# 基于人–物交互图卷积网络的扶梯乘客危险行为识别

邓鑫[1, 2]，谷金晶[1, 2]，赵征鹏[1, 2]，普园媛[1, 2]，徐丹[1, 2]
（1. 云南大学信息学院，云南 昆明 650504；2. 云南省高校物联网技术及应用重点实验室，云南 昆明 650504）

**摘　要：**扶梯乘客的不当行为极易引发公共安全事故和财产损失，基于监控视频准确识别出扶梯乘客危险行为，对于保障公共安全具有重要意义。但现有的行为识别方法鲜有关注扶梯场景下的乘客危险行为，并且缺乏对人与扶梯时空交互的建模分析。因此，提取人体骨架和人–物交互的时空信息，设计了基于距离度量的双流人–物交互图卷积网络来识别扶梯乘客危险行为。首先，分别提取人体骨架和扶梯关键点特征，通过扶梯关键点为人体骨架特征补充场景信息。其次，利用人–扶梯间的距离度量危险行为中人–物关系的动态变化，加强模型对危险行为中时空交互信息的建模。最后，为了填补现有公开数据集中扶梯危险行为视频的空白，构建了一个扶梯乘客危险行为视频数据集 ESC-Danger，该数据集包含倚靠、攀爬、下蹲、伸手、探头、滞留、逆行和奔跑 8 类扶梯乘客危险行为。在 ESC-Danger 数据集上所提模型的识别准确率为 95.06%，相比于其他先进算法，具有较高的识别准确率和良好的泛化性能。

**关键词：**危险行为识别；人–物交互；双流图卷积网络；骨架序列；扶梯
**中图分类号：**TP391
**文献标志码：**A
**doi：**10.11959/j.issn.2096−3750.2025.00470

# Escalator passenger dangerous behavior recognition based on human-object interaction graph convolutional network

DENG Xin[1,2], GU Jinjing[1,2], ZHAO Zhengpeng[1,2], PU Yuanyuan[1,2], XU Dan[1,2]
1. School of Information Science and Engineering, Yunnan University, Kunming 650504, China
2. University Key Laboratory of Internet of Things Technology and Application of Yunnan Province, Kunming 650504, China

**Abstract:** Improper behavior of escalator passengers can easily lead to public safety accidents and property losses. Accurately identifying dangerous behaviors of escalator passengers based on surveillance videos is of great significance for ensuring public safety. However, existing behavior recognition methods rarely focus on the dangerous behaviors of passengers in escalator scenes, and lack modeling and analysis of spatial-temporal interactions between people and escalators. Therefore, spatio-temporal information from human skeleton and human-object interactions were extracted, and a two-stream human-object interaction graph convolutional network considering distance metrics to identify dangerous behaviors of escalator passengers was designed. Firstly, features from both human skeleton and escalator keypoints were extracted, supplementing scene information for human skeleton features using escalator keypoints. Secondly, distance metrics between humans and escalators to dynamically capture changes in human-object relationships within dangerous behaviors was utilized, enhancing the model's modeling of spatio-temporal interaction information in dangerous behaviors. Finally, to fill the gap in existing publicly available datasets regarding videos of dangerous behaviors on escalators, a data-

**收稿日期：**2024−10−15；**修回日期：**2024−11−29
**通信作者：**赵征鹏，zhpzhao@ynu.edu.cn
**基金项目：**国家自然科学基金资助项目（No. 52102382，No. 62362070）；云南省基础研究计划面上项目（No. 202401CF070164）；云南省科技厅科技计划基础研究专项–重点项目（No. 202401AS070149）
**Foundation Items:** National Natural Science Foundation of China (No. 52102382, No. 62362070), Yunnan Provincial Fundamental Research Project (No. 202401CF070164), Yunnan Provincial Department of Science and Technology Basic Research Special Program - Key Project (No. 202401AS070149)

set called ESC-Danger for escalator passenger dangerous behaviors was constructed. This dataset contains eight classes of escalator passenger dangerous behaviors, including lean, climb, crouch, reach out, poke head out, retention, retrograde, and run. The recognition ccuracy of the proposed model on the ESC-Danger dataset is 95.06%, demonstrating higher recognition accuracy and good generalization performance compared to other state-of-the-art algorithms.

**Key words:** dangerous behavior recognition, human-object interaction, two-steam graph convolutional network, skeleton sequence, escalator

## 0　引言

自动扶梯常见于公司商场和地铁站等人员密集场所，扶梯乘客的不当行为极易引发公共安全事故和财产损失。因此，扶梯乘客危险行为识别具有极其重要的意义。相较于其他场景，扶梯空间比较窄长和封闭，乘客与乘客之间极易发生遮挡，并且乘客与扶梯存在多点且长时间的交互，这会大大增加识别和理解扶梯乘客危险行为的难度。

现有识别和理解行为的方法可分为两种，即行为识别[1]和人–物交互检测[2]。主流行为识别方法通过提取轻量化且鲁棒性强的人体骨架序列作为输入，这种数据不易受光照、场景和遮挡的影响。以往的研究者利用循环神经网络（RNN, recurrent neural network）[3]和卷积神经网络（CNN, convolutional neural network）[4]来提取骨架序列中的时空信息以识别不同行为。但由于骨架是非欧几里得式的数据，这些方法识别精度并不高。图卷积网络（GCN, graph convolutional network）可以很好地处理这种数据，基于GCN的方法[5-6]在不断探索如何加强时空信息的建模，以及更好地理解人体关节点间非自然连接关系。但骨架序列只包含人体特征[7]，损失了大量的场景和物体信息。例如，刷牙和喝水两个动作的骨架相似性是非常高的，单纯利用骨架序列信息难以对存在高相似度的行为进行准确分类和识别。

