Journal of Image and Signal Processing 图像与信号处理, 2026, 15(1), 89-101
Published Online January 2026 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2026.151008

Hans汉斯

基于双阶段协同数据增强的暴力行为视频识别算法

文晨曦，杨善铮

赣南师范大学数学与计算机科学学院，江西 赣州

收稿日期：2025年12月12日；录用日期：2025年1月5日；发布日期：2026年1月21日

摘　要

暴力行为视频识别是现代公共安全领域中一项至关重要的技术。合理设计的数据增强方法可以提升暴力行为视频识别精度。针对现有数据增强方法难以全面覆盖时域和空域暴力行为信息的问题，提出双阶段协同数据增强网络。在训练阶段提出时空随机裁剪策略，生成具有时空域暴力行为信息的多样化背景和动作表达，提高模型对时空域暴力行为特征学习的鲁棒性。在测试阶段通过十字区域裁剪策略，扩大裁剪视角，提高暴力行为特征区域覆盖度。在VSD2015数据集上的大量实验验证，双阶段协同数据增强网络仅视觉模态的结果超过先进方法，取得领先性能。本研究通过双阶段协同增强机制，为暴力行为视频识别任务中的数据增强方法提供新的方案。

关键词

暴力行为视频识别，数据增强，双阶段数据增强

Violence Video Recognition Based on Two-Stage Collaborative Data Augmentation

Chenxi Wen, Shanzheng Yang

College of Mathematics and Computer Science, Gannan Normal University, Ganzhou Jiangxi

Received: December 12, 2025; accepted: January 5, 2026; published: January 21, 2026

Abstract

Violence video recognition is a crucial technology in the field of modern public security. Well-designed data augmentation methods can improve the precision of violence recognition. To address the problem that existing data augmentation methods are difficult to fully cover violence information in the temporal and spatial domains, a Two-stage Collaborative Data Augmentation Network (TCDANet) is

文章引用: 文晨曦, 杨善铮. 基于双阶段协同数据增强的暴力行为视频识别算法[J]. 图像与信号处理, 2026, 15(1): 89-101.
DOI: 10.12677/jisp.2026.151008

proposed. In the training phase, a Spatiotemporal Random Crop (STRCrop) strategy is proposed to generate diverse backgrounds and action representations, which containing violence information in the spatiotemporal domains, enhancing the model's robustness in learning spatiotemporal violence features. In the testing phase, a Cross Area Crop (CACrop) strategy is adopted to expand the cropping perspective, improving the coverage of violence feature regions. Extensive experiments are conducted on the VSD2015 dataset. The results of the two-stage collaborative data augmentation network with only visual modality outperform advanced methods, acquiring leading performance. This study provides a new solution for data augmentation methods in violence video recognition tasks through the two-stage collaborative augmentation.

Keywords

Violence Video Recognition, Data Augmentation, Two-Stage Collaborative Data Augmentation

1. 引言

在全球化与信息化深度融合下，暴力行为视频的传播呈现数字化、网络化特征。暴力行为视频作为暴力思想与行为的传播媒介，对社会公共安全构成严重威胁。暴力行为视频识别是公共安全领域暴力行为视频防控的重要技术[1]-[3]，模型识别精度直接影响安全防控的程度。精确地识别暴力行为视频成为防范暴力恐怖袭击和维护社会稳定的重要环节。多数方法[4]-[14]通过设计模型结构提升模型的识别精度。数据增强方法[15] [16]可以设计生成与模型任务适配的多样化数据，是提升模型识别精度的又一途径。暴力行为类视频不仅包含空域上的特征比如武器，还包含时域上连续的暴力动作表达特征比如打架。方法[17]通过随机组合基础变换操作增强空域维度数据多样性。没有考虑时域维度数据增强对模型识别精度的影响，导致在动态暴力行为场景识别中受限。方法[18]对单帧图像进行多尺度空域裁剪，仅能增加多样化的空域暴力行为信息数据，缺乏时域暴力行为信息的数据多样性。暴力行为具有连续性，某一暴力动作的发生可能涉及多个视频帧的动态变化。仅在空域上作数据增强操作，可能使模型无法更好地理解完整的暴力行为事件发展过程。暴力行为视频的场景复杂多变，暴力行为特征可能出现在视频画面的任意位置。方法[19]采用中心视角裁剪方式，裁剪区域可能对暴力行为特征区域覆盖度低，限制模型的性能。

对于上述问题，提出双阶段协同数据增强网络。通过阶段化策略分别提升包含暴力行为信息的时空域数据多样性与暴力行为特征区域覆盖度。在训练阶段，对暴力行为视频进行时空域两个维度的裁剪，增加具有时空域暴力行为信息的多样化场景增强数据，提高模型泛化能力。在测试阶段，设计十字区域裁剪策略。通过十字区域裁剪提高暴力行为特征区域覆盖度，避免中心视角带来的局限性。通过针对性设计，协同提升暴力行为视频识别模型的性能，为公共安全领域提供更可靠的技术支持。

2. 相关工作

2.1. 暴力行为视频识别算法

多数研究[20]-[31]聚焦于模型结构创新提升识别精度。Sjöberg 等人[32]提出 MediaEval 2015 电影情感影响分析任务。明确任务目标、数据标注规范与评价指标，为后续研究提供统一的基准框架。Dai 等人[33]在 MediaEval 2015 任务中通过深度神经网络自动提取影片多模态高阶特征。有效捕捉暴力场景的视

