Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

Knowledge Distillation From VideoMAE To MobileViT For Classroom Student Action Classification Toward Edge Deployment

Raffi Ardhi Naufal[1], Yaya Wihardi[2*], Rani Megasari[3]
[1,2,3]Program Studi Ilmu Komputer, Universitas Pendidikan Indonesia, Kota Bandung, Indonesia
[1]ardhiraffi12@upi.edu, [2*]yayawihardi@upi.edu, [3]megasari@upi.edu

open access

***Corresponding Author**

Article History:
Submitted: 01-11-2026
Accepted: 11-11-2026
Published: 16-11-2026

Keywords:
Action Recognition; Edge Deployment; Knowledge Distillation; MobileViT; VideoMAE.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Background — Monitoring student engagement through automated action recognition can support timely pedagogical intervention, but large transformer models such as Video Masked Autoencoder (VideoMAE) are too computationally expensive for real-time deployment on edge devices. **Objective** — This study evaluates MobileViT with Knowledge Distillation (KD) from a VideoMAE Teacher to identify the most favourable accuracy-speed trade-off for classroom action classification. **Methods** — A 2,160-clip, five-class dataset was compiled from 403 prior and 1,757 newly collected clips. VideoMAE-Base and MobileViT-S were trained, followed by a grid search over 24 KD configurations spanning distillation weight, temperature, and two temporal aggregation methods. Undistilled controls were trained for both aggregation methods, separating the effect of distillation from that of aggregation. Accuracy, Precision, Recall, F1-Score, and frames per second were measured end-to-end and in pure compute on GPU and CPU. **Results** — Across three repetitions the Teacher reached F1 = 86.85 ± 2.20% and the undistilled Students 80.83 ± 3.25% (mean pooling) and 77.46 ± 7.15% (LSTM). The best KD configuration reached 87.84 ± 0.99%, exceeding the Teacher by 0.99 points, smaller than either standard deviation, with 14.7 times fewer parameters. Against matched controls, distillation gained 4.46 to 10.38 points. **Conclusion** — KD raises a lightweight model close to its Teacher, while temporal aggregation alone showed no reliable effect. The Student's roughly twofold pure-compute advantage is largely hidden end-to-end because video decoding consumes most of the GPU processing time, so pipeline optimisation matters as much as model compression for edge deployment.

INTRODUCTION

Student participation and attention are what classroom monitoring systems are built to observe, yet a teacher can attend to only a limited number of students at once. Manual observation is constrained by the attention span of a single observer and is prone to subjective bias, particularly in large classes (Alruwais & Zakariah, 2024). Where observation is incomplete, adaptive instruction and pedagogical decisions rest on partial information; this is stated here as the motivation for automating the observation step, not as an effect measured in this study.

Advances in artificial intelligence (AI) and computer vision have opened opportunities for automated classroom monitoring through video-based action recognition, which allows a system to continuously and objectively identify student behaviors such as active participation (e.g., raising a hand) or disruption (e.g., using a phone) (Heng et al., 2024). Such information can help teachers adjust their instructional strategy in a timely manner (Sassi et al., 2023).

Despite this potential, deploying action recognition systems in real classrooms is constrained by inference speed and computational cost. Large-scale transformer architectures such as Video Masked Autoencoder (VideoMAE) are effective at spatio-temporal representation learning for video understanding (Zhang & Wang, 2025), but they require substantial computational resources and exhibit high inference latency (Atilgan & Mercimek, 2025), making them poorly suited for the edge devices typically used in school surveillance systems.

This gap motivates two related research needs. First, model-side: whether a compact model such as MobileViT can approach VideoMAE-level accuracy on this task once KD is applied, and which combination of distillation weight α, temperature T, and temporal aggregation strategy is most effective. Second, dataset-side: whether extending the training data beyond 403 clips changes the achievable accuracy ceiling for this task. Addressing both gaps is particularly relevant to the Indonesian context, where limited school infrastructure makes efficient edge-AI solutions a strategic alternative to costly server-based deployments, and where an accurate yet lightweight classification module is a prerequisite for deploying the second stage of a two-stage cascaded classroom camera system at scale.

Accordingly, this study aims to (1) implement and evaluate MobileViT for classroom student action classification; (2) apply Knowledge Distillation from a VideoMAE Teacher to optimize MobileViT's accuracy; (3) analyze the computational efficiency trade-off between the distilled Student and the Teacher; and (4) extend the

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

available classroom action dataset by combining newly collected recordings with the 403-clip dataset from (Herawan et al., 2025). The corresponding research questions are: (RQ1) How does MobileViT perform, in terms of Accuracy, Precision, Recall, F1-Score, and inference throughput in frames per second (FPS), relative to VideoMAE as the reference model? (RQ2) How effective is Knowledge Distillation at transferring knowledge from the VideoMAE Teacher to the MobileViT Student in terms of Accuracy and F1-Score? (RQ3) Which configuration — MobileViT without KD, MobileViT with KD, or VideoMAE — offers the most favorable trade-off between classification accuracy and inference speed for edge deployment in an educational setting?

LITERATURE REVIEW

To reconcile accuracy with efficiency, this study adopts MobileViT, a lightweight hybrid Convolutional Neural Network (CNN)–Transformer architecture designed for mobile and resource-constrained devices (Mehta & Rastegari, 2022), (Liu et al., 2025), combined with Knowledge Distillation (KD). KD transfers the dark knowledge encoded in the soft output distribution of a large Teacher Network to a smaller Student Network so that the Student approaches the Teacher's accuracy without inheriting its computational cost (Hinton et al., 2015), (Gan & Xiong, 2025).

VideoMAE addresses the temporal redundancy of raw video through an asymmetric encoder-decoder design with tube masking, in which 90–95% of spatio-temporal patches are masked uniformly along the time axis so that the Vision Transformer encoder processes only the visible tokens; the resulting representation transfers well even when pre-training data is limited (Tong et al., 2022). MobileViT combines convolutional layers for local spatial features with a Transformer encoder applied to the unrolled patches inside each block, modelling global dependencies without the parameter count of a full-scale Vision Transformer (Mehta & Rastegari, 2022). In the Knowledge Distillation formulation used here, the Teacher's logits are softened with a temperature parameter T before being compared against the Student's softened logits via Kullback–Leibler divergence, while the weight α balances this KD loss against a standard cross-entropy loss on the ground-truth label (Hinton et al., 2015), (Ahmadabadi et al., 2023); larger T values smooth the Teacher's output distribution and reveal more of its structural, class-similarity information ("dark knowledge") to the Student.