为了引入场景和物体信息，许多研究者通过外观推理[8]、图解析网络[9]、人–物相对位置[10]来提取交互信息，便于更好地理解人体行为。这些工作大多是基于HICO[11]和CAD-120[12]这两个数据集实现的。HICO虽然规模较大，但是一个图像数据集。CAD-120虽然是一个视频数据集，但侧重于普通行为识别，并且人和人之间的遮挡不严重。然而，扶梯场景中人与扶梯之间存在多点且长时间的交互，人和人之间的遮挡普遍严重。此外，还需要关注普通行为和危险行为的区别。例如，伸手和探头的幅度会影响普通行为和危险行为的区分。

因此，本文创建了一个扶梯乘客危险行为视频数据集ESC-Danger（escalator danger），该数据集覆盖了扶梯使用过程中乘客易出现的8类扶梯危险行为（倚靠、攀爬、下蹲、伸手、探头、滞留、逆行和奔跑）。为了有效提取危险行为中人和扶梯交互信息，本文通过实时性高的姿态估计算法MMPose[13]提取人体骨架数据，并训练一个稳定性强的关键点检测模型RTMpose[14]提取扶梯关键点。然后使用一种并行式预处理方法得到人体骨架图和扶梯关键点图，再利用距离特征量化危险行为中丰富的交互信息，以此加深模型对普通行为和危险行为的区分。并设计了一个双流人–物交互图卷积网络，利用骨骼和距离特征提取人体行为中丰富的时空信息和人–物交互信息。所提模型在ESC-Danger数据集上性能达到了最优。本文有以下3个方面贡献。

1) 创建了一个扶梯乘客危险行为数据集，数据集中包含8类扶梯危险行为（倚靠、攀爬、下蹲、伸手、探头、滞留、逆行和奔跑）。

2) 使用关键点检测模型提取扶梯关键点，构建扶梯关键点图。并利用人–扶梯交互过程中距离的变化来量化人–物交互信息，从而提升模型对人和扶梯之间交互信息的建模能力。

3) 设计了一个双流人–物交互图卷积网络，同时建模骨骼信息和距离信息，捕捉人和扶梯之间多点且长时间的交互。本文提出的方法在构建的扶梯乘客危险行为数据集上性能达到了最优。

## 1　行为识别方法介绍

### 1.1　基于骨架的人体行为识别

基于骨架的人体行为识别由于其鲁棒性和轻量性，逐渐成为行为识别领域的主流方法。有研究采用RNN模拟骨架数据中的上下文依赖关系[15-16]。Liu等[17]和Duan等[18]将人体骨架序列转换成姿态估

计热图进行行为识别，但是，此类方法会破坏骨架序列关节之间的自然连接关系。随着研究的深入，基于GCN的方法取得了出色的表现。

时空图卷积神经网络（ST-GCN, spatial temporal graph convolutional network）[1]可以在不破坏骨架内部拓扑性的情况下，实现高效的行为识别。Shi等[5]用一种有向无环图建模关节和节点之间的依赖关系，BlockGCN[19]引入了拓扑编码和Block图卷积来保持骨架拓扑结构，并有效地捕获多关系信息，提高了性能和效率。Shi[20]提出了一种双流网络，同时提取骨架序列中的一阶和二阶信息。Lin等[21]利用一种动作依赖的对比学习方法，自适应建模人体骨架数据中的静态和动态信息。Lee等[22]提出了一个层次分解图卷积网络来建模时空依赖性，提供自适应时空覆盖。这些工作只关注人体骨架数据中时空信息的建模，鲜有考虑物体信息对行为识别的贡献。在扶梯乘客危险行为识别中，人不断与扶梯产生交互，引入物体信息能帮助模型更好地理解和识别人的行为。

### 1.2　人物交互检测

在扶梯乘客危险行为中，人和扶梯会产生频繁的交互，引入物体信息是必要的。在研究人–物交互时，人的动作和物体信息可以作为彼此的上下文，早期研究者通过贝叶斯模型[23]、分层随机场[24]和支持向量机[25]来学习这种上下文交互。深度神经网络通过探索人–物空间关系[11]、人–谓词–对象三元组[26]、3D人–物对之间的相对位置[10]来建模人–物之间的交互信息。Qi等[27]和Zhou等[9]通过消息传递框架来推断身体部位与周围物体之间的关系。在物体和人体姿态检测器的帮助下，利用坐标特征探索人与物体的关系。Zhang等[28]利用人–物目标框作为空间指导，进而探索谓词视觉语境在人–物交互检测中的应用。Yang[29]利用人与物体之间的交互意图和几何相关性来捕获人–物之间的关系，消除交互不确定性并预测合理的交互元素。Wang等[10]通过感兴趣区（ROI, region of interest）池化提取物体边界框，并将其当作物体节点，计算骨架序列与物体节点之间的相对位置，以此引入人–物交互信息。但这种方法只适用于小物体和人交互的场景。

与以往工作不同的是，扶梯是一个相对静止并且体积巨大的物体。发生危险行为时，乘客往往与扶手产生频繁的交互，并且人体姿态的幅度也会影响危险的判别。可见，如何建立扶梯与人之间的上下文联系显得异常重要。因此，本文引入基于距离度量的人–物交互信息，以提升模型对扶梯乘客危险行为的理解。

## 2　双流人–物交互图卷积网络

本节详细介绍所设计的双流人–物交互图卷积网络。具体包括人体骨架和扶梯关键点的提取和预处理、基于距离度量的人–物交互方式以及双流人–物交互图卷积网络。双流人–物交互图卷积网络框架如图1所示。为了充分提取扶梯乘客危险行为中丰富的时空信息，本文融合人–扶梯距离特征和人体骨架特征搭建了一个双流人–物交互图卷积网络。具体地，两流支路均采用ST-GCN++[30]作为主干网络。为了在确保模型识别性能的同时，避免层数太