觉冲击与情感影响的关联性特征。Trigeorgis 等人[34]通过整合视觉、音频及文本信息多模态特征的互补优势，提升情感影响评估的准确性与鲁棒性。Lam 等人[35]聚焦于特征选择与模型优化。通过精简有效特征、调整模型参数，在保证性能的同时提升算法效率。Marin Vlastelica P.等人[36]通过系统验证低层视觉描述符、运动特征等多种视觉信息的有效性，明确不同视觉特征对电影情感影响预测的贡献差异。Li 等人[37]提出将子类用于视频中的暴力检测，旨在通过将暴力类别划分为子类提高分类准确性。Peixoto 等人[38]-[39]提出基于深度学习的视频暴力建模与分类策略，专注于构建有效的神经网络架构和训练算法。Freire-Obregón 等人[40]、Zheng 等人[41]提出不同的暴力行为检测方法，如使用膨胀 3D 卷积网络上下文分析和时空特征多任务学习提高暴力行为识别精度。Gu 等人[42]通过语义对应性提高暴力行为视频识别精度。Wu 等人[43]提出基于语义嵌入学习的特类视频识别方法。通过构建结构化的语义嵌入空间与高效的度量学习策略，提升模型对视频高层语义的理解与泛化能力。Pu 等人[44]通过局部到全局嵌入的语义多模态方法提高模型性能。Wang 等人[45]提出用于情感视频内容分析的私有－共享子空间学习方法(P2SL)。Savadogo 等人[46]专注于利用卷积神经网络构建老年虐待检测系统。Negre 等人[47]对基于深度学习的视频暴力检测进行文献综述。Vaishy 等人[48]通过将复杂模型的知识迁移至轻量模型，在保证识别精度的前提下提升暴力行为的早期检测效率。Hanief 等人[49]探索基于深度学习技术的视频监控系统进行可疑活动检测。上述研究在视频分析和暴力行为视频识别领域做出重要贡献。多数研究通过模型结构设计提升模型识别精度，同时也会增加一定的时间成本。

2.2. 数据增强方法

通过数据增强方法提升数据多样性，是提升模型泛化能力的重要手段。RandAugment [17]通过基于随机策略的自动数据增强方法，解决传统数据增强中手工调参复杂、策略固定的问题。原始设计针对静态图像，未考虑时域上的数据增强问题。直接应用于暴力行为视频任务不能充分学习复杂多变的暴力行为场景。MultiScaleCrop [18]是一种多尺度裁剪增强方法，广泛应用于图像分类，对时域维度的数据增强不足。这些方法没有针对时域维度进行增强设计。合理的时空域维度数据增强策略可以生成包含复杂暴力行为动作表达的时空域增强数据。试测阶段的数据增强旨在通过合理的区域裁剪方式，提升模型的识别精度。测试阶段常用的方法为 CenterCrop [19]，对视频帧进行中心区域裁剪后再输入模型进行识别。暴力行为视频识别任务中采用 CenterCrop [19]存在一定的缺陷。暴力行为特征可能出现在视频画面的任意位置，尤其是边缘区域。中心视角的裁剪方式无法应对复杂多变的暴力行为场景。当暴力行为特征不在中心区域时，模型识别精度会大幅下降，严重影响实际应用效果。针对训练时 RandAugment [17]、MultiScaleCrop [18]和测试时 CenterCrop [19]裁剪方式存在上述不足，本文提出 TCDANet 弥补上述缺陷。训练时，融合包含暴力行为信息的时域和空域两个维度，提高有暴力行为信息的时空域数据多样性，提升模型对暴力行为视频识别的泛化能力。测试时，采用十字区域裁剪策略，裁剪中心区域和边缘区域提高暴力行为特征区域覆盖度，解决中心视角裁剪对暴力行为特征区域覆盖度低的问题。通过双阶段协同设计，达到全面覆盖时空域暴力行为信息和提高暴力行为识别精度目的。

3. 网络概述

为解决现有数据增强方法难以全面覆盖时域和空域的暴力行为信息问题，针对性提出 TCDANet，整体框架如图 1 所示。本文提出的网络主要包含两个创新性的数据增强方法：STRCrop 和 CACrop。训练时，通过 STRCrop 融合时空域两个维度产生丰富多样包含时空域暴力行为信息的增强数据。增强数据输入到 Backbone [50]进行多样化暴力行为特征学习，增强模型对复杂暴力行为场景的学习能力。测试时，通过 CACrop 裁剪中心及边缘区域，裁剪得到高覆盖暴力行为特征区域的增强数据。随后输入到经过多样化暴力行为特征学习的 Backbone [50]进行推理，协同提升模型对暴力行为视频识别的精度。

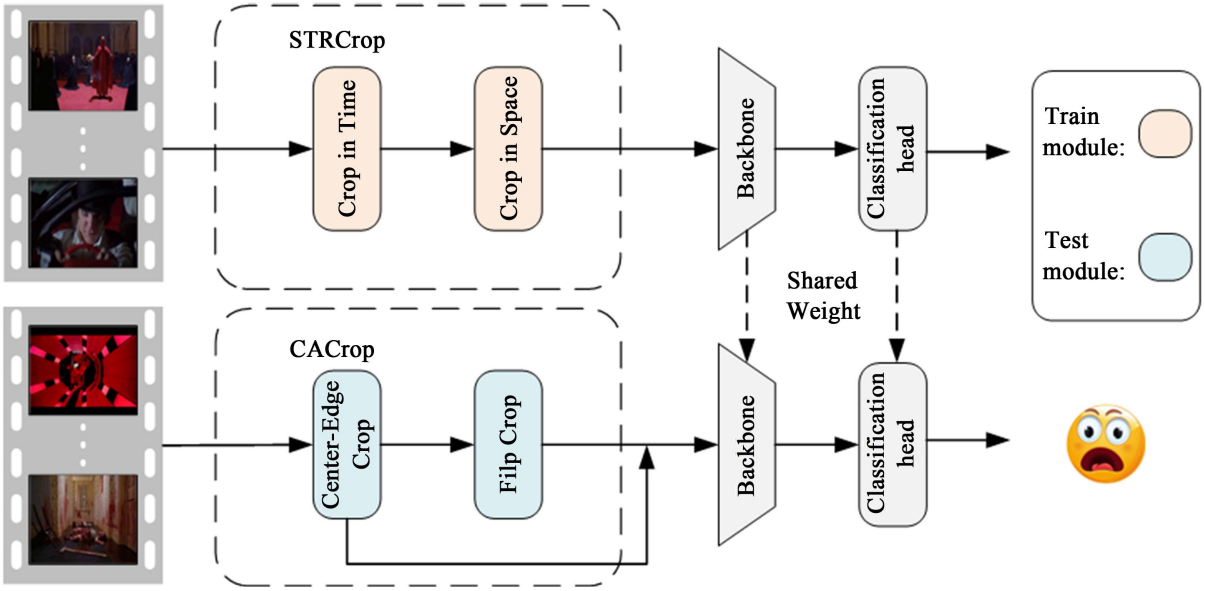

Figure 1. An overview of TCDANet framework
图 1. TCDANet 架构示意图

3.1. 时空随机裁剪

大部分数据增强考虑空域维度的增强，易割裂时域暴力行为信息的内在 AR 关系。通过 STRCrop 裁剪出包含时空域暴力行为信息的增强数据，提升模型的泛化能力。Crop in Time 确保帧间动作连续，Crop in Space 确保随机区域有效性，两者结合避免仅空域维度增强导致的时域暴力行为信息内在联系断裂。接下来 Crop in Time 和 Crop in Space 操作进行介绍。