Because this study targets a multi-person classroom setting, the task is framed as action classification — assigning one of a fixed set of labels to a trimmed clip that already contains a single student — rather than action recognition over an untrimmed, multi-person scene, and it is treated as the second stage of a two-stage cascaded pipeline: a subject detection/localization stage first produces a cropped Region of Interest (ROI) around each student (Zhang & Wang, 2025), (Yan et al., 2025), and the classification stage then recognizes the action of each individual. Two-stage cascaded designs are a common strategy for balancing accuracy and real-time throughput (Meng et al., 2024), and this study specifically targets optimization of the classification stage, treating the detection stage as an upstream pre-processing step whose output is a pre-cropped clip of a single student.

Prior work on classroom action recognition has proposed various label taxonomies and datasets, and the five action classes used in this study are drawn from that body of work. Table 1 lists each class together with the pedagogical state the literature associates with it and the studies that motivate its inclusion.

Table 1. The Five Action Classes, the Pedagogical State Associated with Each, and the Prior Work That Motivates Its Inclusion

Action Class	Pedagogical State	Prior Work
Nodding	Participation	(Herawan et al., 2025)
Hand-raising	Participation	(Lin et al., 2021), (Shou et al., 2024), (Sharma et al., 2021)
Phone use	Disengagement	(Shou et al., 2024), (Yu et al., 2021)
Head-supporting (resting the head on the desk)	Fatigue	(Lin et al., 2021), (Shou et al., 2024)
Looking down	Disengagement	(Lin et al., 2021), (Deshpande & Deshpande, 2023)

Nodding and hand-raising are treated in that literature as participatory signals, phone use and looking down as indicators of disengagement, and resting the head on the desk as a subtler fatigue-related cue requiring fine-grained motion to detect. The set therefore spans both participatory and disengagement-related behaviours; the mapping from each action to a pedagogical state is taken from that prior work and is not itself evaluated here.

However, the direct predecessor of this study (Herawan et al., 2025) trained VideoMAE and TimeSformer on a comparatively small dataset of 403 clips, reporting a VideoMAE accuracy of only 68.08%, and did not investigate whether a lighter model could achieve comparable accuracy through Knowledge Distillation, nor did it examine the trade-off between raw model compute speed and end-to-end inference throughput once a full processing pipeline (including video decoding) is taken into account.

A more recent study by the same research group (Wihardi et al., 2026) addressed the same five classroom activities through a multi-cue fusion framework that combines person detection, head-pose estimation, and an explicit smartphone-related cue, reporting 91.13% accuracy and 91.14% macro F1 under a subject- and session-separated evaluation protocol. That work improved recognition accuracy by adding complementary input modalities and, as its

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

authors state, did not evaluate repeated-run variance. The present study addresses a complementary question: whether a single-stream, lightweight classifier can approach teacher-level accuracy through Knowledge Distillation rather than through additional input cues, and how stable that result is across repeated training runs.

METHOD

This study used a comparative experimental design that trains and evaluates four model scenarios under identical data splits and evaluation protocols: (1) VideoMAE as the Teacher, (2a) MobileViT with mean-pooled temporal aggregation trained without distillation, (2b) MobileViT with LSTM aggregation trained without distillation, and (3) MobileViT trained with Knowledge Distillation from the Teacher across a grid of hyperparameters. Scenario 2b gives every distilled configuration an undistilled control that shares its temporal aggregation, so that the effect of distillation can be separated from the effect of aggregation; without it the best distilled configuration would differ from the baseline in two respects at once.

Participants were students who volunteered to be recorded performing classroom actions for the purpose of dataset construction, not as human subjects of a behavioral intervention. Recordings were collected in two sessions on different days, with 17 different students participating in each session (34 students in total), following the recording protocol of (Herawan et al., 2025): simultaneous capture from three viewpoints — a CCTV camera at the front of the room and two smartphones, one at the front-left and one at the front-right corner — at four different seating positions per session, so that each session captured 24 action executions (6 action sequences × 4 positions). The front-right viewpoint was recorded with a smartphone rather than the handycam used by (Herawan et al., 2025), which was not available. Recording specifications were not uniform across viewpoints: the CCTV camera recorded at 2,560×1,440 and 15 fps and was re-encoded to 1,920×1,080 before clip extraction, three smartphones recorded at 1,920×1,080 and 30 fps, and the front-right smartphone in the first session recorded at 1,280×720 and 30 fps. All clips were subsequently standardised to 1,080×1,080 at 30 fps, so clips originating from the lower-resolution source were upscaled and carry less spatial detail than the rest.

Table 2. Composition of the Combined Dataset by Action Class and Subset

Action class	Train	Validation	Test	Total
Nodding	356	44	45	445
Hand-raising	354	44	46	444
Phone use	337	41	44	422
Head-supporting	355	43	46	444
Looking down	323	40	42	405
Total	1,725	212	223	2,160



Figure 1. Uniformly Sampled Frames for the Five Action Classes (faces blurred to protect participant identity)

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

The dataset combines this newly collected data with the 403-clip dataset from (Herawan et al., 2025), yielding 2,160 pre-cropped video clips across five action classes: nodding, hand-raising, phone use, resting the head on the desk, and looking down. Target clip duration was 3–9 seconds, although a small number of retained clips fell outside this range when the performed action remained sufficiently clear; all clips were manually labeled based on the dominant action performed. The newly collected clips were first stratified into training, validation, and test sets at an intended 80:10:10 ratio, after which they were combined with the corresponding pre-existing subsets of the 403-clip dataset from (Herawan et al., 2025). The resulting combined dataset contained 1,725 training clips (79.86%), 212 validation clips (9.81%), and 223 test clips (10.32%), closely approximating the intended proportion. Table 2 reports the composition per class and subset. Figure 1 presents representative sampled frames for each of the five action classes.

