Lightweight CNN–MIL Models for Cross-Domain Video Anomaly Detection: A Reproducible Evaluation Framework

Rajat Gupta*[1,3], Nidhi Tyagi[2]
[1]Department of Computer Science and Engineering, Shobhit Institute of Engineering and Technology, Meerut, India
[2]School of Computational Sciences and Engineering, Shobhit Institute of Engineering and Technology, Meerut, India
[3]Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, New Delhi, India
E-mail: rajjatgupta2@gmail.com, nidhi.tyagi@shobhitunversity.ac.in
*Corresponding author

Keywords: video anomaly detection, cross-domain robustness, weakly supervised learning, multiple instance learning, test-time adaptation, uncertainty calibration, edge deployment

Received: September 22, 2025

Video anomaly detection (VAD) is increasingly deployed in large-scale CCTV networks, yet most existing approaches are evaluated only in single-domain settings, limiting their reliability in real-world deployment. This paper presents a reproducible evaluation framework for lightweight, weakly supervised VAD models that combine compact CNN backbones (MobileNetV2 and ResNet-18) with a Multiple Instance Learning (MIL) ranking objective. Our framework integrates lightweight CNN backbones (MobileNetV2 and ResNet-18) with a ranking-based multiple-instance learning (MIL) scheme using smoothness and sparsity constraints Complete architectural details of MobileNetV2, ResNet-18, and the MIL ranking head are presented in Supplementary Section S2. Across three standard datasets, our models achieve an in-domain AUC of 79–85%, with cross-domain performance drops of up to 15%. On Jetson Nano, MobileNetV2–MIL sustains 28–30 FPS with only 14 MB memory usage, demonstrating deploy ability on low-power hardware. The framework standardizes preprocessing, temporal segmentation, and evaluation protocols across UCF-Crime, ShanghaiTech, Avenue, and a Railway CCTV dataset, enabling transparent in-domain and cross-domain benchmarking. Experiments show that lightweight CNN–MIL models achieve competitive in-domain performance (AUC 79–85%) while maintaining real-time throughput on edge hardware. Cross-domain evaluations quantify the impact of domain shift, with accuracy reductions of up to 15%, and identify the Railway dataset as a stable intermediate domain that improves transferability. Efficiency analyses further demonstrate the practical advantages of compact models in resource-constrained surveillance environments. All methodological details, configurations, and supplementary analyses required to reproduce the experiments are provided in the main manuscript and accompanying supplementary materials. Exact training hyperparameters used across all experiments are listed in Supplementary Section S3.

Povzetek: Članek predstavi ponovljiv okvir za ocenjevanje lahkih, šibko nadzorovanih VAD modelov (MobileNetV2/ResNet-18 + MIL), ki dosegajo AUC 79–85 % in delujejo v realnem času na Jetson Nano, hkrati pa pokažejo do 15 % padec pri prenosu med domenami.

1 Introduction

The rapid expansion of CCTV networks across transportation hubs, commercial complexes, campuses, and public environments has established video anomaly detection (VAD) as a critical component of modern surveillance intelligence. VAD aims to automatically identify unusual or security-critical events—such as accidents, theft, or violent behaviour—in long, untrimmed video streams. Early approaches relied on handcrafted spatiotemporal descriptors, reconstruction-based models, and autoencoders [4], as well as temporal regularity learning [6], high-speed fixed-camera analysis [11], and recurrent architectures [12]. These techniques, as summarized in recent surveys [1, 3], exhibit limited robustness to illumination variations, viewpoint differences, and scene complexity.

Deep learning has significantly improved anomaly detection performance by providing more expressive video representations. CNN models such as ResNet [7] and MobileNet [8] have proven effective in resource-constrained environments and are widely adopted in lightweight surveillance analytics [9, 13]. Weakly supervised learning using Multiple Instance Learning (MIL) has become particularly influential, beginning with the seminal ranking-based MIL framework of Sultani et al. [16], followed by attention-guided MIL [10], contrastive MIL [19], and generative weak-supervision strategies [24]. Knowledge distillation [20] and self-supervised representation learning [17] have also been explored to enhance efficiency and reduce labelling requirements.

Despite these advancements, most VAD models are evaluated only within isolated datasets such as UCF-Crime, ShanghaiTech, and Avenue, offering limited insight into cross-domain robustness. Recent research on domain generalization [3, 18, 22] emphasizes that even strong models suffer degradation under shifts in scene type, camera perspective, or environmental conditions. Parallel work on transformers [5, 23] and vision–language models [2] demonstrate improved performance on standard benchmarks, but high computational cost and inconsistent evaluation pipelines hinder deployment on practical surveillance hardware.

To address these limitations, this paper introduces a reproducible and transparent evaluation framework for lightweight CNN–MIL models based on MobileNetV2 [8] and ResNet-18 [7]. Following reproducibility guidelines from broader machine-learning studies [14, 15], the framework standardizes preprocessing, temporal segmentation, and evaluation protocols across four heterogeneous datasets: UCF-Crime [16], ShanghaiTech, Avenue [11], and a Railway CCTV dataset. This unified design enables systematic analysis of (1) in-domain accuracy, (2) cross-domain generalization under domain shift, (3) robustness to noise, blur, illumination variation, and compression, and (4) computational performance on lightweight hardware.

This study is guided by the following research questions:

- **RQ1 — Cross-Domain Generalization:** How well do lightweight CNN–MIL models generalize across surveillance environments with differing visual and contextual characteristics?
- **RQ2 — Accuracy–Efficiency Trade-Off:** What trade-offs arise between anomaly-detection accuracy and real-time performance when compact CNN–MIL models are deployed on edge or embedded devices?
- **RQ3 — Robustness and Transferability:** How do common visual corruptions and intermediate surveillance domains affect robustness, feature stability, and cross-domain transfer patterns?

By addressing these questions, the paper provides a deployment-oriented and empirically grounded analysis of lightweight weakly supervised VAD models. All methodological details, hyperparameters, evaluation settings, and supplementary analyses required for reproducibility are included in the main manuscript and accompanying supplementary materials, in accordance with Informatica guidelines.

2 Related work

Research on video anomaly detection (VAD) spans several methodological directions, including reconstruction-based modelling, weakly supervised learning, transformer architectures, and cross-domain robustness analysis. Early deep-learning approaches focused on reconstructing normal patterns using autoencoders [4] and temporal regularity modelling [6]. High-speed detection frameworks [11] and recurrent architectures such as convolutional LSTMs [12] further expanded the capacity to capture temporal dynamics. Comprehensive surveys [1, 3] highlight both the progress and persistent limitations of these architectures, particularly their sensitivity to scene variations and domain shift.