Crop in Time 将原始视频帧序列(长度 N)调整为固定长度 L。当原始帧数量大于目标长度($N \geq L$)，进行均匀采样。为避免随机采样导致的动作跳变，采用线性等间隔采样。采样帧均匀覆盖原始时间轴。设采样后第 i 帧($i \in [0, L-1]$)对应原始帧的索引为 $\text{index}(i)$，计算公式为：

$$\text{index}(i) = \left\lfloor \frac{i \cdot (N-1)}{L-1} \right\rfloor \tag{1}$$

其中 $\lfloor \cdot \rfloor$ 表示向下取整，确保首尾帧分别对应原始序列的第 0 帧和第 $N-1$ 帧，中间帧均匀分布。当原始帧数量小于目标长度($N < L$)，进行帧补全。计算需补充的帧数：$E = L - N$。初始索引序列为 $[0, 1, \cdots, N-1]$，循环在前 $N-1$ 帧后插入重复帧，每插入 1 帧 E 的值减少 1。若仍需补充($E > 0$)，则从最后一帧(索引 $N-1$)开始循环补帧，直至总长度为 L。

Crop in Space 的核心是裁剪满足随机面积比 r 和随机宽高比约束的随机区域，避免裁剪出无意义的极端区域(如过小或宽高比异常的区域)。最多进行 10 次候选操作裁剪出满足条件的随机区域。设原始帧图像尺寸为 (H, W)，面积 $A = H \times W$，通过以下步骤生成有效裁剪区域。通过随机宽高比 AR 生成候选宽高比，在对数空间随机采样 10 个候选宽高比(确保采样比例分布均匀)，公式为：

$$AR_i = f_{opt}\left(U\left(\ln AR_{\min}, \ln AR_{\max}\right)\right), i \in [0, 9] \tag{2}$$

其中 $AR_{\min}$、$AR_{\max}$ 为 AR 的上下限，$U(a,b)$ 表示在 $[a,b]$ 上的均匀分布，$f_{opt}(\,)$ 表示进行候选操作。对数空域采样可避免线性采样导致的小比例值密集问题(如宽高比 0.75~1.33，对数采样后各比例出现概率更均衡)。通过随机面积比 r，随机生成候选面积($r_{\min}$、$r_{\max}$ 为 r 的上下限)：

$$A_i = f_{opt}\left(U\left(r_{\min}, r_{\max}\right)\times A\right), i \in [0,9] \tag{3}$$

根据候选面积与候选宽高比计算候选宽高，确保候选宽高为整数(像素坐标需整数)：

$$W_{crop,i} = \left\lfloor \sqrt{A_i \cdot AR_i} \right\rfloor, H_{crop,i} = \left\lfloor \sqrt{\frac{A_i}{AR_i}} \right\rfloor \tag{4}$$

筛选有效区域，选择第一个满足 $W_{crop,i} \le W$ 且 $H_{crop,i} \le H$ 的区域 i_c。其左上角通过随机采样确定，右下角随左上角和 $W_{crop,i}$、$H_{crop,i}$ 确定，$f_{Crop}(\,)$ 是裁剪操作：

$$x_1 = U\left(0, W - W_{crop,i}\right), y_1 = U\left(0, H - H_{crop,i}\right) \tag{5}$$

$$x_2 = x_1 + W_{crop,i}, y_2 = y_1 + H_{crop,i} \tag{6}$$

$$i_c = f_{Crop}\left(i, x_1, x_2, y_1, y_2\right) \tag{7}$$

候选区域均无效时(如原始图像过小或者候选宽高超出帧图像尺寸)，则采用中心正方形裁剪，确保裁剪有效：取原始图像的最小边长：$S = \min\left(H, W\right)$ 得到最大正方形区域；计算中心坐标：$x_1 = \left(W - S\right)//2$，$y_1 = \left(H - S\right)//2$；边界框为 $\left(x_1, y_1, x_1 + S, y_1 + S\right)$，确保裁剪区域为图像中心的完整正方形。STRCrop 伪代码如下：

Algorithm 1 STRCrop pseudocode

Require: L, $\mathbf{I} \in \mathbf{R}^{N\times H\times W\times C}$, $AR \in \left(AR_{min}, AR_{max}\right)$, $r \in \left(r_{min}, r_{max}\right)$

Ensure: $\mathbf{O} \in \mathbf{R}^{L\times H_c\times W_c\times C}$

function STRCROP(L,**I**,AR,r)
    O = []
    if $N < L$ **then** ▷Crop in Time
        $E = L - N$
        while $E > 0$ **do**
            for $i = 0$ **to** N-2 **do**
                if $E > 0$ **then**
                    I.insert($i + 1$,**I**[i])
                    $E = E - 1$
                    $N = N + 1$
                end if
            end for
        end while
    else if $N \ge L$ **then**
        for $i = 0$ **to** $L - 1$**do**
            $\text{index}(i) = \left\lfloor \frac{i\cdot(N-1)}{L-1} \right\rfloor$
        end for
        I = [**I**[index(i)] | $i \in 0..L - 1$]
    end if
    for each i **in I do** ▷Crop in Space ▷ Crop in Space
        $AR_i = f_{opt}\left(U\left(\ln AR_{min}, \ln AR_{max}\right)\right)$
        $A_i = f_{opt}\left(U\left(r_{min}, r_{max}\right)\times A\right)$
        $W_{crop,i} = \left\lfloor \sqrt{A_i \cdot AR_i} \right\rfloor$
        $H_{crop,i} = \left\lfloor \sqrt{\frac{A_i}{AR_i}} \right\rfloor$