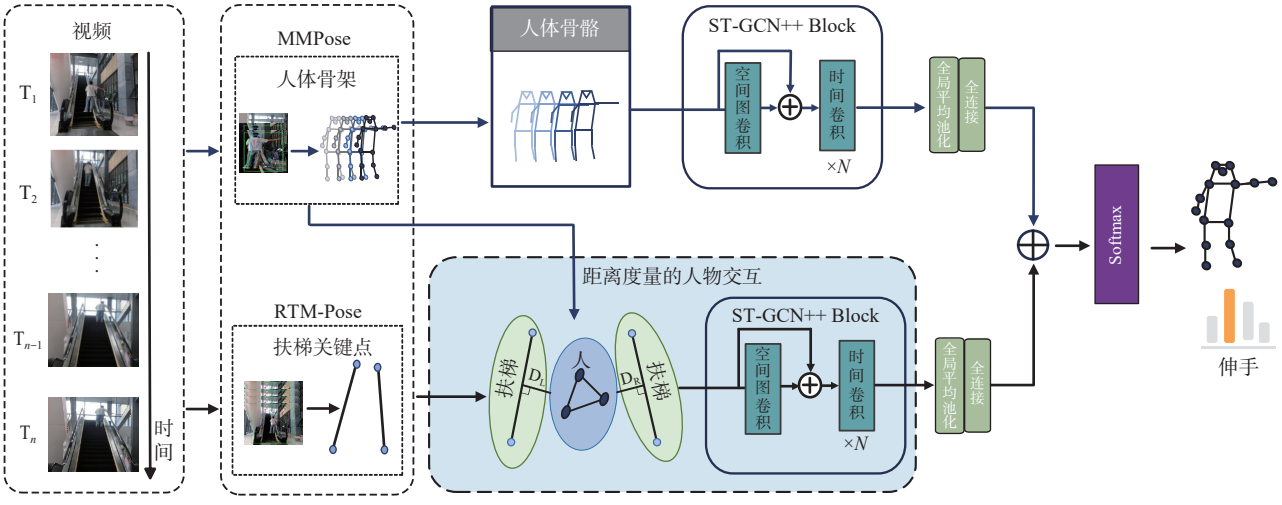



图1　双流人–物交互图卷积网络框架

深而带来过平滑问题[31]，每流支路都使用9层ST-GCN++块。并且将提取特征进行全局平均，并通过全连接层得到每个支路的各类别分数得分。此外，为了更好地融合两条支路的特征，使用双流融合的方式融合特征，并利用Softmax进行最终的分类。

### 2.1　人体骨架和扶梯关键点提取

在扶梯乘客危险行为中，人和人之间遮挡严重并且危险行为发生时，人主要和扶梯的扶手进行交互，将以往工作中使用RoIAlign算法[32]和YOLOv5算法[33]提取的外观特征和目标框作为输入是次优的。因此本文使用预训练的MMPose模型提取人体骨架特征，并基于RTMpose模型提取扶梯关键点特征。RTMpose是一种自上而下的关键点提取模型，先采用FasterRCNN[34]算法检测物体的目标框，再用以CSPNeXt[35]为主干网络的RTMpose模型提取扶梯关键点。以往训练这两个模型的训练集中不包含扶梯场景，因此本文在自建数据集中标注了500张图片，预训练了一个可用于扶梯关键点检测的模型。该模型提取扶梯左上、右上、左下、右下的4个扶梯关键点以及左右两条扶手边，以此引入扶梯特征。为了降低噪声和视频帧数不一致对行为识别的影响，本文设计了一种并行式的数据预处理方法，数据预处理如图2所示。

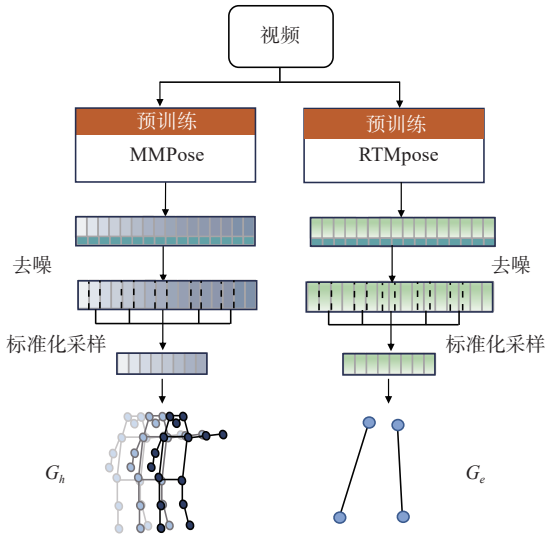



图2　数据预处理

图2首先通过去噪过程，去除人体骨架和扶梯关键点特征中置信度分数小于阈值的噪声数据，再通过标准化采样从每个视频中提取相同数量的帧序列。经过预处理之后可以得到人体$H$的骨架图$G_h=(V_h,E_h)$和扶梯$E$的关键点图$G_e=(V_e,E_e)$，其中$V_h$为骨架图中$T$帧的节点集，$V_h=\{v_{ti_h}|t=1,\cdots,T,i_h=1,\cdots,N_h\}$，人体骨架关节点总数$E_h=17$，$E_h$表示人体骨架图的边，它包含两个部分，如式(1)~式(2)所示

$$E_{s_h}=\{v_{ti_h}v_{tj_h}|(i_h,j_h)\in H\} \tag{1}$$

$$E_{F_h}=\{v_{ti_h}v_{(t+1)i_h}\} \tag{2}$$

其中，$E_{s_h}$代表第$i_h$个关节点$v_{ti_h}$和第$j_h$个关节点$v_{tj_h}$自然连接而成的骨骼边，$E_{F_h}$代表同一人体关节点相邻帧之间连接而成的边。在实际监控视频中，扶梯关键点可能随着镜头抖动而发生位置改变。因此，扶梯关键点图$G_e=(V_e,E_e)$也采用类似方式进行构建，其节点集$V_e=\{v_{ti_e}|t=1,\cdots,T,i_e=1,\cdots,N_e\}$，扶梯关键点总数$N_e=4$，$E_e$是扶梯扶手边，它同样包含两个部分，分别如式(3)~式(4)所示

$$E_{s_e}=\{v_{ti_e}v_{tj_e}|(i_e,j_e)\in E\} \tag{3}$$

$$E_{F_e}=\{v_{ti_e}v_{(t+1)i_e}\} \tag{4}$$

其中，$E_{s_e}$代表第$i_e$个关键点$v_{ti_e}$和第$j_e$个关键点$v_{tj_e}$连接而成的扶梯扶手边。$E_{F_e}$代表相邻帧之间扶梯关键点连接而成的边。