The instrumentation consisted of two deep learning models. The Teacher, VideoMAE-Base (86.24M parameters, pretrained on Kinetics-400 with 90% tube masking), was fine-tuned with a randomly initialized 768→5 linear classification head, applying a lower learning rate to the pretrained encoder than to the new head (5×10^{-6} and 1×10^{-4} respectively), with the AdamW optimizer and a ReduceLROnPlateau scheduler. The Student, MobileViT-S (4.94M trainable parameters with mean pooling and 5.86M with LSTM, ImageNet-pretrained), processes 16 uniformly sampled frames per clip independently through its backbone before aggregating the resulting 640-dimensional per-frame features via either mean pooling or a single-layer Long Short-Term Memory (LSTM) network (256 hidden units), following the recurrent aggregation strategy introduced for video sequence modelling (Donahue et al., 2015), prior to classification, trained with AdamW and CosineAnnealingLR. For Knowledge Distillation, the Student's total loss combined a standard cross-entropy term with a temperature-scaled Kullback–Leibler divergence term between the softened Teacher and Student output distributions (Hinton et al., 2015), weighted by α and scaled by T^2 (Ahmadabadi et al., 2023): $L = (1 - \alpha)\cdot L_CE + \alpha\cdot T^2\cdot L_KD$. All three scenarios used early stopping with a patience of 30 epochs (maximum 800 epochs for the Teacher and 300 for the Student, following (Ahmadabadi et al., 2023) for the initial KD hyperparameters), and Random Horizontal Flip and Random Resized Crop augmentation were applied only to the training split, given the still-limited dataset size. Every scenario and every KD configuration was trained three times with independent random seeds, and all accuracy metrics in this article are reported as the mean and standard deviation across those repetitions; the best configuration was selected by mean validation accuracy rather than by test performance. Seven of the 24 KD configurations completed only two repetitions because the GPU used for training failed irrecoverably partway through the second pass, and those configurations are identified in Table 4. Inference speed, by contrast, was measured once per configuration in a separate run under uniform system conditions and is therefore reported without a standard deviation. Figure 2 summarizes the resulting distillation architecture.

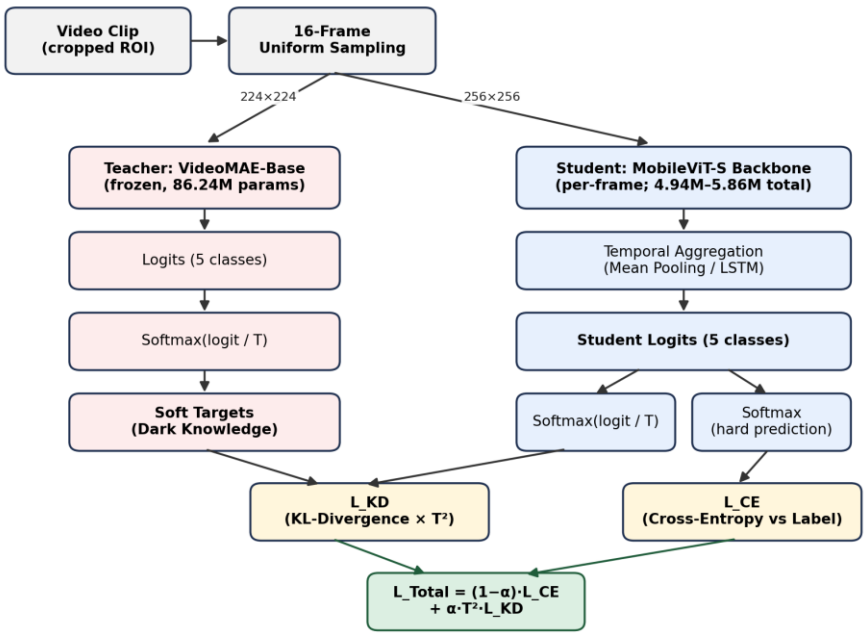


Figure 2. Knowledge Distillation Architecture from VideoMAE (Teacher) to MobileViT (Student)

Procedures followed a two-environment workflow: code and data pipelines were first validated on a laptop using its central processing unit (CPU) only, with a small data subset and 1–2 epochs to confirm the data loader, model instantiation, and forward pass were error-free, before full-scale training was executed on a High-Performance Computing (HPC) workstation equipped with a graphics processing unit (GPU; NVIDIA RTX 5090, 32 GB VRAM). Data collection took place over two sessions in April 2026 at Universitas Pendidikan Indonesia.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

For analysis, the three scenarios were compared using Accuracy, macro-averaged Precision, Recall, and F1-Score on the held-out test set, computed from the standard confusion-matrix counts (TP, TN, FP, FN). Inference speed was measured separately from training, using FPS under four conditions — end-to-end (E2E, including file reading, decoding, and pre-processing) and pure-compute (forward pass only) — each on both GPU and CPU, with a uniform batch size (4) and worker configuration across all 26 evaluated configurations (Teacher, Student baseline, and 24 KD combinations) to ensure a fair comparison. Each end-to-end measurement timed 20 consecutive batches on GPU and 10 on CPU, after 5 and 2 warm-up batches respectively, at a batch size of 4; each pure-compute measurement timed 50 forward passes over a single cached batch after 10 warm-up passes, so that no file reading or decoding entered the timing. Every reported FPS value is a single aggregate throughput, total frames divided by total elapsed time, rather than an average of per-batch timings, and the speed measurements were performed once rather than repeated across the three training seeds. FPS measurement was intentionally decoupled from the training runs and executed within a single session to avoid bias from system-load variation across the days-long grid-search training process. Consequently the accuracy metrics reported in this article carry standard deviations while the FPS values do not; the dispersion of the FPS measurements is instead estimated in the Result section from the configurations that share an identical computation graph. Model selection relied exclusively on the validation split: both early stopping within each training run and the selection of the best distillation configuration among the 24 grid combinations were based on validation performance. The test split was used once, at the end, to report the performance of the already-selected configurations, so that the test data did not influence which configuration was reported as best. Statistical testing was performed after training had finished, on the stored per-repetition F1-Scores and without any retraining. Differences between scenarios were tested with Welch's t-test, which does not assume equal variances, and the effect of the three grid-search factors (temporal aggregation, distillation weight α, and temperature T) was tested with a Type II analysis of variance over all per-repetition F1-Scores, Type II being appropriate because the grid is unbalanced: seven configurations have only two repetitions. The significance level was 0.05.

This study is limited to the classification stage of a two-stage cascaded pipeline; subject detection/localization was treated as an upstream pre-processing step and was not evaluated. The dataset split preserved camera-simultaneous recordings rather than enforcing a subject-independent split, meaning the same individual may appear in both the training and test sets; this is discussed further in the Limitations subsection below.

RESULT

The combined dataset comprised 2,160 video clips: 403 from (Herawan et al., 2025) and 1,757 newly collected across two recording sessions with 17 different students each, preserving camera-angle diversity (one CCTV camera and two smartphones) in the newly collected portion. Class representation is relatively balanced at 405–445 clips per class, and the per-class split is given in Table 2.

