全局聚合. 直接聚合部分, 采用全局平局池化 (global average pooling, GAP) 将骨架特征 S_{out} 转换为长度为 C_S 的特征向量、将 RGB 特征 R_{out} 转换为长度为 C_R 的特征向量, 通过 1×1 卷积降低特征通道维数并执行信道交互, 得到侧重整体表达的全局特征 (global feature); 特征引导部分, 设计引导策略 (guided strategy), 利用骨架特征对 RGB 特征进行语义约束, 强化重点局部区域的关注度, 并采用全局平均池化获得引导特征 (guided feature); 全局聚合部分, 通过全连接操作聚合引导特征与全局特征, 得到全局、局部并重的多模态聚合特征.

引导策略结构如图 9 所示, 首先将骨架特征 S_{out} 与 RGB 特征 R_{out} 沿时间和空间维度进行分割, 以通道拼接方式对多模态特征进行元素级联; 然后执行特征学习关联运算 (feature learning correlation, FLC), 实现两种模态数据语义对齐; 最后利用 1×1 卷积强化局部区域语义信息, 完成语义特征引导.

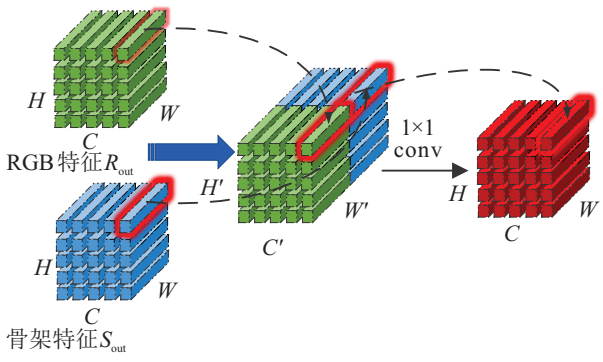


图 9 特征引导策略

2 实验结果与分析

本文算法均在型号为 GTX1080Ti 的 GPU 中运行, 实验选择 Ubuntu 系统, 软件环境为 CUDA 10.2+anaconda3.7+Python3.7+pytorch1.8. 设置模型输入大小为 32×3×224×224, RGB 模态分支的 ResNet50 网络选择 ImageNet 预训练权重作为初始化参数, 初始学习率为 1e-5, 迭代次数为 1 000.

本节将从分类精度、模型运行效率两方面对算法性能进行综合评估. 由参考文献可知, 行为识别多采用准确率 (accuary) 衡量分类性能的优劣, 模型的复杂度与运行效率可通过浮点运算次数 (giga floating-point operations per second, GFLOPs)、模型推理速度 (frames per second, FPS) 进行度量. 为便于多算法对比, 本文亦如此.

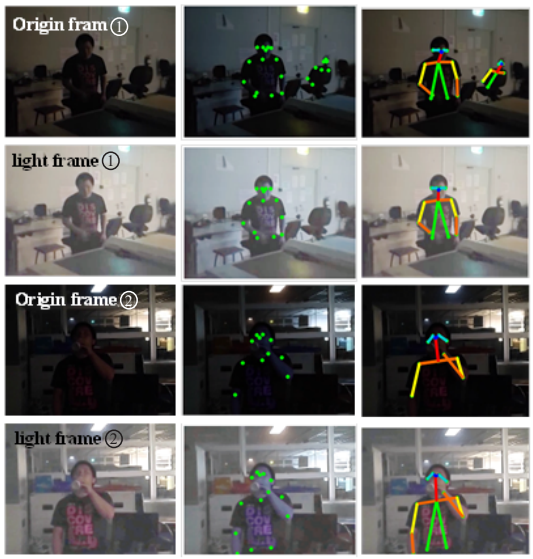
2.1 低光环境实验

为了更好地评价所提出算法, 选择由南洋理工大学于 2020 年发布的 ARID 夜间行为数据集进行低光环境实验, 并对实验结果进行分析.