### 2.2　基于距离度量的人–物交互检测

现有基于骨骼的行为识别模型的输入只包含人体信息，鲜少考虑物体和场景的信息。这不仅无法引入人–物交互信息，还无法显性地表征人体行为幅度和物体之间的上下文信息，导致模型无法有效判别普通行为和危险行为。鉴于此，本文基于上节提取出的扶梯关键点特征，分别度量人体17个骨架关节点与扶梯两条扶手边的距离，以此量化人与扶梯交互过程中丰富的时空信息。以第$t$帧为例，基于距离度量的人–物交互检测如图3所示。

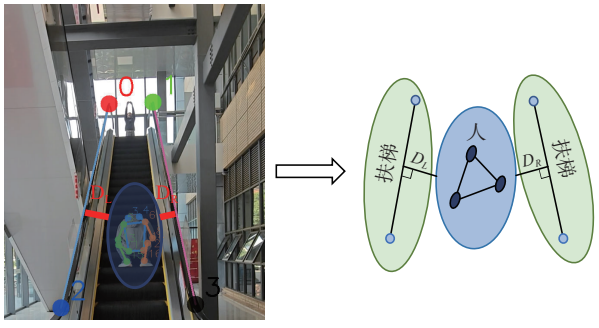



图3　基于距离度量的人–物交互检测

图 3 中，$D_L$ 与 $D_R$ 分别表示人体 17 个关键点和扶梯左右扶手边距离的集合。将扶梯的 4 个关键点坐标输入式(5)~式(7)可以计算出扶梯两条扶手边的直线方程系数

$$A = Y_u - Y_b \tag{5}$$

$$B = X_b - X_u \tag{6}$$

$$C = X_b \cdot Y_u - X_u \cdot Y_b \tag{7}$$

其中，$(X_b, Y_b)$ 和 $(X_u, Y_u)$ 均代表扶梯关键点坐标，$b \neq u$ 且 $b, u \in N_e$。

根据式(5)~式(6)，可以计算出 $A_r, B_r, C_r$ 以及 $A_l, B_l, C_l$ 进而获得右扶手边 $A_r x + B_r y + C_r = 0$ 和左扶手边 $A_l x + B_l y + C_l = 0$ 的表达式。接着，通过距离计算式(8)~式(9)分别计算人体骨架图中每个关节点到扶梯左、右扶手的距离

$$d_{li_h} = \frac{|A_l x_{i_h} + B_l y_{i_h} + C_l|}{\sqrt{A_l{}^2 + B_l{}^2}} \tag{8}$$

$$d_{ri_h} = \frac{|A_r x_{i_h} + B_r y_{i_h} + C_r|}{\sqrt{A_r{}^2 + B_r{}^2}} \tag{9}$$

这种方法不需要多余的 RGB 特征，不会增加计算量，并且能更好地识别所发生行为是否危险。例如，在伸手和探头行为中，头和手与较远的扶手边距离会达到峰值，通过学习这种特征，可以很好地区分行为是否危险，并且具有较高的可解释性。

### 2.3　双流人–物交互图卷积网络

本节将介绍双流人–物交互图卷积网络中的各个模块。

#### 2.3.1　空间图卷积网络

ST-GCN++模型中，空间图卷积（SGC, spatial graph convolution）可以定义为

$$f_{\text{out}(v_{ti})} = \sum_{v_{tj} \in N(v_{ti})} \frac{1}{\mathrm{Z}_{ti}(v_{tj})} f_{\text{in}}(v_{tj}).w(l_{ti}(v_{tj})) \tag{10}$$

其中，在骨骼支路中，$v_{ti}$ 表示第 $t$ 帧的第 $i$ 个人体关节点，$f_{\text{in}}(\cdot)$ 和 $f_{\text{out}}(\cdot)$ 分别代表对应关节点的输入及输出特征，$N(v_{ti})$ 代表 $v_{ti}$ 的相邻节点集，设置 $Z_{ti}$ 来归一化不同相邻节点对 $v_{ti}$ 的重要性，$w(\cdot)$ 是权重函数，用于分配由标签函数 $l_{ti}(\cdot)$ 索引的权重。本文采用图中关节点之间的距离 $d_g$ 来表示 $l_{ti}(\cdot)$，即 $l_{ti}(\cdot) = d_g(v_{ti}, v_{tj})$，具有相同距离的关节将形成一个子集，并共享一个可学习的权重函数。

然而只对人体关节点进行建模，无法捕捉人体与周围物体（如扶梯扶手）之间的空间关系。在现实场景中，乘客在扶梯上移动时，人体关节点与扶手之间的距离会发生显著变化。这些动态变化在捕捉和分析乘客行为时具有重要意义。例如，当乘客在扶梯上伸手去抓扶手，或者探头查看前方情况时，这些动作都会导致关节点与扶手之间的距离发生变化。如果仅依赖于人体关节点的时空信息，很难充分捕捉这些变化，会丢失大量乘客和扶梯之间的交互信息。引入关节点与扶手之间的距离信息，可以更有效地捕捉这些危险行为，提高模型对危险行为的识别准确性。因此在距离流中，利用 SGC 处理人体关节点和扶梯扶手间的距离特征 $d_{ti}$，$d_{ti}$ 包含第 2.2 节中计算得到 $d_{ri_h}$ 和 $d_{li_h}$。在式(10)中人体关节点 $v_{ti} = (x_{i_h}, y_{i_h})$，为了保证形状的一致，在距离流中使用 4 种不同构造方式的距离特征 $d_{ti}$ 替换 $v_{ti}$（具体可见实验部分第 3.4.2 节），距离流中的图卷积如式(11)所示

$$f_{\text{dout}(d_{ti})} = \sum_{d_{tj} \in N(d_{ti})} \frac{1}{\mathrm{Z}_{ti}(d_{tj})} f_{\text{din}}(d_{tj}).w(l_{d_{ti}}(d_{tj})) \tag{11}$$

