文章编号:1674-2974(2024)12-0129-10 DOI:10.16339/j.cnki.hdxbzkb.2024290

基于特征交互模块增强RGB-骨骼动作识别鲁棒性研究

侯永宏 ,刘超 ,刘鑫†,岳焕景 ,杨敬钰
(天津大学 电气自动化与信息工程学院, 天津 300072)

摘 要:恶意攻击者可以通过在自然样本中添加人类无法察觉的对抗噪声轻易地欺骗神经网络,从而导致分类错误.为了增强模型对抗扰动的鲁棒性,先前的研究大多关注单模态任务,对多模态场景的研究相对匮乏.为了提升多模态RGB-骨骼动作识别的鲁棒性,提出了一个基于特征交互模块(FIM)的鲁棒动作识别框架,提取对抗样本的全局信息并学习模态间的联合表征,以此来校准多模态特征.实验结果表明,面对CW攻击,该动作识别框架在NTURGB+D数据集上进行鲁棒性评估,其RI值达到25.14%,平均鲁棒准确率也达到48.99%,比最新的MinSim+ExFMem方法分别提高了8.55和23.79个百分点,表明其在增强鲁棒性及平衡准确率方面都优于其他方法.

关键词:计算机视觉;多模态;RGB-骨骼动作识别;对抗训练
中图分类号:TP391.4 **文献标志码**:A

Study on Enhancing the Robustness of RGB-skeleton Action Recognition Based on the Feature Interaction Module

HOU Yonghong, LIU Chao, LIU Xin†, YUE Huanjing, YANG Jingyu
(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: Malicious attackers can easily deceive neural networks by adding human-imperceptible adversarial noise to natural samples, leading to misclassification. To enhance the model's robustness against such adversarial perturbations, previous research has predominantly concentrated on the robustness of single-modal tasks, with insufficient exploration of multimodal scenarios. Therefore, this paper aims to improve the robustness of multimodal RGB-skeleton action recognition and introduces a robust action recognition framework based on a Feature Interaction Module (FIM), which extracts global information from adversarial samples to learn inter-modal joint representations for calibrating multi-modal features. A corresponding loss function tailored to this framework is also developed. Experimental results demonstrate that against CW attack, our method achieves a RI of 25.14% and an

* **收稿日期**:2023-11-30
基金项目:国家自然科学基金资助项目(62072331), National Natural Science Foundation of China(62072331);国家自然科学基金资助项目(62231018), National Natural Science Foundation of China(62231018);国家自然科学基金资助项目(62171309), National Natural Science Foundation of China(62171309)
作者简介:侯永宏(1968—),男,天津人,天津大学教授,博士生导师,博士
† 通信联系人,E-mail: linuxsino@gmail.com

average robust accuracy of 48.99% on the NTURGB+D dataset, outperforming the latest SimMin+ExFMem method by 8.55 and 23.79 percentage points, respectively. These findings confirm that our approach surpasses others in enhancing robustness and balancing accuracy rates.

Key words: computer vision; multimodal; RGB-skeleton action recognition; adversarial training

深度神经网络(DNN)在视觉任务如分割[1]、检测[2]和超分辨率[3]中取得显著进展,但对于微小的对抗扰动极为敏感,容易受到恶意攻击的影响[4],尤其在视频动作识别[5]、人的再识别[6]和自然语言处理[7]等领域.鉴于DNN在自动驾驶[8-9]和医疗诊断[10-12]等安全关键领域的应用,增强其鲁棒性成为聚焦点.

为此,学界探索了对抗攻击策略[13-15]和防御策略[16-25].一方面,提出了多种生成对抗样本的方法[13-15].另一方面,提出了诸如模型修剪[16]、量化[17]、特征去噪[18]、输入净化[19]、对抗训练(adversarial training, AT)[15]及其变体[20-22]、通道激活抑制[23]、防御蒸馏[24]和模型压缩[25]等多种防御方法.AT是提升鲁棒性的有效方法.它通过在训练中引入l_p范数球内的微小扰动,将对抗样本纳入训练过程中,有效平滑了自然样本周围的损失曲面,提高了模型在特征空间边界区域的分类能力.

随着人工智能的发展,多模态技术在视觉-语言导航[26]、图像-文本匹配[27],医疗图像及诊断[28-29]和人机交互[30]等领域取得了巨大成功.然而,与单模态领域(图像分类[31]、视频理解[32])的研究相比,多模态模型鲁棒性的研究相对滞后,因此加强多模态模型鲁棒性的研究迫在眉睫.最近,学界开始关注vision-language(VL)模型的对抗攻击和鲁棒性评估的研究.Zhang等[33]针对流行的VL预训练模型提出了协同多模态对抗性攻击.Zhao等[34]评估了大型VL模型的对抗鲁棒性,揭示了操纵视觉输入相对于文本输入在欺骗整个VL系统方面的有效性.

本文将鲁棒性研究扩展至多模态,重点探索RGB-骨骼动作识别[35-37],其作为一种代表性多模态任务,亟待鲁棒性提升.通过分析RGB-骨骼数据的特点,本文提出了增强策略.

文献[38]研究了多模态攻击下视觉-听觉模型的鲁棒性.实验表明多模态攻击下视觉-听觉模型鲁棒性未增强,甚至反而降低.因此,提出了一种基于视听不相似性约束(MinSim)和外部特征记忆库(ExFMem)的防御方法,尽管这是目前多模态防御领域最先进的方法,但存在磁盘空间需求大和未充分考虑模态间鲁棒性差异的问题.本文提出一种创新的多模态防御策略,克服这些限制,提高对抗环境下的性能.

