

Article

Campus Abnormal Behavior Detection with a Spatio-Temporal Fusion–Temporal Difference Network

Fupeng Wei ¹, Yibo Jiao ¹, Nan Wang ^{2,*}, Kai Zheng ³, Ge Shi ¹, Mengfan Yang ¹ and Wen Zhao ⁴¹ School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China² School of Information Science and Technology, Hainan Normal University, Haikou 571158, China³ School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China⁴ Measurement Center of Guangdong Power Grid Co., Ltd., Guangzhou 510062, China

* Correspondence: nanwang.ac@hainnu.edu.cn

Abstract

The detection of abnormal behavior has consistently garnered significant attention. Conventional methods employ vision-based dual-stream networks or 3D convolutions to represent spatio-temporal information in video sequences to identify normal and pathological behaviors. Nonetheless, these methodologies generally employ datasets balanced across data categories and consist solely of two classifications. In actuality, anomalous behaviors frequently display multi-category characteristics, with each category's distribution demonstrating a pronounced long-tail phenomenon. This paper presents a video-based technique for detecting multi-category abnormal behavior, termed the Spatio-Temporal Fusion–Temporal Difference Network (STF-TDN). The system first employs a temporal difference network (TDN) model to encapsulate movie temporal dynamics via local and global modeling. To enhance recognition performance, this study develops a feature fusion module—Spatial-Temporal Fusion (STF)—which augments the model's representational capacity by amalgamating spatial and temporal data. Furthermore, given the long-tailed distribution characteristics of the datasets, this study employs focused loss rather than the conventional cross-entropy loss function to enhance the model's recognition capability for under-represented categories. We perform comprehensive experiments and ablation studies on two datasets. Precision is 96.3% for the Violence5 dataset and 87.5% for the RWF-2000 dataset. The results of the experiment indicate the enhanced efficacy of the proposed strategy in detecting anomalous behavior.

Keywords: campus violence; behavior detection; temporal difference network; feature fusion

Academic Editor: George A. Tsihrintzis

Received: 26 September 2025

Revised: 23 October 2025

Accepted: 27 October 2025

Published: 29 October 2025

Citation: Wei, F.; Jiao, Y.; Wang, N.; Zheng, K.; Shi, G.; Yang, M.; Zhao, W. Campus Abnormal Behavior Detection with a Spatio-Temporal Fusion–Temporal Difference Network. *Electronics* **2025**, *14*, 4221. <https://doi.org/10.3390/electronics14214221>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Bullying and violence in educational institutions is a significant global topic in the realm of education [1–6]. Research disseminated by the United Nations Educational, Scientific and Cultural Organization (UNESCO) has elucidated the gravity of the bullying phenomenon in educational institutions, with statistics indicating that approximately one-third of children globally have encountered at least one bullying incident in the preceding month [1]. In this context, we focus on detecting specific anomalous behaviors—including physical altercations, unauthorized boundary crossing, sudden falls, and emergency scenarios—that deviate from normal campus activities and may pose potential safety risks. This study's results indicated that children who experienced regular bullying were considerably more prone to feelings of ostracism at school compared to their non-bullied

counterparts, and their rates of absence were markedly elevated. Moreover, these children generally underperform academically compared to their peers and have a considerably elevated risk of school dropout following the completion of secondary schooling. Because violence in educational institutions, in all its manifestations, represents a grave infringement of the right to education and the physical and mental well-being of children and adolescents, UNESCO Member States have collectively established the first Thursday of November annually as the “International Day Against Violence and Bullying in Schools.” This initiative aims to globally raise awareness regarding the issues of violence and bullying in schools, including cyberbullying. The objective is to enhance global awareness of violence and bullying, including cyberbullying, in educational institutions and to implement effective strategies for prevention and response.

In the initial phases of bullying prevention at educational institutions, a prevalent intervention method is to urge bystanders to report the occurrence to school personnel. This method has considerable problems regarding timeliness, and witnesses may neglect their reporting duties due to fear, social pressure, or other psychological influences. Meanwhile, while video surveillance technology can aid in identifying school bullying to some degree [7], real-time manual monitoring requires substantial human resources and is susceptible to oversights and errors when managing multiple surveillance feeds concurrently. This method, which is dependent on individual active participation, has significant shortcomings in the prevention and quick response to school bullying due to its intrinsic limits. Consequently, an automated detection system is critically required to efficiently mitigate anomalous activities in educational institutions.

Recently, machine learning (ML)-based automated detection systems have emerged as a prominent research focus, designed to identify and predict potential bullying behaviors through advanced algorithms that conduct in-depth analyses of school surveillance footage, thereby addressing the shortcomings of traditional eyewitness-reliant methods [8,9]. Traditional ML algorithms depend on manually crafted features for feature extraction, hindering their ability to effectively capture the varied and evolving patterns of bullying behaviors in educational environments. Moreover, in complicated situations characterized by lighting variations, occlusion, and background interference, conventional machine learning approaches struggle to reliably extract and differentiate essential behavioral variables. Consequently, the advent of deep learning techniques, characterized by their robust feature learning skills, may autonomously train and extract intricate spatio-temporal features straight from unprocessed video data, offering a novel option for the automated identification of school bullying [10,11].

However, common deep learning architectures such as Two-Stream networks and 3D CNNs face specific challenges in campus settings. Two-Stream pipelines are computationally expensive due to optical-flow estimation, limiting real-time analysis in multi-camera school environments, while 3D CNNs impose heavy compute/memory budgets and may under-represent fine-grained temporal cues under practical constraints [12,13]. Furthermore, many methods are developed or validated under class-balanced assumptions, overlooking the inherently long-tailed nature of campus surveillance where abnormal events are rare amid predominant normal activities [14–16]. These limitations motivate our design that emphasizes efficient Spatio-Temporal Fusion and long-tailed learning.

The article proposes a technique for detecting abnormal behavior on campus based on video sequences, as illustrated in Figure 1. It processes video streaming data from campus surveillance cameras to enable real-time classification and monitoring of campus activities through a video analysis engine. For recognized routine campus activities, the system will remain inactive to prevent false alarms. Upon detection of anomalous behaviors, such as suspected bullying, the system will immediately activate the alarm mechanism and alert

management for an immediate response. The model has a parameter inheritance method (shown by a dotted line for Retrain), enabling succeeding researchers to immediately employ the stored node parameter values during training with new data, thus obviating the requirement to re-initiate the parameter optimization process. The system would offer organized storage for unusual behavioral films to enhance the development of a comprehensive and high-quality dataset.

