electronics

MDPI

Article

Classroom Behavior Recognition: A Spatiotemporal Frequency-Domain Approach with Imbalanced Data Handling

Linrunjia Liu [1,2], Yinfei Ma [2], Shuai Wu [2], Bingwen Jia [2] and Qiguang Miao [2,*]

[1] College of Cyber Security, Tarim University, Aral 843300, China; liulinrunjia@xidian.edu.cn
[2] School of Computer Science and Technology, Xidian University, Xi'an 710071, China; 21031211449@stu.xidian.edu.cn (Y.M.); wushuai@xidian.edu.cn (S.W.); bwjia_1@stu.xidian.edu.cn (B.J.)
* Correspondence: qgmiao@xidian.edu.cn

Abstract

Existing video action recognition methods face two challenges when applied to classroom surveillance: first, they struggle to balance local spatial and global temporal information; second, they are affected by imbalanced category distribution where routine behaviors dominate while educationally significant rare behaviors are often missed. To address these issues, this paper proposes two components. A Regional Global-Time Featurizer (RGTF) extracts compact local spatial and global temporal features via frequency-domain transformation with low computational overhead. An Exponential Focal Loss (EFL) adaptively reweights hard samples to mitigate the impact of imbalanced data. Experiments on XDCR, AVA2.2, Kinetics-400, UCF101, and Something-Something V2 show that RGTF improves baseline models by up to 1.8% and EFL outperforms standard losses by up to 0.92%. An optional LLM-assisted module is additionally provided as an application example to illustrate one possible way of using the recognition outputs for generating qualitative classroom feedback. This module is independent of the core recognition pipeline.

Keywords: action recognition; classroom behavior analysis; spatiotemporal feature extraction; imbalanced data

1. Introduction

With the rapid development of computer vision and deep learning, video-based human action recognition has been widely applied in intelligent surveillance, human–computer interaction, and intelligent education. In classroom scenarios, automatic recognition of teacher and student behavior can help analyze learning engagement, optimize teaching strategies, and improve teaching quality. However, real-world classroom environments bring three severe challenges that existing methods struggle to address: temporal redundancy, multi-label classification, and imbalanced category distribution.

In classroom scenarios, two major challenges arise. The first lies in balancing local spatial and global temporal information, as many classroom actions are spatially localized but temporally sustained. The second challenge is the imbalanced distribution of behavior categories, where frequent behaviors dominate while rare but educationally significant behaviors are underrepresented.

A comprehensive survey by Stergiou and Poppe [1] systematically reviewed advances in video action understanding, highlighting that current systems still face challenges in fine-grained temporal modeling and long-tailed distribution handling. Meanwhile, most existing frequency-domain methods apply transformation on spatial dimensions or for

Check for updates

Academic Editor: George A. Tsihrintzis

Received: 14 April 2026
Revised: 3 May 2026
Accepted: 5 May 2026
Published: 9 May 2026

Copyright: © 2026 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.

attention acceleration, ignoring the intrinsic characteristic of human actions: spatially local but temporally global. In addition, the imbalance of categories in real classroom data further degrades recognition performance. Moreover, most existing methods stop at behavior recognition and fail to provide high-level semantic interpretation for teaching analysis, limiting their practical value in educational scenarios.

Privacy-preserving video action recognition has also attracted increasing attention. Dinh et al. [2] designed FlowerAction, a federated learning framework that trains models without sharing raw video data. Zhu et al. [3] tackled client heterogeneity and motion preservation in federated skeletal action recognition. Lin et al. [4] proposed a chaos-based image encryption scheme for visual privacy in consumer electronics, which is directly relevant to classroom monitoring. In line with these efforts, our system uses only aggregated behavior statistics rather than raw video feeds, meeting privacy requirements in smart classroom environments.

Unlike existing frequency-domain methods that focus on spatial transformations, our RGTF applies DCT exclusively on the temporal axis. Wave-ViT [5] uses spatial DWT and leaves temporal modeling to attention. AFNO [6] performs global Fourier mixing in space, not time. F2D-SIFPNet [7] exploits DCT coefficients from compressed frames but still operates spatially. By contrast, our RGTF captures global temporal frequency while preserving spatial locality, which is essential for actions that are spatially limited but temporally continuous. Specifically, classroom actions such as writing, looking down at a phone, or raising a hand are spatially localized (involving only a small region) yet temporally sustained (requiring multiple frames to recognize). Existing spatial-domain frequency transforms cannot capture such temporal dependencies, while 3D convolutions have limited receptive fields. Applying DCT directly on the temporal axis provides a lightweight solution to this problem.

To address these practical challenges, this paper explores two techniques within existing action recognition frameworks:

- We introduce RGTF, a lightweight module that applies DCT on the temporal axis to capture global temporal information while preserving spatial locality. Experiments show that adding RGTF to existing backbones (SlowFast, VideoMAE) improves performance on classroom data by up to 1.8%.
- We adopt EFL to mitigate the impact of imbalanced category distribution. Experiments on our classroom data show that EFL outperforms CE and FL, with a maximum improvement of 0.92%.

We evaluate these techniques on a classroom dataset (XDCR) collected from real surveillance cameras, as well as public benchmarks. The results show consistent improvements. To demonstrate a potential use case of the recognition results, we additionally provide an optional LLM-assisted module that converts behavior statistics into natural-language classroom descriptions. This module serves as an application example and is not the main contribution of this work.

2. Related Works

2.1. Video Action Recognition and Frequency-Domain Methods

Video action recognition has evolved from handcrafted feature-based approaches to modern deep learning architectures that explicitly model spatiotemporal structures in videos [1,8]. Early two-stream architectures disentangle static appearance and dynamic motion cues, while 3D convolutional networks directly model videos as four-dimensional spatiotemporal tensors. Among representative architectures, the SlowFast network employs

a dual-branch design, consisting of a slow pathway for capturing semantic content at low frame rates and a fast pathway for modeling fine-grained motion at high frame rates [9].

Recent progress in long-form video understanding has focused on efficient and memory-efficient modeling. VideoLLaMB introduces recurrent memory bridges to support long streaming video understanding with linear GPU memory consumption [10]. For online action detection, the Context-Enhanced Memory-Refined Transformer (CMeRT) aggregates historical observations and anticipates future frames in a streaming setting [11]. HERMES further improves long-video efficiency by using episodic compression and semantic retrieval to maintain temporal coherence while reducing inference costs [12]. On the task of action detection, AdaTAD++ enhances scalability via transformer-based spatiotemporal adaptation [13].