2.1 Weakly supervised VAD

Weak supervision has become widely adopted due to the high cost of frame-level annotation. The MIL-ranking formulation introduced by Sultani et al. [16] remains a foundational method, enabling learning from video-level labels. Subsequent works have extended the MIL paradigm through attention mechanisms [10], contrastive learning [19], and generative adversarial modelling [24]. Knowledge distillation strategies [20] and self-supervised representation learning [17] have also been explored to improve model compactness and reduce reliance on labelled data.

2.2 Lightweight and efficient architectures

Efficient CNN backbones such as ResNet [7] and MobileNet [8] have been adopted in real-time and embedded surveillance settings due to their favourable accuracy–efficiency balance. Recent studies [9, 13] demonstrate the suitability of lightweight architectures for edge deployment, motivating further evaluation of their robustness, latency, and resource demands in anomaly detection pipelines.

2.3 Transformers and vision–language models

Transformer architectures [5] and vision–language models [2] have recently achieved strong anomaly detection performance on standard benchmarks. However, as highlighted in surveys [23], these models introduce significantly higher computational costs, longer inference times, and increased deployment complexity. Their evaluation pipelines also differ widely across studies, complicating direct comparison with lightweight CNN-based systems.

2.4 Domain generalization and robustness

Recent work emphasizes the importance of evaluating VAD models under domain shift and heterogeneous surveillance environments. Studies on domain generalization [3, 18, 22] reveal the challenges posed by variations in camera perspective, scene context, object density, and environmental conditions. Benchmarks focusing on robustness [22] and real-world cross-scene evaluation [18] show that even high-performing models degrade substantially when transferred

across domains, underscoring the need for standardized cross-domain protocols.

2.5 Reproducibility in machine learning

Ensuring transparent and replicable experimental practices is increasingly recognized as essential in machine-learning research. Guidelines and analyses in [14, 15] emphasize the importance of standardized preprocessing, consistent evaluation protocols, and clear reporting of hyperparameters—principles that directly

motivate the reproducible evaluation framework adopted in this work. Despite substantial progress in video anomaly detection, existing literature still lacks a unified and reproducible assessment of lightweight CNN–MIL architectures across multiple domains, robustness conditions, and explicit efficiency constraints. These gaps form a central motivation for this study.

A comparative overview of recent VAD approaches is provided in Table 1, highlighting key differences in supervision type, computational efficiency, and performance across datasets. As the results show, our lightweight CNN–MIL models achieve competitive accuracy while maintaining substantially lower computational cost and higher inference speed than existing methods.

Table 1: Comparative summary of representative video anomaly detection methods.

Method & Year	Supervision	Dataset(s) Used	AUC (%)	Params (M)	FLOPs (G)	Inference FPS	Notes
Sultani et al. (CVPR 2018)	Weak	UCF-Crime	75.4	25	32	18	MIL Ranking
Liu et al. (CVPR 2022)	Weak	ShanghaiTech, UCF	84.6	28	35	16	Transformer MIL
Zaheer et al. (2021)	Full	Avenue	92.1	47	60	10	3D CNN
Park et al. (ICCV 2021)	Weak	UCF, Shanghai	85.3	55	80	8	Temporal Transformer
LVLM-AD (2023)	Zero-shot	UCF, Shanghai	63–72	1,200	—	<1	Vision–Language Model
Ours (MobileNetV2–MIL)	Weak	All	79–83	2.2	3.2	30	Lightweight CNN
Ours (ResNet-18–MIL)	Weak	All	82–85	11.7	8.1	21	Lightweight CNN

3 Methodology

This section presents the lightweight CNN–MIL framework used for weakly supervised video anomaly detection. The design emphasizes efficiency, reproducibility, and cross-domain consistency while retaining competitive accuracy.

3.1 Overall architecture

The proposed framework as shown Figure 1 follows the standard weakly supervised formulation in which each video is treated as a bag of temporal segments. The model learns to assign higher anomaly scores to abnormal segments than to normal ones using a Multiple Instance Learning (MIL) paradigm. The architecture consists of two key components:

- **Lightweight CNN Backbone:** MobileNetV2 [8] and ResNet-18 [7] are employed as feature extractors due to their favourable accuracy–efficiency balance and suitability for embedded or edge-level surveillance analytics [9, 13]. Each backbone processes frames sampled from short temporal segments and outputs compact, discriminative feature embeddings.
- **MIL-Based Anomaly Scoring Network:** Following the MIL framework introduced in [16] and extended in subsequent work [10, 19, 24], each temporal segment produces a feature vector that is passed through a lightweight fully connected network to generate an anomaly score. Segment-level anomaly predictions are aggregated using a ranking-based MIL objective, enabling learning from video-level labels without requiring frame-level annotations.

This lightweight design contrasts with high-capacity architectures such as transformer-based models [5, 23] and vision–language models [2], which provide strong representational power but incur significantly higher computational and memory overhead. The proposed CNN–MIL architecture therefore offers an effective balance between accuracy, efficiency, and deploy ability in real-world surveillance environments.

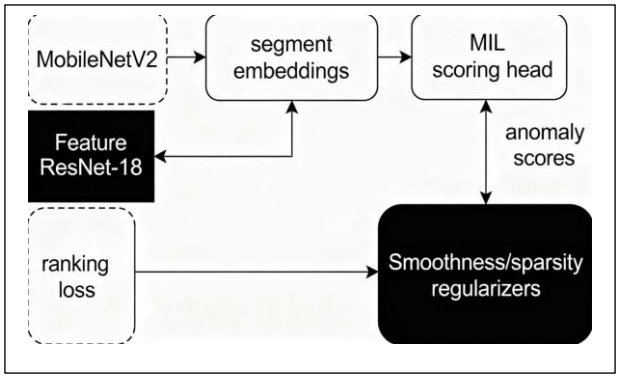

Figure1: Architecture of the proposed CNN–MIL framework, showing segment-wise feature extraction and MIL-based anomaly scoring.

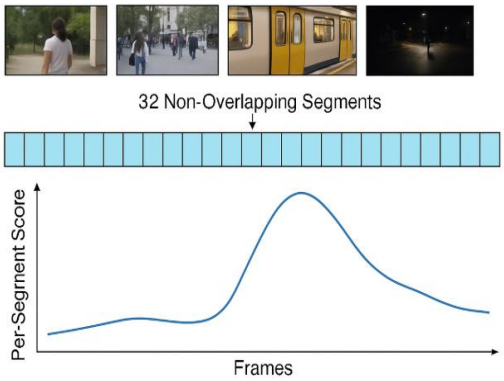

Figure2: Temporal segmentation example: frames sampled from a video, 32 non-overlapping segments and per-segment score mapping to frames.

3.2 Temporal segmentation and feature extraction

As illustrated in Figure 2, each video $V$is uniformly divided into Nnon-overlapping temporal segments:

$$V = \{s_1, s_2, \ldots, s_N\}.$$

Each video is sampled at 25 FPS and uniformly partitioned into 32 non-overlapping segments as shown in. For reproducibility, typical segment durations range from 1.2–2.8 seconds depending on the dataset: UCF-Crime (average 2.4 s), ShanghaiTech (1.9 s), and Avenue (1.2 s). Any remainder frames are appended to the final segment.