其中，$f_{\text{din}}(\cdot)$ 和 $f_{\text{dout}}(\cdot)$ 分别代表对应距离流的输入及输出特征，标签函数 $l_{d_{ti}}(\cdot) = d_g(d_{ti}, d_{tj})$。

图卷积中通常使用系数矩阵 $\boldsymbol{A}$ 来表示权重函数，相应地，式(10)可改写为式(12)，式(11)可改写为式(13)

$$f_{\text{out}} = \sum_{d_g=0}^{D_g} W_{d_g} f_{\text{in}}(\boldsymbol{A} \odot M_{d_g}) \tag{12}$$

$$f_{\text{dout}} = \sum_{d_g=0}^{D_g} W'_{d_g} f_{\text{din}}(\boldsymbol{A}' \odot M'_{d_g}) \tag{13}$$

其中，$D_g$ 为预定义的最大图距离，$f_{\text{in}}$ 和 $f_{\text{out}}$ 表示骨骼流的输入和输出特征映射，$f_{\text{din}}$ 和 $f_{\text{dout}}$ 表示距离流的输入和输出特征映射，$\odot$ 表示逐元素相乘，系数矩阵 $\boldsymbol{A} = \boldsymbol{\Lambda}_{d_g}^{-\frac{1}{2}} \boldsymbol{A}_{d_g} \boldsymbol{\Lambda}_{d_g}^{-\frac{1}{2}}$，$\boldsymbol{A}' = \boldsymbol{\Lambda}'^{-\frac{1}{2}}_{d_g} \boldsymbol{A}'_{d_g} \boldsymbol{\Lambda}'^{-\frac{1}{2}}_{d_g}$。$\boldsymbol{A}_{d_g}$ 表示骨骼流中图距离为 $d_g$ 的节点对的第 $d_g$ 阶邻接矩阵，$\boldsymbol{A}'_{d_g}$ 表示距离流中图距离为 $d_g$ 的节点对的第 $d_g$ 阶邻接矩阵。$\boldsymbol{\Lambda}_{d_g}$ 用于对 $\boldsymbol{A}_{d_g}$ 进行归一化，$\boldsymbol{\Lambda}'_{d_g}$ 用于对 $\boldsymbol{A}'_{d_g}$ 进行归一化。$W_d$、$M_d$、$W'_{d_g}$ 和 $M'_{d_g}$ 都是可学习的参数，分别用于骨骼流和距离流实现卷积运算和调整每条边的重要性。

SGC 网络中利用不同的 $\boldsymbol{A}$ 将不同的节点特征融合在一起，ST-GCN 用一组可学习的权重对系数矩阵中的每个元素进行重新加权。然而，在 ST-

GCN++中，只使用预定义的联合拓扑来初始系数矩阵。在训练过程中采用梯度下降法迭代更新系数矩阵，不需要任何稀疏约束。

#### 2.3.2　多分支时间卷积网络

传统的时空卷积模块中，只使用一个内核较大（一般设置为9）的1维Conv作为时空卷积模块[1]。获得一个较大的感受野，但会带来很大的参数损耗。受Liu等[6]的启发，STGCN++[32]中使用多分支的时间卷积网络，取代传统单分支的时间卷积网络（TCN, temporal convolutional network）。该网络由1个1 × 1 Conv支路、1个Max-Pooling支路和4个1维Conv分支组成。这4个1维Conv的内核大小均为3，膨胀系数为1~4。这种设计方式不仅能够提高时间建模能力，而且由于减小了每个分支的内核大小，能够节省计算参数量。

#### 2.3.3　多特征融合

文献[10]和文献[20]表明，多特征融合可以提高行为识别精度。其中，2s-AGCN[20]设计了两条支路分别处理关节点和骨骼信息，并且沿用文献[36]中的双流融合方式融合两个支路的信息。本文基于这种双流网络的思想，设计了一个双流人–物交互图卷积网络，融合距离流和骨骼流，在分数层融合距离通道的特征$H_d$和骨骼通道的特征$H_b$得到双流网络的输出特征再利用分类器获得最终的扶梯乘客危险行为识别结果。