续表

if $W_{crop,i} < W$ and $H_{crop,i} < H$ **then**
$x_1 = U\left(0, W - W_{crop,i}\right)$
$y_1 = U\left(0, H - H_{crop,i}\right)$
else
$S = \min(W, H)$
$x_1 = \left|\frac{W - S}{2}\right|$
$y_1 = \left|\frac{H - S}{2}\right|$
end if
$x_2 = x_1 + W_{crop,i}$
$y_2 = y_1 + H_{crop,i}$
$i_c = f_{Crop}\left(i, x_1, x_2, y_1, y_2\right)$
O.append(i_c)
end for
return O
end function

3.2. 十字区域裁剪

针对暴力行为特征可能在帧图像的任意区域的特点，仅中心视角裁剪对暴力行为特征区域覆盖度低。通过 CACrop 对中心及边缘区域裁剪提高暴力行为特征区域覆盖度。结合 Center-Edge Crop、Flip Crop 操作，扩大裁剪视角，提升模型识别精度。

Center-Edge Crop 将输入宽高为(H, W)视频帧，裁剪出宽高为(H_c, W_c)的中心区域和根据中心区域裁剪的边缘区域。根据中心区域的宽高计算中心偏移(后续可以通过中心偏移计算边缘区域宽高)，水平中心偏移为 $center\_x = (W - W_c)//2$ (确保裁剪中心区域水平居中时的 x 坐标)，垂直中心偏移为 $center\_y = (H - H_c)//2$ (确保裁剪中心区域垂直居中时的 y 坐标)。基于中心区域的宽高和中心偏移，定义十字区域的坐标：视频帧左上角坐标$(0,0)$、顶部区域左上角坐标$(center_x, 0)$、底部区域左上角坐标$(center_x, H - center_y)$、左侧区域左上角坐标$(0, center_y)$、右侧区域左上角坐标$(center_x, center_y)$、中心区域左上角坐标$(center_x, center_y)$。

设输入视频帧集为$\mathrm{I} = \{i_1, i_2, \cdots, i_n\}$，每个视频帧$\mathrm{I}_i$是一个三维张量，定义为：$\mathrm{I}_i \in R^{H \times W \times C}$，其中 H 为帧图像高度，W 为帧图像宽度，C 为通道数(如 RGB 图像的 3 个通道)。裁剪操作根据中心区域的宽高、中心偏移量和指定区域左上角坐标裁剪区域。在视频帧中从水平方向 x_{offset} 到 $x_{offset} + w$，垂直方向 y_{offset} 到 $y_{offset} + h$ 裁剪出指定区域。(x_{offset}, y_{offset})表示指定区域左上角坐标，(w, h)表示指定区域宽高，裁剪后的视频帧集I_c可表示为：

$$\mathrm{I}_c = \left\{ \mathrm{I}_i \left[x_{offset} : x_{offset} + w, y_{offset} : y_{offset} + h, : \right] \mid i = 1, 2, \cdots, n \right\} \tag{8}$$

其中左侧和右侧区域(w, h)为$(center_x, H_c)$，顶部和底部区域(w, h)为$(W_c, center_y)$。最后的:表示保留所有通道(不改变通道维度)。

每个裁剪后的图像 $\mathrm{crop}_i \in \mathrm{I}_c$ 的维度为 $h \times w \times C$。Flip Crop 对裁剪后的集合进行水平翻转后得到的新集合 $\mathrm{flip\_I}_c$ 定义为：

$$\text{flip\_I}_c = \left\{ \widehat{\text{crop}_i} \mid \widehat{\text{crop}_i}[j,k,l] = \text{crop}_i[h-1-j,k,l], i = 1,2,\cdots,m \right\} \tag{9}$$

其中 $j \in [0, h-1]$ 表示水平方向坐标(行索引)，$k \in [0, w-1]$ 表示垂直方向坐标(列引)，$l \in [0, C-1]$ 表示通道索引。

4. 实验及结果分析

4.1. 实验数据集

本文实验采用由 MediaEval 2015 比赛公开发布的大型暴力行为视频识别数据集 VSD2015 [32]。一共有 10,900 个片段，包含 502 个暴力行为视频样本和 10,398 个非暴力行为视频样本。暴力行为视频与非暴力行为视频数量比例约为 1:20，呈现典型的类不平衡状态。每个片段都是截取自 YouTube 视频或者 Hollywood 电影，时长为 8~12 秒。数据集每个片段都被标记相应的暴力行为或者非暴力行为标签。暴力行为视频和非暴力行为视频被打乱分到训练集和测试集，训练集包含 6144 个片段，测试集为 4756 个片段。如图 2 所示暴力行为视频样本和非暴力行为视频样本。

Figure 2. Examples of the VSD2015 dataset
图 2. VSD2015 数据集示意图

在前面提到暴力行为视频和非暴力行为视频的比例约为 1:20，属于类不平衡。为更加贴合暴力行为视频识别的任务，使用对少数类(暴力行为类)更为敏感的评估指标。这里使用官方指定的 average precision (AP)值[51]作为评价指标。

4.2. 实验设置

本实验基于 Pytorch 架构，通过 Backbone [50]和两块 Tesla V100 GPU 在 VSD2015 暴力行为视频二分类任务中(暴力行为/非暴力行为)展开训练与评估。训练过程设置 30 个 epochs、批次大小 2、帧大小 16、1clip，使用 AdamW [52]优化器。初始学习率为 2e−5、$\beta = (0.9, 0.999)$、权重衰减 0.01)，搭配 LinearLR (前 5 轮线性升温)与 CosineAnnealingLR (后续 50 轮余弦退火)的学习率调度策略。测试过程设置批次大小 2、帧大小 16、2clips。基准数据增强方法训练采用 RandAugment [17]、MultiScaleCrop [18]和测试采用 CenterCrop [19]。