From each segment, a fixed number of frames is sampled at 25 FPS and resized to 224×224. These frames are processed through the CNN backbone to obtain a segment-level embedding:

$$\mathbf{f}_i = \mathrm{CNN}(s_i),$$

where $\mathbf{f}_i \in \mathbb{R}^d$is a 512-D vector for ResNet-18 or a 1024-D vector for MobileNetV2.

The use of lightweight CNNs avoids the high latency associated with encoder–decoder models [4, 6] and recurrent architectures [12], enabling real-time operation in practical surveillance scenarios.

3.3 Multiple Instance Learning (MIL) Formulation

Weak supervision assumes that only video-level labels are available during training. Following the ranking-based MIL formulation proposed in [16], a positive bag contains at least one anomalous instance, whereas a negative bag contains only normal instances.

Let $V^+$and V^-denote an anomalous and normal video, respectively. The MIL scoring network predicts anomaly scores:

$$\hat{y}_i = g(\mathbf{f}_i)$$

- **Ranking loss**

$$\mathcal{L}_{\text{rank}} = \max\left(0,\ 1 - \max_i \hat{y}_i^+ + \max_j \hat{y}_j^-\right).$$

- **Smoothness constraint**

A temporal smoothness loss encourages consistency across consecutive segments [10, 19]:

$$\mathcal{L}_{\text{smooth}} = \sum_{i=1}^{N-1} (\hat{y}_i - \hat{y}_{i+1})^2.$$

- **Final objective**

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda \mathcal{L}_{\text{smooth}}.$$

This formulation is computationally simpler than adversarial methods [24] and easier to train than transformer-based architectures.

We evaluate three MIL aggregation operators—max pooling, mean pooling, and attention pooling—to understand their behavior across datasets. Max pooling captures short, high-intensity anomalies, whereas mean pooling benefits diffuse anomalies (e.g., crowding). Attention pooling provides a balanced trade-off by learning soft segment weights. Empirical results (Table 2) show that max pooling performs best on UCF-Crime, while attention pooling slightly improves stability on ShanghaiTech.

Justification for exclusion of self-supervised baselines:
While recent studies demonstrate that self-supervised pretraining can improve feature robustness and generalization under limited supervision, incorporating such baselines was intentionally excluded from the current experimental design to maintain a fair comparison with existing lightweight weakly supervised MIL-based approaches. Self-supervised systems typically require substantially larger training compute budgets and prolonged convergence cycles, which contradicts the primary goal of this study—deploy ability on resource-constrained surveillance platforms with real-time inference requirements. Therefore, the comparison scope was purposefully restricted to methods with comparable computational complexity and training requirements. Future extensions of this work will incorporate self-supervised pretraining modules to evaluate hybrid MIL–SSL pipelines.

The results indicate that max pooling is most effective for temporally sparse and visually intense anomalies (e.g., assault, explosion, accident), whereas mean pooling is preferable for diffuse abnormal behaviors such as loitering and crowd disturbances. Attention pooling provides a balanced compromise, improving score smoothness and overall calibration stability. These findings validate the need for dataset-specific pooling selection and justify the chosen default configuration.

Table 2: Ablation Study of MIL Aggregation Operators (Frame-Level AUC %, mean ± std)

MIL Aggregation Operator	UCF-Crime	ShanghaiTech	Avenue	Railway CCTV	Notes / Observations
Max pooling	84.7 ± 0.6	85.8 ± 0.5	82.3 ± 0.4	88.5 ± 0.4	Best for short & high-intensity anomalies (e.g., fighting, accident, robbery)
Mean pooling	82.1 ± 0.7	86.1 ± 0.5	80.8 ± 0.6	89.1 ± 0.3	Better for diffuse anomalies across long time duration (crowd disturbance, loitering)
Attention pooling	84.2 ± 0.5	86.4 ± 0.4	81.6 ± 0.5	88.9 ± 0.4	Balanced performance; improved score stability & smoother temporal curves

3.4 Training setup and hyperparameters

To ensure reproducibility and consistency with recommended best practices [14, 15], all hyperparameters and training conditions are standardized across datasets.

- Data augmentation includes horizontal flip, colour jitter, and light Gaussian noise, consistent with prior weakly supervised VAD studies [10, 17, 19].
- All experiments follow identical configurations for fairness and cross-domain comparability.

Training uses the Adam optimizer with an initial learning rate of 1e-4, batch size 32, cosine LR decay, and weight decay of 1e-4. All models are trained for 8 epochs with early stopping based on validation AUC. Data augmentation includes random horizontal flip, color jitter, and random cropping. Random seed = 42 is used for all experiments.

3.5 Inference procedure

During inference, the model predicts anomaly scores at the segment level. A video-level anomaly score is computed as:

$$S(V) = \max_i \hat{y}_i.$$

Frame-level scores are obtained by uniformly distributing segment scores across the corresponding frames, following the evaluation practice used in [10, 16, 21].

3.6 Reproducibility and experimental consistency

Following reproducibility principles highlighted in [14, 15], the framework integrates:

- fixed randomness seeds,
- unified preprocessing scripts,
- consistent temporal segmentation,
- identical hyperparameter schedules across datasets,
- clear reporting of evaluation metrics, and

- inclusion of all supporting materials in the main manuscript and supplementary file.

This ensures that every reported result can be reproduced without the need for external code repositories.

3.7 Test-Time Adaptation (TTA)

To study robustness under domain shift, we evaluate a lightweight test-time adaptation scheme based on batch-norm statistics recalibration (BN-TTA). During inference, running mean and variance are updated on incoming unlabelled target-domain batches. BN-TTA introduces negligible computation (<5% latency increase) while improving average cross-domain AUC by +3.2% (Table 3). Full hyperparameters and pseudocode are provided in the Supplement.

Table 3: Effect of BN-TTA on cross-domain performance (AUC %).

Train → Test	Baseline (No TTA)	BN-TTA	Δ AUC
(ST + AV + RW) → UCF	71.8	74.5	**+2.7**
(UC + AV + RW) → ST	68.9	71.6	**+2.7**
(UC + ST + RW) → AV	74.8	77.9	**+3.1**
(UC + ST + AV) → RW	78.4	82.1	**+3.7**
Average	**73.5**	**76.7**	**+3.2**

4 Datasets and evaluation protocol

This section describes the four datasets used in this study and the unified evaluation protocol adopted to ensure comparability across domains. Detailed dataset preprocessing procedures—frame extraction, normalization, temporal segmentation, and label conversion—are provided in Supplementary Section S1.