Frequency-domain modeling has emerged as an efficient alternative to dense spatial–temporal computation. Adaptive Fourier Neural Operators (AFNOs) learn global interactions via Fourier transforms, serving as efficient token mixers in vision transformers [6]. Wave-ViT integrates wavelet decomposition into transformer architectures to improve multi-scale visual representation [5]. For compressed video understanding, F2D-SIFPNet directly exploits DCT coefficients to accelerate feature extraction without full frame decoding [7]. In infrared small target detection, triple-domain feature learning with frequency-aware memory enhancement has also been proposed to suppress background clutter [14].

Despite these advances, a key limitation remains. 3D-CNNs excel at local spatiotemporal modeling but lack long-range temporal coverage, whereas vision transformers capture global dependencies at a substantially higher computational cost. Existing frequency-based methods mostly operate on spatial dimensions or approximate global attention, rather than explicitly modeling temporal dynamics in the frequency-domain. Thus, how to efficiently extract representations that are spatially local yet temporally global remains an open challenge for action recognition.

2.2. Long-Tailed Learning

Real-world visual datasets often exhibit imbalanced class distributions, where a small number of head classes occupy most samples while tail classes are severely underrepresented [1]. This imbalance can degrade classification performance on rare categories. To address this issue, various methods have been proposed, including probabilistic contrastive learning for learning more balanced representations [15].

In classroom scenarios, this issue is particularly prominent. As shown in our XDCR dataset, routine behaviors such as sitting and listening appear frequently, while atypical behaviors such as using mobile phones or sleeping are rare. This imbalance causes models to bias toward frequent classes and underperform on rare but educationally significant behaviors.

To mitigate this issue, we experiment with loss functions designed for imbalanced data. Focal Loss (FL) [16] reweights standard classification loss to emphasize hard and underrepresented examples. However, we find that FL's polynomial weighting scheme suffers from vanishing gradients on extremely hard samples in our classroom data. We therefore adopt Exponential Focal Loss (EFL), which replaces the polynomial weight with an exponential function to maintain gradient magnitude on hard samples. EFL is not claimed as a general solution to imbalanced recognition, but as an effective choice for our specific classroom data.

2.3. Classroom Behavior Analysis

With the development of smart classrooms, deep learning-based classroom behavior analysis has gained increasing attention. A systematic review of 80 studies shows that vision-based classroom behavior recognition mainly focuses on four categories: physical actions, learning engagement, attention status, and emotions [17]. Among these, physical action recognition and learning engagement assessment are the most widely studied, with deep learning models as the dominant paradigm [17]. Video analytics has also been applied to surveillance-oriented tasks such as examination malpractice detection, using face detection, tracking, and gesture recognition [18].

However, real-world classroom systems still face substantial challenges: large-scale variations caused by different seating positions, frequent occlusions, and inherent long-tailed behavior distributions [8]. Furthermore, most existing methods only perform low-level behavior classification and do not provide interpretable, pedagogically meaningful feedback. This gap between visual recognition and high-level educational interpretation restricts the practical value of automated classroom analysis [1].

3. Proposed Method

3.1. Overall Architecture

The proposed framework forms a complete pipeline from raw video input to behavior recognition outputs. The RGTF and EFL modules improve recognition accuracy and class balance. An optional LLM-assisted module is additionally provided as an application example. All modules are designed to be lightweight and deployable in real-world classroom environments. The overall pipeline is illustrated in Figure 1.

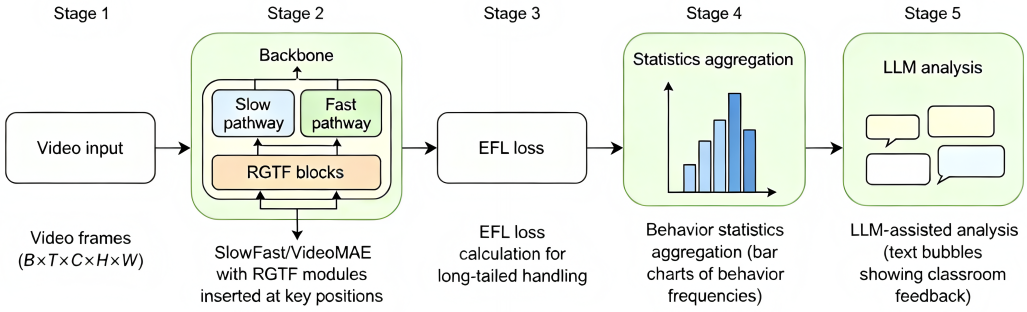


Figure 1. Overall framework architecture. The pipeline consists of: (1) video input, (2) backbone (SlowFast/VideoMAE) with RGTF modules inserted at key positions, (3) EFL loss for long-tailed handling, (4) behavior statistics aggregation, and (5) LLM-assisted analysis for classroom feedback.

3.2. RGTF: Design and Formulation

The Regional Global-Time Featurizer (RGTF) is a plug-and-play deep learning module designed to extract global temporal and local spatial information from input features, thereby obtaining the minimal but complete information to analyze behavioral categories for each action. This characteristic of the RGTF simultaneously compensates for the respective shortcomings of the CNN and ViT architectures in action recognition tasks.

DCT and Its Inverse

DCT converts time-domain signals—such as image signals expressed in RGB—to frequency-domain signals. If a time-domain signal is decomposed into a linear combination of cosine or sine functions with different frequencies, the corresponding frequency-domain signal can be expressed as the amplitude coefficient of these functions. With DCT, widespread features in an image are converted into coefficients of low-frequency functions, whereas detailed features are transformed into coefficients of high-frequency functions.

Thus, the frequency-domain signal converted from DCT can encode the original image with fewer coefficients but negligible information loss, thereby effectively counteracting noise perturbation.

Let the video feature tensor be of shape $B \times T \times C \times H \times W$, where B is the batch size, C the number of channels, T the number of frames, H the frame height, and W the frame width. For each spatial location and each channel, a temporal signal $x(n)$ of length $N = T$ is extracted from the time axis, where $n = 0, 1, \ldots, N-1$ indexes the frame position. The Type-2 DCT transforms this time-domain signal into the frequency domain:

$$y(k) = \sqrt{\frac{2-\delta_{k0}}{N}} \sum_{n=0}^{N-1} x(n) \cos\left(\frac{\pi k(2n+1)}{2N}\right), \tag{1}$$

where $y(k)$ is the frequency-domain coefficient at frequency index k, δ_{k0} is the Kronecker delta ($\delta_{k0} = 1$ if $k = 0$,and 0 otherwise), and $k = 0, 1, \ldots, N-1$.