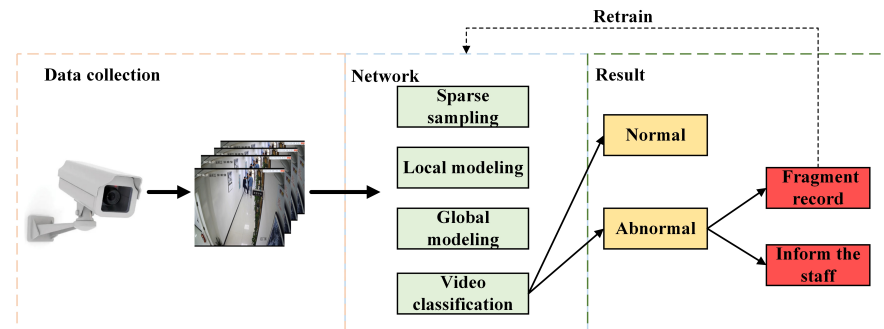


Figure 1. The campus abnormal behavior detection method's general structure.

2. Related Work

2.1. Conventional Anomaly Detection Techniques

The detection of violence in educational institutions has garnered significant interest within the academic sphere as a crucial intervention strategy to safeguard the physical and psychological well-being of students. The conventional detection paradigm during technological advancements primarily depends on the implementation of video surveillance systems and real-time monitoring by security professionals, augmented by a definitive behavioral code system as a preventive strategy. Nonetheless, this manual monitoring and rule-based detection framework possesses considerable limitations, and its subjective nature may result in identification inaccuracies, including false alarms and omissions. With the swift advancement of computing technology, machine learning methods may autonomously extract features from intricate data using multi-dimensional data analysis and pattern recognition techniques, thereby identifying anomalous behavioral characteristics that markedly deviate from standard behavioral patterns. In this area, Sanjay Singla et al. [17] employed natural language processing techniques and several machine learning algorithms to examine the linguistic characteristics of Hinglish texts and detect occurrences of cyberbullying, aiming to mitigate false positives and under-reporting. Similarly, Al-Khowarizmi et al. [18] employed a support vector machine algorithm to effectively identify cyberbullying behaviors associated with Indonesia's "Cipta Kerja" policy on Twitter, minimizing subjective errors inherent in manual monitoring and addressing the constraints of conventional approaches. ML approaches are more successful in identifying probable patterns of violence by minimizing dependence on human resources. Nonetheless, it is constrained by the laborious procedure of manual feature creation and extraction and is heavily reliant on domain expertise. Consequently, researchers are dedicated to investigating deep learning (DL) techniques by utilizing multimodal data [19–22] (such as audio and sensor data) for the automatic identification of anomalous behaviors to address the intrinsic shortcomings of conventional machine learning approaches in feature engineering.

2.2. Deep Learning-Based Anomaly Detection Methods

As data volume increases and feature engineering becomes more complex, deep learning methods can extract and analyze rich semantic information in videos via multi-layer neural architectures, thereby reducing reliance on manually crafted features. In a prior paper, Liang Ye et al. [23] proposed a sensor-based system for violence detection that gathers

data via motion sensors, applies an enhanced Relief-F algorithm for feature extraction, and then uses a two-stage classifier to improve discrimination. Batyrkhan et al. [19] proposed a skeleton-based violence detection system that identifies hostile situations by analyzing human postures with neural networks. Simone et al. [24] introduced a 3D convolution-based technique for video violence detection, using pre-trained 3D convolutions as a feature extractor to improve generalization without a priori assumptions. Ullah et al. [25] presented an industrial IoT system that adopts a lightweight CNN and ConvLSTF to improve computational efficiency. Romas Vijeikis et al. [26] proposed a lightweight model based on U-Net, with MobileNetV2 as an encoder and LSTM for temporal feature extraction and classification, effectively capturing dynamic information in videos.

On the other hand, vision-only pipelines predominantly extract cues from RGB video streams to identify abnormal behaviors by analyzing motion, pose, or scene changes [16,27–31]. These methods often rely on binary classification or category-balanced datasets for training. However, the predominance of image/video data in most surveillance systems reduces the practicality of multimodal designs, and in campus scenarios the data distribution is highly skewed, which poses challenges for robust generalization when models are re-trained.

Despite the above progress, prior deep-learning approaches exhibit two recurring gaps that are critical in campus deployments: (i) insufficient deep fusion of spatial and temporal cues, where fusion is often deferred to late-stage weighted averaging or shallow concatenation, limiting the modeling of subtle, long-term interactions [14,15,24], and (ii) a mismatch between training assumptions and real-world prevalence, as many methods are trained/evaluated on balanced or binary datasets, overlooking the inherently long-tailed distribution of campus anomalies [16,27,29]. These gaps motivate our design choices toward stronger Spatio-Temporal Fusion and learning under long-tailed distributions.

Inspired by architectural innovations in other domains—for example, the CNN-LVQ model of Mohd Anul Haq [32], which integrates convolutional networks with learning vector quantization under rigorous hyperparameter optimization—we adopt a similar principle of task-tailored architectural fusion for our problem. Accordingly, this paper proposes a video sequence-based method for detecting anomalous actions on campus. First, we construct an imbalanced dataset, Violence5, to mirror the data distribution commonly observed in college settings. To address the long-tailed distribution, we employ focal loss to mitigate the adverse effects of class imbalance. Second, because many existing studies model either spatial or temporal information in isolation, we propose a novel Spatio-Temporal Fusion module (STF) to integrate both modalities. The experimental results indicate that the proposed algorithm attains an accuracy of 96.3% and a precision of 96.2% on our dataset, demonstrating clear advantages over current comparative models.