## 3　实验

本节将通过实验展开验证，包括数据集、评估指标、消融实验和结果分析等部分。

### 3.1　数据集

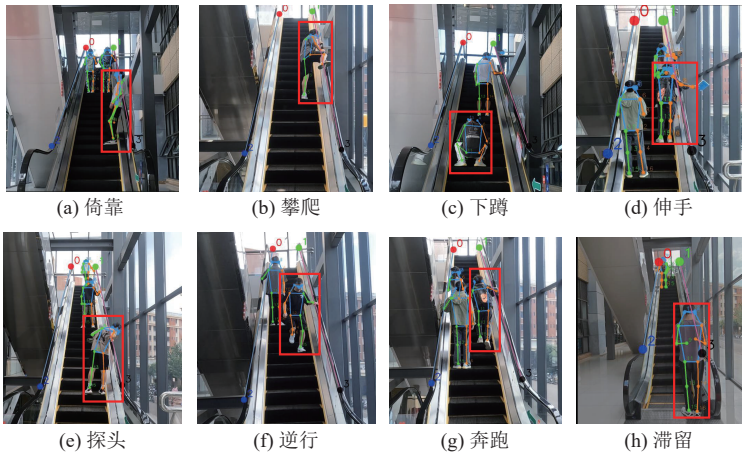
本文创建了一个扶梯乘客危险行为视频数据集ESC-Danger。这是一项具有挑战性的任务，因为扶梯是一个窄长的大型设备，乘客之间的遮挡比其他场景更严重，并且扶梯上人的各类行为与扶梯交互频繁。数据集中包含5名男性和1名女性，年龄在22~24岁，被随机分组并执行8种不同的危险行为。这8种危险行为包括：倚靠、攀爬、下蹲、伸手、探头、滞留、逆行和奔跑。每种行为都包含单人、双人、多人3种情况。根据实际应用中扶梯监控头的安放位置，视频从两个角度拍摄（自上而下和自下而上）。数据集中总共有312个视频，帧率为每秒30帧。总帧数为93 600，平均持续时间为10 s。通过MMpose提取人体骨架和RTMpose提取扶梯关键点，最终每个视频包含RGB画面、人体骨架数据和扶梯关键点数据3类特征。数据集部分实例图如图4所示。在实验中，训练集和测试集的比例为3∶1。

### 3.2　评价指标与实验细节

在实验中，本文使用Top-1精度作为评价指标，使用动量为0.9、权重衰减为0.000 5的SGD优化器。初始学习率为0.1，批次处理大小设置为16，并训练200个周期。所有视频在送入模型之前都只抽取其中的100帧，以保证数据大小一致。本文所有实验都在一块24 GB显存的RTX 4090 GPU上进行。

### 3.3　实验结果分析

本节比较了本文所提方法和不同基线模型在ESC-Danger数据集上的效果，在ESC-Danger数据集上的对比实验结果见表1。ST-GCN首次将图卷



(a) 倚靠　(b) 攀爬　(c) 下蹲　(d) 伸手

(e) 探头　(f) 逆行　(g) 奔跑　(h) 滞留

图4　数据集部分实例图

积网络应用于行为识别领域。本文模型的Top-1精度比ST-GCN精度提升了11.11%。2s-AGCN设计了一种双流自适应图卷积网络进行行为识别，本文方法也延续了这种双流设计的思想，并且最终效果优于2s-AGCN。PoseC3D、CTR-GCN和ST-GCN++都是目前先进的基线模型，本文方法也优于这3种模型。并且实验证明本文方法在较低计算复杂度的前提下，也能获得较好的效果。在实时性分析中，由于本文所使用和对比的方法，基本都是基于轻量化的骨架特征实现的，在实时性方面，均有不错的表现。

**表1　在ESC-Danger数据集上的对比实验结果**

| 对比模型 | Top-1精度 | GFLOs | MParms | 检测时间/ms |
|---|---|---|---|---|
| ST-GCN[1] | 83.95% | 5.34 | 3.08 | 19.2 |
| PoseC3D[18] | 85.19% | 15.9 | 2.0 | **17.8** |
| 2s-AGCN[20] | 87.65% | 6.07 | 3.77 | 19.0 |
| CTR-GCN[37] | 88.89% | 2.82 | 1.43 | 18.7 |
| ST-GCN++[32] | 93.82% | 2.80 | 1.39 | 18.6 |
| 本文方法 | **95.06%** | **2.80** | **1.39** | 18.6 |

## 3.4　消融实验

为了验证模型整体框架的有效性，在ESC-Danger数据集上进行了一系列的消融实验。

### 3.4.1　不同主干网络和数据预处理模块消融

本文引入了距离信息度量危险行为中的人–物交互信息。为了验证其有效性，本节选择不同的主干网络搭建模型进行消融实验。不同主干网络消融结果见表2。消融实验中使用了目前行为识别领域常用的3种主干网络：ST-GCN，CTR-GCN和ST-GCN++。结果表明，无论是哪种主干网络，引入距离信息和对数据进行预处理都能提升模型性能，这说明本文引入的距离特征能够有效地提高危险行为识别模型的性能。不同输入帧数消融结果如图5所示，使用STGCN++作为主干网络搭建模型，对不同输入帧数量进行了消融。实验证明当输入帧数量为100时，模型性能最好，并且在输入帧数量不同时，引入距离信息依旧能提升模型性能。

**表2　不同主干网络消融结果**

| 主干网络 | 精度 | | |
|---|---|---|---|
| | Top-1 | w/o距离信息 Top-1 | w/o数据预处理Top-1 |
| ST-GCN | 86.42% | 83.95% | 86.42% |
| CTR-GCN | 90.12% | 88.89% | 87.65% |
| ST-GCN++ | 95.06% | 93.82% | 91.36% |

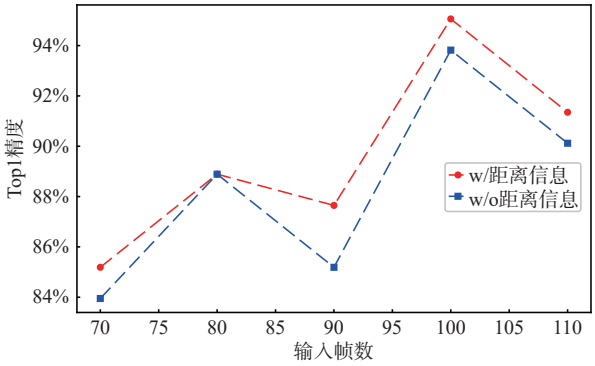



图5　不同输入帧数消融结果

### 3.4.2　距离特征消融

在引入距离特征时，本文使用人体关节点和扶梯两条扶手边的距离代替关节点坐标。具体使用了4种替换方式。①只使用$d_{li_h}$代替$x_{i_h}$、$y_{i_h}$赋零。②只使用$d_{ri_h}$代替$x_{i_h}$、$y_{i_h}$赋零。这两种方法只使用一边的距离，是因为在实际场景中，扶梯左右两边危险性是不一样的。③使用$(d_{li_h}, d_{ri_h})$代替$(x_{i_h}, y_{i_h})$。这种方法充分考虑人体关节点和左右扶手之间距离的变换。④选择人体关节点到扶梯扶手的最大的距离Max代替$x_{i_h}$、$y_{i_h}$赋零。这是因为人和扶梯交互的过程中，人和扶梯之间的距离会达到峰值。例如，伸手行为中，伸出的手和扶梯的距离会达到峰值，然后收回。不同距离替换方式实验结果见表3，其中使用距离最大值做替换的效果最好。