The inverse DCT (IDCT) reconstructs the time-domain signal:

$$x(n) = \sum_{k=0}^{N-1} \sqrt{\frac{2-\delta_{k0}}{N}} y(k) \cos\left(\frac{\pi k(2n+1)}{2N}\right) \tag{2}$$

The computational flow of the RGTF module is illustrated in Figure 2. Video data is abstracted as a tensor of shape $B \times T \times C \times H \times W$. The RGTF acquires essential action information by sequentially extracting global and local features. It first uses a 3×3 convolution on H and W axes to capture local spatial information, then performs DCT on the T-axis to calculate frequency coefficients for global temporal information. Another 3×3 spatial convolution on the H and W axes captures global temporal information in adjacent spatial regions; convolution on frequency coefficients minimize redundant data impact, as these coefficients cover all T-axis information without omitting temporal details.

The extracted information is fed into the hard threshold function (Equation (3)) to filter insignificant coefficients, enhancing model generalizability. After IDCT, new spatiotemporal information is concatenated with the original and sequentially convolved along spatial and temporal axes to strengthen locality modeling, without global information loss due to prior DCT feature extraction.

$$f_\sigma(x) = \begin{cases} x, & |x| > \sigma \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

3.3. EFL Function

Similar to FL, EFL varies loss values according to sample confidence scores (Equation (4))

$$p = \begin{cases} \hat{y}, & y = 1 \\ 1 - \hat{y}, & y = 0 \end{cases} \tag{4}$$

where y and $\hat{y}$ are true and predicted labels. FL and EFL are expressed as Equations (5) and (6):

$$\mathrm{FL}(p) = -\alpha \sum_j (1 - p_j)^\gamma \log(p_j) \tag{5}$$

$$\mathrm{EFL}_\beta(p(y, \hat{y})) = -\alpha \sum_j \left(\frac{e^{-\beta p} - e^{-\beta}}{1 - e^{-\beta}}\right)^\gamma \log(p_j) \tag{6}$$

where α, β, and γ are user-defined constants. Unlike FL's polynomial function, EFL's exponential function avoids derivative vanishing, strengthening the model's ability to adjust sample weights.

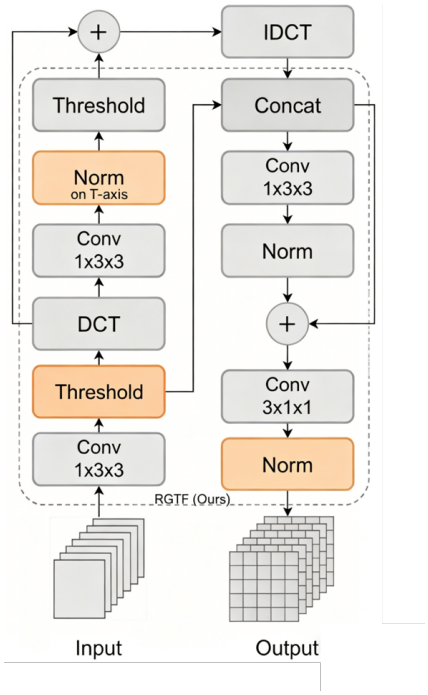


Figure 2. The architecture of the proposed RGTF module. Standard operations (Conv, Norm, DCT, IDCT, Concat, and residual addition) are shown in gray. Our contributions, including the application of DCT on the temporal axis and the hard threshold function, are highlighted in orange. The module sequentially extracts local spatial features via 3 × 3 convolution, captures global temporal information via temporal DCT, filters insignificant coefficients via hard thresholding, and fuses features after IDCT through concatenation and residual connections.

As shown in Table 1, EFL assigns higher weights to difficult samples (p = 0.4) than to simple samples (p = 0.8), with a loss ratio that is more favorable for long-tailed learning compared to BCE loss and FL. This verifies EFL's effectiveness for imbalanced data.

Table 1. Comparison of loss values.

	BCE	**FL**	**EFL ($\beta = 1$)**	**EFL ($\beta = 2$)**
$\hat{y}_1/\hat{y}_2$	4.11	12.22	15.10	19.59

Note: $\hat{y}_1$ = loss value of difficult samples (p = 0.4); $\hat{y}_2$ = loss value of simple samples (p = 0.8).

3.4. Optional Application Example: LLM-Assisted Feedback Generation

In classroom monitoring, the end-users (teachers) may benefit from natural-language summaries rather than raw behavior statistics. To illustrate how the recognition outputs of our framework could be potentially used in practice, we provide an optional LLM-assisted module as an application example.

Given aggregated behavior statistics (e.g., frequency counts of listening, mobile phone use, and peer chatting), the LLM is prompted to generate structured descriptions across three dimensions: student learning status, classroom condition, and interpretive notes. The module is implemented as an asynchronous process that does not affect the real-time recognition pipeline.

We emphasize that this module is not a core component of our proposed framework. The main technical contributions remain the RGTF and EFL modules. This LLM-assisted

example is provided solely to demonstrate a possible direction for extending behavior recognition results toward pedagogical reporting.

4. Experimental Setup

4.1. Datasets

In this study, we selected the AVA2.2 dataset [19] and the XDCR dataset (our own dataset) for classroom scenarios to evaluate the performance of the RGTF in generic and classroom-based action recognition under given spatiotemporal regions. The same datasets were also used to compare the EFL, CE, and FL functions as they are characterized by long-tailed distribution. In addition, the Kinetics-400 (K400) [20], Something-Something V2 (Sthv2) [21], and UCF101 [22] were selected as three additional datasets with balanced class distribution to analyze the performance of the RGTF module on generic video understanding tasks.

The XDCR dataset, which encompasses classroom surveillance footage, was tagged with behavioral categories only in specific spatiotemporal regions and with multiple labels. Specifically, XDCR labels 15 behaviors in 120 classroom videos of 15 min each, including the following categories:

- Basic behaviors: Sit, stand, walk.
- Restricted single-person behaviors: Read, write, lie/sleep, text on/looking at a cellphone, work on a computer, answer phone.
- Unrestricted single-person behaviors: Carry/hold (an object), point to (an object), touch (an object).
- Interactive behaviors: Listen to (a person), talk to (e.g., self, a person, a group), watch (a person).