4.3. 与其他方法的比较

实验为证明 TCDANet 在暴力行为视频识别任务中的性能，与其它一些比较先进的模型进行比较。为

能够具有对比性，同时选择在暴力行为数据集 VSD2015 上进行实验结果比较。

TCDANet 通过多阶段协同优化策略提高包含暴力行为信息的时空域维度数据多样性和暴力行为特征区域覆盖度，弥补先进模型对暴力行为场景识别的局限。如表 1 所示，TCDANet 仅视觉模态的结果与这些先进方法结果比较是最佳的，验证 TCDANet 的有效性，为暴力行为视频识别任务提供新方向。

Table 1. Comparison with other methods on the VSD2015 dataset
表 1. 在 VSD2015 数据集上与其他方法的对比

方法	Modality	VSD2015 (%)
Fudan-Huawei [33]	V + A	29.59
Zheng *et al.* [41]	V + A	32.42
Gu *et al.* [42]	V + A	41.31
Wu *et al.* [43]	V + A	44.55
Pu *et al.* [44]	V + A	47.39
TCDANet	V	48.89

4.4. 消融实验

为验证前面所提到的针对性设计策略在覆盖时空域暴力行为信息的协同效果，分别探究 STRCrop 方法、CACrop 方法以及两者结合对模型性能的影响，设计三组方法的消融实验。为保证实验的公平性，使用相同的实验设置。

STRCrop 和 CACrop 是针对提升覆盖时空域暴力行为信息不同方面提出的方法。通过在空域维度融入时域维度信息，STRCrop 增强模型对时空域的暴力行为场景和动作表达的学习能力。CACrop 裁剪中心及边缘区域，通过结构化裁剪提升暴力行为特征区域覆盖度。如表 2 所示，两者结合显著优于单个数据增强方法效果。前者通过时空域关联打破静态局限，让模型学习更全面的暴力行为模式。后者通过扩大中心视角的裁剪，避免暴力行为特征遗漏，提高模型识别精度。验证 STRCrop 与 CACrop 结合对模型性能的提升形成协同效果。

Table 2. Ablation evaluation of proposed methods on VSD2015 dataset
表 2. 在 VSD2015 数据集上提出方法的消融评估

方法	VSD2015 (%)
STRCrop	47.31
CACrop	47.40
STRCrop + CACrop	48.89

4.5. 与基准方法的比较

通过 STRCrop 对模型精度提升的综合表现验证设计的合理性。选择训练基准方法 RandAugment [17] 和 MultiScaleCrop [18]做比较。通过训练时间和推理结果综合评估训练阶段提出的 STRCrop 方法。训练时间根据不同方法在相同训练周期的平均训练时间计算得到。为使实验公平，实验设置保持一致。

如表 3 所示，在相同实验条件下 STRCrop 的结果优于 RandAugment [17]和 MultiScaleCrop [18]，证明时空随机裁剪策略生成丰富的包含暴力行为信息的时空域增强数据，为模型提供更佳的暴力行为场景学习环境，提升模型泛化能力。STRCrop 增加对时域维度的增强操作，需要更多的计算处理，导致训练时间的增加。推理结果和训练时间的综合表现证明 STRCrop 设计的合理性。

Table 3. Comparison with training benchmark methods on VSD2015 dataset
表 3. 在 VSD2015 数据集上与训练基准方法的比较

方法	训练时间(分钟)	VSD2015(%)
RandAugment [17]	15.79	46.18
MultiScaleCrop [18]	15.28	46.89
STRCrop	25.28	47.31

通过测试效果和时长综合验证 CACrop 和 CACrop + STRCrop 的合理性，选择测试基准方法 CenterCrop [19]作比较。测试时间是测试单样本的平均时间，即测试一次总时间对应的总测试样本的平均计算时间。为确保实验公正，采用统一的实验设置进行测试。

如表 4 所示，CACrop 打破 CenterCrop [19]“仅聚焦中心”的局限性，比测试基准效果好证明其对暴力行为特征覆盖能力优于传统“中心固定裁剪”方法。CACrop + STRCrop 方法形成互补，比 CACrop 的效果更好，形成协同提高模型识别精度。CACrop 扩大裁剪视角，增加模型除中心区域的计算，故时长比 CenterCrop 长。STRCrop + CACrop 通过 STRCrop 增加时空域暴力行为信息的数据多样性，提升模型对暴力行为场景的学习能力，不会对测试时间有太大影响。故 STRCrop + CACrop 和 CACrop 的推理时间相差甚微。测试效果和时长形成良好的平衡，证明 CACrop 和 STRCrop + CACrop 的合理设计。

Table 4. Comparison with testing benchmark methods on VSD2015 dataset
表 4. 在 VSD2015 数据集上与测试基准方法的比较

方法	测试时间(秒)	VSD2015 (%)
CenterCrop [19]	0.11	46.18
CACrop	0.72	47.40
STRCrop + CACrop	0.72	48.89

4.6. 结果分析

为直观呈现 TCDANet 在暴力行为视频识别任务中的分类性能，本节通过对比 TCDANet 与基准模型混淆矩阵对模型预测结果进行可视化解析。混淆矩阵的行代表真实类别(Non-Violence 为非暴力行为，Violence 为暴力行为)，列代表预测类别，矩阵元素为对应类别的预测概率，右侧颜色条用于映射概率数值与颜色的对应关系(颜色越深，概率越高)。

如图 3 反映出在 VSD2015 数据集上的暴力行为视频识别任务本身存在一定的挑战性，可能由于类不平衡和暴力行为场景的多样性、复杂性(如不同的暴力行为手段、场景背景等)，导致模型对暴力行为视频的特征提取和分类存在一定难度。TCDANet 通过针对性数据增强，提高训练时包含暴力行为信息的时空域数据多样性和测试时的暴力行为特征区域覆盖度。在训练阶段 STRCrop 让模型学习到暴力行为行为在不同时空片段的呈现方式。在测试阶段 CACrop 确保暴力行为行为的关键特征(如肢体冲突的边缘区域)不