本文深入分析了两种模态数据的鲁棒性.尽管RGB视频丰富的视觉信息(颜色、背景等)有助于动作识别,但其对环境变化和对抗攻击的敏感性也增加了模型的脆弱性.相比之下,Yan等人[39]指出,骨骼数据因其简洁性,在抵御光照和场景变化方面表现出更强的鲁棒性,专注于人体核心动作特征,不易受外部环境影响.然而,多模态融合后的鲁棒性并未如预期般提高,反而有所下降,这与文献[38]中的发现一致,表明简单的模态融合并不足以增强模型的鲁棒性.

基于以上分析,本文提出了基于特征交互模块(feature interaction module, FIM)的增强RGB-骨骼动作识别方法.FIM通过全局平均池化层提取两个模态对抗样本的全局信息,生成包含这些信息的联合表征,以此产生多模态激活信号并重新校准各模态的特征.此外,本文还设计了一种新型适配损失函数,以适应本文的方法并训练出一个鲁棒性更强的RGB-骨骼动作识别模型.

本文贡献包括:

1)创新地提出FIM,通过通道维度的特征校准与多模态对抗训练结合,提升模型鲁棒性.

2)设计新的损失函数,融合多个预测结果,优化训练.

3)在主流数据集上进行实验,实验结果表明,在不增加任何存储开销的情况下,本文提出的方法能显著提升RGB-骨骼动作识别鲁棒性,平衡模态

间差异.

1　概述

1.1　标准训练

给定一个标准的训练数据集$D=\{(x_i,y_i)\}_{i=1}^{n}$，其具有$n$个样本，$N$个类别，其中$x_i\in\mathbb{R}^d$为自然样本(没有加入对抗扰动的干净的样本)，$y_i\in\{0,1,\cdots,N-1\}$为对应的真实标签.标准训练通过最小化训练数据集上的经验风险来获得良好的分类性能，如式(1)所示:

$$\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}l(f_\theta(x_i),y_i) \tag{1}$$

式中，l为分类任务中广泛使用的交叉熵损失函数;$f_\theta(x_i):\mathbb{R}^d\to\mathbb{R}^C$是由$\theta$参数化的神经网络;$\Theta$是$\theta$的参数空间.

标准训练中，通过减少预测结果$f_\theta(x_i)$与y_i的交叉熵损失，逐步优化网络模型参数直至收敛.然而，此方法未考虑对抗样本，导致模型面对难以察觉扰动的对抗性样本时表现不佳.图1展示了标准训练与对抗训练的区别.图1(a)展示的一组点代表自然样本，它们可以通过简单的线性决策边界轻易分离，这是标准训练产生的结果.图1(b)表明简单的决策边界无法有效区分数据点周围的"l_p范数球"(此处表示为正方形)，一些对抗性样本(标记为五角星形状)被错误分类.图1(c)意味着分离"l_p范数球"需要一个更加复杂的决策边界.由此得到的分类器对具有有界l_p范数扰动的对抗样本是鲁棒的.

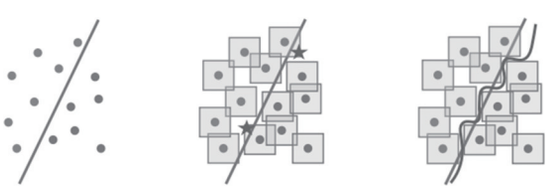


图1　标准训练与对抗训练的决策边界的概念图

Fig.1　Conceptual illustration of standard vs. adversarial decision boundaries

标准训练产生的简单决策边界只能正确分类自然样本.由于对抗样本通常位于自然样本的"l_p范数球"内，标准训练得到的模型无法正确分类这些对抗样本，需要更复杂的决策边界来应对.

1.2　对抗训练(AT)

为了解决此问题，Madry等[15]使用对抗样本来训练神经网络，即对抗训练.对抗训练的基本思想是将对抗样本纳入训练过程中.可以将其表述为最小-最大优化问题:

$$\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\left\{\max_{x'\in\mathcal{B}_\varepsilon[x_i]}l(f(x'),y_i)\right\} \tag{2}$$

式(2)可以看作内部最大化和外部最小化优化的目标函数.为了实现概念上的一致性以及实现更清晰的表达，也可以改写为:

$$\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}l(f_\theta(x_i'),y_i) \tag{3}$$

其中:

$$x_i'=\arg\max_{x'\in\mathcal{B}_\varepsilon[x_i]}l(f_\theta(x'),y_i) \tag{4}$$

式中，$\mathcal{B}_\varepsilon[x_i]$是对抗样本$x_i'$的采样空间，它在自然样本$x_i$的$l_p$范数邻近空间中有界，即$\mathcal{B}_\varepsilon[x_i]=\{x'\in\mathbb{R}^d\,|\,\|x'-x_i\|_p\leqslant\varepsilon\}$，其为半径$\varepsilon$，圆心为$x_i$的闭球.$\delta=x_i'-x_i$为对抗性扰动，本文只考虑$p=\infty$.

最小-最大公式将对抗鲁棒网络优化分为两步:首先最大化对抗样本损失以生成攻击性较强的样本，然后最小化这些样本在网络参数θ下的损失.

对抗训练包括将对抗样本加入训练集，改变模型的决策边界，使其能正确识别这些样本.如图1(c)所示，当模型在某邻域内能正确分类所有点，即具备对抗稳健性，任何在该邻域内的扰动都无法改变分类结果.

为了生成对抗样本，对抗训练通常采用快速梯度符号法[4](fast gradient sign method，FGSM)或者投影梯度下降法[15](projected gradient descent，PGD).FGSM通过增大分类损失来定向增强扰动，目标是找到在特定范围内能使模型误分类的样本.FGSM的数学表达式如下:

$$x'=x+\alpha\,\mathrm{sign}(\nabla_x l(f_\theta(x),y)) \tag{5}$$

式中:$x\in\mathbb{R}^d$，为自然样本;x'是相应的对抗样本;对抗扰动为$\delta=\alpha\,\mathrm{sign}(\nabla_x l(f_\theta(x),y))$，它的维度大小和输入$x$相同;$y$是真实标签;$\alpha>0$，为超参数，表示梯度的步长;$\mathrm{sign}(\cdot)$为符号函数.