ARID 数据集中的视频片段完全在夜间拍摄, 由 8 名男性和 3 名女性在 9 个室外场景、9 个室内场景中拍摄, 具体包括 11 类动作, 其中跳跃、奔跑、转身、行走和挥手为单人动作, 饮酒、采摘、倒酒、推搡、坐、站为人与物体交互动作. 每个场景的照明条件均不同, 数据集总共包含 3 784 个视频, 分辨率均为 320×240, 拍摄帧率为 30 帧每秒. 本文手动标注后进行实验, 其中 70 %用作训练, 30 %用作测试.

2.1.1　光照增强分析

GIC 光照增强前后骨架提取效果对比实验结果如图 10 所示. 可以看出,经过光照增强后的骨架提取效果更好. 例如: 未引入光照增强时,图 10①存在“椅子”骨架错误识别; 引入光照增强后,图 10②骨架提取更完整. 光照增强后的低光图像如图 11 示,图像亮度有所增加,目标清晰度上升明显,但从灰度直方图结果来看,图像也随之增加了噪声. 可以看出,光照增强的过程并非是完全正面的,选择均衡的增强程度是至关重要的.



(a) RGB 图像　(b) 人体关键点　(c) 人体骨架

图 10　GIC 增强前后骨架提取效果对比

(a) RGB 图像　　(b) 灰度直方图

图 11　GIC 增强效果

γ 值是决定 GIC 增强效果的关键参数,本文将 γ 值作为超参数,通过改变 γ 值大小分析其对识别精度的影响. 如图 12 所示, γ 由 1 至 5 不断增长,平均精度先增高后变低, γ 为 3.5 时精度达到最大,故选择 $\gamma = 3.5$ 作为 GIC 光照增强算法超参数值.

为进一步验证 Gamma 变换方法的有效性,将 GIC 与 HE[21]、LIME[22]、BIMEF[23]、KinD[24] 等光照增强算法进行对比分析,实验结果如表 1 所示. HE

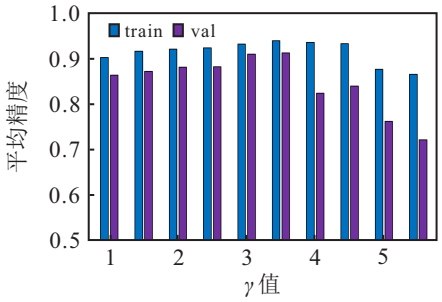


图 12　γ 值对精度的影响

方法主要以扩大图像动态范围的方式增强整幅图像的对比度; LIME 与 BIMEF 方法均将图像看作由照度图像和反射图像组成,基于 Retinex 理论改善图像亮度; KinD 则是一种基于深度学习的方法,其利用双流结构同时进行反射率恢复和光照调整来进行增强. 由实验结果可以看出, GIC 光照增强算法优势显著. 分析原因可知, GIC 方法侧重于调整视频帧的亮度,对像素值的分布特性影响较小,因此产生的噪声干扰相对较低.

表 1　光照增强多算法对比与消融实验结果

	方法	准确率/%
光照增强多算法对比	ours-HE	90.12
	ours-LIME	93.23
	ours-BIMEF	88.74
	ours-KinD	80.14
	ours-GIC	**9**4.09
光照增强消融实验	without GIC	89.21
	skeleton with GIC	**94.09**
	RGB with GIC	87.32
	all with GIC	90.54

此外,本文进行了 GIC 模块嵌入模态验证实验,对不进行光照增强(without GIC)、仅对骨架模态进行光照增强(skeleton with GIC)、仅对 RGB 模态进行光照增强(RGB with GIC)、在两个模态同时增强(all with GIC) 4 种情况进行对比分析,实验结果如表 1 所示. 可以看出,仅在骨架模态引入光照增强模块效果最佳,在 RGB 模态单独引入效果最差. 分析主要原因如下: 光照增强模块在提升人体目标可见度的同时,也相应带来了部分环境噪声干扰,骨架模态关注动作主体,对环境噪声的鲁棒性较强,故引入光照增强识别精度有所提升; RGB 模态的 Resnet 网络对图像整体特征进行了较深层次的挖掘,残差结构也多次融合不同层间的特征,加剧了环境噪声积累与叠加,导致引入光照增强识别精度下降; 两个模态共同增强,

精度优于RGB模态单独引入,弱于骨架模态单独引入. 因此,本文最终选择仅在骨架模态进行光照增强处理.