The Teacher (VideoMAE) was trained three times with independent random seeds; early stopping triggered after 49 epochs on average out of a maximum of 800, and the model reached 87.00 ± 2.24% test accuracy with F1 = 86.85 ± 2.20% (Precision = 87.27 ± 1.93%, Recall = 86.91 ± 2.23%). This exceeds the 68.08% accuracy previously reported for VideoMAE on a smaller 403-clip version of this task (Herawan et al., 2025), plausibly reflecting the larger and more diverse training set as well as the discriminative learning-rate fine-tuning scheme used here, although this study did not run a controlled ablation to isolate the individual contribution of each factor.

The undistilled Student baseline (MobileViT-S with mean-pooled temporal aggregation) stopped after 98 epochs on average out of a maximum of 300, reaching 80.87 ± 3.60% test accuracy and F1 = 80.83 ± 3.25% — 6.02 percentage points below the Teacher, motivating the subsequent Knowledge Distillation experiments. Table 3 sets the four main scenarios side by side on every reported metric.

Table 3. Comparison of the Four Main Scenarios (Mean ± Standard Deviation over Three Repetitions). Validation accuracy is computed on the validation split; test accuracy, precision, recall, and F1-Score are computed on the held-out test split.

Scenario	Val. Acc. (%)	Test Acc. (%)	Precision (%)	Recall (%)	F1 Macro (%)
Teacher (VideoMAE-Base)	91.19 ± 0.72	87.00 ± 2.24	87.27 ± 1.93	86.91 ± 2.23	86.85 ± 2.20
Student baseline (MobileViT-S, mean pooling, no KD)	87.89 ± 2.33	80.87 ± 3.60	82.53 ± 1.85	80.75 ± 3.49	80.83 ± 3.25
Student baseline (MobileViT-S, LSTM, no KD)	83.49 ± 2.36	77.43 ± 7.30	78.65 ± 6.19	77.27 ± 7.44	77.46 ± 7.15
Student + KD (LSTM, α=0.45, T=7)	94.03 ± 0.72	88.04 ± 0.93	88.24 ± 1.07	87.87 ± 0.96	87.84 ± 0.99

The second undistilled control (MobileViT-S with LSTM aggregation) was trained under identical hyperparameters and data splits, differing from the mean-pooling baseline only in temporal aggregation. Across three repetitions it stopped after 70, 86, and 34 epochs and reached F1 = 77.46 ± 7.15%. The standard deviation is markedly

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

larger than in any other scenario, and the cause is the third repetition: it attained its highest validation F1 of 81.29% at epoch 4 and never surpassed it over the following 30 epochs, so early stopping retained the epoch-4 weights even though training loss kept falling from 0.8865 to 0.2177. The early-stopping rule was identical across all scenarios. This control was trained on an NVIDIA Tesla T4 with PyTorch 2.11 because the HPC GPU had failed by then; the architecture was verified identical by its 5,858,469 trainable parameters, and inference speed was not remeasured because α and T enter only the loss function and never the model definition.

Table 4. Full Grid Search Results for 24 Knowledge Distillation Configurations (Accuracy as Mean ± Standard Deviation over n Repetitions; FPS as a Single End-to-End GPU Measurement)

Configuration	Temporal	Test Acc. (%)	F1 Macro (%)	FPS (GPU)	n
StandardKD α=0.30 T=7	Mean	84.45 ± 1.13	84.21 ± 1.02	136.56	3
StandardKD α=0.30 T=11	Mean	84.75 ± 1.62	84.36 ± 1.66	125.93	3
StandardKD α=0.30 T=15	Mean	85.50 ± 1.87	85.29 ± 1.94	122.18	3
StandardKD α=0.45 T=7	Mean	84.45 ± 2.47	84.13 ± 2.59	100.18	3
StandardKD α=0.45 T=11	Mean	85.05 ± 1.13	84.77 ± 1.15	121.23	3
StandardKD α=0.45 T=15	Mean	85.95 ± 0.93	85.68 ± 0.89	125.75	3
StandardKD α=0.60 T=7	Mean	84.01 ± 1.57	83.78 ± 1.47	114.90	3
StandardKD α=0.60 T=11	Mean	84.45 ± 1.81	84.10 ± 1.86	115.29	3
StandardKD α=0.60 T=15	Mean	84.60 ± 3.15	84.28 ± 3.05	118.75	3
PureKD α=1.00 T=7	Mean	85.35 ± 2.47	85.00 ± 2.57	127.06	3
PureKD α=1.00 T=11	Mean	84.75 ± 1.62	84.42 ± 1.64	117.15	3
PureKD α=1.00 T=15	Mean	86.40 ± 1.13	86.13 ± 1.08	136.10	3
StandardKD α=0.30 T=7	LSTM	84.45 ± 2.63	84.25 ± 2.54	120.32	3
StandardKD α=0.30 T=11	LSTM	85.65 ± 3.56	85.47 ± 3.38	121.25	3
StandardKD α=0.30 T=15	LSTM	85.95 ± 1.04	85.73 ± 1.09	110.83	3
★ StandardKD α=0.45 T=7	LSTM	88.04 ± 0.93	87.84 ± 0.99	126.81	3
StandardKD α=0.45 T=11	LSTM	84.75 ± 3.83	84.51 ± 3.92	136.96	3
StandardKD α=0.45 T=15	LSTM	86.32 ± 4.12	86.16 ± 4.01	123.00	2
StandardKD α=0.60 T=7	LSTM	84.98 ± 0.95	84.60 ± 1.23	130.50	2
StandardKD α=0.60 T=11	LSTM	86.55 ± 2.54	86.31 ± 2.62	131.62	2
StandardKD α=0.60 T=15	LSTM	85.43 ± 0.32	85.17 ± 0.40	124.32	2
PureKD α=1.00 T=7	LSTM	85.20 ± 1.27	84.93 ± 1.19	123.73	2
PureKD α=1.00 T=11	LSTM	86.10 ± 1.27	85.83 ± 1.34	126.97	2
PureKD α=1.00 T=15	LSTM	86.32 ± 1.59	86.09 ± 1.65	145.69	2