PGD是FGSM的迭代版本，与仅执行一次迭代的FGSM不同，PGD进行多次迭代，每次迭代都进

行小幅更新，并将扰动限制在规定范围内. 给定起始点 $x'_0 \in \mathbb{R}^d$，步长 $\alpha > 0$，PGD的工作原理如下：

$$x'_{t+1} = \prod_{\mathcal{B}_\varepsilon[x_0']} (x'_t + \alpha \operatorname{sign}(\nabla_{x'_t} l(f_\theta(x'_t), y))), \forall t \geqslant 0 \tag{6}$$

持续迭代，直到满足某一停止准则，例如固定的迭代次数 K，即 PGD-K 算法[15]. x'_0 为自然数据或受高斯小噪声或均匀随机噪声破坏的自然数据；x'_t 是第 t 步的对抗数据；$\prod_{\mathcal{B}_\varepsilon[x_0']}(\cdot)$ 是投影函数，必要时将对抗性数据投影回以 x'_0 为中心，以 $\varepsilon > 0$ 为半径的球. PGD是目前公认为最强的攻击方法，也是用于评估模型鲁棒性的基准测试算法之一.

1.3 多模态对抗训练

设 $x_{\mathrm{R},i}$ 为输入视频帧，$x_{\mathrm{S},i}$ 为输入骨骼数据，y_i 为多模态输入 $\{x_{\mathrm{R},i}, x_{\mathrm{S},i}\}$ 对应的真值标签. 由于存在多个输入，可以将攻击分为两类：只生成RGB视频对抗样本 $x'_{\mathrm{R},i}$ 或骨骼对抗样本 $x'_{\mathrm{S},i}$ 的单模态攻击，以及同时生成RGB视频和骨骼对抗样本 $\{x'_{\mathrm{R},i}, x'_{\mathrm{S},i}\}$ 的RGB-骨骼多模态攻击. 对抗训练最小-最大公式可以推广到多模态数据：

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} l(f_\theta(x'_{\mathrm{R},i}, x'_{\mathrm{S},i}), y_i) \tag{7}$$

其中：

$$\{x'_{\mathrm{R},i}, x'_{\mathrm{S},i}\} = \underset{\substack{x'_{\mathrm{R}} \in \mathcal{B}_{\varepsilon_{\mathrm{R}}}[x_{\mathrm{R},i}],\\ x'_{\mathrm{S}} \in \mathcal{B}_{\varepsilon_{\mathrm{S}}}[x_{\mathrm{S},i}]}}{\arg\max}\ l(f(x'_{\mathrm{R}}, x'_{\mathrm{S}}), y_i) \tag{8}$$

式中，$\mathcal{B}_{\varepsilon_{\mathrm{R}}}[x_{\mathrm{R},i}] = \{x'_{\mathrm{R}} \in \mathbb{R}^{d_{\mathrm{R}}} \mid \|x'_{\mathrm{R}} - x_{\mathrm{R},i}\|_p \leqslant \varepsilon_{\mathrm{R}}\}$，$\mathcal{B}_{\varepsilon_{\mathrm{S}}}[x_{\mathrm{S},i}] = \{x'_{\mathrm{S}} \in \mathbb{R}^{d_{\mathrm{S}}} \mid \|x'_{\mathrm{S}} - x_{\mathrm{S},i}\|_p \leqslant \varepsilon_{\mathrm{S}}\}$. 视频对抗扰动和骨骼对抗扰动分别是：$\delta_{\mathrm{R}} = x'_{\mathrm{R},i} - x_{\mathrm{R},i}$，$\delta_{\mathrm{S}} = x'_{\mathrm{S},i} - x_{\mathrm{S},i}$. 对于单模态攻击，$\delta_{\mathrm{R}}$ 或 δ_{S} 的值为0. 同理，可以将式(6)进行推广，以生成多模态对抗样本：

$$\begin{cases} x'_{\mathrm{R},t+1} = \prod_{\mathcal{B}_{\varepsilon_{\mathrm{R}}}[x_{\mathrm{R},0}']} (x'_{\mathrm{R},t} + \alpha \operatorname{sign}(\nabla_{x'_{\mathrm{R},t}} l(f_\theta(x'_{\mathrm{R},t}, x'_{\mathrm{S},t}), y))) \\ x'_{\mathrm{S},t+1} = \prod_{\mathcal{B}_{\varepsilon_{\mathrm{S}}}[x_{\mathrm{S},0}']} (x'_{\mathrm{S},t} + \alpha \operatorname{sign}(\nabla_{x'_{\mathrm{S},t}} l(f_\theta(x'_{\mathrm{R},t}, x'_{\mathrm{S},t}), y))) \end{cases} \tag{9}$$

2 方法描述

2.1 特征交互模块(FIM)

根据分析，骨骼模态通常比RGB模态鲁棒性更强. 因此，本文提出特征交互模块(FIM)使模型能显式地建立模态特征间的鲁棒性关系，强调鲁棒性较强的特征，同时抑制较弱的特征.

FIM的结构如图2所示，为了更好地可视化，本文将其时空维度表示在一个轴上. 具体来说，本文使用式(9)来从多模态自然样本 $\{x_{\mathrm{R}}, x_{\mathrm{S}}\}$(batch size大小为 B)中获得相应的对抗性样本 $\{x'_{\mathrm{R}}, x'_{\mathrm{S}}\}$.