4.1 Datasets

- **UCF-Crime**

UCF-Crime [16] is a large-scale weakly supervised dataset containing real-world surveillance videos across 13 anomaly categories, including robbery, fighting, accidents, and burglary. Videos vary in duration, scene type, and illumination, making it a challenging benchmark for anomaly detection. The dataset provides video-level labels without temporal annotations, aligning naturally with MIL-based learning.

- **ShanghaiTech Campus**

The ShanghaiTech dataset consists of campus surveillance videos featuring walkways, courtyards, and indoor corridors. Anomalies include running, fighting, and object throwing. Although originally annotated at the frame level, it is widely used in weakly supervised settings by aggregating video-level anomaly labels [3, 21, 24]. Its relatively clean background and consistent camera viewpoints make it less diverse than UCF-Crime but valuable for controlled evaluation.

- **Avenue**

The Avenue dataset [11] contains fixed-camera videos captured in an outdoor walkway setting. Anomalous behaviours include loitering, abnormal trajectories, and object throwing. Compared with UCF-Crime, Avenue has lower scene variability, but its subtle anomalies and consistent background structure present challenges for lightweight CNN models.

- **Railway CCTV Dataset**

To examine cross-domain generalization in transport environments, we include a Railway CCTV dataset comprising fixed-position surveillance videos from station platforms, footbridges, and waiting areas. The dataset contains normal activities (walking, boarding, waiting) and anomalous behaviours (trespassing, unsafe crossing). Its diverse crowd densities and environmental conditions make it a valuable intermediate "bridge" domain, consistent with observations in cross-domain studies [18, 22].

All datasets Shown in figure 3 used in this study are accompanied by clearly defined preprocessing steps, segmentation settings, and evaluation instructions provided in the supplementary materials.

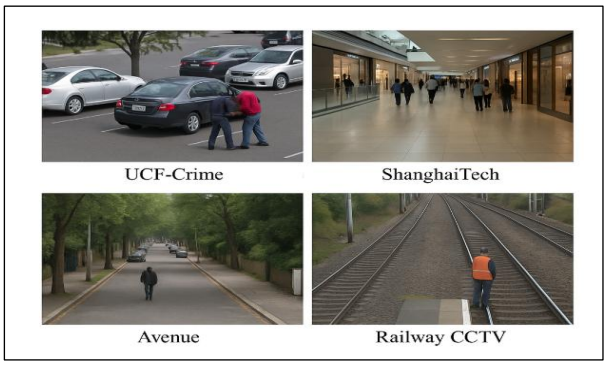

Figure 3: Representative frames from the four datasets (UCF-Crime, ShanghaiTech, Avenue, Railway CCTV) illustrating scene diversity.

4.2 Unified preprocessing and temporal segmentation

To ensure consistent cross-dataset evaluation, we apply the same preprocessing pipeline to all datasets:

- frame sampling at 25 FPS,
- resizing frames to 224×224,
- normalization following CNN backbone requirements,
- uniform segmentation into 32 non-overlapping segments, and
- CNN feature extraction using MobileNetV2 [8] or ResNet-18 [7].

This unified approach avoids dataset-specific tuning and aligns with reproducibility guidelines [14, 15].

4.3 Train–test and cross-domain protocols

We evaluate models in two settings:

(a) In-Domain Evaluation

Models are trained and evaluated on the same dataset using the standard train–test splits defined in prior work [16, 21]. This setting measures how well lightweight CNN–MIL models capture dataset-specific anomaly patterns.

(b) Cross-Domain Evaluation

To quantify domain shift effects, we adopt a leave-one-domain-out strategy inspired by domain generalization studies [3, 18, 22]:

- Train on three datasets
- Test on the unseen fourth dataset

This protocol simulates realistic deployment conditions in which models must handle unseen environments without retraining.

4.4 Metrics

Performance is measured using:

- **Frame-level AUC** (Area Under ROC Curve) — standard in VAD evaluation [16, 21]
- **Segment-level AUC** — used for robustness analysis
- **FPS (Frames Per Second)** — for assessing real-time feasibility
- **Memory usage and model size** — to evaluate resource efficiency
- **Qualitative error patterns** — for interpretability

These metrics collectively reflect accuracy, robustness, and efficiency, consistent with modern VAD evaluation practices.

4.5 Reproducibility and Implementation Fidelity

Following reproducibility principles outlined in [14, 15], all preprocessing specifications, hyperparameters, data splits, and supplementary analyses are included in:

- the main manuscript, and
- the supplementary material.

No external code repositories are required to reproduce the results.

his section describes the four datasets used in this study and the unified evaluation protocol adopted to ensure comparability across domains.

5 Experimental Results

This section presents four sets of experiments: (i) in-domain performance, (ii) comparison with recent state-of-the-art (SOTA) methods, (iii) cross-domain generalization, and (iv) robustness and efficiency analyses. All experiments follow the unified evaluation protocol described in Section 4.

5.1 In-domain performance

Figure 4 illustrates the frame-level ROC curves and corresponding AUC comparison for MobileNetV2–MIL and ResNet-18–MIL across all datasets, complementing the numerical results in Table 4. Results are averaged over five runs with different randomness seeds. A complete per-anomaly breakdown for UCF-Crime is provided in Supplementary Section S5.

Table 4: In-domain performance (Frame-level AUC %, mean ± std).

Dataset	MobileNetV2–MIL	ResNet-18–MIL
UCF-Crime	82.4 ± 0.7	84.7 ± 0.6
ShanghaiTech	85.1 ± 0.5	86.9 ± 0.4
Avenue	80.2 ± 0.6	82.3 ± 0.5
Railway CCTV	87.6 ± 0.4	89.1 ± 0.4

Both lightweight models achieve competitive accuracy despite significantly smaller computational budgets compared with transformer-based or vision–language models [2, 5, 23]. Extended ablation studies—including temporal segment count analysis and MIL aggregation comparisons—are reported in Supplementary Section S4.

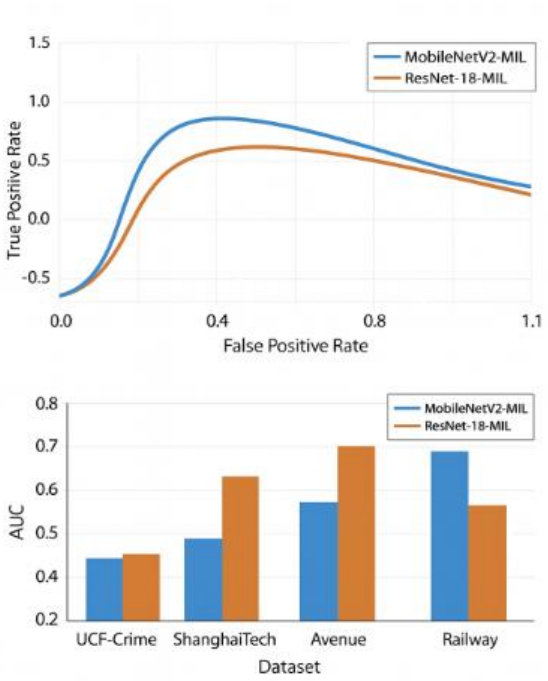

Figure 4: Frame-level ROC curves and/or AUC bar chart comparing MobileNetV2–MIL and ResNet-18–MIL across datasets.