The 24-configuration KD grid search (Table 4) showed that all 24 configurations improved mean F1-Score over the 80.83 ± 3.25% baseline, so the benefit of distillation was not confined to a narrow region of the hyperparameter space. The best configuration — StandardKD with LSTM aggregation, α=0.45, T=7 — reached test accuracy of 88.04 ± 0.93% and F1 = 87.84 ± 0.99%, a 7.01-point improvement over the Student baseline and 0.99 points above the Teacher, despite MobileViT-S having roughly 14.7 times fewer parameters (≈5.86M vs. 86.24M); because that 0.99-point difference is smaller than the standard deviation of either model, the two are better described as close than as ranked. The direction of that difference does not even hold across repetitions: in the first repetition the Teacher led by 0.38 points (89.06% against 88.68%), and the ordering reverses only once the three repetitions are averaged. Configurations using LSTM aggregation spanned a mean F1 range of 84.25–87.84% against 83.78–86.13% for mean pooling. Pure-KD configurations (α=1.00, no cross-entropy term) exceeded the baseline at every temperature and peaked at F1 = 86.13 ± 1.08% (T=15), showing that the Teacher's soft output distribution alone carries sufficient signal for the Student to learn the task.

Table 5 reports inference speed under the four conditions defined in the Method: end-to-end and pure compute, each on GPU and CPU.

Table 5. Inference Speed of the Three Main Scenarios (Frames per Second; Single Measurement per Condition, Not Repeated across Seeds)

Scenario	E2E GPU	E2E CPU	Pure GPU	Pure CPU
Teacher (VideoMAE)	122.16	18.48	1,673.79	26.69
Student baseline (mean pooling)	126.69	19.98	3,350.54	31.65
Best KD (LSTM α=0.45 T=7)	126.81	19.61	3,329.94	31.37

Figure 3 plots the same measurements on a logarithmic scale, placing the pure-compute and end-to-end channels side by side.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

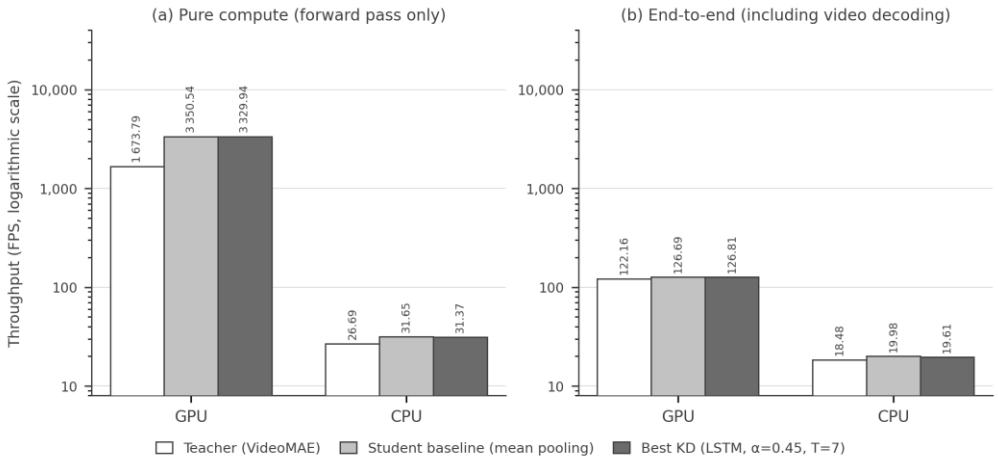


Figure 3. Inference Throughput under Pure-Compute and End-to-End Measurement (Logarithmic Scale)

On pure-compute GPU, the Student was about 2.00× faster than the Teacher (3,350.54 vs. 1,673.79 FPS for the baseline Student and Teacher respectively), but this advantage nearly disappeared under end-to-end (E2E) GPU measurement (126.69 vs. 122.16 FPS for the same pair), because 93–96% of E2E GPU processing time was consumed by the pipeline (chiefly video decoding) rather than the forward pass itself. Figure 3 makes that contrast direct: the roughly twofold gap visible under pure compute is absent once decoding is included. On CPU, where the pipeline occupies a smaller share of total time (31–37%), the pure-compute speed-up of the Student over the Teacher was smaller (≈1.18×), and the E2E CPU throughput accordingly showed only a modest difference (19.98 vs. 18.48 FPS). The best KD configuration achieved 126.81 E2E GPU FPS against the Teacher's 122.16 FPS. That 3.8% difference is smaller than the 6.95–8.11% spread measured across configurations that run an identical computation graph, so the two are reported as comparable end-to-end rather than ranked. The distilled model nonetheless requires roughly 22.4 MB of storage versus the Teacher's 329 MB — an approximately 14.7× reduction in storage footprint alongside near-parity accuracy.

Two separate effects stand between the 14.7-fold parameter reduction and the throughput actually observed, and they act in sequence. The first is that parameter count is not a measure of work per clip, because the two models consume a clip differently: the Student applies a 2D backbone once per frame, so a single 16-frame clip costs 16 forward passes at 256×256, whereas the Teacher processes the whole 16-frame volume in one pass at 224×224 through tubelet embedding. For the LSTM Student, a 14.7-fold reduction in parameters therefore yields only a 1.99-fold gain in pure-compute throughput. This study did not profile floating-point operations, so this account of the first effect rests on the structure of the two models rather than on a measurement. The second effect is the pipeline, and it can be checked numerically. If video decoding and pre-processing occupy a share p of total end-to-end time and the model itself becomes s times faster, the end-to-end gain is $1/(p + (1-p)/s)$. With the measured p of 0.93–0.96 on GPU and s = 1.99, that expression predicts an end-to-end gain of 1.02–1.04×, against 1.04× observed (126.81 versus 122.16 FPS). On CPU, where p falls to 0.31–0.37 and s is only 1.18, it predicts 1.11× against 1.08× observed, slightly over-predicting the gain. The near-parity end-to-end throughput reported above is therefore the expected consequence of these two attenuations rather than an anomaly, and it means a further reduction in model size would move end-to-end throughput very little while decoding retains its present share of the budget.

The per-configuration end-to-end FPS values in Table 4 vary considerably, but that variation does not originate in the models, and the following analysis establishes why they should not be compared row by row. The distillation weight and the temperature enter only the loss function and never the model definition, so all twelve configurations within one temporal-aggregation group execute an identical computation graph at inference time, with identical tensor shapes and no data-dependent branching. Those twelve configurations can therefore be treated as twelve repeated measurements of a single workload, which makes their spread an estimate of measurement uncertainty rather than a difference between models. Within the mean-pooling group the twelve end-to-end GPU values have a mean of 121.76 FPS and a standard deviation of 9.87 FPS, a coefficient of variation of 8.11%; within the LSTM group the corresponding figures are 126.83 and 8.82 FPS, a coefficient of variation of 6.95%. That 6.95–8.11% band is the measurement uncertainty against which the end-to-end differences reported above were judged.