3. Contribution

This paper introduces the Spatio-Temporal Fusion–Temporal Difference Network (STF-TDN), a novel video classification framework that systematically addresses two critical challenges in real-world campus anomaly detection: severe class imbalance and ineffective Spatio-Temporal Fusion. Our contributions are threefold:

1. **A Novel Spatio-Temporal Fusion Module:** We design a dedicated STF block to replace the rudimentary weighted averaging in standard TDNs. By employing a SENet-inspired bottleneck with GELU activation and residual connections, it enables dynamic, channel-wise refinement for a deeper integration of spatial and temporal features, thereby strengthening the representation of complex behaviors.
2. **Robust Learning under Long-Tailed Distributions:** We introduce the use of focal loss to replace cross-entropy, directly tackling the natural class imbalance in surveillance data. This strategy recalibrates the learning focus towards hard and minority examples

without resorting to artificial re-balancing, significantly enhancing model robustness and generalization.

3. A Realistic Benchmark for Fine-Grained Analysis: We construct the Violence5 dataset, a multi-category, naturally imbalanced collection of campus scenes. It emphasizes subtle anomalous behaviors and supports model evaluation under real-world long-tailed distributions.

4. Methods

4.1. Improved TDN Network

In recent years, deep neural networks have achieved substantial advancements in video action recognition [12–15,33–37]. Temporal modeling, a crucial element in collecting action information in videos, is primarily executed through two fundamental approaches. The dual-stream network architecture utilizes a parallel processing approach, wherein one branch pulls static visual elements from the RGB imagined space, while the other captures temporal motion information from the optical flow field. Thereafter, the outputs from the two branches are amalgamated via a feature fusion strategy to enhance action recognition accuracy; however, this method incurs high computational complexity, primarily due to the extra resources needed for optical flow extraction. The second approach utilizes 3D convolution [38] or temporal convolution [39–41] to directly extract spatio-temporal information from RGB frame sequences. Nonetheless, the modeling capacity of 3D convolutions in the temporal dimension is constrained, along with its significant processing demands, which somewhat restricts its efficacy in practical applications.

To overcome the shortcomings of the aforementioned methodologies, other experts have proposed enhancement strategies. Wang L et al. [42] proposed a network architecture utilizing time slicing, which markedly diminishes computational complexity via a sparse sampling method, whereas Carreira J et al. [43] enhanced the temporal feature extraction efficacy of 3D convolutions by implicit modeling. Furthermore, Lin J et al. [14] proposed the time offset model, allowing 2D convolutions to attain performance comparable to 3D convolutions in temporal feature extraction.

The TDN model integrates the benefits of the aforementioned methodologies and demonstrates that the RGB difference method effectively extracts action information. The model employs multiscale temporal modeling via two primary modules: the S-TDM, which concentrates on local temporal feature extraction, and the L-TDM, which manages global temporal modeling at the video level. The local features obtained from the S-TDM are inputted into the L-TDM, hence facilitating the integration of temporal data from the local level to the global level.

This work employs the optimized TDN model [15] to identify anomalous behaviors on campus (as shown in Figure 2), concentrating on the enhancement of two fundamental modules:

1. To address the issue of data distribution imbalance, the Multi-Class Focus Loss Function (MFLF) is developed to supplant the conventional Cross-Entropy Loss Function (CELF), therefore markedly improving the model's capacity to learn from sparse category samples.
2. We revise the model architecture to enhance overall performance by upgrading the local temporal modeling module (S-TDM) of the original model to the spatio-temporal differential fusion module (S-TDF). The two optimization procedures collaboratively enhance the system's detection efficacy through the design of the loss function and the fusion of spatio-temporal features, respectively.

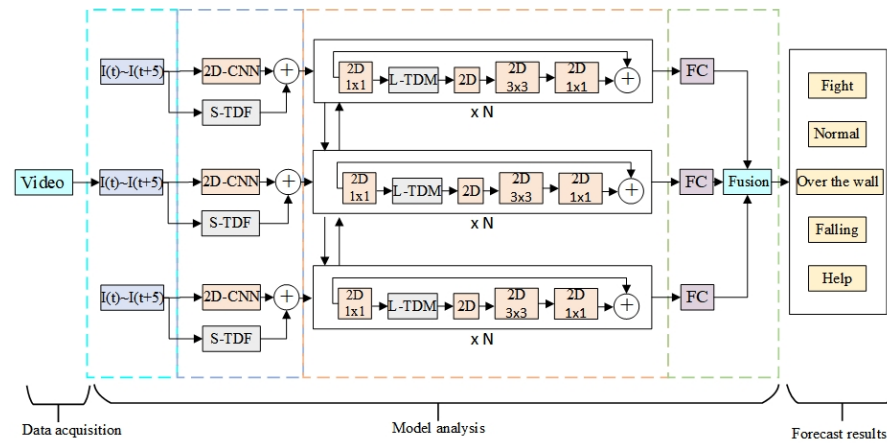


Figure 2. Structure diagram of the improved IDN model, with the number of layers of the back three networks in the ResNet50 dataset.

4.1.1. S-TDF Module

This study introduces the Spatio-Temporal Fusion (STF) module to enhance feature fusion efficiency and effectiveness, as illustrated in Figure 3. Integrated into the S-TDF module (highlighted in green), the STF module strengthens the model's ability to capture spatio-temporal correlations in anomalous behaviors by aligning and fusing spatial features (extracted from the central frame) with temporal dynamics (derived from RGB inter-frame differences). The video input is divided into eight segments, each containing 90 frames. From each segment, five frames—the central frame and its two adjacent frames—are sampled, forming a frame sequence denoted as $I = [I_1, \dots, I_5]$, where each I is a 224×224 RGB image with dimensions $[T, C, H, W]$. Spatial features are obtained from the central frame I_3 , while temporal features are generated from RGB differences across the five frames. These features are then synchronously fused via the STF module. Unlike the original TDN, which relies on weighted averaging for Spatio-Temporal Fusion, the STF module employs a structured alignment and integration mechanism, improving the cohesion and discriminability of the resulting feature representations.