被遗漏，对暴力行为视频的判别更精确。TCDANet 在正确识别暴力行为类视频的概率和漏检率都比基准高。验证 TCDANet 针对性设计策略对模型分类性能的提升。

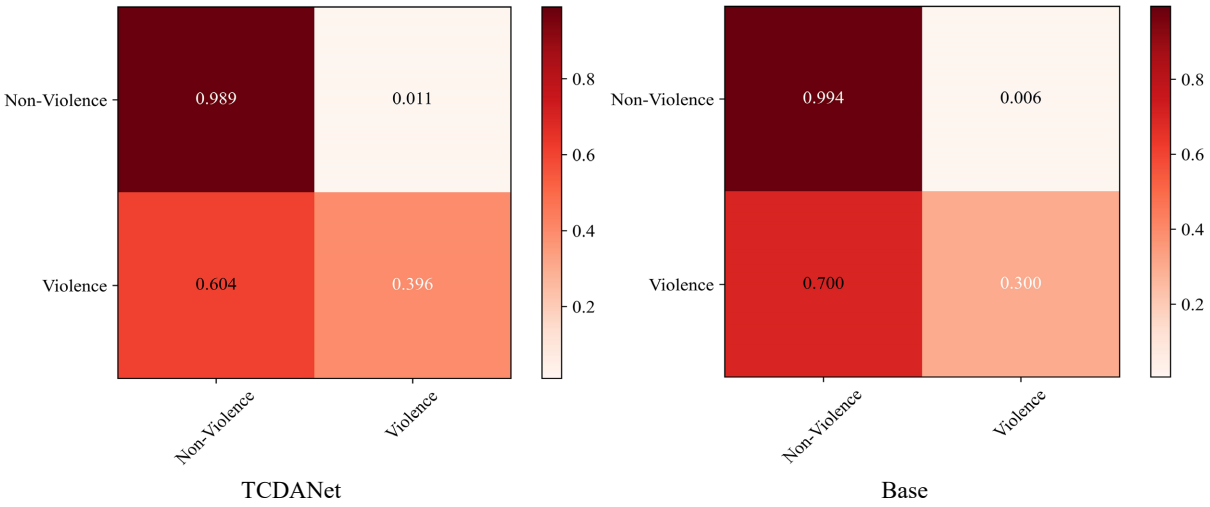

Figure 3. Confusion matrix comparison of different methods on VSD2015 dataset
图 3. 不同方法在 VSD2015 数据集上的混淆矩阵对比

通过可视化预测结果比较 TCDANet 和基准方法的效果。用“×”表示预测标签与真实标签不一致，“√”表示预测标签与真实标签一致。基准方法和 TCDANet 后面的标签表示预测标签，真实标签放在最后。在相同的实验参数下，TCDANet 和基准方法同时进行三个相同视频预测结果示例。

图 4 所示，TCDANet 展现出比基准方法更好的效果，预测结果都正确。TCDANet 在训练阶段学习更加多样化的暴力行为背景和动作表达，在测试阶段扩大视野避免遗漏关键特征，协同提升暴力行为类视频识别的效果。

Figure 4. Examples of TCDANet and benchmark method prediction results
图 4. TCDANet 和基准方法预测结果示意图

5. 结束语

对于暴力行为视频识别任务，提出的 TCDANet 通过针对性的双阶段数据增强方法设计，提升暴力行为视频识别精度。训练阶段通过 STRCrop 在时域和空域维度裁剪包含暴力行为信息的视频帧序列，融合生成复杂多样化的暴力行为场景，提高模型的泛化能力。测试阶段通过 CACrop 裁剪中心及边缘区域，扩大裁剪视角，提升对暴力行为特征区域覆盖度。VSD2015 数据集的实验结果表明，TCDANet 仅在视觉模态下超过先进方法的结果，取得领先性能。TCDANet 的协同优化策略，为提升暴力行为视频识别精度提供新的思路。未来对公共安全领域的智能视频分析也有实践价值。

基金项目

江西省自然科学基金(No. 20232BAB202017)。

参考文献

[1] Garcia-Cobo, G. and SanMiguel, J.C. (2023) Human Skeletons and Change Detection for Efficient Violence Detection in Surveillance Videos. *Computer Vision and Image Understanding*, **233**, Article ID: 103739. https://doi.org/10.1016/j.cviu.2023.103739

[2] Li, C., Yang, X. and Liang, G. (2023) Keyframe-Guided Video Swin Transformer with Multi-Path Excitation for Violence Detection. *The Computer Journal*, **67**, 1826-1837. https://doi.org/10.1093/comjnl/bxad103

[3] Hachiuma, R., Sato, F. and Sekii, T. (2023) Unified Keypoint-Based Action Recognition Framework via Structured Keypoint Pooling. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 22962-22971. https://doi.org/10.1109/cvpr52729.2023.02199

[4] Asad, M., Yang, J., He, J., Shamsolmoali, P. and He, X. (2020) Multi-Frame Feature-Fusion-Based Model for Violence Detection. *The Visual Computer*, **37**, 1415-1431. https://doi.org/10.1007/s00371-020-01878-6

[5] Contardo, P., Tomassini, S., Falcionelli, N., *et al*. (2023) Combining a Mobile Deep Neural Network and a Recurrent Layer for Violence Detection in Videos. *CEUR Workshop Proceedings*. *CEUR-WS*, Vol. 3402, 35-43.