Statistical testing was applied to the accuracy results. A Type II analysis of variance over the 65 per-repetition F1-Scores found no effect of temporal aggregation ($F(1,58) = 3.53$, $p = 0.065$), weight α ($F(3,58) = 0.77$, $p = 0.518$), or temperature T ($F(2,58) = 0.97$, $p = 0.386$), the three factors accounting for 11.6% of the variance against a within-configuration standard deviation of 1.93 points. Welch's t-tests separate neither the best configuration from the five next-highest ($p = 0.114$–0.658) nor the Teacher from it ($p = 0.529$). No scenario-level comparison reached 0.05 either, including the 7.01-point gain over the undistilled baseline ($p = 0.054$); with three repetitions per scenario and baseline

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

standard deviations of 3.25 and 7.15 points this indicates limited power rather than absence of effect, and the direction of the distillation effect rests instead on all 24 configurations exceeding the baseline mean and 44 of 65 distilled repetitions exceeding even the best baseline repetition (84.15%).

Table 6 breaks the test-set F1-Score down by action class for the first repetition, locating where distillation actually helps.

Table 6. Per-Class F1-Score on the Test Set for the Three Main Scenarios (First Repetition)

Action class	Teacher (%)	Student baseline (%)	Student + KD (%)
Nodding	88.17	76.19	80.85
Hand-raising	96.84	94.74	97.83
Phone use	84.34	75.00	90.48
Head-supporting	89.89	83.50	90.72
Looking down	86.05	73.91	83.54

Relative to the undistilled baseline, the largest gain appears on phone use, whose F1-Score rises from 75.00% to 90.48%; recall for that class rises from 0.614 to 0.864, meaning the number of correctly recognised clips increases from 27 to 38 out of 44. Nodding improves from 76.19% to 80.85% and head-supporting from 83.50% to 90.72%, while looking down improves more modestly and hand-raising is already near-saturated in every scenario. On phone use the distilled Student even exceeds its own Teacher (90.48% against 84.34%). These values derive from the confusion matrices of the first repetition only and are therefore reported without standard deviations.

DISCUSSION

These results directly address the three research questions. For RQ1, MobileViT without KD trails VideoMAE by 6.02 F1 points, showing that the evaluated MobileViT-S configuration did not match VideoMAE's accuracy on this task without KD. For RQ2, Knowledge Distillation is effective on this task: in the best configuration it removes the entire 6.02-point deficit and places the Student 0.99 points above the Teacher, a margin that lies within the standard deviation of both and is therefore best read as parity rather than superiority; the improvement is broad rather than isolated to a single hyperparameter setting, since all 24 grid-search configurations beat the baseline. For RQ3, when accuracy, model size, and end-to-end throughput are considered jointly, the best KD configuration (LSTM, α=0.45, T=7) is the configuration selected here for deployment on edge devices equipped with a GPU or a neural processing unit (NPU), reaching accuracy close to the Teacher with a smaller footprint and E2E throughput that is comparable under the GPU conditions tested here; on CPU-only hardware, the advantage is more limited because the pure-compute speed-up of the smaller model is largely absorbed by pipeline overhead. As noted in the Limitations, these GPU and CPU measurements were obtained on HPC hardware rather than on a physical edge device, so this conclusion should be read as an identified candidate configuration rather than a validated edge-deployment result.

With both undistilled controls available, the effect of distillation can be read off three comparisons that each change one thing. Temporal aggregation without distillation: 80.83 ± 3.25% for mean pooling against 77.46 ± 7.15% for LSTM, a 3.37-point difference far smaller than either standard deviation and therefore not interpretable as a difference. Distillation with aggregation held at mean pooling: the best mean-pooling KD configuration, selected by the same criterion of highest mean validation accuracy among configurations with three repetitions, reached 85.29 ± 1.94% against 80.83 ± 3.25%, a gain of 4.46 points. Distillation with aggregation held at LSTM: 87.84 ± 0.99% against 77.46 ± 7.15%, a gain of 10.38 points. Neither comparison reaches the 0.05 level on three repetitions, as reported above. The 7.01-point figure obtained by comparing the mean-pooling baseline with the LSTM-based best configuration therefore conflates two changes; once separated, the gain attributable to distillation lies between 4.46 and 10.38 points depending on which aggregation is held fixed, while aggregation alone shows no effect that survives its own variance. The dispersion itself follows the same direction: 7.15 and 3.25 points for the undistilled scenarios against 1.94 and 0.99 points for the two best distilled configurations, consistent with the Teacher distribution providing a more stable target than hard labels alone, though three repetitions per group cannot establish this.

The advantage of LSTM aggregation over mean pooling at the best configuration is also worth interpreting carefully. Because LSTM preserves the order of the 16 sampled frames while mean pooling discards it, LSTM aggregation is plausibly better suited to actions with a clear temporal progression, such as nodding or hand-raising, whereas mean pooling may be equally adequate for more static-appearing actions such as phone use or resting the head on the desk. The per-class results in Table 6 are partly consistent with this reading, since the largest distillation gains occur on phone use and head-supporting rather than on the most clearly sequential class, so this remains a plausible explanation rather than a confirmed mechanism. Table 7 positions these results against prior work on classroom action recognition.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

Table 7. Comparison with Prior Work on Classroom Action Recognition

Study	Model	Dataset	Split protocol	Acc. (%)	F1 (%)
Herawan et al. (2025)	VideoMAE	403 clips	stratified random	68.08	n/r
Herawan et al. (2025)	TimeSformer	403 clips	stratified random	78.84	n/r
Wihardi et al. (2026)	Multi-cue fusion	tracklets	subject + session sep.	91.13	91.14
This study	VideoMAE (Teacher)	2,160 clips	stratified random	87.00 ± 2.24	86.85 ± 2.20
This study	MobileViT + KD	2,160 clips	stratified random	88.04 ± 0.93	87.84 ± 0.99

The rows in Table 7 are not directly comparable, and the table is intended to position this study rather than to rank the approaches. The studies differ in dataset size, prediction unit, and above all in split protocol: (Wihardi et al., 2026) enforces subject- and session-separated evaluation, a stricter setting than the stratified random split used in this study and in (Herawan et al., 2025). A higher number under a more permissive protocol does not indicate a stronger model. The comparison that is meaningful here is with (Herawan et al., 2025), whose dataset this study extends and whose split arrangement it preserves: on the same task and the same label set, the Teacher trained on the extended 2,160-clip dataset reaches 87.00 ± 2.24% accuracy against the 68.08% reported for VideoMAE on the original 403-clip dataset, a difference of 18.92 points. The two settings differ in dataset size and in fine-tuning configuration, and this study did not run an ablation separating those factors, so the difference is reported as a difference in reported values rather than attributed to any single cause.