Fine-grained distinctions such as “reading” vs. “writing” carry pedagogical significance that would be lost under coarse-grained merging (e.g., grouping both into a single “desk work” category). Differentiating “listening” from “looking at a cellphone” is also essential for engagement assessment. For these reasons, the original 15 atomic categories are preserved in this study. The resulting class imbalance is addressed by the proposed EFL loss (Section 3.3).

These categorical behaviors belong to a subset of category labels in the AVA2.2 dataset that are prevalent in classroom scenarios. They are sufficiently atomistic to provide meaningful references for more advanced and complex behavioral analysis. For example, in the case of the category of “basic behaviors”, these behaviors are highly correlated with each individual’s position in the classroom; e.g., “sit” generally implies that the person is in the position of listening to the teacher, “stand” generally implies that the person is speaking at the podium or being asked to speak by the teacher, and “walk” generally implies that the teacher is walking around the classroom or talking to the students.

For the category of “restricted single-person behaviors”, these behaviors usually show the students’ and teacher’s listening states, such as “reading” and “writing,” and the frequent occurrence of these behaviors may imply effective teaching and feedback between the teacher and the students; in contrast, “sleeping” and “writing” may imply the inefficiency or even absence of teaching activities. The category of “unrestricted single-person behaviors” addresses human–object interactions and does not qualify the types or properties of objects. These behaviors are often associated with multifaceted instructional activities, such as a teacher demonstrating the procedure of an experiment on the podium or a student being asked to complete a handmade task. Interactive behaviors reflect both teacher–student and student–student interactions, thereby representing the educator’s overall teaching style.

Figure 3 shows a typical long-tailed distribution: when the categories are sorted according to the decreasing number of samples from left to right, the leftmost category has many more samples than the rightmost one. More specifically, the three categories "sit," "watch (a person)," and "stand" can be considered the head categories, with the other categories can be considered the tail categories, because Figure 4 shows that the three head categories account for more than half of samples in the 15 total categories.

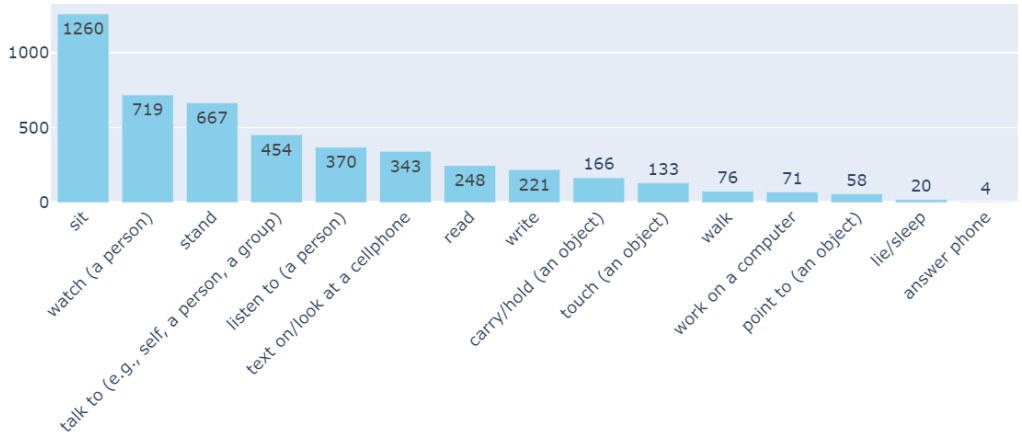


Figure 3. Long-tailed category distribution in the XDCR dataset.

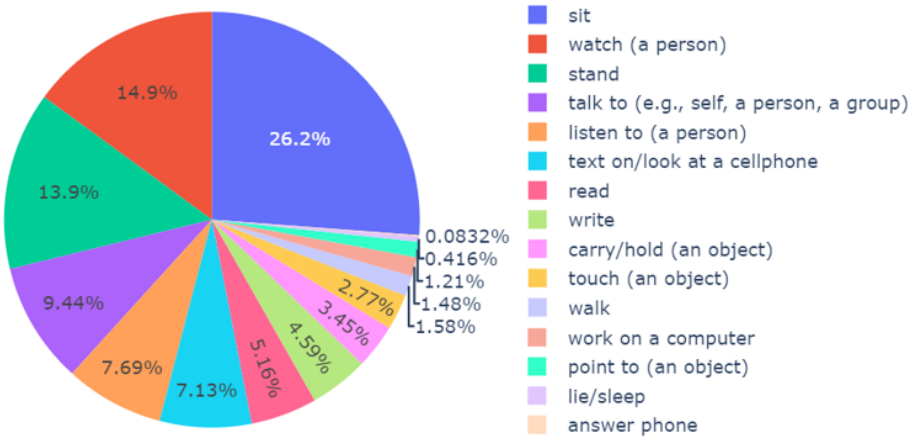


Figure 4. Sample proportion of the 15 categories. Three head categories account for over half of all samples. (All proportions are normalized to 100% in theory. The slight deviation of 99.9992% originates from decimal rounding of statistical data. The item "answer phone" occupies only 0.0832% with an ultra-small sector that cannot display a separate color, and its value is marked outside the pie chart).

AVA2.2. The AVA (Atomic Visual Actions) dataset contains 437 videos that are 15 min each, with frame-level annotations for 80 atomic action categories. Each action is annotated within specific spatiotemporal regions, and the labeling is multi-label (an action can belong to multiple categories simultaneously). We use AVA2.2 to evaluate multi-label action recognition performance in generic scenarios.

Kinetics-400. The Kinetics-400 dataset consists of approximately 306,000 video clips covering 400 human action classes. Each clip lasts around 10 s and is sourced from YouTube videos. The dataset is widely used for pre-training and benchmarking video action recognition models. It has a relatively balanced class distribution by design.

Something-Something V2. The Something-Something V2 dataset contains 220,847 video clips of humans performing pre-defined basic actions with everyday objects, covering

174 action classes. Actions in this dataset are defined by temporal transitions (e.g., "pulling something from left to right"), making it particularly suitable for evaluating temporal modeling capability. The dataset is split into 168,913 training videos, 24,777 validation videos, and 27,157 test videos.

UCF101. The UCF101 dataset contains 13,320 video clips distributed across 101 action categories, with a resolution of 320 × 240 pixels. The videos are collected from YouTube and feature mostly coarse sports activities with strong camera motion.

The detailed statistics and characteristics of all datasets used in this study are summarized in Table 2.

Table 2. Summary of datasets used in this study.