[6] Mumtaz, N., Ejaz, N., Aladhadh, S., Habib, S. and Lee, M.Y. (2022) Deep Multi-Scale Features Fusion for Effective Violence Detection and Control Charts Visualization. *Sensors*, **22**, Article No. 9383. https://doi.org/10.3390/s22239383

[7] Aarthy, K. and Nithya, A.A. (2022) Crowd Violence Detection in Videos Using Deep Learning Architecture. 2022 *IEEE 2nd Mysore Sub Section International Conference* (*MysuruCon*), Mysuru, 16-17 October 2022, 1-6. https://doi.org/10.1109/mysurucon55714.2022.9972624

[8] Gupta, H. and Ali, S.T. (2022) Violence Detection Using Deep Learning Techniques. 2022 *International Conference on Emerging Techniques in Computational Intelligence* (*ICETCI*), Hyderabad, 25-27 August 2022, 121-124. https://doi.org/10.1109/icetci55171.2022.9921388

[9] Islam, M.S., Hasan, M.M., Abdullah, S., Akbar, J.U.M., Arafat, N.H.M. and Murad, S.A. (2021) A Deep Spatio-Temporal Network for Vision-Based Sexual Harassment Detection. 2021 *Emerging Technology in Computing, Communication and Electronics* (*ETCCE*), Dhaka, 21-23 December 2021, 1-6. https://doi.org/10.1109/etcce54784.2021.9689891

[10] Jahlan, H.M.B. and Elrefaei, L.A. (2021) Mobile Neural Architecture Search Network and Convolutional Long Short-Term Memory-Based Deep Features toward Detecting Violence from Video. *Arabian Journal for Science and Engineering*, **46**, 8549-8563. https://doi.org/10.1007/s13369-021-05589-5

[11] Singh, N., Prasad, O. and Sujithra, T. (2022) Deep Learning-Based Violence Detection from Videos. In: Satapathy, S.C., et al., Eds., *Intelligent Data Engineering and Analytics*, Springer, 323-332. https://doi.org/10.1007/978-981-16-6624-7_32

[12] Srivastava, A., Badal, T., Saxena, P., Vidyarthi, A. and Singh, R. (2022) UAV Surveillance for Violence Detection and Individual Identification. *Automated Software Engineering*, **29**, Article No. 28. https://doi.org/10.1007/s10515-022-00323-3

[13] Jeevan, R. and Avanthika, B. (2025) Intelligent Video Surveillance Systems with Violence Detection. 2025 *International Conference on Data Science, Agents & Artificial Intelligence* (*ICDSAAI*), Chennai, 28-29 March 2025, 1-6. https://doi.org/10.1109/icdsaai65575.2025.11011829

[14] Chandane, S., Nadar, A.T., Lokhande, M., Kanthakumar, D. and Shaikh, R. (2024) Violence Detection Using Deep Learning. 2024 *International Conference on Innovations and Challenges in Emerging Technologies* (*ICICET*), Nagpur, 7-8 June 2024, 1-6. https://doi.org/10.1109/icicet59348.2024.10616304

[15] Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T., Shlens, J. and Le, Q.V. (2020) Learning Data Augmentation Strategies for Object Detection. In: Vedaldi, A., *et al*., Eds., *Computer Vision—ECCV* 2020, Springer International Publishing, 566-583. https://doi.org/10.1007/978-3-030-58583-9_34

[16] Senadeera, D.C., Yang, X., Kollias, D. and Slabaugh, G. (2024) CUE-Net: Violence Detection Video Analytics with Spatial Cropping, Enhanced UniformerV2 and Modified Efficient Additive Attention. 2024 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 17-18 June 2024, 4888-4897. https://doi.org/10.1109/cvprw63382.2024.00493

[17] Cubuk, E.D., Zoph, B., Shlens, J. and Le, Q.V. (2020) Randaugment: Practical Automated Data Augmentation with a Reduced Search Space. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 14-19 June 2020, 702-703. https://doi.org/10.1109/cvprw50498.2020.00359

[18] Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., *et al*. (2023) VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 14549-14560. https://doi.org/10.1109/cvpr52729.2023.01398

[19] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[20] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[21] Ahmed, M., Ramzan, M., Ullah Khan, H., Iqbal, S., Attique Khan, M., Choi, J., *et al*. (2021) Real-Time Violent Action Recognition Using Key Frames Extraction and Deep Learning. *Computers*, *Materials & Continua*, **69**, 2217-2230. https://doi.org/10.32604/cmc.2021.018103

[22] Sharma, S., Sudharsan, B., Naraharisetti, S., Trehan, V. and Jayavel, K. (2021) A Fully Integrated Violence Detection System Using CNN and LSTM. *International Journal of Electrical and Computer Engineering* (*IJECE*), **11**, 3374-3380. https://doi.org/10.11591/ijece.v11i4.pp3374-3380

[23] de Oliveira Lima, J.P. and Figueiredo, C.M.S. (2021) Temporal Fusion Approach for Video Classification with Convolutional and LSTM Neural Networks Applied to Violence Detection. *Inteligencia Artificial*, **24**, 40-50. https://doi.org/10.4114/intartif.vol24iss67pp40-50

[24] Traoré, A. and Akhloufi, M.A. (2020) 2D Bidirectional Gated Recurrent Unit Convolutional Neural Networks for End-To-End Violence Detection in Videos. In: Campilho, A., *et al*., Eds., *Image Analysis and Recognition*, Springer International Publishing, 152-160. https://doi.org/10.1007/978-3-030-50347-5_14

[25] Rendón-Segador, F.J., Álvarez-García, J.A., Enríquez, F. and Deniz, O. (2021) ViolenceNet: Dense Multi-Head Self-Attention with Bidirectional Convolutional LSTM for Detecting Violence. *Electronics*, **10**, 1601. https://doi.org/10.3390/electronics10131601

[26] Abdali, A.R. (2021) Data Efficient Video Transformer for Violence Detection. 2021 *IEEE International Conference on Communication*, *Networks and Satellite* (*COMNETSAT*), Purwokerto, 17-18 July 2021, 195-199. https://doi.org/10.1109/comnetsat53002.2021.9530829

[27] Dosovitskiy, A. (2020) An Image Is Worth 16 x 16 Words: Transformers for Image Recognition at Scale.

[28] Li, K., Wang, Y., Gao, P., *et al*. (2022) Uniformer: Unified Transformer for Efficient Spatiotemporal Representation Learning.