**表3　不同距离替换方式实验结果**

| 距离替换方式 | Top-1精度 |
|---|---|
| $(x_{i_h}, y_{i_h}) \rightarrow (d_{li_h}, 0)$ | 91.98% |
| $(x_{i_h}, y_{i_h}) \rightarrow (d_{ri_h}, 0)$ | 92.59% |
| $(x_{i_h}, y_{i_h}) \rightarrow (d_{li_h}, d_{ri_h})$ | 93.21% |
| $(x_{i_h}, y_{i_h}) \rightarrow (\mathrm{Max}, 0)$ | 95.06% |

### 3.4.3　不同支路的消融

首先对双流图卷积网络的不同支路进行了消融实验，双流网络中不同支路消融结果见表4。虽然距离流（Distance）的效果没有骨骼流（Bone）好，但是优于关节点流（Joint），说明距离特征对于扶梯乘客危险行为识别是有效的。在本文双流人–物交互图卷积网络中，使用骨骼流作为基准流，再融合其他流的特征。在使用Bone+Distance搭建双流网络时，能够同时建模人体骨架数据中的时空

特征和人与扶梯之间的交互特征，并能达到最好的精度95.06%。

**表4　双流网络中不同支路消融结果**

| 双流图卷积网络支路 | Top-1精度 |
|---|---|
| Joint | 71.60% |
| Distance | 74.07% |
| Bone | 92.59% |
| Bone + Joint | 93.82% |
| Bone + Distance | 95.06% |

### 3.5　交叉验证实验

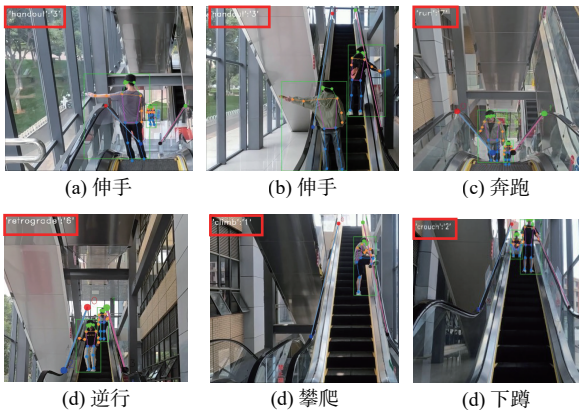
为了进一步验证所提方法的泛化性，本节按照拍摄主体，从6个演员中选择1个演员的视频作为测试集，其他人的视频作为训练集。交叉验证实验结果见表5。从表5可以看出，引入距离信息可以提高模型性能，同时证明了本文所提方法的泛化性。

**表5　交叉验证实验结果**

| 交叉验证 | Top-1精度 |
|---|---|
| w/o距离信息 | 92.55% |
| w/距离信息 | 93.62% |

### 3.6　可视化结果及各行为精度对比

扶梯乘客危险行为可视化结果如图6所示，每张结果图的左上角均显示了识别出的扶梯乘客危险行为及类别编号。结果表明无论在单人还是多人场景中，所提模型都能有效识别出危险行为。各行为精度如图7所示，可以看出，倚靠、下蹲、伸手、滞留和奔跑几类在引入距离信息之后，识别准确率都上升了，这也证明了本文所提方法的有效性。



(a) 伸手　(b) 伸手　(c) 奔跑

(d) 逆行　(d) 攀爬　(d) 下蹲

图6　扶梯乘客危险行为可视化结果

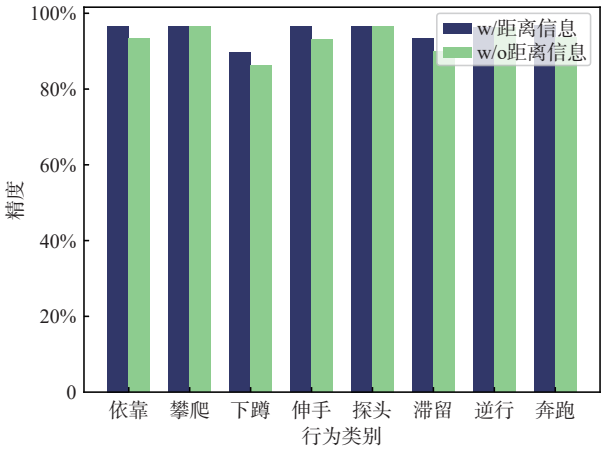



图7　各行为精度

## 4　结束语

本文设计了一个双流人–物交互图卷积网络进行扶梯乘客危险行为识别。具体工作包括：①搭建了一个包含8类危险行为的扶梯乘客危险行为数据集，通过深入研究这些数据，可以揭示扶梯乘客危险行为的规律；②通过关键点检测模型提取扶梯关键点信息，能够引入轻量化的物体特征，并利用距离度量的人–物交互特征加深模型对危险行为的理解。结合这两点，本文提出的双流人–物交互图卷积网络，能够同时提取人体骨架中的时空信息和人与扶梯间的人–物交互信息。并且该方法在本文自建的扶梯乘客危险行为数据集上性能达到最优。在未来的工作中，将进一步研究人–物交互在扶梯乘客危险行为中的应用，以满足模型在扶梯危险行为实际应用场景中的应用需求。

**参考文献：**

[1] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2018, 32(1): 7444-7452.