Dataset	Year	Samples	Duration	Classes	Distribution
XDCR (Ours)	2023	120 videos	15 min	15	Long-tailed
AVA2.2	2018	437 videos	15 min	80	Multi-label
Kinetics-400	2017	∼306,000 clips	10 s	400	Balanced
Something-Something V2	2017	220,847 clips	∼4 s	174	Balanced
UCF101	2012	13,320 clips	1–30 s	101	Balanced

4.2. Evaluation Metrics

For multi-label action recognition tasks (AVA2.2), the mean Average Precision (mAP) is used as the main evaluation metric, which can comprehensively reflect the model's performance on all categories, especially tail categories. For general video classification tasks (Kinetics400, Something-Something V2, UCF101), Top-1 and Top-5 accuracy are used as evaluation metrics, which are the most commonly used metrics in video classification tasks.

$$\text{mAP} = \frac{1}{C}\sum_{i=1}^{C}\text{AP}_i$$

where AP_i is the area under the precision–recall curve for class i, and C is the total number of classes.

For single-label video classification (Kinetics-400, Something-Something V2, UCF101), we adopt Top-1 and Top-5 accuracy:

$$\text{Top-}k\text{ Acc} = \frac{1}{N}\sum_{i=1}^{N}1\{y_i \in \text{top-}k(\hat{y}_i)\}$$

where N is the total number of samples, y_i is the ground-truth label of the i-th sample, top-$k(\hat{y}_i)$ denotes the set of k class indices with the highest predicted scores, and $\mathbf{1}\{\cdot\}$ is the indicator function (1 if the true class is among the top-k predictions, and 0 otherwise). We report Top-1 accuracy ($k = 1$) for exact match and Top-5 accuracy ($k = 5$) to tolerate ranking errors in fine-grained classification.

4.3. Implementation Details

RGTF is inserted after each bottleneck in SlowFast and after each multi-head self-attention (MHSA) in VideoMAE. The threshold σ is set to 0.01, which filters out frequency coefficients with negligible magnitudes (considered as noise or redundant information) while preserving informative signals. This setting stabilizes the hard thresholding operation and generalizes well across different video data. A systematic sensitivity analysis of the threshold σ in the hard threshold function is not provided in this study. The value σ = 0.01 is empirically set.

Computational cost. We report the computational overhead of RGTF in terms of FLOPs (Floating-Point Operations Per Second) for a single forward pass on one sample. As shown in Table 3, adding RGTF introduces a minor computational burden.

Table 3. Computational cost (FLOPs in G) of baseline models and their RGTF-enhanced versions.

Model	FLOPs (G)
SlowFast	1.40
SlowFast + RGTF	1.68
VideoMAE	29.3
VideoMAE + RGTF	29.5

5. Results

This section presents the experimental results that validate the two proposed contributions (RGTF and EFL). Section 5.1 evaluates RGTF and EFL on the classroom dataset XDCR. Section 5.2 evaluates the generalization of RGTF on public benchmarks AVA2.2, Kinetics-400, Sthv2, and UCF101. Section 5.3 presents ablation studies. Section 5.4 compares the proposed method with state-of-the-art approaches.

5.1. Action Recognition in Classroom Scenarios

Table 4 reports mAP values on the XDCR dataset. Based on the data, three points are noted.

Table 4. mAP (%) values obtained from the XDCR dataset for each model based on different loss functions.

Model	SlowFast	SlowFast + RGTF	VideoMAE	VideoMAE + RGTF
CE	**51.19**	53.01	**53.05**	53.82
FL	51.45	53.01	53.77	54.36
EFL	51.46	**53.04**	53.79	**54.74**

Note: bold values highlight the baseline and our proposed results for clear comparison.

Adding RGTF improves performance. For SlowFast, adding RGTF increases mAP from 51.19% to 53.01% under CE, from 51.45% to 53.01% under FL, and from 51.46% to 53.04% under EFL. For VideoMAE, adding RGTF increases mAP from 53.05% to 53.82% under CE, from 53.77% to 54.36% under FL, and from 53.79% to 54.74% under EFL. In all six comparisons, the version with RGTF achieves higher mAP.

EFL achieves higher mAP than CE and FL. In all four model configurations (SlowFast, SlowFast + RGTF, VideoMAE, VideoMAE + RGTF), EFL yields the highest mAP among the three loss functions. The difference between EFL and CE ranges from 0.21% to 0.92%, with the largest difference (0.92%) on VideoMAE + RGTF.

The best result is achieved by combining RGTF and EFL. The highest mAP in Table 4 is 54.74%, obtained by VideoMAE + RGTF + EFL.

5.2. General-Purpose Action Recognition

We select the AVA dataset to evaluate RGTF performance on generic action recognition tasks, as these tasks require multi-label classification under bounded spatiotemporal regions, similar to XDCR. Three video classification datasets (K400, Sthv2, and UCF101) are also used to examine RGTF performance on complete video clips. The results are shown in Tables 5 and 6.

Table 5. Performance of each model on AVA dataset.

Model	Backbone	Frame Rate	mAP (AVA)
SlowFast	ResNet50	8 × 8	25.9
SlowFast+RGTF	ResNet50	8 × 8	26.7
VideoMAE	ViT-B	16 × 4	33.6
VideoMAE + RGTF	ViT-B	16 × 4	33.9

Table 6. Performance of each model on video classification datasets.

Model	Backbone	Acc (K400)	Acc (STHv2)	Acc (UCF101)
SlowFast	ResNet50	76.8	59.4	92.5
SlowFast + RGTF	ResNet50	76.9	60.9	92.8
VideoMAE	ViT-B	81.3	61.3	97.4
VideoMAE + RGTF	ViT-B	81.5	61.4	98.2

On AVA, adding RGTF improves SlowFast from 25.9% to 26.7% and VideoMAE from 33.6% to 33.9%. The gains are smaller than those on XDCR, as AVA actions are annotated at the frame level with short temporal windows, reducing the benefit of global temporal modeling.

On the video classification datasets, the gains vary. The largest improvement is on Sthv2 (SlowFast + RGTF: +1.5%), where actions are defined by temporal transitions (e.g., "pulling something from left to right") and cannot be recognized from single frames. On UCF101, VideoMAE + RGTF improves by 0.8%. On Kinetics-400, the improvements are smaller (+0.1% for SlowFast, +0.2% for VideoMAE), as many actions in this dataset can be recognized from static appearance.

5.3. Ablation Study

To examine the role of each component in RGTF, we compare three variants: (1) removing DCT and IDCT, (2) replacing DCT/IDCT with fully connected layers, and (3) removing the hard threshold function. The results are shown in Table 7.