A key implication for research and practice is that the pipeline surrounding a model, not merely its architecture, determines real-world throughput. Because video decoding dominates E2E GPU latency in this study, architectural compression (moving from VideoMAE to a distilled MobileViT) delivers a smaller net speed-up than the nearly 2× raw compute advantage would suggest. This points to hardware-accelerated decoding or selective frame extraction as a complementary optimization target alongside model-level compression for deployment in real classroom surveillance systems built on a two-stage cascaded architecture (Meng et al., 2024).

A second implication concerns model selection under hardware constraints: because the Student's throughput advantage over the Teacher was far smaller on CPU (≈1.18×) than on GPU (≈1.99× in pure compute), the practical benefit of distillation for real-time deployment depends heavily on whether the target edge device has a GPU or NPU accelerator available, and this should be assessed explicitly before committing to a specific model for a given deployment target.

The main research contribution of this study is twofold. First, it provides a systematic grid search over KD weight, temperature, and temporal-aggregation strategy specifically for classroom action recognition, rather than adopting a single fixed KD configuration as in prior distillation work on action recognition (Ahmadabadi et al., 2023). Second, it extends the classroom action dataset introduced by (Herawan et al., 2025) more than fivefold (403 → 2,160 clips) using a consistent multi-camera recording protocol, which — while not the primary object of evaluation — plausibly contributed to the higher Teacher accuracy obtained relative to the original 403-clip study, and provides a larger resource for future classroom action-recognition research on the same five action classes. Together, these contributions give practitioners a concrete, empirically grounded configuration recommendation rather than a purely architectural proof of concept.

Several limitations should be noted. The most consequential concerns the data split. The recording protocol makes train-test overlap at the individual level near-certain rather than merely possible: each participant performed all five target actions from four seating positions, and every execution was captured simultaneously by three cameras, so one participant can contribute up to twelve clips of the same action class. Under an 80:10:10 random split, those clips are almost certain to be distributed across the training and test sets, differing only in viewpoint and seating position. This yields two forms of overlap, one at the individual level and one at the event level, where clips of the same execution recorded from different angles are near-duplicates. Reported accuracy and F1 are therefore optimistic with respect to generalization to unseen students and unseen recording sessions, and are not directly comparable with studies that enforce subject- and session-separated protocols such as (Wihardi et al., 2026). This limitation applies uniformly to every compared scenario, since the Teacher, the undistilled Student, and all distilled configurations were trained and evaluated on exactly the same split; the relative comparison between scenarios therefore remains valid, and it is the absolute values that should be read with caution.

The remaining limitations are narrower in scope, and each touches a specific part of the results. Table 8 lists them together with the part of the study they affect and the way they constrain interpretation.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

Table 8. Narrower Limitations of This Study, the Part of the Results Each Affects, and Its Effect on Interpretation

Limitation	Part Affected	Effect on Interpretation
Incomplete repetitions	Seven of the 24 KD configurations in Table 4	The training GPU failed irrecoverably partway through the second pass, so these seven ran twice instead of three times. Their standard deviations rest on two values and are correspondingly less stable, and because all seven fall within the LSTM block the shortfall is concentrated rather than evenly distributed.
Limited statistical power	Comparisons between scenarios	No scenario-level comparison reaches the 0.05 level on three repetitions, including the 7.01-point gain over the undistilled baseline ($p = 0.054$). This reflects the power available at $n = 3$ rather than an absence of effect, so the direction of the effect is supported descriptively by all 24 configurations exceeding the baseline.
Per-frame augmentation	LSTM aggregation	Augmentation was applied independently to each of the 16 sampled frames. This is unproblematic for mean pooling but may introduce unintended temporal inconsistency across the sequence the LSTM receives.
Single backbone and aggregation pair	Choice of architecture	MobileViT-S and LSTM were not compared against alternative lightweight backbones (e.g., MobileNetV3, EfficientNet-Lite) or aggregation methods (e.g., GRU, attention pooling), so neither can be claimed globally optimal among lightweight options.
CPU timing on server hardware	CPU throughput in Table 5	CPU-side FPS was measured on server-class HPC hardware in CPU-only mode rather than on an actual edge device such as a Raspberry Pi or Jetson. Absolute CPU FPS should not be read as a proxy for edge-device performance, although the relative GPU-versus-CPU comparison remains informative.
Early-stopping rule	Weight selection in every scenario	The retained weights are those with the highest validation F1-Score and training stops after 30 epochs without improvement, so a run that peaks very early retains undertrained weights, as happened in the third repetition of the LSTM control. Across the 26 trainings recorded in the experiment logs the validation peak fell within the first ten epochs once. The rule applied uniformly to every scenario.
Hardware difference for the LSTM control	Scenario 2b	The LSTM control was trained on different hardware and a different PyTorch version from the other scenarios, although the architecture was verified to be identical.

Future work should therefore consider a subject-independent data split to obtain a more rigorous estimate of generalization; a larger number of repetitions per configuration, so that the statistical tests reported here carry adequate power; clip-consistent (rather than per-frame) augmentation decisions for sequential aggregation methods such as LSTM; direct benchmarking on physical edge hardware such as Raspberry Pi, Jetson, or NPU-equipped smartphones; and pipeline-level optimizations such as hardware-accelerated video decoding or selective frame extraction, complementing the model-level optimization explored in this study. Comparing MobileViT-S against other lightweight backbones and temporal-aggregation strategies would also help establish whether the trade-off identified here generalizes beyond the specific architecture pair evaluated. Finally, integrating the distilled classifier with an upstream subject-detection stage into a complete two-stage cascaded pipeline, and evaluating the two stages jointly on live classroom video, would be a natural next step toward real-world deployment.