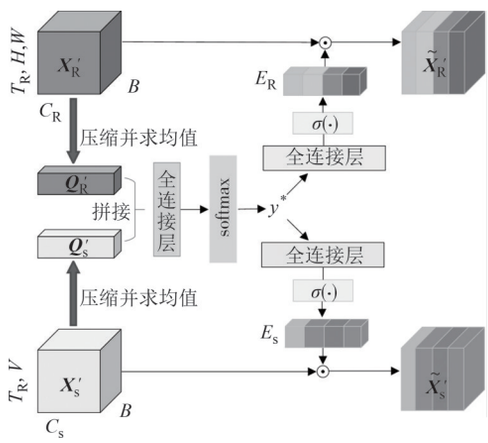


图2 FIM的结构

Fig.2 The architecture of FIM

本文从视频和骨骼模态的隐藏层提取特征，分别表示为 $\boldsymbol{X}'_{\mathrm{R}} \in \mathbb{R}^{B \times C_{\mathrm{R}} \times T_{\mathrm{R}} \times H \times W}$ 和 $\boldsymbol{X}'_{\mathrm{S}} \in \mathbb{R}^{B \times C_{\mathrm{S}} \times T_{\mathrm{S}} \times V}$，其中 C_{R} 和 C_{S} 为相应的输出通道数，T_{R} 和 T_{S} 是视频和骨骼序列的帧数. H, W 是视频的空间维度. V 为骨骼数据每帧的关键点数. 由于特征的维度较多，且本文重点关注的是通道维度的变化. 为了更好地可视化结构图，本文将RGB的时空维度 T_{R}, H, W 和骨骼的时空维度 T_{S}, V 都表示在了同一个轴上，将 B 和通道维度 $C_{\mathrm{R}}, C_{\mathrm{S}}$ 单独表示为一个轴. FIM接受两个模态的对抗样本的特征 $\boldsymbol{X}'_{\mathrm{R}}, \boldsymbol{X}'_{\mathrm{S}}$ 作为输入，从中学习全局的多模态嵌入，并使用该嵌入重新校准RGB视频和骨骼的输入特征. 通过下面两个主要步骤，完成模态的重新校准.

1)压缩(squeeze)并求均值：为弥补卷积层输出特征中因感受野大小限制而缺乏的全局上下文信息，本文通过对输入特征进行压缩并求均值，将全局时空信息压缩到通道描述符中：

$$\begin{cases} \boldsymbol{Q}'_{\mathrm{R}} = \dfrac{1}{T_{\mathrm{R}} \times H \times W} \sum_{t=1}^{T_{\mathrm{R}}} \sum_{h=1}^{H} \sum_{w=1}^{W} \boldsymbol{X}'_{\mathrm{R}}, \\ \boldsymbol{Q}'_{\mathrm{S}} = \dfrac{1}{T_{\mathrm{S}} \times V} \sum_{t=1}^{T_{\mathrm{S}}} \sum_{v=1}^{V} \boldsymbol{X}'_{\mathrm{S}} \end{cases} \tag{10}$$

其中:

$$\begin{cases} \boldsymbol{Q}'_{\mathrm{R}} \in \mathbb{R}^{B \times C_{\mathrm{R}}} \\ \boldsymbol{Q}'_{\mathrm{S}} \in \mathbb{R}^{B \times C_{\mathrm{S}}} \end{cases} \tag{11}$$

$\boldsymbol{Q}'_{\mathrm{R}}, \boldsymbol{Q}'_{\mathrm{S}}$的功能是产生通道相关特征响应的全局分布的嵌入,允许来自网络的全局感受野的信息被其所有层使用.同时,压缩操作可以实现具有任意时空维度特征的模态之间的融合.

2)多模态激活:这一步旨在生成两个模态各自的激活门控信号,对每个模态应用不同的校准权重.首先,将$\boldsymbol{Q}'_{\mathrm{R}}, \boldsymbol{Q}'_{\mathrm{S}}$沿着各自的通道维度拼接在一起得到$\boldsymbol{Q}' \in \mathbb{R}^{B \times (C_{\mathrm{R}} + C_{\mathrm{S}})}$,它结合了两个模态特征的全局信息,可以进行辅助分类.本文将$\boldsymbol{Q}'$输入一个全连接层中,再经过Softmax操作,公式如下所示:

$$\boldsymbol{y}^{*} = \mathrm{Softmax}(\boldsymbol{W}\boldsymbol{Q}' + \boldsymbol{b}) \tag{12}$$

式中:$\boldsymbol{W} \in \mathbb{R}^{N \times (C_{\mathrm{R}} + C_{\mathrm{S}})}$,是全连接层的权重;$\boldsymbol{b} \in \mathbb{R}^{N}$,是全连接层的偏置;联合表征$\boldsymbol{y}^{*} \in \mathbb{R}^{B \times N}$,$N$为类别数.本文将输出的联合表征$\boldsymbol{y}^{*}$的大小控制为类别数,因此也可以将它看作中间特征的辅助预测结果,在损失函数中使用这个结果.

然后通过两个独立的全连接层生成每个模态的激活控制信号:

$$\begin{cases} \boldsymbol{E}_{\mathrm{R}} = \boldsymbol{W}_{\mathrm{R}}\boldsymbol{Q}'_{\mathrm{R}} + \boldsymbol{b}_{\mathrm{R}} \\ \boldsymbol{E}_{\mathrm{S}} = \boldsymbol{W}_{\mathrm{S}}\boldsymbol{Q}'_{\mathrm{S}} + \boldsymbol{b}_{\mathrm{S}} \end{cases} \tag{13}$$

式中:$\boldsymbol{W}_{\mathrm{R}} \in \mathbb{R}^{C_{\mathrm{R}} \times N}, \boldsymbol{W}_{\mathrm{S}} \in \mathbb{R}^{C_{\mathrm{S}} \times N}$,是全连接层的权重;$\boldsymbol{b}_{\mathrm{R}} \in \mathbb{R}^{C_{\mathrm{R}}}, \boldsymbol{b}_{\mathrm{S}} \in \mathbb{R}^{C_{\mathrm{S}}}$,是对应的偏置.