2.1.2 层次化时空特征融合策略消融研究

本文通过将全局层次(global level, GL)、两侧层次(side level, SL)、四肢层次(limb levels, LL)进行不同的融合实验验证分层思想的有效性,并使用CNN、RNN、LSTM、Bi-LSTM四种时空特征提取网络进行对比分析,验证Bi-LSTM的独特优势. 实验结果如表2所示. 可以看出:相较于仅使用整体骨架的ours with GL,融合两侧层次、四肢层次均对算法识别精度有所帮助;同时将三者融合,精度则进一步提升. 这充分说明了层次化结构挖掘思想的有效性以及各层之间时空特征的互补性,本文算法在行为特征显著性表达方面具有优势. 另外,由不同时空特征提取网络的对比结果可以看出,不能捕捉时间变化的CNN效果最差,3种时序网络中,双向捕获全局特征的Bi-LSTM取得了最高的精度,验证了所提出算法在行为特征上优异的全局表达能力.

表2 HSFTS与特征融合消融实验结果

	方法	准确率/%
HSFTS消融实验	ours with GL	86.52
	ours with GL+SL	88.54
	ours with GL+LL	90.08
	ours with CNN	89.88
	ours with RNN	91.43
	ours with LSTM	92.62
	HSFTN	94.09
特征融合消融实验	ours only Skeleton	87.39
	ours only RGB	83.74
	ours without guided block	91.62
	ours	**94.09**

2.1.3 特征聚合消融研究

本文聚合两个不同模态特征实现夜间行为识别. 为验证特征聚合效果以及特征引导策略的有效性,设置4组对比实验:单骨架模态(ours only skeleton)、单RGB模态(ours only RGB)、直接聚合(ours without guided block)以及带有引导策略(ours)的识别,结果如表2所示. 可以看出,单骨架模态与单RGB模态识别精度均低于聚合网络,表明骨架数据与RGB数据具有互补性;加入特征引导策略后,网络识别精度相较于直接聚合上升2.47%,表明引导策略优势明显.

2.1.4 多算法对比实验

为进一步验证本文算法的有效性,将其与经典two-stream、3D-CNN类方法、文献[12]、文献[25]等先进算法进行对比实验,结果如表3所示. 表中“–”表示原文献源码未给出,无法对其复杂度进行精准衡量.

表3 多算法性能指标结果对比

VGGG	模型	准确率/%	GFLOPs	FPS
two-stream[10]	VGG-TS	32.08	5.6	1
	TSN	57.96	33	–
	I3D-TS	72.78	–	–
3D-CNN[10]	C3D	40.34	38.5	26
	3D-ResNet-18	54.68	–	–
	pseudo-3D-199	71.93	–	–
	3D-ResNext-101	74.73	–	–
with image enhancement	delta sampling strategy[25]	90.46	–	–
	R(2+1)D-34-darklight-SA[12]	94.04	48	20
	ours	**94.09**	50.5	33

在准确性方面,经典two-stream类与3D-CNN类的方法在ARID数据集上表现并不理想,即便是网络深度较大的3D-ResNext-101也仅取得了74.73%的平均准确率. 相较于文献[12, 25],本文算法分类性能表现最好.

在模型运行效率方面,相较于对比方法,本文算法虽有较高的浮点运算次数(GFLOPs),但在算法推理速度(FPS)上具有优势. 这是由于RGB分支采用了Resnet50网络配合时间位移模块进行时空特征提取,类似C3D模型的批处理方式,实现在所需运算次数较高的前提下仍具备快速的推理能力;同时,骨架分支、RGB分支的并行处理也在一定程度上提升了算法的推理速度.

2.2 普适性实验

为验证本文特征聚合方法的普适性,在公开数据集下做进一步分析研究,选择经典的UCF101数据集进行算法验证.