Table 7. Results of the ablation experiments on RGTF.

Model	mAP (AVA)	mAP (XDCR)
SlowFast	25.9	51.2
SlowFast + RGTF	26.7	53.0
SlowFast + RGTF(1)	26.4	51.7
SlowFast + RGTF(2)	26.4	51.3
SlowFast + RGTF(3)	26.4	51.4
VideoMAE	33.6	53.1
VideoMAE + RGTF	33.9	53.8
VideoMAE + RGTF(1)	33.7	53.4
VideoMAE + RGTF(2)	33.8	53.5
VideoMAE + RGTF(3)	33.8	52.8

Removing DCT and IDCT (variant 1) reduces mAP on both datasets. On XDCR, SlowFast+RGTF(1) drops from 53.0% to 51.7%, and VideoMAE + RGTF(1) drops from 53.8%

to 53.4%. This indicates that the frequency-domain operation contributes to the performance gain.

Replacing DCT/IDCT with fully connected layers (variant 2) also reduces mAP. On XDCR, VideoMAE+RGTF(2) drops from 53.8% to 53.5%. The fact that DCT outperforms learnable weights suggests that frequency-domain representation provides more compact information encoding for this task.

Removing the hard threshold function (variant 3) leads to a performance drop, particularly for VideoMAE on XDCR (from 53.8% to 52.8%). This suggests that filtering out small frequency coefficients helps remove noise and improves generalization.

The mAP values on XDCR are higher than those on AVA across all models. This is because XDCR has fewer behavior categories (15 vs. 80 in AVA) and models are first trained on AVA before fine-tuning on XDCR.

5.4. Comparison with State-of-the-Art Methods

To further validate the effectiveness of our proposed method, we compare it with recent state-of-the-art action recognition methods on public benchmarks. Table 8 summarizes the results on the AVA2.2, Kinetics-400, UCF101, and Something-Something V2 datasets.

Table 8. Comparison with state-of-the-art methods on public datasets (results are reported in mAP for AVA2.2 and Top-1 accuracy for others).

Method	AVA	K400	UCF101	STHv2	Params	FLOPs
RAAG (2025) [23]	40.5	–	–	–	——	768
YOWOv3 (2025) [24]	20.93	–	–	–	54.7 M	37.2 G
DarkActionNet (2025) [25]	–	80.8	–	–	30.8 M	50.2 G
CAN (2025) [26]	–	74.9	86.9	63.9	25.3 M	35 G
Graph-based HAR (2024) [27]	–	88.9	89.6	–	–	–
VideoMAE (2022) [28]	33.6	81.3	97.4	61.3	86 M	29.3 G
SlowFast (2019) [9]	25.9	76.8	92.5	59.4	34.5 M	1.4 G
Ours: VideoMAE + RGTF	33.9	81.5	98.2	61.4	86 M	29.5
Ours: SlowFast + RGTF	26.7	76.9	92.8	60.9	34.5 M	1.68 G

Note: '–' indicates that the corresponding metric is not reported in the original paper.

Our method achieves moderate performance and does not outperform the best existing methods on most metrics. However, the advantage of our method is computational efficiency. VideoMAE + RGTF achieves 33.9% mAP on AVA2.2, close to the VideoMAE baseline (33.6%), with only 0.10G MACs added. On UCF101, our method achieves 98.2% Top-1 accuracy. On Kinetics-400, SlowFast + RGTF achieves 76.9% Top-1 accuracy, similar to the baseline, while requiring 1.68G FLOPs.

Compared to recent methods, RAAG achieves higher mAP (40.5%) on AVA2.2 but uses a ViT-L backbone with approximately 307M parameters. YOWOv3 is a lightweight model (54.7M parameters, 37.2G FLOPs) but achieves lower accuracy (20.93%). DarkActionNet achieves 80.8% Top-1 accuracy on Kinetics-400 but requires 50.2G FLOPs. Our method performs at a comparable level to these methods with lower computational cost.

The contribution of our method is not pursuing the highest recognition accuracy, but achieving acceptable performance with low computational cost. This makes it suitable for deployment in resource-constrained classroom scenarios, where multiple video streams need to be processed simultaneously. The consistent performance across multiple public datasets suggests that our method generalizes to different domains.

5.5. Qualitative Classroom Analysis with an Optional LLM Example

To demonstrate the practical applicability of the proposed framework, we apply the trained model to real classroom surveillance videos collected from a typical high school setting. The videos capture both student seating areas and the teacher's podium region, covering a full 90-min class session. The recognition results are visualized at three levels: individual student behavior, teacher behavior, and classroom-level aggregated statistics.

Figure 5 compares recognition results before and after model optimization on a representative classroom frame. The baseline model (Figure 5a) mostly outputs "sit" or incorrectly predicts "stand" for front-row students, failing to capture fine-grained behaviors such as reading, looking at a mobile phone, or listening. This is because the baseline model relies heavily on static posture cues and lacks the ability to model temporally sustained but spatially localized actions. In contrast, the optimized model (Figure 5b) produces more accurate and diverse predictions for the same front-row students, including "watch", "listen", "look at a cellphone", and "read". These improvements are attributed to the RGTF module, which captures global temporal information via DCT on the temporal axis, and the EFL loss, which better handles rare behavior categories. For back-row students, where occlusion and small spatial resolution are severe, both models perform poorly. However, the optimized model yields slightly better results, occasionally recognizing "look at a cellphone" in cases where the baseline outputs only "sit". This suggests that the proposed method provides some robustness to challenging conditions, though occlusion remains an open problem.



(**a**) Results before model optimization



(**b**) Results after model optimization

Figure 5. Visualization of classroom student behavior analysis.

Figure 6 shows teacher behavior analysis on the podium. Detection accuracy is high due to the open space and minimal occlusion. Actions such as "writing on the blackboard", "pointing to the screen", and "interacting with students" are correctly recognized by both the baseline and optimized models. However, the optimized model shows improved temporal consistency, reducing flickering predictions between consecutive frames. This is particularly noticeable during transitions between writing and lecturing, where the baseline sometimes misclassifies the teacher's pose as "standing idle" while the optimized model correctly maintains the "writing" or "interacting" label.

Figure 7 visualizes classroom-level behavior analysis, integrating frame-level recognition results with aggregated behavior statistics over an entire class session. The distribution of detected behaviors is summarized as a bar chart, providing an intuitive overview of dominant classroom activity patterns. In the example shown, listening and looking behaviors account for the majority of student actions, while mobile phone-related behaviors and peer chatting occur at lower frequencies. Such aggregated statistics can help instructors quickly assess overall classroom engagement without reviewing raw video footage.