最后,将$\boldsymbol{E}_{\mathrm{R}}, \boldsymbol{E}_{\mathrm{S}}$经过sigmoid激活函数,再对输入的多模态特征进行重新校准:

$$\begin{cases} \tilde{\boldsymbol{X}} = \sigma(\boldsymbol{E}_{\mathrm{R}}) \odot \boldsymbol{X}'_{\mathrm{R}} \\ \tilde{\boldsymbol{X}} = \sigma(\boldsymbol{E}_{\mathrm{S}}) \odot \boldsymbol{X}'_{\mathrm{S}} \end{cases} \tag{14}$$

式中:$\sigma(\cdot)$表示sigmoid函数,它将激活信号$\boldsymbol{E}_{\mathrm{R}}, \boldsymbol{E}_{\mathrm{S}}$的值缩放到(0,1)之间;$\odot$表示在通道维度上相乘,允许在每个分支中抑制或激励不同的特征通道.

通过联合表征,FIM使用多模态对抗样本的全局信息来重新校准特征.同时受到攻击时,如果RGB模态较弱而骨骼模态较强,FIM进行跨模态重新校准.

2.2　基于FIM的RGB-骨骼动作识别网络

本文的整体框架如图3所示.框架由I3D[40]网络和HCN[41]两个分支网络构成,I3D通过3D卷积提取RGB视频特征,HCN则用普通卷积处理骨架流.I3D和HCN的简洁设计符合本文的需求,旨在强化模型鲁棒性,提升动作识别率.通过简化网络结构,可以减少不相关干扰.

网络深层处的FIM单元连接两个网络,并且负责处理高级语义特征.FIM的数量可以根据需要增减,位置如图3所指示.FIM处理双模态输入,输出经过校准的特征和一个用于计算损失函数的预测值$\boldsymbol{y}^{*}$.校准后的特征继续前向传播进行进一步提取.

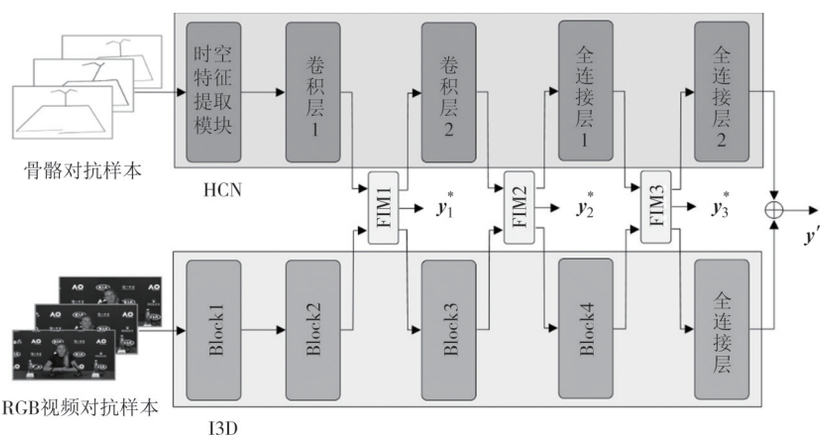


图3　框架的结构

Fig.3　The structure of the framework

2.3　损失函数和训练流程

FIM与骨干网络一起进行对抗训练.本文设计了一个新的损失函数.以插入网络s个FIM为例,损失函数可定义为:

$$L = l(\boldsymbol{y}', \boldsymbol{y}) + \beta \frac{1}{s} \sum_{i=1}^{s} l(\boldsymbol{y}_i^{*}, \boldsymbol{y}) \tag{15}$$

式中:$\boldsymbol{y}'$是对抗样本最终预测输出;$\boldsymbol{y}$为多模态输入$\{x_{\mathrm{R}}, x_{\mathrm{S}}\}$对应的真值标签;$l(\cdot)$是交叉熵损失函数;$\beta$

作为超参数,用于平衡两项损失.

首先,使用PGD攻击在自然样本上生成对抗样本[参见式(9)].将这些对抗样本输入网络,计算总损失,然后通过随机梯度下降更新网络参数.重复此过程直到模型收敛,最终得到一个鲁棒的多模态网络.在推理阶段,FIM继续参与前向传播,使用其在训练期间更新的参数进行特征校准,而不引入额外参数.

3 实验与分析

3.1 数据集

NTURGB+D[42]是一个流行的大型多模态动作识别数据集,含有40名参与者在80个视角下完成的60种活动,共56 880个样本.骨骼数据包含25个身体关节在3D空间中的坐标.

iMiGUE[43]是针对人工智能情感研究的新数据集,包含18 419个视频剪辑,分为32种动作类别.相应的骨骼数据由17个身体关节点坐标构成.

3.2 实验设置

训练设置:对于NTURGB+D数据集采用三个FIM,对于iMiGUE数据集,只用一个FIM.使用SGD优化器(动量为0.9,batch size大小为16),在iMiGUE上训练40个epoch,在NTURGB+D上训练80个epoch,权重衰减为5×10^{-4},初始学习率为0.01,采用余弦退火率策略调整学习率.

测试设置:本文报告了在自然样本上的干净准确率和在对抗样本上的准确率,使用两种流行的攻击方法:PGD20[15]和CW[14](通过PGD30优化),其中步长$\alpha=2/255$,扰动大小为$\varepsilon_R=8/255,\varepsilon_S=8/255$,超参数$\beta=1$.

所有实验均在两个NVIDIA GeForce RTX 3090 GPU上运行.环境为Python 3.8.8和Pytorch 1.10.0.

3.3 鲁棒性评估

为验证FIM的有效性,本文将其与baseline和最先进的(SOTA)方法进行了比较.所比较的SOTA方法也基于多模态防御,而非单模态.

符号定义:√表示使用干净的自然样本;×表示使用算法产生的对抗样本;AT with×R/×S/×RS分别表示使用RGB、骨骼或二者结合的对抗样本进行的对抗训练防御策略;None表示不使用任何防御方法.