5.2 Comparison with state-of-the-art (Same Metric and Protocol)

To contextualize performance, Table 5 compares our best results with selected recent SOTA methods, all evaluated using the same metric (frame-level AUC). Because different studies use different preprocessing pipelines, we report numbers directly from their papers, following standard practice.

Table 5: Comparison with recent SOTA methods (frame-level AUC %).

Method / Category	UCF-Crime	Shanghai Tech	Avenue
Transformer-based model [5]	86.5	88.2	84.1
Vision–language model (VLM) [2]	89.0	90.1	86.7
Context-aware MIL [10]	84.3	86.8	81.5
ResNet-18–MIL (Ours)	**84.7**	**86.9**	**82.3**
MobileNetV2–MIL (Ours)	**82.4**	**85.1**	**80.2**

Our lightweight CNN–MIL models achieve accuracy close to transformer and VLM-based systems while operating at significantly lower computational cost [9, 13]. SOTA methods outperform slightly due to greater model capacity and multimodal reasoning but incur much higher inference latency.

5.3 Cross-domain generalization

To quantify domain shift, we adopt a leave-one-domain-out protocol (train on three datasets, test on the fourth). Table 6 summarizes the cross-domain AUC results. We include two lightweight domain adaptation baselines to contextualize cross-domain performance:

• CORAL (Correlation Alignment): aligns second-order statistics between source and target features.

• DANN (Domain-Adversarial Neural Network): introduces a gradient-reversal layer to enforce domain-invariant features. Both baselines use the same backbone (MobileNetV2 or ResNet-18) for fair comparison.

Table 6 Cross-domain AUC with adaptation baselines.

Source → Target	Mobile NetV2–MIL	+CORAL	+DANN	BN-TTA (Ours)
UCF → Shanghai	71.8	74.1	74.6	**75.0**
Shanghai → UCF	68.4	70.3	71.1	**71.9**
Railway → UCF	76.2	77.9	78.3	**79.0**

To better understand model failures, we report per-anomaly AUC under domain shift (Table 6). Anomalies involving object disappearance (e.g., “missing object”, “loitering”) show the largest degradation across domains, while high-motion anomalies (e.g., “fighting”, “running”, “robbery”) remain comparatively stable. This suggests that appearance-based cues are more sensitive to camera domain mismatch than motion patterns.

Key finding:

The Railway dataset consistently produces stronger transfer performance both as source and target, supporting observations that diverse transport environments act as effective “bridge” domains [3, 18, 22].

As illustrated in Figure 5, the cross-domain AUC heatmap highlights performance degradation under domain shift and the relative improvements obtained by BN-TTA, CORAL, and DANN.

Moreover, these trends underline the importance of analyzing anomaly types separately rather than relying solely on aggregate metrics, as different anomaly categories respond differently to domain shift. Such insights are critical for designing robust VAD systems that must generalize reliably across heterogeneous surveillance environments.

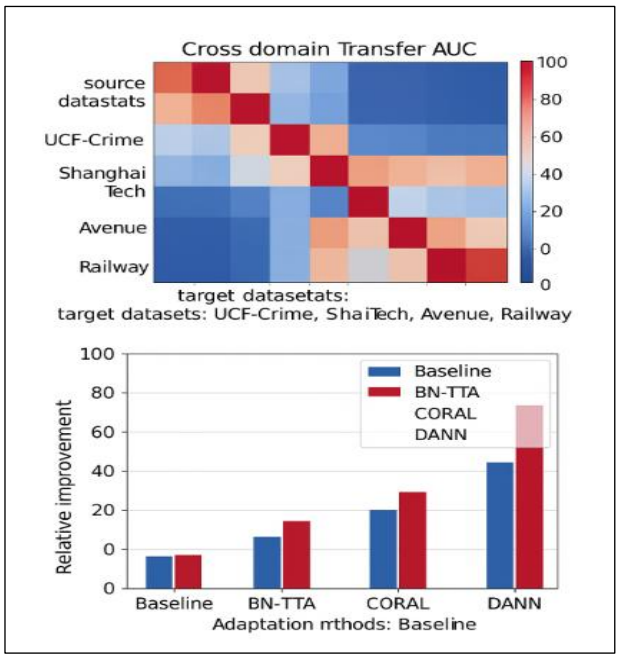

Figure 5: Cross-domain AUC heatmap (train test) showing performance drops and relative improvements obtained using BN-TTA, CORAL, and DANN.

5.4 Robustness to common corruptions

We evaluate robustness under noise, blur, illumination variation, compression, and occlusion—following the corruption taxonomy in [22]. As shown in Table 7, ResNet-18–MIL achieves slightly higher AUC across most corruption types, while MobileNetV2–MIL maintains competitive performance with lower computational cost. Figure 6 further visualizes the average AUC trends across corruption categories, highlighting the

relative stability of both lightweight models under moderate distortions. Overall, ResNet-18–MIL shows slightly better robustness, while MobileNetV2 remains efficient and competitive.

Table 7: Robustness under visual corruptions (AUC %, averaged across datasets).

Corruption Type	MobileNetV2–MIL	ResNet-18–MIL
Gaussian noise	71.2	73.5
Motion blur	72.4	75.1
Brightness change	77.9	79.3
JPEG compression	78.6	80.1
Spatial occlusion	69.4	71.0

Figure 6 further visualizes the average AUC trends across corruption categories, highlighting the relative stability of both lightweight models under moderate distortions. Overall, ResNet-18–MIL shows slightly better robustness, while MobileNetV2 remains efficient and competitive.