[29] Zumerle, F., Comanducci, L., Zanoni, M., Bernardini, A., Antonacci, F. and Sarti, A. (2023) Procedural Music Generation for Videogames Conditioned through Video Emotion Recognition. 2023 4*th International Symposium on the Internet of Sounds*, Pisa, 26-27 October 2023, 1-8. https://doi.org/10.1109/ieeeconf59510.2023.10335439

[30] Huynh, V.T., Yang, H., Lee, G. and Kim, S. (2023) Prediction of Evoked Expression from Videos with Temporal Position Fusion. *Pattern Recognition Letters*, **172**, 245-251. https://doi.org/10.1016/j.patrec.2023.07.002

[31] Duja, K.U., Khan, I.A. and Alsuhaibani, M. (2024) Video Surveillance Anomaly Detection: A Review on Deep Learning Benchmarks. *IEEE Access*, **12**, 164811-164842. https://doi.org/10.1109/access.2024.3491868

[32] Sjöberg, M., Baveye, Y., Wang, H., *et al*. (2015) The MediaEval 2015 Affective Impact of Movies Task. *MediaEval*, Wurzen, 14-15 September 2015, 1436.

[33] Dai, Q., Zhao, R.W., Wu, Z., *et al*. (2015) Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. *MediaEval*, Wurzen, 14-15 September 2015, 1436.

[34] Trigeorgis, G., Ringeval, F., Marchi, E., *et al*. (2015) The ICL-TUM-PASSAU Approach for the MediaEval 2015 "Affective Impact of Movies" Task.

[35] Lam, V., Le, S.P., Le, D.D., *et al*. (2015) NII-UIT at MediaEval 2015 Affective Impact of Movies Task. *MediaEval*, Wurzen, 14-15 September 2015, 1436.

[36] Marin Vlastelica, P., Hayrapetyan, S., Tapaswi, M., *et al*. (2015) KIT at MediaEval 2015-Evaluating Visual Cues for Affective Impact of Movies Task. *MediaEval*, Wurzen, 14-15 September 2015.

[37] Li, X., Huo, Y., Jin, Q. and Xu, J. (2016) Detecting Violence in Video Using Subclasses. *Proceedings of the* 24*th ACM International Conference on Multimedia*, Amsterdam, 15-19 October 2016, 586-590. https://doi.org/10.1145/2964284.2967289

[38] Peixoto, B.M., Avila, S., Dias, Z. and Rocha, A. (2018) Breaking down Violence: A Deep-Learning Strategy to Model and Classify Violence in Videos. *Proceedings of the* 13*th International Conference on Availability*, *Reliability and Security*, Hamburg, 27-30 August 2018, 1-7. https://doi.org/10.1145/3230833.3232809

[39] Peixoto, B., Lavi, B., Pereira Martin, J.P., Avila, S., Dias, Z. and Rocha, A. (2019) Toward Subjective Violence Detection in Videos. *ICASSP* 2019-2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 8276-8280. https://doi.org/10.1109/icassp.2019.8682833

[40] Freire-Obregón, D., Barra, P., Castrillón-Santana, M. and Marsico, M.D. (2021) Inflated 3D Convnet Context Analysis for Violence Detection. *Machine Vision and Applications*, **33**, 15. https://doi.org/10.1007/s00138-021-01264-9

[41] Zheng, Z., Zhong, W., Ye, L., Fang, L. and Zhang, Q. (2021) Violent Scene Detection of Film Videos Based on Multi-Task Learning of Temporal-Spatial Features. 2021 *IEEE* 4*th International Conference on Multimedia Information Processing and Retrieval* (*MIPR*), Tokyo, 8-10 September 2021, 360-365. https://doi.org/10.1109/mipr51284.2021.00067

[42] Gu, C., Wu, X. and Wang, S. (2020) Violent Video Detection Based on Semantic Correspondence. *IEEE Access*, **8**, 85958-85967. https://doi.org/10.1109/access.2020.2992617

[43] 吴晓雨, 蒲禹江, 王生进, 刘子豪. 基于语义嵌入学习的特类视频识别[J]. 电子学报, 2023, 51(11): 3225-3237.

[44] Pu, Y., Wu, X., Wang, S., Huang, Y., Liu, Z. and Gu, C. (2022) Semantic Multimodal Violence Detection Based on Local-to-Global Embedding. *Neurocomputing*, **514**, 148-161. https://doi.org/10.1016/j.neucom.2022.09.090

[45] Wang, Q., Xiang, X., Zhao, J. and Deng, X. (2022) P2SL: Private-Shared Subspaces Learning for Affective Video Content Analysis. 2022 *IEEE International Conference on Multimedia and Expo* (*ICME*), 18-22 July 2022, 1-6. https://doi.org/10.1109/icme52920.2022.9859902

[46] Savadogo, W.A.R., Lin, C., Hung, C., Chen, C., Liu, Z. and Liu, T. (2023) A Study on Constructing an Elderly Abuse Detection System by Convolutional Neural Networks. *Journal of the Chinese Institute of Engineers*, **46**, 118-127. https://doi.org/10.1080/02533839.2022.2161941

[47] Negre, P., Alonso, R.S., González-Briones, A., Prieto, J. and Rodríguez-González, S. (2024) Literature Review of Deep-Learning-Based Detection of Violence in Video. *Sensors*, **24**, Article No. 4016. https://doi.org/10.3390/s24124016

[48] Vaishy, A., Basak, S. and Gautam, A. (2025) Early Violence Recognition Using Knowledge Distillation. In: Kakarla, J., *et al*., Eds., *Computer Vision and Image Processing*, Springer, 57-70. https://doi.org/10.1007/978-3-031-93697-5_5

[49] Hanief Wani, M. and Faridi, A.R. (2024) Deep Learning-Based Video Surveillance System for Suspicious Activity Detection. *Journal of Intelligent & Fuzzy Systems*, **47**, 71-82. https://doi.org/10.3233/jifs-234365

[50] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., *et al*. (2023) UniFormerV2: Unlocking the Potential of Image Vits for Video Understanding. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 1632-1643. https://doi.org/10.1109/iccv51070.2023.00157

[51] Padilla, R., Netto, S.L. and da Silva, E.A.B. (2020) A Survey on Performance Metrics for Object-Detection Algorithms. 2020 *International Conference on Systems*, *Signals and Image Processing* (*IWSSIP*), Niteroi, 1-3 July 2020, 237-242. https://doi.org/10.1109/iwssip48289.2020.9145130

[52] Loshchilov, I. and Hutter, F. (2017) Fixing Weight Decay Regularization in Adam.