评估标准:本文遵循文献[38]中的评估协议,使用识别准确率作为度量.√表示在RGB和骨骼数据均为自然样本时的准确率(干净准确率);×R、×S、×RS分别表示在仅RGB、仅骨骼或两者均遭受攻击时的准确率(反应模型鲁棒性).为综合评估不同防御策略的鲁棒性,本文还计算了平均鲁棒性作为总体性能指标:

$$\mathrm{Avg}=\frac{1}{3}(\times\mathrm{R}+\times\mathrm{S}+\times\mathrm{RS}) \tag{16}$$

为了综合评估不同方法的干净准确率和鲁棒性,Tian等[38]提出了Relative Improvement(RI)指标:

$$\mathrm{RI}=(\sqrt{\mathrm{RS}_m}+\mathrm{Avg}_m)-(\sqrt{\mathrm{RS}_n}+\mathrm{Avg}_n) \tag{17}$$

式中同时考虑了干净准确率和对抗样本的结果.m代表一种防御方法;n代表baseline,即本文实验中的None.如果一个防御方法降低干净数据的性能,RI将相应地惩罚它.

3.4 结果分析

表1和表2展示了在CW攻击下,各防御策略的性能指标,每一项最好的结果用黑体加粗.在代表模型整体鲁棒性的×RS和Avg上,本文的FIM几乎优于其他所有方法.例如,FIM在iMiGUE上的×RS和Avg达到38.00%和38.76%,比次高值分别提升7.03和5.12个百分点.×R和×S分别衡量模型在RGB和骨骼单模态上的鲁棒性.实验显示,使用对应模态的对抗样本训练可获得该模态下最佳鲁棒性,但在其他模态下表现不佳.

MinSim等方法虽可提升单一模态鲁棒性,但在另一模态上效果不佳,而FIM兼顾了所有模态的鲁棒性.例如,MinSim在NTURGB+D数据集上的×S达

表1 iMiGUE数据集上不同的防御方法的准确率

Tab.1 Accuracy rates with different defense strategies on the iMiGUE dataset %

防御策略	√RS	×R	×S	×RS	Avg	RI
None	**62.37**	0.13	6.67	0.02	2.27	0
AT with ×R	47.32	**47.32**	0.00	0.00	15.77	−1.55
AT with ×S	60.20	0.00	**59.79**	0.00	19.93	**15.49**
AT with ×RS	39.01	39.01	30.94	30.97	33.64	8.01
MinSim	60.60	0.00	53.74	0.02	17.92	13.88
ExFMem	60.53	0.00	45.19	0.00	15.06	10.95
MinSim+ExFMem	60.62	0.00	54.72	0.00	18.24	14.20
FIM(Ours)	40.30	40.30	37.98	**38.00**	**38.76**	14.42

表2　NTURGB+D数据集上不同的防御方法的准确率

Tab.2　Accuracy rates with different defense strategies on the NTURGB+D dataset　%

防御策略	√RS	×R	×S	×RS	Avg	RI
None	**88.09**	0.98	9.00	0.00	3.33	0
AT with ×R	75.59	**75.59**	2.80	2.89	27.09	11.26
AT with ×S	78.44	1.41	78.31	1.41	27.04	14.06
AT with ×RS	61.42	61.41	38.74	38.82	46.32	16.32
MinSim	80.85	0.50	**80.84**	0.48	27.27	16.70
ExFMem	84.19	0.00	60.98	0.00	20.33	13.10
MinSim+ExFMem	82.81	0.00	75.60	0.00	25.20	16.59
FIM(Ours)	67.57	67.57	39.71	**39.70**	**48.99**	**25.14**

到了80.84%,但在×R和×RS上仅为0.50%和0.48%.

多数防御策略会降低干净准确率√RS,因此无防御方法时干净准确率最高,这符合对抗防御理论.防御的目标是在保持干净准确率的同时提高鲁棒性.引入防御策略通常不可避免导致干净准确率下降,因此引入了RI值来综合评估干净准确率和鲁棒性.FIM的√RS相比于MinSim等方法有所下降,但鲁棒性有所提高.FIM的RI值都是最高或者接近最高,表明本文的FIM做到了干净准确率和鲁棒性的平衡,例如FIM在NTURGB+D上的RI达到25.14%,比MinSim+ExFMem方法的16.59%提高了8.55个百分点.

总之,大量实验结果证明了FIM在提升模型鲁棒性、平衡不同模态鲁棒性以及兼顾干净准确率和鲁棒性方面优于其他方法.

3.5　消融实验

3.5.1　FIM的数量

本文的FIM方法是基于AT的.为了评估FIM的有效性及其数量对网络性能的影响,本文进行了消融实验.AT with ×RS代表只使用对抗训练而不加入FIM,即0个FIM.从两个分支网络的最后一个全连接层前开始,即以FIM3为起始,向前插入1~3个FIM.在这里使用PGD20攻击进行测试,其他设置与3.2节一致.实验结果如表3和表4所示.虽然单纯的对抗训练能提升鲁棒性,但加入FIM后,无论数量多少,各项准确率都有了明显的提升,证明FIM的加入是鲁棒性提升的关键,验证了FIM的关键作用.

在iMiGUE数据集上,1个FIM表现最佳,几乎所有指标均为最高,如RI比次优值高出0.89个百分点.在NTURGB+D数据集上,3个FIM的综合性能更优,

表3　iMiGUE数据集上FIM数量对准确率的影响

Tab.3　The impact of the number of FIMs on accuracy rates on the iMiGUE dataset　%

FIM数量	√RS	×R	×S	×RS	Avg	RI
None	**62.37**	0.11	6.97	0.07	2.38	0
0 FIM	39.01	39.01	31.89	31.95	34.28	8.54
1 FIM	40.30	**40.30**	**37.85**	**37.83**	**38.66**	**14.21**
2 FIM	39.95	39.95	36.51	36.53	37.66	12.86
3 FIM	39.95	39.95	37.21	37.19	38.12	13.32

表4　NTURGB+D数据集上FIM数量对准确率的影响

Tab.4　The impact of the number of FIMs on accuracy rates on the NTURGB+D dataset　%

FIM数量	√RS	×R	×S	×RS	Avg	RI
None	**88.09**	2.09	11.84	0.15	4.69	0
0 FIM	61.42	61.41	40.41	40.32	47.38	16.02
1 FIM	66.86	66.86	**40.45**	**40.45**	**49.25**	23.33
2 FIM	64.06	64.06	38.92	38.95	47.31	18.59
3 FIM	67.57	**67.57**	40.14	40.00	49.24	**24.03**

RI值高出次优值0.70个百分点.