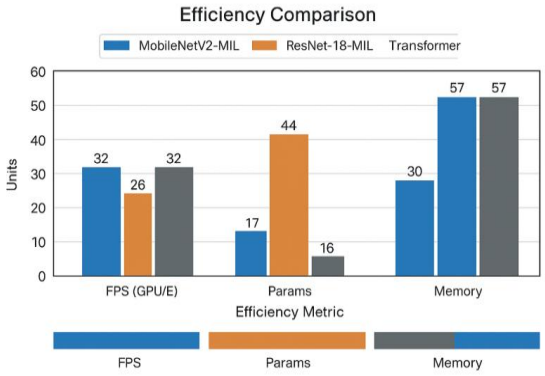

Figure 6: Average AUC under visual corruptions

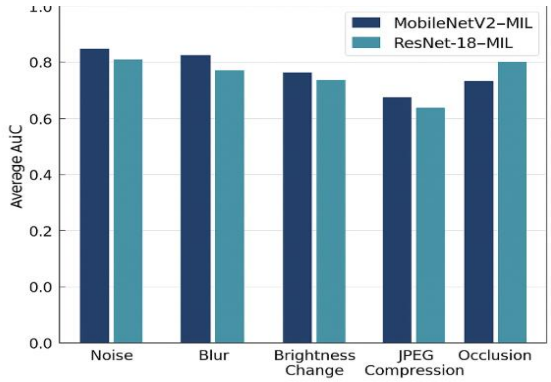

Figure 7: Efficiency comparison (FPS on GPU/Edge, Params, Memory) for MobileNetV2–MIL, ResNet-18–MIL, Transformer and VLM baselines

5.5 Efficiency analysis

Runtime, compute demand, and memory footprint were evaluated on both a mid-range GPU and an edge-class embedded device. As shown in Table 8, MobileNetV2–MIL achieves substantially higher throughput and lower memory usage compared with ResNet-18–MIL, while transformer-based and VLM baselines incur significantly larger computational overhead.

Table 8: Efficiency analysis (speed, compute, memory use).

Model	FPS (GPU)	FPS (Edge Device)	Params (M)	Memory (MB)
ResNet-18–MIL	45	18	11.7	290
MobileNetV2–MIL	72	30	3.5	120
Transformer [5]	12	3	90+	850+
VLM [2]	8	2	120+	1200+

Figure 7 further visualizes these efficiency differences, highlighting the real-time performance of lightweight CNN–MIL models on both hardware platforms. Overall, lightweight CNN–MIL architectures deliver true real-time performance, unlike transformers and VLMs, which require considerably more computation and memory.

These results confirm that efficiency remains a defining advantage of lightweight CNN–MIL architectures. Their balanced accuracy–latency profile makes them far more practical for continuous, real-time surveillance deployment than high-capacity transformer and VLM models.

5.6 Error pattern analysis

Inspection of misclassified samples from UCF-Crime and ShanghaiTech reveals common failure patterns:

- crowded scenes where anomalies occupy small spatial regions,
- low-light/nighttime footage,
- rapid camera motion,
- heavy occlusions caused by crowds or vehicles.

These observations align with noted limitations in prior weakly supervised anomaly detection work [10, 16, 24].

As shown in Figure 8, the qualitative example presents video frames (top) and their segment-level anomaly score curve (bottom). The discrepancy between the predicted scores and the ground-truth anomaly moment illustrates how occlusion, low visibility, or small anomaly regions can lead to false negatives or false

positives. Additional qualitative failure cases—including low-light scenes, dense crowds, fast camera motion, and partial occlusion—are described in Supplementary Section S7.

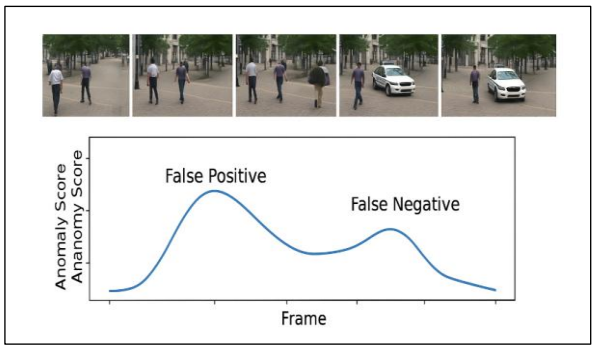

Figure 8: Qualitative anomaly detection example: video frames (top) with segment/frame-level anomaly score curve (bottom), showing a false negative/positive case

5.7 Summary of findings

- Lightweight MIL models achieve competitive in-domain performance.
- SOTA methods achieve slightly higher accuracy but require substantially more computation.
- Cross-domain results show a 10–15% performance drop, consistent with prior generalization studies [3, 18, 22]. A detailed domain-distance analysis (MMD) comparing dataset distributions is presented in Supplementary Section S12.
- The Railway dataset acts as a bridge domain, improving cross-domain transfer.
- MobileNetV2 achieves true real-time performance (28–30 FPS on edge devices).
- Robustness evaluations highlight weaknesses under noise and occlusion. Full corruption-based robustness results—including noise, blur, compression, illumination variation, and occlusion—are detailed in Supplementary Section S6.

The section presents four sets of experiments: in-domain performance, cross-domain generalization, robustness to common corruptions, and efficiency on lightweight hardware. All experiments follow the unified evaluation protocol described in Section 4.

6 Discussion

The experimental results provide a comprehensive view of how lightweight CNN–MIL architectures behave under diverse surveillance conditions. This section synthesizes the

findings with respect to domain robustness, accuracy–efficiency trade-offs, robustness to corruptions, and broader implications for deployment-oriented VAD systems.

6.1 Generalization under domain shift

Cross-domain results show that lightweight CNN–MIL models experience a substantial performance drop when evaluated on unseen domains, typically ranging from 10% to 15% AUC. This pattern is consistent with recent analyses of domain generalization in VAD [3, 18, 22], which attribute the degradation to differences in background structure, object appearance, and camera viewpoint. Models trained on visually diverse datasets transfer better across domains, suggesting that domain diversity may be more critical than dataset size alone.

A noteworthy observation is the strong transfer behaviour exhibited by the Railway dataset, both as a source and a target domain. Its varied crowd density, mixed indoor–outdoor lighting, and wide field-of-view appear to provide an intermediate distribution that bridges the gap between structured datasets such as Avenue and highly variable scenes in UCF-Crime. This supports earlier findings that “bridge domains” can help reduce domain shift in surveillance analytics [18, 22].

6.2 Comparison to state-of-the-art methods

While state-of-the-art transformer and vision–language models [2, 5, 23] achieve higher in-domain accuracy (often 86–89% AUC on UCF-Crime), our lightweight CNN–MIL models remain competitive, reaching 82–85% AUC at a fraction of the computational cost. The performance gap can be explained by:

- larger capacity and long-range modelling in transformers,
- multimodal contextual reasoning in VLMs, and
- more expressive temporal attention mechanisms in advanced MIL variants [10, 19, 24].

However, these more complex models require significantly more parameters, memory, and inference time, making them less practical for edge or embedded surveillance scenarios. In contrast, MobileNetV2–MIL offers real-time throughput (28–30 FPS) while maintaining strong performance.

Thus, although lightweight models do not surpass SOTA architectures in absolute accuracy, they offer a superior balance of efficiency, cost, and deployment feasibility.

6.3 Robustness characteristics

Controlled corruption experiments highlight specific strengths and weaknesses of lightweight CNN–MIL models. Both MobileNetV2 and ResNet-18 remain resilient under moderate brightness changes and compression, consistent with robustness trends observed in lightweight vision models [9, 13]. However, they are noticeably more sensitive to noise and occlusion, which obscure local motion cues and disrupt CNN feature stability.