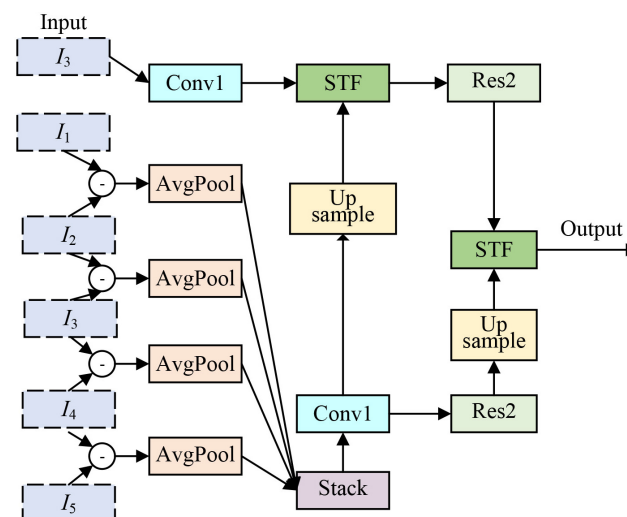


Figure 3. S-TDF module (Res2 illustrates the use of the second-layer network in ResNet50 to extract features and “-” denotes frame-wise subtraction ($I_t - I_{t-1}$) to highlight motion.).

The article aims to rectify the inadequate modeling capacity stemming from the early S-TDM module's dependence on the rudimentary weighted average fusion of spatio-temporal features by introducing a two-tier STF module [44]. The initial STF is implemented in the

shallow network for the primary fusion of fundamental spatio-temporal features (e.g., limb movement contours); the subsequent STF is integrated into the deep network to execute the secondary refinement fusion of high-level semantic features (e.g., complex behavioral patterns) produced by the first STF. The second STF is integrated within the deep network to execute secondary fine-grained fusion of the high-level semantic features (e.g., intricate behavioral patterns) produced by the first STF, thereby markedly enhancing spatio-temporal information integration through this hierarchical progressive architecture (the shallow layer emphasizes local motion, while the deep layer correlates global behavior).

The STF module is designed for adaptive Spatio-Temporal Feature Fusion. Its three-step process begins with concatenating spatial (S) and temporal (T) features along the channel dimension. This combined tensor then passes through a SENet-inspired bottleneck—comprising a linear layer and GELU activation—that performs a nonlinear transformation and squeeze–excitation operations to dynamically highlight the most informative interactions. A residual connection adding the simple average of the original S and T features to the refined output preserves the original information and enriches the final representation.

Furthermore, drawing inspiration from the squeeze–excitation (SE) network, a dynamic channel weighting mechanism is incorporated into the feature fusion process (as illustrated in Figure 4). Initially, spatial features (n channel) and temporal features (n channel) are combined to create two n -channel features. Subsequently, these features are projected into a low-dimensional space via a linear layer, followed by the application of the GELU activation function [45], resulting in a reduction in n -channels. To mitigate feature degradation, the weighted average of the original spatio-temporal features is residually connected to the processed features. The mathematical representation of this STF module is

$$H(i) = \text{avg}(S + T) + W^T(\text{cat}(S, T)) \quad (1)$$

where S and T represent spatial and temporal features, respectively, and avg means the average, which we superimpose in the channel dimension using cat .

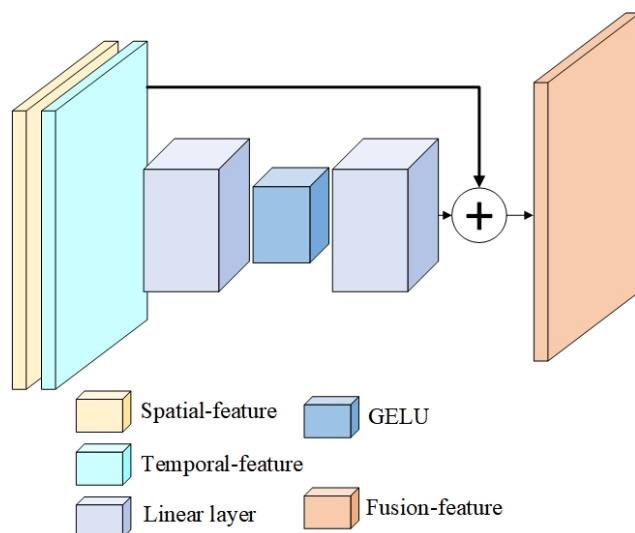


Figure 4. Fusion of aligned spatial and temporal features through the STF module.

4.1.2. L-TDM Module

L-TDM leverages S-TDF features for global temporal modeling through a bidirectional multiscale velocity difference mechanism. As shown in Figure 5, this module captures behavioral dynamics by computing oriented differences between adjacent segment features F_i and F_{i+1} , rather than relying on instantaneous states. The bidirectional design integrates contextual information from both past and future segments, while the multiscale archi-

texture utilizes parallel convolutional branches with varying receptive fields to capture motion patterns from local limb movements to global displacement. The input feature F is first compressed via Conv1 and split into adjacent segment representations F_i (from slices 1–7) and F_{i+1} (from slices 2–8). Spatial misalignment is mitigated using a channel-wise Conv2 layer and average pooling. The alignment difference $C(F_i, F_{i+1})$ is defined as

$$C(F_i, F_{i+1}) = F_i - \text{Conv}(F_{i+1}) \quad (2)$$

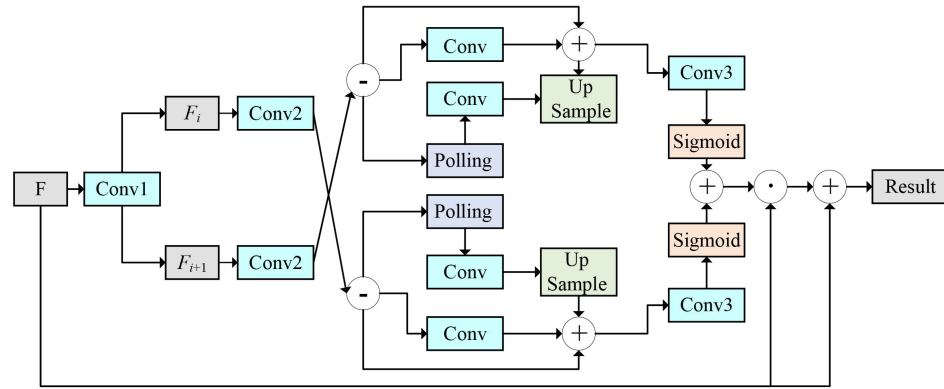


Figure 5. L-TDM module.