[2] LI Y L, LIU X P, LU H, et al. Detailed 2D-3D joint representation for human-object interaction[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10163-10172.

[3] JIANG X H, XU K, SUN T F. Action recognition scheme based on skeleton representation with DS-LSTM network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(7): 2129-2140.

[4] YANG Z Y, LI Y C, YANG J C, et al. Action recognition with spatio-temporal visual attention on skeleton image sequences[J].

IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(8): 2405-2415.

[5] SHI L, ZHANG Y F, CHENG J, et al. Skeleton-based action recognition with directed graph neural networks[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 7904-7913.

[6] LIU Z Y, ZHANG H W, CHEN Z H, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 140-149.

[7] 杨学存, 李杰华, 陈丽媛, 等. 基于人体骨架的扶梯乘客异常行为识别方法[J]. 安全与环境学报, 2024, 24(2): 636-643.
YANG X C, LI J H, CHEN L Y, et al. An abnormal behavior recognition method of escalator passengers based on human skeletons[J]. Journal of Safety and Environment, 2024, 24(2): 636-643.

[8] GKIOXARI G, GIRSHICK R, DOLLÁR P, et al. Detecting and recognizing human-object interactions[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8359-8367.

[9] ZHOU P H, CHI M M. Relation parsing neural network for human-object interaction detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 843-851.

[10] WANG H R, YU B S, LI J Q, et al. Multi-stream interaction networks for human action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 3050-3060.

[11] CHAO Y W, LIU Y F, LIU X Y, et al. Learning to detect human-object interactions[C]//Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2018: 381-389.

[12] KOPPULA H S, SAXENA A. Anticipating human activities using object affordances for reactive robotic response[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 14-29.

[13] SENGUPTA A, JIN F, ZHANG R Y, et al. Mm-pose: real-time human skeletal posture estimation using mmWave radars and CNNs[J]. IEEE Sensors Journal, 2020, 20(17): 10032-10044.

[14] JIANG T, LU P, ZHANG L, et al. RTMPose: real-time multi-person pose estimation based on MMPose[EB]. 2023.

[15] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]//Proceedings of the Computer Vision-ECCV 2016. Cham: Springer, 2016: 816-833.

[16] ZHU W T, LAN C L, XING J L, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2016: 3697-3703.

[17] LIU M Y, YUAN J S. Recognizing human actions as the evolution of pose estimation maps[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1159-1168.

[18] DUAN H D, ZHAO Y, CHEN K, et al. Revisiting skeleton-based action recognition[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 2959-2968.

[19] ZHOU Y X, YAN X D, CHENG Z Q, et al. BlockGCN: redefine topology awareness for skeleton-based action recognition[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 2049-2058.

[20] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 12018-12027.

[21] LIN L L, ZHANG J H, LIU J Y. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 2363-2372.

[22] LEE J, LEE M, LEE D, et al. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2024: 10410-10419.

[23] GUPTA A, KEMBHAVI A, DAVIS L S. Observing human-object interactions: using spatial and functional compatibility for recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1775-1789.

[24] YAO B P, LI F F. Modeling mutual context of object and human pose in human-object interaction activities[C]//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2010: 17-24.

[25] DESAI C, RAMANAN D, FOWLKES C. Discriminative models for static human-object interactions[C]//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. Piscataway: IEEE Press, 2010: 9-16.

[26] LIU Y, YUAN J S, CHEN C W. ConsNet: learning consistency graph for zero-shot human-object interaction detection[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 4235-4243.

[27] QI S Y, WANG W G, JIA B X, et al. Learning human-object interactions by graph parsing neural networks[C]//Proceedings of the Computer Vision - ECCV 2018. Cham: Springer, 2018: 407-423.

[28] ZHANG F Z, YUAN Y H, CAMPBELL D, et al. Exploring predicate visual context in detecting of human-object interactions[C]//Proceedings of the 2023 IEEE/CVF International Conference on

Computer Vision (ICCV). Piscataway: IEEE Press, 2024: 10377-10387.

[29] YANG Y H, ZHAI W, LUO H C, et al. LEMON: learning 3D human-object interaction relation from 2D images[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 16284-16295.

[30] DUAN H D, WANG J Q, CHEN K, et al. PYSKL: towards good practices for skeleton action recognition[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 7351-7354.

[31] CHEN M, WEI Z W, HUANG Z F, et al. Simple and deep graph convolutional networks[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM Press, 2020: 1725-1735.

[32] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2980-2988.

[33] 王源鹏, 万海斌, 黄凯, 等. 基于YOLOv5s的自动扶梯乘客异常行为实时检测算法[J]. 激光与光电子学进展, 2024, 61(8): 0812004.
WANG Y P, WAN H B, HUANG K, et al. Real-time detection of abnormal behavior of escalator passengers based on YOLOv5s[J]. Laser & Optoelectronics Progress, 2024, 61(8): 0812004.

[34] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[35] LYU C Q, ZHANG W W, HUANG H A, et al. RTMDet: an empirical study of designing real-time object detectors[EB]. 2022.

[36] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. New York: ACM Press, 2014: 568-576.

[37] CHEN Y X, ZHANG Z Q, YUAN C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 13339-13348.

[作者简介]

**邓鑫**（1999– ），男，云南大学信息学院硕士生，主要研究方向为计算机视觉、行为识别。

**谷金晶**（1990– ），女，博士，云南大学信息学院讲师、硕士生导师，主要研究方向为跨媒体语义分析与理解、视频异常行为识别、视频描述。

**赵征鹏**（1972– ），男，云南大学信息学院副教授、硕士生导师，主要研究方向为信号与信息处理、计算机系统及应用。

**普园媛**（1972– ），女，博士，云南大学信息学院教授，主要研究方向为图像风格迁移、多模态情感分析、视觉媒体计算。

**徐丹**（1968– ），女，博士，云南大学信息学院教授，主要研究方向为图形绘制技术、图像融合、虚拟现实、视觉计算及认知。