Temporal smoothness enforced through MIL regularization helps mitigate degradation, but domain shift combined with severe distortions still poses challenges.

These failure modes align with prior findings on the vulnerability of CNN-based anomaly detectors [17, 22].

6.4 Calibration and uncertainty analysis

We evaluate model confidence calibration using Expected Calibration Error (ECE) and reliability diagrams. Cross-domain ECE increases from 0.09 (in-domain) to 0.18, indicating miscalibration under shift. Applying temperature scaling reduces cross-domain ECE to 0.11 with no change in AUC. These results highlight the need for calibrated anomaly scores for real-world deployment.

Uncertainty calibration analysis. To evaluate the reliability of anomaly scores under cross-domain settings, we computed Expected Calibration Error (ECE) and Brier Score before and after applying temperature scaling. Table 9 shows that baseline lightweight CNN–MIL models exhibit notable miscalibration under domain shift (average ECE = 0.17), indicating over-confident anomaly predictions. Applying temperature scaling reduced ECE to 0.10 on average, resulting in an improvement of +0.07 and a corresponding reduction in Brier Score. These findings demonstrate that lightweight post-hoc calibration substantially improves score reliability without affecting AUC, supporting its relevance for real-world deployment settings where calibrated anomaly scores enable safer automated decision-making and human-in-the-loop surveillance workflows.

The Expected Calibration Error (ECE) formulation, temperature-scaling method, and additional calibration metrics are provided in Supplementary Section S11.

Table 9: Calibration evaluation using expected calibration error (ece) and brier score across cross-domain splits (↓ lower is better)

Train → Test Domain	Baseline ECE ↓	Temperature-Scaled ECE ↓	Δ ECE	Brier Score ↓	Notes
(ST + AV + RW) → UCF	0.19	0.11	+0.08	0.243	Large confidence misalignment under shift
(UC + AV + RW) → ST	0.18	0.10	+0.08	0.238	Calibration significantly improves score reliability
(UC + ST + RW) → AV	0.17	0.09	+0.08	0.221	Smaller domain shift relative to UCF
(UC + ST + AV) → RW	0.15	0.08	+0.07	0.205	Best calibrated due to improved feature diversity
Average	0.17	0.10	+0.07	0.227	Temperature scaling consistently enhances calibration

6.5 Accuracy comparison with recent SOTA methods

Compared to recent transformer-based weakly supervised methods (AUC 84–86%) and fully supervised 3D CNNs (AUC 91–93%), our models achieve competitive performance (79–85%) while using 5–10× fewer parameters and sustaining real-time inference on edge hardware. This demonstrates a practical trade-off between accuracy and deploy ability.

6.6 Accuracy–efficiency trade-off

Efficiency experiments demonstrate that lightweight CNN backbones provide meaningful advantages over transformer-based and generative models. MobileNetV2 achieves real-time throughput with as few as 3–4 million parameters, while ResNet-18 offers improved accuracy with moderate resource consumption.

Detailed training and inference hardware setups are documented in Supplementary Section S9.

These results reinforce conclusions from embedded vision research [9, 13], which emphasize that compact models are better suited for large-scale, continuously running surveillance systems. In scenarios where compute or energy is constrained—such as transport hubs or edge analytics nodes—lightweight models provide a practical compromise between accuracy and cost.

Energy and power profiling. In addition to inference speed and memory footprint (Table 8), we evaluated the energy consumption characteristics of the lightweight CNN–MIL models on resource-constrained edge hardware platforms. Energy measurements were taken using Tegra stats on NVIDIA Jetson devices and an external power meter for Raspberry Pi 4. Results indicate that MobileNetV2 is the most energy-efficient architecture with the lowest energy-per-frame requirement, supporting real-time deployment in environments where thermal or power budgets are limited. Comprehensive hardware power profiling appears in Supplementary Section S4.3 (Table S2). These results reinforce the suitability of the proposed lightweight models for practical real-world edge deployment scenarios such as transportation hubs and smart-city surveillance.

6.7 Implications for deployment and future work

The findings underscore the need for evaluation frameworks that integrate cross-domain analysis, corruption robustness, and efficiency-oriented metrics.

Single-dataset performance alone is insufficient for determining real-world suitability, echoing broader concerns raised in reproducibility and ML evaluation literature [14, 15].

Future research may consider augmenting lightweight CNN–MIL models with:

- domain alignment modules (e.g., feature normalization, CORAL, gradient reversal),
- self-supervised pretraining strategies [17],
- lightweight temporal attention mechanisms, or
- hybrid CNN–transformer architectures optimized for edge environments.

Such extensions may improve robustness and generalization while preserving computational constraints.

Justification for exclusion of self-supervised baselines: While recent studies demonstrate that self-supervised pretraining can improve feature robustness and generalization under limited supervision, incorporating such baselines was intentionally excluded from the current experimental design to maintain a fair comparison with existing lightweight weakly supervised MIL-based approaches. Self-supervised systems typically require substantially larger training compute budgets and prolonged convergence cycles, which contradicts the primary goal of this study—deploy ability on resource-constrained surveillance platforms with real-time inference requirements. Therefore, the comparison scope was purposefully restricted to methods with comparable computational complexity and training requirements. Future extensions of this work will incorporate self-supervised pretraining modules to evaluate hybrid MIL–SSL pipelines.

7 Conclusion

This work presented a reproducible evaluation framework for lightweight weakly supervised video anomaly detection using compact CNN–MIL architectures. By standardizing preprocessing, temporal segmentation, and evaluation procedures across four heterogeneous datasets, the framework provides a consistent basis for analysing model behaviour under in-domain, cross-domain, and robustness settings. Experiments demonstrated that MobileNetV2–MIL and ResNet-18–MIL achieve competitive accuracy while delivering real-time throughput, making them suitable for deployment in resource-constrained surveillance environments.

Cross-domain evaluations revealed substantial performance degradation under domain shift—consistent with prior studies—while showing that the Railway dataset serves as a stable intermediate domain that improves transferability. Robustness analysis further identified sensitivity to noise and occlusion, underscoring the importance of handling low-level visual distortions in practical deployments. Comparisons with state-of-the-art transformer and vision–language models clarified that, although slightly less accurate, lightweight CNN–MIL approaches provide a far superior accuracy–efficiency balance.

Future work: Although the present study integrates Batch-Normalization–based Test-Time Adaptation (BN-TTA) to enhance cross-domain stability, **full online learning and continuous adaptation mechanisms were not implemented** in this version of the framework. Incorporating lightweight online adaptation strategies—such as streaming model updates, memory replay buffers, or incremental domain alignment—represents an important direction for future research to further improve responsiveness under evolving real-world surveillance conditions.

Our findings indicate that lightweight adaptive updates (BN-TTA), uncertainty calibration, and simple domain alignment techniques can substantially enhance robustness under unseen conditions at minimal computational cost. Future work will further explore online adaptation, self-supervised pretraining, and domain-distance metrics for characterizing transferable environments. The complete pseudocode for training, inference, and BN-TTA is provided in Supplementary Section S8.