这一差异可能因为NTURGB+D数据集规模(57k样本)大于iMiGUE(18k样本)而导致.因此,对于大型数据集,更多FIM能带来更佳性能.总的来说,使用FIM的性能显著优于不使用FIM的情况.

3.5.2　超参数β

在这一节中,本文专注于评估提出的新损失函数——即式(15)中的超参数β对模型性能的影响.为此,本文在iMiGUE数据集上进行了实验,用5个不同的值$\beta=\{0.1, 0.5, 1, 2, 5\}$,训练了包含1个FIM的模型,并评估了它们在PGD20攻击下的准确率.在这里,β控制了辅助预测损失在损失函数中的比例.实验结果如表5所示,随着β的增加,模型的鲁棒准确率Avg和RI值先增大再减小,在$\beta=2$时达到最高值.

表5　β对准确率的影响(PGD20)

Tab.5　The impact of β on accuracy rates (PGD20)　%

β	√RS	×R	×S	×RS	Avg	RI
None	**62.37**	0.11	6.97	0.07	2.38	0
0.1	40.81	**40.81**	35.90	35.94	37.55	13.61
0.5	40.39	40.39	37.30	37.30	38.33	13.97
1	40.30	40.30	37.85	37.83	38.66	14.21
2	40.50	40.50	**38.29**	**38.26**	**39.02**	**14.77**
5	40.43	40.43	38.18	38.18	38.93	14.61

3.5.3 特征损失

在FIM中，首先将两个对抗样本的特征进行压缩并求均值，从而将时空特征压缩到通道维度.这种做法虽常见,但可能导致某些特征信息的丢失(仅限于这一步骤，最终的校准特征未受影响).文献[44]也采用了类似方法，并被后续众多研究采纳.为了探究这种特征损失是否影响性能，本文设计了一系列实验进行验证.

与FIM表示的方法相比,对比方法在FIM第一步中不进行特征压缩和求均值,而是通过特征拉伸保留时空信息(用FIM*表示).在iMiGUE数据集上，本文使用单个FIM进行测试,超参数$\beta = 1$，使用PGD20攻击.如表6所示,FIM*显示出轻微的性能提升,其Avg和RI值分别比FIM高出0.24和1.16个百分点.然而,表7显示FIM*参数量却是FIM的1.7倍,这是由于没有压缩降维操作,而只是对特征进行拉伸,故全连接层参数量剧增,同时也增加了过拟合的风险.为了避免过拟合,本文在模型中引入了dropout操作.

表6 两种方法的性能比较

Tab.6 Comparison of the performance of two methods

%

对比方法	√RS	×R	×S	×RS	Avg	RI
None	**62.37**	0.11	6.97	0.07	2.38	0
FIM	40.30	40.30	**37.85**	**37.83**	38.66	14.21
FIM*	41.22	**41.22**	37.72	37.76	**38.90**	**15.37**

表7 两种方法的参数量比较

Tab.7 Comparison of parameter quantities of two methods

对比方法	参数量/10^6
FIM	47.42
FIM*	80.91

总体来看,尽管特征拉伸保留了时空信息从而略微提升了性能,但同时导致了参数量的大幅增加和过拟合风险的提高.因此,原始的特征压缩并求均值方法在效率和性能平衡方面表现更佳.

4 结 论

尽管多模态技术发展迅速，其鲁棒性研究却相对较少.本文针对RGB-骨骼多模态动作识别模型在面临对抗攻击时性能缺乏鲁棒性的问题，提出了基于特征交互模块(FIM)增强RGB-骨骼动作识别鲁棒性的方法.FIM通过全局平均池化提取两种模态对抗样本的全局信息,并学习联合表征,进而利用这些表征生成多模态激活信号,以重新校准各模态特征.本文还创新性地设计了一个损失函数,用于训练更鲁棒的模型.在NTURGB+D和iMiGUE两个公开数据集上的广泛实验表明,本文的方法在提高模型鲁棒性、平衡各模态鲁棒性以及维持干净数据准确率方面，相比其他多模态防御方法有显著优势.

参考文献

[1] PAL N R, PAL S K. A review on image segmentation techniques[J]. Pattern Recognition, 1993, 26(9): 1277-1294.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA: IEEE, 2016: 779-788.

[3] GLASNER D, BAGON S, IRANI M. Super-resolution from a single image [C]//2009 IEEE 12th International Conference on Computer Vision. September 29-October 2, 2009, Kyoto, Japan: IEEE, 2009: 349-356.

[4] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, 2015: 32-40.

[5] WEI X X, ZHU J, YUAN S, et al. Sparse adversarial perturbations for videos[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8973-8980.

[6] WANG H J, WANG G R, LI Y, et al. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA: IEEE, 2020: 339-348.

[7] LI L Y, MA R T, GUO Q P, et al. BERT-ATTACK: adversarial attack against BERT using BERT [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 6193-6202.

[8] CHEN J Y, YUAN B D, TOMIZUKA M. Model-free deep reinforcement learning for urban autonomous driving [C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). October 27-30, 2019, Auckland, New Zealand: IEEE, 2019: 2765-2771.

[9] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018. Salt Lake City, UT, USA: IEEE,2018:1625-1634.

[10] BUCH V H, AHMED I, MARUTHAPPU M. Artificial intelligence in medicine: current trends and future possibilities[J]. British Journal of General Practice,2018,68(668):143-144.