The feature representation is then enhanced by a three-branch structure in which one branch performs convolution directly on the existing size of the feature map: one branch is a residual join for feature representation enhancement. The third branch pools the image to obtain a smaller image; then, convolution is performed by Conv, and finally, the three are size-aligned and directly summed by upsampling. The enhanced features obtained from the three-branch structure are denoted by $M(F_i, F_{i+1})$. Ultimately, the module outputs a set of refined, temporally discriminative attention weights. These weights adaptively enhance the motion feature channels most relevant to anomalous behaviors through channel-wise weighting while suppressing static or irrelevant background interference, thereby significantly improving the model's discriminative capacity for complex anomalous behaviors. The formula for finding $M(F_i, F_{i+1})$ is as follows:

$$M(F_i, F_{i+1}) = \sigma(\text{Conv} \sum_{j=1}^N \text{CNN}_j(C(F_i, F_{i+1}))) \quad (3)$$

where N represents the number of branches, while σ denotes the sigmoid activation function.

Finally, the bidirectional streams are extracted to the spatio-temporal features and weighted together by a factor of 0.5 to form an attention map with the input features F for the elemental dot product, and the output is obtained after connecting with the residuals of F . The formula is as follows:

$$F_i \odot \Theta(F_i, F_{i+1}) = F_i \odot \frac{1}{2} [M(F_i, F_{i+1}) + M(F_{i+1}, F_i)] \quad (4)$$

where Θ represents the L-TDM:

4.2. Focal Loss Function

The distribution of training data in classical classification and identification tasks is frequently subject to artificial equalization, and there is no substantial difference in the number of samples from various categories. A balanced dataset simplifies the necessity for algorithm resilience and, to some extent, guarantees the trustworthiness of the final model. However, as the number of categories of interest grows, maintaining a balance among them would result in exponentially growing acquisition costs. The natural collection of data from

each category of campus abnormal behavior detection ranges in number, and this type of data is also known as long-tail data. When training classification and recognition systems on long-tail data, the models tend to focus on the head data and disregard the tail data. The efficient utilization of unbalanced long-tail data to train a balanced classifier improves data sampling speed while dramatically lowering acquisition costs. The focal Loss function is a loss function that can effectively counteract the long-tailed distribution of data.

Focal Loss is a loss function that adjusts the weights to reduce the contribution of easily categorized samples to the total loss, focusing the model's attention on the difficult-to-classify samples. Kai-Ming He first presented it to overcome the model performance problem caused by data imbalance [46].

The focal loss in the binary classification issue is determined as follows:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (5)$$

where p_t represents the anticipated value of the target, often after softmax, and denotes the model's predicted probability for the target category. α_t is a variable that regulates the weighting of positive and negative samples in the overall loss. When the data are uneven, changing α_t can equalize the contribution of positive and negative samples to the total loss. γ is also a controllable parameter that controls the contribution of simple and difficult-to-classify samples to the overall loss. $\log(p_t)$ is a component of the cross-entropy loss function that represents the model's loss when the target category is properly predicted.

In comparison to the classic cross-entropy loss function, focal loss reduces the contribution of readily categorized and majority-category data to total loss by modifying the weights so that the model focuses more on difficult-to-categorize and minority-category samples. This increases the model's performance on these samples and, as a result, the overall performance.

5. Experiment

5.1. Environment Setup and Dataset Construction

This experiment utilized an NVIDIA RTX3090 graphics processor (GPU) with 24GB of RAM, PyTorch 2.1 deep learning framework, and the Python 3.8 programming language. This study constructed Violence5, a dataset of campus deviant behavior, which was compiled from several sources like YouTube, Bilibili videos, self-recorded surveillance films, and public information. The dataset comprises 1098 brief video clips (2–5 s in duration), categorized into a training set (732 instances) and a test set (366 instances) in a 7:3 ratio, with each clip labeled independently. To preserve the authentic distribution characteristics of anomalous behaviors on campus, no artificial category equalization is used in the dataset, resulting in a naturally distributed sample size for each category. Table 1 shows the number of video categories. Some examples from the collected video frames are displayed in Figure 6. For the in-house Violence5 dataset curated to support this study, we enable face anonymization (blurring) by default at deployment. Our system targets human action recognition; it models body pose, motion patterns, and spatio-temporal dynamics and does not perform identity recognition or re-identification. Proactively obfuscating faces—the most sensitive biometric identifier—mitigates student-privacy risk at the source and is fully aligned with our technical design.

Table 1. Video category list.

Fight	Normal	Over the Wall	Help	Falling
319	274	261	126	118



Figure 6. Four abnormal samples in the collected video frames.

To precisely define the behavioral categories within the Violence5 dataset, clear distinctions are established for the “falling” and “help” classes against normal behaviors. The “falling” category is defined as a sudden, involuntary, and uncontrolled descending motion (e.g., slipping and fainting), which is fundamentally distinct from voluntary and controlled actions such as “sitting down”. The “help” category specifically refers to the provision of immediate physical assistance (e.g., helping someone up and intervening in a fight) in response to an ongoing anomalous event. The key distinguishing factor for “help” is its occurrence within an urgent context, thereby differentiating it from routine, non-emergency social interactions categorized as “normal”.

Figure 7 depicts the video quality distribution in the Violence5 dataset. The figure reveals numerous prominent clusters centered on the usual resolutions of 1080P, 1280P, and 1920P, which are largely from surveillance cameras or video sites with greater resolutions. There is also a small collection of videos of approximately 500P, which are more likely to be from the public dataset’s lower-definition videos.

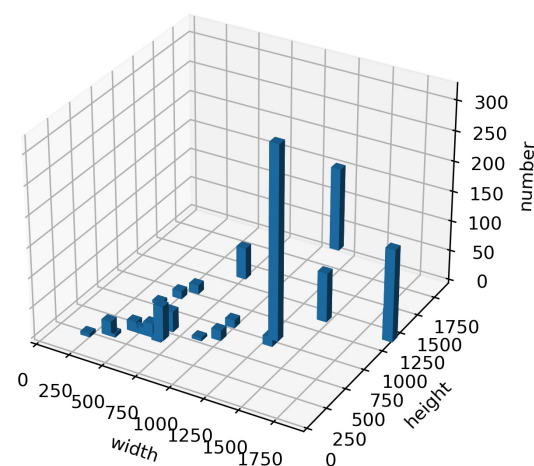


Figure 7. Resolution distribution of the Violence5 database. The unit is in pixels.