CONCLUSION

This study evaluated MobileViT and Knowledge Distillation as an efficient alternative to VideoMAE for classroom student action classification. The VideoMAE Teacher achieved 87.00 ± 2.24% test accuracy (F1 = 86.85 ± 2.20%) across three repetitions on an extended 2,160-clip five-class dataset, higher than the 68.08% VideoMAE accuracy reported in the predecessor 403-clip study, although the two were trained on different datasets and are therefore not directly comparable. Knowledge Distillation closed the gap between the compact MobileViT Student and the Teacher: the best configuration (LSTM aggregation, α=0.45, T=7) reached F1 = 87.84 ± 0.99%, a 7.01-point gain over the undistilled Student baseline and 0.99 points above the Teacher, while using about 14.7 times fewer parameters;

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

since that margin is smaller than either standard deviation, the two models are reported as close rather than ranked. In terms of speed, the best KD configuration achieved an end-to-end GPU throughput of 126.81 FPS against the Teacher's 122.16 FPS, a difference smaller than the spread measured across configurations running an identical computation graph and therefore treated as parity; the underlying pure-compute advantage of the Student (≈1.99×) was largely masked by video-decoding overhead in the processing pipeline, and the advantage was smaller still on CPU-only hardware (≈1.18×). Taken together, these findings indicate that Knowledge Distillation can bring a lightweight model's accuracy close to that of a large video transformer while reducing the parameter count by about 14.7 times, but the realized speed benefit for real-time edge deployment depends as much on pipeline efficiency as on model architecture - a consideration future two-stage cascaded classroom monitoring systems should account for explicitly.

ACKNOWLEDGMENT

The authors thank the AI Laboratory, Universitas Pendidikan Indonesia, for providing access to the HPC workstation used for model training and evaluation. The authors also thank Roshan Syalwan Nurilham and Defrizal Yahdiyan Risyad for their assistance during the additional data collection sessions, and all student participants who took part in those sessions.

REFERENCES

Ahmadabadi, H., Manzari, O. N., & Ayatollahi, A. (2023). Distilling knowledge from CNN-Transformer models for enhanced human action recognition. In 2023 13th International Conference on Computer and Knowledge Engineering (ICCKE) (pp. 180–184). https://doi.org/10.1109/ICCKE60553.2023.10326272

Alruwais, N. M., & Zakariah, M. (2024). Student recognition and activity monitoring in e-classes using deep learning in higher education. IEEE Access, 12, 66110–66128. https://doi.org/10.1109/ACCESS.2024.3354981

Atilgan, C., & Mercimek, M. (2025). Balancing precision and speed: Introducing the performance efficiency evaluation ratio (PEER) in visual odometry. IEEE Access, 13, 89874–89904. https://doi.org/10.1109/ACCESS.2025.3571921

Deshpande, A., & Deshpande, V. (2023). Student activity recognition in classroom environments using transfer learning. arXiv. https://arxiv.org/abs/2312.00348

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2625–2634).

Gan, J., & Xiong, J. (2025). Masked autoencoder of multi-scale convolution strategy combined with knowledge distillation for facial beauty prediction. Scientific Reports, 15(1), 2784. https://doi.org/10.1038/s41598-025-86831-0

Heng, C. H., Toyoura, M., Leow, C. S., & Nishizaki, H. (2024). Analysis of classroom processes based on deep learning with video and audio features. IEEE Access, 12, 110705–110712. https://doi.org/10.1109/ACCESS.2024.3434742

Herawan, M. J. A., Megasari, R., & Wihardi, Y. (2025). Klasifikasi aksi peserta didik di dalam ruang kelas menggunakan video mask auto encoder. Journal of Computer and Information Systems Ampera, 6(3), 160–172. https://doi.org/10.51519/journalcisa.v6i3.692

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv. https://arxiv.org/abs/1503.02531

Lin, F. C., Ngo, H. H., Dow, C. R., Lam, K. H., & Le, H. L. (2021). Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. Sensors, 21(16), 5314. https://doi.org/10.3390/s21165314

Liu, X., Sui, Q., & Chen, Z. (2025). Real time weed identification with enhanced MobileViT model for mobile devices. Scientific Reports, 15(1), 27323. https://doi.org/10.1038/s41598-025-12036-0

Mehta, S., & Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In Proceedings of the 10th International Conference on Learning Representations (ICLR).

Meng, S., Chen, W., & Jiang, Y. (2024). A novel strategy of two-stage cascaded CNN and overhaul knowledge distillation for fast railway foreign objects intrusion detection. IEEE Access, 12, 127825–127835. https://doi.org/10.1109/ACCESS.2024.3457515

Sassi, A., Jaafar, W., Cherif, S., Ben Abderrazak, J., & Yanikomeroglu, H. (2023). Video traffic analysis for real-time emotion recognition and visualization in online learning. IEEE Access, 11, 99376–99386. https://doi.org/10.1109/ACCESS.2023.3313973

Sharma, V., Gupta, M., Kumar, A., & Mishra, D. (2021). EduNet: A new video dataset for understanding human activity in the classroom environment. Sensors, 21(17), 5699. https://doi.org/10.3390/s21175699

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9915

Shou, Z., Yuan, X., Li, D., Mo, J., Zhang, H., Zhang, J., & Wu, Z. (2024). A dynamic position embedding-based model for student classroom complete meta-action recognition. Sensors, 24(16), 5371. https://doi.org/10.3390/s24165371

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Advances in Neural Information Processing Systems (Vol. 35). Curran Associates.

Wihardi, Y., Naufal, R. A., Herawan, M. J. A., & Megasari, R. (2026). Fine-grained classroom activity recognition via detection, head-pose, and appearance fusion. Brilliance: Research of Artificial Intelligence, 6(3), 511–524. https://doi.org/10.47709/brilliance.v6i3.9453

Yan, H., Han, R., Feng, W., Zhao, J., & Wang, S. (2025). Weakly supervised instance action recognition. Computational Visual Media, 11(3), 603–618. https://doi.org/10.26599/CVM.2025.9450390

Yu, B. X. B., Liu, Y., & Chan, K. C. C. (2021). Multimodal fusion via teacher-student network for indoor action recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 35(4), 3199–3207. https://doi.org/10.1609/aaai.v35i4.16430

Zhang, Y., & Wang, Y. (2025). A comprehensive survey on RGB-D-based human action recognition: Algorithms, datasets, and popular applications. EURASIP Journal on Image and Video Processing, 2025(1), 15. https://doi.org/10.1186/s13640-025-00677-0

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.