[11] KONG B, WANG X, LI Z Y, et al. Cancer metastasis detection via spatially structured deep network [M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017:236-248.

[12] MA X J, NIU Y H, GU L, et al. Understanding adversarial attacks on deep learning based medical image analysis systems[J]. Pattern Recognition,2021,110:107332.

[13] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [EB/OL]. 2013: 1312.6199. https://arxiv.org/abs/1312.6199v4

[14] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C]//2017 IEEE Symposium on Security and Privacy (SP). May 22-26, 2017, San Jose, CA, USA. IEEE, 2017:39-57.

[15] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C]// 6th International Conference on Learning Representations, ICLR 2018. April 30-May 3, 2018. Vancouver, BC, Canada: OpenReview.net, 2018.

[16] MADAAN D, SHIN J, HWANG S J. Adversarial neural pruning with latent vulnerability suppression[C]//International Conference on Machine Learning, PMLR, 2021: 6575-6585.

[17] LIN J, GAN C, HAN S. Defensive quantization: when efficiency meets robustness[C]// 7th International Conference on Learning Representations, ICLR 2019. May 6-9, 2019. New Orleans, LA, USA: OpenReview.net, 2019.

[18] XIE C H, WU Y X, VAN DER MAATEN L, et al. Feature denoising for improving adversarial robustness[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA: IEEE, 2019:501-509.

[19] NASEER M, KHAN S, HAYAT M, et al. A self-supervised approach for adversarial robustness [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19,2020,Seattle, WA, USA: IEEE, 2020: 259-268.

[20] ZHANG J F, XU X L, HAN B, et al. Attacks which do not kill training make adversarial learning stronger [EB/OL]. 2020: 2002.11242. https://arxiv. org/abs/2002.11242v2.

[21] ZHANG J, ZHU J, NIU G, et al. Geometry-aware instance-reweighted adversarial training [EB/OL]. 2021: 2010.01736. https://arxiv.org/abs/2010.01736.

[22] LIU X, KUANG H, LIN X, et al. CAT: Collaborative adversarial training [EB/OL]. 2023:/2205.11156v1. https://arxiv. org/abs/2205.11156v1.

[23] BAI Y, ZENG Y, JIANG Y, et al. Improving adversarial robustness via channel-wise activation suppressing [C]// 9th International Conference on Learning Representations, ICLR 2021. May 3-7, 2021. Virtual Event., Austria: OpenReview.net, 2021.

[24] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C]//2016 IEEE Symposium on Security and Privacy (SP). May 22-26,2016,San Jose,CA,USA:IEEE,2016:582-597.

[25] DAS N, SHANBHOGUE M, CHEN S T, et al. Compression to the rescue: defending from adversarial attacks across modalities [C]//ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2018.

[26] WANG L, HE Z, TANG J, et al. A dual semantic-aware recurrent global-adaptive network for vision-and-language navigation[C]// Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. August 19-25, 2023. Macao: International Joint Conferences on Artificial Intelligence Organization, 2023: 1479-1487.

[27] FU Z, MAO Z, SONG Y, et al. Learning semantic relationship among instances for image-text matching[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 17-24,2023,Vancouver,BC,Canada: IEEE,2023:15159-15168.

[28] KHADER F, MUELLER-FRANZES G, WANG T, et al. Medical diagnosis with large scale multimodal transformers: leveraging diverse data for more accurate diagnosis [EB/OL]. 2022: 2212.09162.https: arxiv.org/abs/2212.09162.

[29] MOON J H, LEE H, SHIN W, et al. Multi-modal understanding and generation for medical images and text *via* vision-language pre-training [J]. IEEE Journal of Biomedical and Health Informatics,2022,26(12):6070-6080.

[30] JAIMES A, SEBE N. Multimodal human-computer interaction: a survey[J]. Computer Vision and Image Understanding, 2007, 108(1/2): 116-134.

[31] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Dec,2020. ACM,2020: 125-136.

[32] KINFU K A, VIDAL R. Analysis and extensions of adversarial training for video classification[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). June 19-20, 2022, New Orleans, LA, USA. IEEE,

2022:3415-3424.

[33] ZHANG J M, YI Q, SANG J T. Towards adversarial attack on vision-language pre-training models[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa Portugal: ACM, 2022: 5005-5013.

[34] ZHAO Y Q, PANG T Y, DU C, et al. On evaluating adversarial robustness of large vision-language models [EB/OL]. 2023: 2305.16934. https://arxiv. org/abs/2305.16934v2.

[35] DUAN H D, ZHAO Y, CHEN K, et al. Revisiting skeleton-based action recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-24, 2022, New Orleans, LA, USA: IEEE, 2022: 2959-2968.

[36] YU B X B, LIU Y, ZHANG X, et al. MMNet: a model-based multimodal network for human action recognition in RGB-D videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3522-3538.

[37] VAEZI JOZE H R, SHABAN A, IUZZOLINO M L, et al. MMTM: multimodal transfer module for CNN fusion[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA: IEEE, 2020: 13286-13296.

[38] TIAN Y P, XU C L. Can audio-visual integration strengthen robustness under multimodal attacks? [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 20-25, 2021, Nashville, TN, USA: IEEE, 2021: 5597-5607.

[39] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 9-18.

[40] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA: IEEE, 2017: 4724-4733.

[41] LI C, ZHONG Q Y, XIE D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. July 13-19, 2018. Stockholm, Sweden. California: International Joint Conferences on Artificial Intelligence Organization, 2018: 786-792.

[42] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB D: a large scale dataset for 3D human activity analysis[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA: IEEE, 2016: 1010-1019.

[43] LIU X, SHI H L, CHEN H Y, et al. iMiGUE: an identity-free video dataset for micro-gesture understanding and emotion analysis[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 20-25, 2021, Nashville, TN, USA: IEEE, 2021: 10626-10637.

[44] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.