5.2. Experimental Configuration

We normalize the size of all videos to 224×224 . The size of 224×224 was chosen since we employ ResNet50 as the backbone network, and this size corresponds to ResNet50’s input size for pretraining on ImageNet. According to the results, the 224×224 size also performs better. We utilize sparse sampling to break each video into eight uniform segments and take five frames of images in each segment. This sampling strategy ensures that the

model remains consistent across varying lengths of video input and is more computationally efficient than dense sampling. In the S-TDF model, we employ the conv1_layers of ResNet50 to extract features and the STF module to fuse spatial and temporal information. In the L-TDM model, we extract features using ResNet50's conv2_layers, conv3_layers, and conv4_layers. On both the Violence5 and RWF-2000 datasets, the number of training cycles is 60, with learning rates decaying at 30, 45, and 55 rounds, respectively.

5.3. Evaluation Metrics

We defined nonviolent behavior as positive and violent behavior as negative in the RWF-2000 dataset, so *TP* (true positive) means nonviolent behavior was identified as nonviolent; *FP* (false positive) means violent behavior was identified as nonviolent; *TN* (true negative) means violent behavior was identified as violent; and *FN* (false negative) means nonviolent behavior was identified as violent. We evaluate the model's performance in the Violence5 dataset by obtaining the average scores for each category. The following four metrics were used to evaluate classification performance:

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \quad (6)$$

$$P = \frac{TP}{TP + FP} \quad (7)$$

$$R = \frac{TP}{TP + FN} \quad (8)$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (9)$$

5.4. Introduction to Experimental Methods

In total, we tested five different models for the RWF-2000 and Violence5 datasets, namely [42], TSM [14], H2CN [47], TDN [15], and our own STF-TDN. All comparative models were trained under identical settings (a learning rate of 0.01, the SGD optimizer, and a batch size of 64), ensuring a fair comparison. Among these, the TSN network addresses the issue that the Two-Stream network [48] cannot manage the long-time structure, as well as the fact that dense sampling takes up a lot of memory space and requires a long time when compared to sparse sampling. TSM enhances information in the time dimension via channel shifting in the time dimension and improves video classification accuracy. H2CN offers a unique temporal convolution, HgC, with an hourglass shape for HgC's temporal receptive field, in which the spatial receptive field is enhanced in the anterior and posterior time frames, allowing substantial shifts to be captured. TDN combines local modeling with global modeling for video-level temporal modeling of information stacking across segments. Our proposed STF-TDN model improves feature fusion by optimizing the TDN's model feature fusion module with the STF structure. Meanwhile, the focal loss function is utilized instead of the cross-entropy loss function for datasets with unbalanced data types to increase the model's accuracy on unbalanced datasets.

5.5. Experimental Comparison on the RWF-2000 Dataset

The RWF-2000 dataset [16] has 2000 video clips, with 1600 instances allocated to the training set and 400 instances to the test set, following an 8:2 ratio. The dataset comprises videos of two categories, violent and nonviolent behaviors, each representing 50% of the total. All clips are directly recorded by surveillance cameras without multimedia enhancement, thereby accurately reflecting the characteristics of violent events in real scenarios. The data size substantially surpasses that of prior comparable datasets. For the RWF-2000 dataset, the experimental data are shown in Table 2.

Table 2. Experimental comparison of the RWF-2000 dataset.

Model	Acc (%)	P (%)	R (%)	F ₁ (%)	Param (M)
TSN	77.3	75.8	80.0	77.9	23.5
TSM	82.8	80.8	86.0	83.3	23.5
H2CN	80.5	82.1	78.0	80.0	23.7
TDN	84.3	87.0	80.5	83.6	24.0
Ours	87.5	89.9	84.5	87.1	24.2

Our STF-TDN model obtains an accuracy of 87.5% and 89.9% precision on the RWF2000 dataset. When compared to the original TDN model, it improves accuracy by 3.2% and precision by 2.9% without significantly increasing the number of parameters. Overall, the upgraded TDN model outperforms the previous TDN model. On RWF-2000, the best results were obtained when the learning rate was set to 0.024, the momentum was set to 0.6, and the number of hidden layer parameters in the two STFs was set to 16 and 256.

5.6. Experimental Comparison on the Violence5 Dataset

Table 3 compares our model to the other four models based on *Acc*, *P*, *R*, the *F₁* value, and the number of parameters. As shown in the table, our model outperforms the other models in the four metrics assessed, achieving 96.3% accuracy and 96.2% precision, which are greater than those of the original TDN model. Although the number of parameters is slightly more than in the other models, it is still within acceptable bounds, and we believe that the overhead is worth it considering the new model's performance gains. We increased the number of parameters of the STF module's hidden layer on the Violence5 dataset, so it has slightly more overall parameters than the RWF-2000 dataset. On Violence5, the best results were obtained when the learning rate was set to 0.013, the momentum was set to 0.64, and the number of hidden-layer parameters in the two STF modules was set to 256.

Table 3. Experimental comparison of the Violence5 dataset.

Model	Acc (%)	P (%)	R (%)	F ₁ (%)	Param (M)
TSN	91.5	91.0	91.2	91.1	23.5
TSM	91.5	91.5	90.8	91.0	23.5
H2CN	87.8	88.5	86.7	87.0	23.7
TDN	94.5	94.8	94.5	94.6	24.0
Ours	96.3	96.2	96.4	96.3	24.3

5.7. Ablation Experiments

We conducted systematic ablation experiments on the Violence5 dataset to validate the effectiveness of the focal loss + STF approach for campus abnormal behavior detection. The experimental setup included four comparative groups: the original TDN method, the standalone STF module method, the focal loss method, and the combined focal loss + STF approach. Specifically, the STF method enhances feature representation by embedding our proposed Spatio-Temporal Feature Fusion module (STF) into the original framework; the focal loss method addresses long-tailed data distribution by replacing cross-entropy loss with focal loss; while the focal loss + STF method further incorporates the STF module for collaborative optimization.