Figure 6. Visualization of classroom teacher behavior analysis.



Figure 7. Visualization of classroom teacher–student behavior analysis. The Chinese labels in the pie chart are auto-generated by the proposed system, representing attendance of 78%, absence of 17%, and take leave of 5%, respectively.

To illustrate a possible application of the recognition results, we provide an optional LLM-assisted example. Table 9 shows representative outputs generated by the LLM module described in Section 3.4. Given a structured prompt containing behavior statistics (e.g., frequencies of listening, looking, mobile phone use, and chatting), the LLM generates qualitative descriptions across three dimensions: student learning status, classroom condition, and interpretive notes.

These examples are provided for illustration only. Evaluating the pedagogical effectiveness of these descriptions is beyond the scope of this paper, as the main focus remains on the action recognition task. The descriptions are intended as auxiliary references for post-class reflection, not as automated evaluation or decision-making. In Case 1 (mobile phone-related behaviors: 441 times among 40 students), the LLM notes "noticeable variation in attention states". In Case 2 (mobile phone use: 196 times among 52 students; interaction: 70 times), the LLM describes a "relatively orderly classroom environment". These examples demonstrate one potential way of extending behavior recognition outputs toward pedagogical reporting.

Table 9. Qualitative descriptions generated from classroom behavior statistics.

Case	Student Learning Status Description	Classroom Condition Description	Interpretive Notes
Case 1: Behavior statistics: 1. 40 students in the class exhibited mobile phone-related behaviors 441 times during a 90-min session. 2. Chatting or discussion behaviors occurred 26 times. 3. Looking-up behaviors occurred 3466 times. 4. Listening behaviors occurred 4266 times.	1. Attention-related patterns: A relatively high frequency of mobile phone-related behaviors was observed, indicating noticeable variation in attention states. 2. Learning engagement patterns: Students were oriented toward the teacher for a large proportion of the session, suggesting sustained visual engagement with instructional activities. 3. Classroom interaction patterns: A limited number of peer interaction behaviors were detected during the session.	1. Classroom organization characteristics: The observed behavior distribution suggests fluctuations in student attention during the class. 2. Instructional process characteristics: The coexistence of attentive behaviors and mobile phone use reflects heterogeneous engagement responses. 3. Classroom atmosphere: Frequent listening and looking behaviors indicate an overall active learning environment, accompanied by intermittent off-task behaviors.	The descriptive results may serve as supplementary references for post-class reflection. They can assist instructors in reviewing classroom dynamics and considering potential adjustments to instructional organization and interaction design.

Table 9. *Cont.*

Case	Student Learning Status Description	Classroom Condition Description	Interpretive Notes
Case 2: Behavior statistics: 1. 52 students in the class exhibited mobile phone-related behaviors 196 times during a 90-min session. 2. Chatting or discussion behaviors occurred 70 times. 3. Looking-up behaviors occurred 4532 times. 4. Listening behaviors occurred 5345 times.	1. Attention-related patterns: Mobile phone-related behaviors occurred less frequently, suggesting relatively stable attention patterns during the session. 2. Learning engagement patterns: Students were frequently oriented toward the teacher, reflecting sustained engagement with classroom instruction. 3. Classroom interaction patterns: Moderate levels of peer interaction were observed during the class.	1. Classroom organization characteristics: The overall behavior distribution suggests a relatively orderly classroom environment. 2. Instructional process characteristics: Observed engagement-related behaviors are consistent with continuous instructional activities. 3. Classroom atmosphere: The combination of attentive behaviors and moderate interaction reflects an active and balanced learning environment.	These descriptive results may provide supportive information for instructors during post-class review. They can be used as auxiliary references for understanding classroom behavior patterns across different sessions.

6. Discussion and Conclusions

In this work, we focus on practical challenges in recognizing teacher and student behaviors from classroom surveillance videos, particularly in terms of temporal modeling and class imbalance. Based on an analysis of real classroom data, two key issues emerge. One is the need to capture both local spatial and global temporal information. The other is the imbalanced distribution of behavior categories, where routine behaviors dominate. To address these issues, we introduce two techniques into existing action recognition frameworks. RGTF extracts compact spatiotemporal features via frequency-domain transformation, improving the temporal modeling capability of baseline models. EFL adaptively reweights hard samples, mitigating the impact of imbalanced data. Experiments on classroom and public datasets demonstrate consistent improvements over baseline methods.

The proposed methods are not claimed as breakthroughs in action recognition research. Instead, they are presented as effective, lightweight solutions for the specific domain of classroom behavior analysis. In addition, an optional LLM-assisted module is provided as an application example to demonstrate one possible way of generating qualitative classroom feedback from recognition results. This module is independent of the core recognition framework.

Future work includes improving robustness under occlusion and exploring privacy-preserving mechanisms. In addition, while our ST-PPM module incorporates multi-scale pooling, detecting extremely small targets (e.g., back-row students) remains challenging; more advanced multi-scale feature extraction strategies (e.g., feature pyramid networks) could be explored to improve small-target recognition in classroom scenes. Furthermore, as suggested by a reviewer of this paper, the trade-off between class granularity and recognition stability deserves systematic investigation. While we preserve 15 fine-grained categories for pedagogical reasons, functionally integrating similar low-sample classes (e.g., merging “read” and “write” into “desk work”) may reduce class imbalance and improve statistical robustness. Future research may explore hybrid approaches that combine fine-grained recognition with coarse-grained aggregation for different analytical purposes.

In this study, we report FLOPs as an algorithmic efficiency metric. Hardware-level measurements such as FPS, latency, memory usage, and power consumption are not provided. Evaluating the proposed method on edge devices and reporting such metrics is also an important direction for future work.