首先,人为降低UCF101数据集图像光强,形成低光环境模拟数据集Dark-UCF101. 由表4所示多算法对比实验结果可以看出,本文算法准确率表现最佳,表明其在低光场景具有较好的鲁棒性. 其次,在原始UCF101数据集上进行实验,分析算法在正常光照下的效果,实验结果如表5所示. 可以看出,当不采用光照增强模块时,无论是单独的RGB模态分支(ours only RGB)还是骨架模态分支(ours only skeleton),识别精度均低于模态融合结果,同时本文方法在正常照明环境下仍保持着最佳识别精度,表明RGB与骨架两种行为特征描述之间的互补作用适用于各类光照

环境. 另外,由于UCF101数据集拍摄于正常照明环境且视频图像人体区域占比较大,环境噪声相对较少,有利于获得完整骨架信息表征,因此RGB分支特征引导对改善骨架特征品质贡献相对较小,加入特征引导策略后,算法精度尽管有所提升,但提升幅度低于低光环境.

表4　Dark-UCF101多算法对比实验结果

方法	准确率/%
two-stream	50.23
TSN	61.17
C3D	45.12
I3D	64.30
ours	**98.52**

表5　UCF101数据集多算法对比实验结果

方法	光照增强	准确率/%
two-stream	–	88.00
TSN	–	94.24
C3D	–	85.20
I3D	–	95.60
ours only skeleton	–	90.20
ours only RGB	–	93.66
ours without guided block	–	97.60
ours	–	**98.79**
ours	✓	**98.65**

若采用光照增强模块,则对于正常光照图像而言,光照增强操作在一定程度上影响了行为识别准确率,不过准确率下降幅度并不大. 究其原因,骨架分支的层次化时空特征融合策略可以较好地发挥骨架数据的表征效果,RGB分支通过特征引导策略也进一步提升了时空特征的完整性表达,二者弱化了曝光过度带来的影响. 因此,本文算法在低光与正常照明环境下均呈现出较好的环境适应能力. 但同样可以看出,光照增强算法带来的影响仍无法完全消除,设计更加灵活的光照增强算法是进一步需要提升的方向.

3　结　论

本文提出一种基于特征引导的多模态聚合低光环境行为识别方法,通过对骨架、RGB两个模态时空关系的抽取与聚合,实现了行为特征的完整表达. 从能见度较低、噪声干扰较多的低光视频中分别提取侧重动作本身表达的骨架时空特征及侧重于场景表观信息表达的RGB时空特征,解决低光场景下由单一模态数据提取的时空特征表征能力较弱的问题. 对两种模态之间互补能力进行深入探讨,在直接融合两类特征的基础上,利用骨架模态得到的特征对RGB模态获取特征进行引导,进一步增强行为特征表达,提升算法在低光环境下的行为识别精度. 低光和正常光照数据集实验结果表明,所提出方法优于对比算法,可以较好地完成低光环境下的行为识别任务.

参考文献(References)

[1] 陈莹, 龚苏明. 改进通道注意力机制下的人体行为识别网络[J]. 电子与信息学报, 2021, 43(12): 3538-3545. (Chen Y, Gong S M. Human action recognition network based on improved channel attention mechanism[J]. Journal of Electronics & Information Technology, 2021, 43(12): 3538-3545.)

[2] Hao X, Li J, Guo Y, et al. Hypergraph neural network for skeleton-based action recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 2263-2275.

[3] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 1725-1732.

[4] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]. IEEE International Conference on Computer Vision. Santiago, 2016: 4489-4497.

[5] Bilen H, Fernando B, Gavves E, et al. Action recognition with dynamic image networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2799-2813.

[6] Ke Q H, Bennamoun M, An S J, et al. A new representation of skeleton sequences for 3D action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 4570-4579.

[7] Ulhaq A. Action recognition in the dark via deep representation learning[C]. IEEE International Conference on Image Processing, Applications and Systems. Sophia Antipolis, 2019: 131-136.