As shown in Table 4, the focal loss + STF approach surpasses all baseline and individual improved methods across all metrics, validating the efficacy of this combined strategy for boosting abnormal behavior detection performance. Notably, the significant superiority of the focal loss method over the original baseline confirms its capability to counteract the long-tail effect caused by class imbalance. Furthermore, the accuracy improvement

of the focal loss + STF method compared to standalone focal loss demonstrates that the STF module optimizes detection precision by enhancing the model's capability to extract spatio-temporal features of abnormal behaviors.

Table 4. Violence5 dataset ablation experiment.

Model	Acc (%)	P (%)	R (%)	F ₁ (%)
Original	94.5	94.8	94.5	94.6
STF	95.4	95.2	94.8	94.9
Focal loss	95.4	95.7	95.3	95.5
Focal loss + STF	96.3	96.2	96.4	96.3

We further performed parameter sensitivity analysis to validate our design choices. For the STF module, experimental results demonstrated that a hidden layer dimension of 256 achieved optimal performance while maintaining computational efficiency. Regarding focal Loss, an evaluation of different γ values (one, two, and three) showed minimal performance variation, confirming that its core reweighting mechanism effectively handles class imbalance across different parameter settings.

5.8. Applications

We utilized the STF-TDN model to detect abnormal behavior in real-world scenarios to further evaluate its performance. The video collection cameras are placed throughout the campus in classrooms, walls, squares, and other areas. When students exhibit abnormal behavior, STF-TDN will recognize the corresponding behavior category from the video in real time and send an alarm message to security staff. At the same time, abnormal video snippets will be recorded for archival purposes. Figure 8 depicts multiple video frames of our school campus's outside and indoor inspection results. The results indicate that STF-TDN can efficiently identify students' abnormal behavior in real-world circumstances. Our detection algorithm can assist security staff in intervening promptly and preventing abnormal behaviors such as student bullying. At the same time, the monitoring algorithm significantly reduces the effort of security staff in maintaining campus security, which has considerable practical application value for the construction of a safe campus.



Figure 8. Examples of video frame detection results: (a) help—a person calling for assistance; (b) over the wall—a person climbing over a wall; (c) fight—two individuals in a physical altercation; (d) falling—a person accidentally falling; (e) normal—daily indoor behavior; (f) normal—ordinary outdoor activity. The red labels in each frame denote the predicted behavior category.

6. Conclusions

This work advances violence detection under real-world, long-tailed conditions by coupling TDN with the proposed STF module and focal loss. The resulting STF–TDN model delivers consistent improvements on in-house and public datasets while remaining lightweight for practical deployment. Extensive experimental comparisons demonstrate that STF–TDN outperforms several state-of-the-art methods across all evaluation metrics. On Violence5, it achieves 96.3% accuracy and 96.2% precision, with comparable gains on RWF-2000, underscoring its effectiveness and practical value for campus security applications. Future work will further optimize computational efficiency, extend coverage to subtler anomalous behaviors, and address privacy, fairness, and governance considerations in educational settings.

Author Contributions: Conceptualization, F.W., Y.J. and N.W.; Methodology, F.W., Y.J. and N.W.; Software, Y.J. and K.Z.; Validation, F.W., G.S. and M.Y.; Formal analysis, F.W., Y.J. and W.Z.; Investigation, Y.J. and K.Z.; Resources, N.W.; Data curation, F.W., Y.J., G.S., M.Y. and W.Z.; Writing—original draft, F.W., Y.J., G.S. and M.Y.; Writing—review and editing, N.W. and W.Z.; Visualization, K.Z.; Supervision, N.W.; Project administration, N.W.; Funding acquisition, F.W., N.W. and G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the following projects: the Key Research Projects of Henan Higher Education Institutions (Grant No. 23A520031), the International Science and Technology Cooperation Project of Henan Province (Grant No. 242102520040), Hainan’s Provincial Natural Science Foundation of China under Grant 625QN303, Hainan’s Provincial Education and Teaching Reform Project of Colleges and Universities under Grant Hnjg2025-58, and the Program for Scientific Research Start-up Funds of Hainan Normal University under Grant HSZK-KYQD-202431.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: Wen Zhao was employed by the company Measurement Center of Guangdong Power Grid Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Salmivalli, C.; Laninga-Wijnen, L.; Malamut, S.T.; Garandeau, C.F. Bullying prevention in adolescence: Solutions and new challenges from the past decade. *J. Res. Adolesc.* **2021**, *31*, 1023–1046. [\[CrossRef\]](#)
2. Thornberg, R.; Delby, H. How do secondary school students explain bullying? *Educ. Res.* **2019**, *61*, 142–160. [\[CrossRef\]](#)
3. Maunder, R.E.; Crafter, S. School bullying from a sociocultural perspective. *Aggress. Violent Behav.* **2018**, *38*, 13–20. [\[CrossRef\]](#)
4. Gaffney, H.; Ttofi, M.M.; Farrington, D.P. What works in anti-bullying programs? Analysis of effective intervention components. *J. Sch. Psychol.* **2021**, *85*, 37–56. [\[CrossRef\]](#)
5. Zych, I.; Viejo, C.; Vila, E.; Farrington, D.P. School bullying and dating violence in adolescents: A systematic review and meta-analysis. *Trauma Violence Abus.* **2021**, *22*, 397–412. [\[CrossRef\]](#) [\[PubMed\]](#)
6. López, D.P.; Llor-Esteban, B.; Ruiz-Hernández, J.A.; Luna-Maldonado, A.; Puente-López, E. Attitudes towards school violence: A qualitative study with Spanish children. *J. Interpers. Violence* **2022**, *37*, NP10782–NP10809. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Suski, E.F. Beyond the schoolhouse gates: The unprecedented expansion of school surveillance authority under cyberbullying laws. *Case West. Reserve Law Rev.* **2014**, *65*, 63.
8. Ptaszynski, M.; Dybala, P.; Matsuba, T.; Masui, F.; Rzepka, R.; Araki, K. Machine learning and affect analysis against cyber-bullying. In Proceedings of the 36th AISB, Leicester, UK, 29 March–1 April 2010; pp. 7–16.
9. Raisi, E.; Huang, B. Cyberbullying detection with weakly supervised machine learning. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, NSW, Australia, 31 July–3 August 2017; pp. 409–416.
10. Zaib, M.H.; Bashir, F.; Qureshi, K.N.; Kausar, S.; Rizwan, M.; Jeon, G. Deep learning based cyber bullying early detection using distributed denial of service flow. *Multimed. Syst.* **2022**, *28*, 1905–1924. [\[CrossRef\]](#)
11. Iwendi, C.; Srivastava, G.; Khan, S.; Maddikunta, P.K.R. Cyberbullying detection solutions based on deep learning architectures. *Multimed. Syst.* **2023**, *29*, 1839–1852. [\[CrossRef\]](#)

12. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
13. Feichtenhofer, C. X3d: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 203–213.
14. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7083–7093.
15. Wang, L.; Tong, Z.; Ji, B.; Wu, G. Tdn: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1895–1904.
16. Cheng, M.; Cai, K.; Li, M. RWF-2000: An Open Large Scale Video Database for Violence Detection. *arXiv* **2019**, arXiv:1911.05913.
17. Singla, S.; Lal, R.; Sharma, K.; Solanki, A.; Kumar, J. Machine learning techniques to detect cyber-bullying. In Proceedings of the 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 3–5 August 2023; pp. 639–643.
18. Dedeepya, P.; Sowmya, P.; Saketh, T.D.; Sruthi, P.; Abhijit, P.; Praveen, S.P. Detecting cyber bullying on twitter using support vector machine. In Proceedings of the 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2–4 February 2023; pp. 817–822.
19. Omarov, B.; Narynov, S.; Zhumanov, Z.; Gumar, A.; Khassanova, M. A Skeleton-based Approach for Campus Violence Detection. *Comput. Mater. Contin.* **2022**, *72*, 315–331. [\[CrossRef\]](#)
20. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXX 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 322–339.
21. Dang, L.M.; Min, K.; Wang, H.; Piran, M.J.; Lee, C.H.; Moon, H. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognit.* **2020**, *108*, 107561. [\[CrossRef\]](#)
22. Ye, L.; Wang, L.; Ferdinando, H.; Seppänen, T.; Alasaarela, E. A video-based DT-SVM school violence detecting algorithm. *Sensors* **2020**, *20*, 2018. [\[CrossRef\]](#)
23. Ye, L.; Shi, J.; Ferdinando, H.; Seppänen, T.; Alasaarela, E. School violence detection based on multi-sensor fusion and improved Relief-F algorithms. In Proceedings of the Artificial Intelligence for Communications and Networks: First EAI International Conference, AICON 2019, Harbin, China, 25–26 May 2019; Proceedings, Part II 1; Springer: Berlin/Heidelberg, Germany, 2019; pp. 261–269.
24. Accattoli, S.; Sernani, P.; Falcionelli, N.; Mekuria, D.N.; Dragoni, A.F. Violence detection in videos by combining 3D convolutional neural networks and support vector machines. *Appl. Artif. Intell.* **2020**, *34*, 329–344. [\[CrossRef\]](#)
25. Ullah, F.U.M.; Muhammad, K.; Haq, I.U.; Khan, N.; Heidari, A.A.; Baik, S.W.; de Albuquerque, V.H.C. AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. *IEEE Trans. Ind. Inform.* **2021**, *18*, 5359–5370. [\[CrossRef\]](#)
26. Vijeikis, R.; Raudonis, V.; Dervinis, G. Efficient violence detection in surveillance. *Sensors* **2022**, *22*, 2216. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Gao, Y.; Liu, H.; Sun, X.; Wang, C.; Liu, Y. Violence detection using oriented violent flows. *Image Vis. Comput.* **2016**, *48*, 37–41. [\[CrossRef\]](#)
28. Ullah, W.; Hussain, T.; Ullah, F.U.M.; Lee, M.Y.; Baik, S.W. TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106173. [\[CrossRef\]](#)
29. Serrano, I.; Deniz, O.; Espinosa-Aranda, J.L.; Bueno, G. Fight recognition in video using hough forests and 2D convolutional neural network. *IEEE Trans. Image Process.* **2018**, *27*, 4787–4797. [\[CrossRef\]](#)
30. Ullah, F.U.M.; Ullah, A.; Muhammad, K.; Haq, I.U.; Baik, S.W. Violence detection using spatiotemporal features with 3D convolutional neural network. *Sensors* **2019**, *19*, 2472. [\[CrossRef\]](#)
31. Zhang, Y.; Xiao, Y.; Zhang, Y.; Zhang, T. Video saliency prediction via single feature enhancement and temporal recurrence. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111840. [\[CrossRef\]](#)
32. Haq, M.A. CNN based automated weed detection system using UAV imagery. *Comput. Syst. Sci. Eng.* **2022**, *42*, 837–849. [\[CrossRef\]](#)
33. Yang, T.; Zhu, Y.; Xie, Y.; Zhang, A.; Chen, C.; Li, M. Aim: Adapting image models for efficient video action recognition. *arXiv* **2023**, arXiv:2302.03024. [\[CrossRef\]](#)
34. Hao, Y.; Wang, S.; Cao, P.; Gao, X.; Xu, T.; Wu, J.; He, X. Attention in attention: Modeling context correlation for efficient video classification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7120–7132. [\[CrossRef\]](#)
35. Su, H.; Li, K.; Feng, J.; Wang, D.; Gan, W.; Wu, W.; Qiao, Y. TSI: Temporal saliency integration for video action recognition. *arXiv* **2021**, arXiv:2106.01088. [\[CrossRef\]](#)
36. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3202–3211.
37. Neimark, D.; Bar, O.; Zohar, M.; Asselmann, D. Video transformer network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3163–3172.
38. Huang, Y.; Guo, Y.; Gao, C. Efficient parallel inflated 3D convolution architecture for action recognition. *IEEE Access* **2020**, *8*, 45753–45765. [\[CrossRef\]](#)

39. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541.
40. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 305–321.
41. Tran, D.; Wang, H.; Torresani, L.; Feiszli, M. Video classification with channel-separated convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5552–5561.
42. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2740–2755. [[CrossRef](#)]
43. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
44. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
45. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
46. Lin, T. Focal Loss for Dense Object Detection. *arXiv* **2017**, arXiv:1708.02002.
47. Tan, Y.; Hao, Y.; Zhang, H.; Wang, S.; He, X. Hierarchical Hourglass Convolutional Network for Efficient Video Classification. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 5880–5891.
48. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.