Author Contributions: Methodolody and Formal Analysis, L.L.; Software and Validation, Y.M.; Visualization, B.J.; Data Curation, S.W.; Resources and Conceptualization, Q.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific Research Program of Shaanxi Provincial Department of Education under Grant No. 24JK0627.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author. The XDCR dataset used in this study is a private dataset collected with informed consent from all participants for research use. The remaining datasets (AVA2.2, Kinetics400, Something-Something V2, UCF101) are publicly available from their original sources.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Stergiou, A.; Poppe, R. About Time: Advances, Challenges, and Outlooks of Action Understanding. *Int. J. Comput. Vis.* **2025**, *133*, 6251–6315. [CrossRef]
2. Dinh, T.Q.K.; Tran, T.-H.; Tran, T.-K.; Le, T.-L. FlowerAction: A Federated Deep Learning Framework for Video-Based Human Action Recognition. *J. Ambient Intell. Hum. Comput.* **2025**, *16*, 459–470. [CrossRef]
3. Zhu, X.; Shu, X.; Tang, J. Client-Unbiased Skeletal Action Recognizer in Federated Learning. *IEEE Trans. Image Process.* **2025**, *34*, 4908–4922. [CrossRef] [PubMed]
4. Lin, Y.; Liao, Y.; Zeng, W.; Wei, Y.; Chen, D.; Yuan, X.; Li, Y.; Erkan, U.; Toktas, A.; Zhang, C.; et al. 3D Non-degenerate Hyperchaos: Design, Analysis, and Application in Image Encryption. *IEEE Trans. Consum. Electron.* **2026**, 1-1. [CrossRef]
5. Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; Mei, T. Wave-ViT: Unifying Wavelet and Transformers for Visual Representation Learning. In *Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 Octber 2022* ; Springer: Cham, Switzerland, 2022; pp. 328–345. [CrossRef]
6. Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; An kumar, A.; Catanzaro, B. Efficient Token Mixing for Transformers via Adaptive Fourier Neural Operators. In Proceedings of the Tenth International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022. Available online: https://openreview.net/forum?id=EXHG-A3jlM (accessed on 15 April 2026).
7. Ming, Y.; Zhou, J.; Jia, X.; Zheng, Q.; Xiong, L.; Feng, F.; Hu, N. F2D-SIFPNet: A Frequency 2D Slow-I-Fast-P Network for Faster Compressed Video Action Recognition. *Appl. Intell.* **2024**, *54*, 5197–5215. [CrossRef]
8. Wanyan, Y.; Yang, X.; Dong, W.; Xu, C. A Comprehensive Review of Few-Shot Action Recognition. *Int. J. Comput. Vis.* **2025**, *133*, 6832–6859. [CrossRef]
9. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 2 October–2 November 2019; pp. 6201–6210. [CrossRef]
10. Wang, Y.; Song, Y.; Xie, C.; Liu, Y.; Zheng, Z. VideoLLaMB: Long Streaming Video Understanding with Recurrent Memory Bridges. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025. Available online: https://openaccess.thecvf.com/content/ICCV2025/html/Wang_VideoLLaMB_Long_Streaming_Video_Understanding_with_Recurrent_Memory_Bridges_ICCV_2025_paper.html (accessed on 15 April 2026).
11. Pang, Z.; Sener, F.; Yao, A. Context-Enhanced Memory-Refined Transformer for Online Action Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025. [CrossRef]
12. Faure, G.J.; Yeh, J.F.; Chen, M.H.; Su, H.T.; Lai, S.H.; Hsu, W.H. HERMES: Temporal-coHERent Long-forM Understanding with Episodes and Semantics. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025. Available online: https://openaccess.thecvf.com/content/ICCV2025/html/Faure_HERMES_temporal-coHERent_long-forM_understanding_with_Episodes_and_Semantics_ICCV_2025_paper.html (accessed on 15 April 2026).
13. Agrawal, T.; Ali, A.; Dantcheva, A.; Bremond, F. Scaling Action Detection: AdaTAD++ with Transformer-Enhanced Temporal-Spatial Adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025. Available online: https://openaccess.thecvf.com/content/ICCV2025/html/Agrawal_Scaling_Action_Detection_AdaTAD_with_Transformer-Enhanced_Temporal-Spatial_Adaptation_ICCV_2025_paper.html (accessed on 15 April 2026).
14. Duan, W.; Ji, L.; Chen, S.; Zhu, S.; Ye, M. Triple-domain Feature Learning with Frequency-aware Memory Enhancement for Moving Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5006014. [CrossRef]

15. Du, C.; Wang, Y.; Song, S.; Huang, G. Probabilistic Contrastive Learning for Long-Tailed Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 3122–3135. [CrossRef] [PubMed]
16. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
17. Liu, Q.; Jiang, X.; Jiang, R. Classroom Behavior Recognition Using Computer Vision: A Systematic Review. *Sensors* **2025**, *25*, 373. [CrossRef] [PubMed]
18. Sreekar, K.; Jeyakumar, G. Digital Guardians: Exploring Video Analytics in Autonomous Classroom Surveillance Systems. In Proceedings of the IEEE International Conference on Machine Learning and Autonomous Systems (ICMLAS), Prawet, Thailand, 10–12 March 2025. [CrossRef]
19. Gu, C.; Sun, C.; Ross, D.A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6047–6056. [CrossRef]
20. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The Kinetics Human Action Video Dataset. *arXiv* **2017**, arXiv:1705.06950. Available online: https://arxiv.org/abs/1705.06950 (accessed on 15 April 2026). [CrossRef]
21. Goyal, R.; Ebrahimi, Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5843–5851. [CrossRef]
22. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv* **2012**, arXiv:1212.0402. Available online: https://arxiv.org/abs/1212.0402 (accessed on 15 April 2026). [CrossRef]
23. Chen, J.; Deng, S.; Yu, W.; Wei, L. RAAG: Redundancy-Adaptive and Attention-Guided Token Pruning for Efficient Video Action Detection. *Neurocomputing* **2025**, *658*, 131615. [CrossRef]
24. Nguyen, D.D.M.; Nhan, B.D.; Wang, J.C.; Duong, V.H. YOWOv3: An Efficient and Generalized Framework for Spatio-Temporal Action Detection. *IEEE Intell. Syst.* **2025**, *41*, 75–85. [CrossRef]
25. Chen, Y.; Jiao, T.; Guo, C.; Zhang, Z.; Song, J.DarkActionNet: An efficient neural network for low-light animal behavior recognition. *Signal Image Video Process.* **2025**, *19*, 807. [CrossRef]
26. Li, X.; Yang, W.; Wang, K.; Wang, T.; Fei, Q. Context-Aware Network Based on Multi-scale Spatio-temporal Attention for Action Recognition in Videos. *arXiv* **2025**, arXiv:2512.18750. Available online: https://arxiv.org/abs/2512.18750 (accessed on 15 April 2026).
27. Bsoul, A.A.R.K. Human Activity Recognition Using Graph Structures and Deep Neural Networks. *Computers* **2024**, *14*, 9. [CrossRef]
28. Song, Y.; Tong, Z.; Wang, J.; Wang, L. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 28 November–9 December 2022; pp. 10078–10093. [CrossRef]

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.