[8] Ma X X, Chau L P, Yap K H, et al. Convolutional three-stream network fusion for driver fatigue detection from infrared videos[C]. IEEE International Symposium on Circuits and Systems. Sapporo, 2019: 1-5.

[9] Chen X, Gao C, Li C, et al. Infrared action detection in the dark via cross-stream attention mechanism[J]. IEEE Transactions on Multimedia, 2022(24): 24.

[10] Xu Y, Yang J, Cao H, et al. ARID: A comprehensive study on recognizing actions in the dark and a new benchmark dataset[J/OL]. 2020, arXiv: 2006.03876.

[11] Patel H R, Doshi J T. Human action recognition in

dark videos[C]. International Conference on Artificial Intelligence and Machine Vision (AIMV). Piscataway: IEEE, 2021: 1-5.

[12] Chen R, Chen J, Liang Z, et al. Dark light networks for action recognition in the dark[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 846-852.

[13] 薛盼盼, 刘云, 李辉, 等. 基于时域扩张残差网络和双分支结构的人体行为识别[J]. 控制与决策, 2022, 37(11): 2993-3002.
(Xue P P, Liu Y, Li H, et al. Human behavior recognition based on time domain extended residual network and dual branching structure[J]. Control and Decision, 2022, 37(11): 2993-3002.)

[14] Li J N, Xie X M, Pan Q Z, et al. SGM-Net: Skeleton-guided multimodal network for action recognition[J]. Pattern Recognition, 2020, 104: 107356.

[15] Zhao R, Ali H, van der Smagt P. Two-stream RNN/CNN for action recognition in 3D videos[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Vancouver, 2017: 4260-4267.

[16] Liu G Y, Qian J C, Wen F, et al. Action recognition based on 3D skeleton and RGB frame fusion[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2020: 258-264.

[17] Franco A, Magnani A, Maio D. A multimodal approach for human activity recognition based on skeleton and RGB data[J]. Pattern Recognition Letters, 2020, 131: 293-299.

[18] Shahroudy A, Wang G, Ng T T. Multi-modal feature fusion for action recognition in RGB-D sequences[C]. The 6th International Symposium on Communications, Control and Signal Processing. Athens, 2014: 1-4.

[19] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 6450-6459.

[20] Yang Z, An G Y. STSM: Spatio-temporal shift module for efficient action recognition[J/OL]. 2021, arXiv: 2112.02523.

[21] Trahanias P E, Venetsanopoulos A N. Color image enhancement through 3-D histogram equalization[C]. Proceedings 11th IAPR International Conference on Pattern Recognition. The Hague, 2002: 545-548.

[22] Guo X J, Li Y, Ling H B. LIME: Low-light image enhancement via illumination map estimation[J]. IEEE Transactions on Image Processing, 2017, 26(2): 982-993.

[23] Ying Z Q, Li G, Ren Y R, et al. A new image contrast enhancement algorithm using exposure fusion framework[C]. International Conference on Computer Analysis of Images and Patterns. Cham: Springer, 2017: 36-46.

[24] Zhang Y H, Zhang J W, Guo X J. Kindling the darkness: A practical low-light image enhancer[C]. Proceedings of the 27th ACM International Conference on Multimedia. Nice, 2019: 1632-1640.

[25] Hira S, Das R, Modi A, et al. Delta Sampling R-BERT for limited data and low-light action recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, 2021: 853-862.

作者简介

刘光辉 (1976−), 男, 副教授, 博士, 从事计算机视觉感知与理解、建筑智能化技术等研究, E-mail: guanghuil@163.com;

王秦蒙 (1996−), 男, 硕士生, 从事基于视觉行为识别的研究, E-mail: 1913313992@qq.com;

孟月波 (1979−), 女, 教授, 博士, 从事计算机视觉理解、建筑环境智能感知与调控、建筑智能化技术等研究, E-mail: mengyuebo@163.com;

陈廷廷 (2000−), 女, 硕士生, 从事计算机视觉感知与理解的研究, E-mail: 2270822886@qq.com;

张娅琳 (1998−), 女, 硕士生, 从事建筑环境智能感知的研究, E-mail: 1243697118@qq.com.