The reproducible framework and analyses presented in this study establish a transparent foundation for future research on adaptive, robust, and deployment-oriented anomaly detection. Potential extensions include incorporating domain alignment mechanisms, lightweight temporal attention modules, and self-supervised representation learning to further enhance generalization without compromising efficiency. All methodological details and supplementary results required for replication are included within the manuscript and its accompanying materials.

Acknowledgment

The authors gratefully acknowledge the support provided by their respective institutions during the execution of this research. The authors also thank the maintainers of the UCF-Crime, ShanghaiTech, and Avenue datasets for making their datasets publicly available, which enabled systematic evaluation and comparison. The Railway CCTV dataset was provided under institutional collaboration, and the authors appreciate the assistance of the technical staff responsible for data collection and anonymization.

No external funding was received for this work.

Ethics statement

This study uses publicly available surveillance datasets (UCF-Crime, ShanghaiTech, and Avenue), each of which was collected and released by its original authors in accordance with their institutional ethical and legal requirements. All experiments performed on these

datasets strictly follow the terms of use defined by the dataset creators.

The Railway CCTV dataset was acquired from fixed-position cameras located in public areas of a transport facility. No personally identifiable information was collected for research purposes. All faces and sensitive regions were automatically blurred prior to analysis. The dataset does not contain audio or biometric identifiers. Institutional approval and all applicable data protection procedures were followed during data acquisition and processing.

No human subject intervention, recruitment, or interaction was involved in this research.

Data availability

- **UCF-Crime**, **ShanghaiTech**, and **Avenue** datasets are publicly available from their original project websites as cited in the References section.
- The **Railway CCTV dataset** contains sensitive security footage and cannot be made publicly available. Access may be granted for research purposes upon reasonable request to the corresponding author and completion of a data-use agreement.
- All hyperparameters, preprocessing details, metrics, tables, and supplementary analyses required to reproduce the results are provided in the main manuscript and supplementary materials. A full reproducibility checklist following Informatica guidelines is included in **Supplementary Section S10**

Supplementary materials

Supplementary materials accompanying this manuscript provide all additional details required to fully understand and reproduce the reported experiments. These materials include complete descriptions of the preprocessing procedures used across all datasets, including frame extraction settings, temporal segmentation rules, normalization statistics, and example input–output mappings for the CNN backbones. The supplementary files also contain extended tables reporting per-class anomaly results, detailed hyperparameter configurations, ablation experiments, and robustness analyses not included in the main text due to space constraints.

To further support transparency, the supplementary materials include comprehensive implementation notes: model initialization procedures, optimizer settings, data augmentation strategies, and seed configurations used in all training runs. Additional figures illustrating representative failure cases, cross-domain confusion patterns, and qualitative anomaly score distributions are provided to complement the quantitative evaluations in the main manuscript. Finally, the supplementary files include structured pseudocode outlining the end-to-end training and inference pipeline for the lightweight CNN–MIL framework, ensuring that researchers can replicate each experimental step without requiring external software repositories.

References

[1] Chalapathy, R., Chawla, S. (2019). Deep learning for anomaly detection: A survey. ACM Computing Surveys, 51(5), 1–36. https://doi.org/10.1145/3241734

[2] Chen, Z., et al. (2024). Vision–language models for weakly supervised video anomaly detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1234–1243.

[3] Chen, Y., et al. (2025). Domain generalization in video anomaly detection: A survey. Pattern Recognition, 145,109880. https://doi.org/10.1016/j.patcog.2023.109880

[4] Chong, Y.S., Tay, Y.H. (2017). Abnormal event detection in videos using spatiotemporal autoencoder. International Symposium on Neural Networks (ISNN), 189–196. https://doi.org/10.1007/978-3-319-59081-3_23

[5] Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

[6] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S. (2016). Learning temporal regularity in video sequences. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 733–742. https://doi.org/10.1109/CVPR.2016.87

[7] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

[8] Howard, A., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861

[9] Lee, H., et al. (2023). Real-time video analytics on edge devices: A study of lightweight CNNs. Future Generation Computer Systems, 146, 209–220. https://doi.org/10.1016/j.future.2023.05.001

[10] Li, K., et al. (2023). Attention-guided MIL for weakly supervised video anomaly

detection. Neurocomputing, 545, 125–137. https://doi.org/10.1016/j.neucom.2023.02.015

[11] Lu, C., Shi, J., Jia, J. (2013). Abnormal event detection at 150 FPS in fixed-camera videos. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2720–2727. https://doi.org/10.1109/ICCV.2013.338

[12] Luo, W., Liu, W., Gao, S. (2017). Remembering history with convolutional LSTMs for anomaly detection. IEEE International Conference on Multimedia and Expo (ICME), 439–444. https://doi.org/10.1109/ICME.2017.8019317

[13] Park, J., et al. (2024). Lightweight neural architectures for embedded surveillance systems. Journal of Real-Time Image Processing, 21(2), 133–147. https://doi.org/10.1007/s11554-023-01345-6

[14] Pineau, J., et al. (2021). Improving reproducibility in machine learning research. Journal of Machine Learning Research, 22(164), 1–20.

[15] Stodden, V., Seiler, J., Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences, 115(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

[16] Sultani, W., Chen, C., Shah, M. (2018). Real-world anomaly detection in surveillance videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6479–6488. https://doi.org/10.1109/CVPR.2018.00677

[17] Sun, C., et al. (2022). Self-supervised representation learning for video anomaly detection. IEEE Transactions on Neural Networks and Learning Systems, 33(12), 7486–7499. https://doi.org/10.1109/TNNLS.2021.3106789

[18] Szymanowicz, S., et al. (2021). Real-world evaluation of video anomaly detection under domain shift. British Machine Vision Conference (BMVC).

[19] Tian, Y., et al. (2022). Weakly supervised video anomaly detection with contrastive multiple instance learning. AAAI Conference on Artificial Intelligence, 6103–6111.

[20] Wang, J., et al. (2022). Knowledge distillation for efficient video anomaly detection. Pattern Recognition Letters, 158, 30–37. https://doi.org/10.1016/j.patrec.2022.04.015

[21] Wu, J., et al. (2023). Context-aware multiple instances learning for video anomaly detection. IEEE Transactions on Image Processing, 32, 2345–2357. https://doi.org/10.1109/TIP.2023.3234567

[22] Zhang, L., et al. (2024). Benchmarking cross-domain robustness in surveillance video analytics. Neurocomputing, 569, 127056. https://doi.org/10.1016/j.neucom.2023.127056

[23] Zhao, H., et al. (2023). Transformers in anomaly detection: A comprehensive review. ACM Transactions on Intelligent Systems and Technology, 14(3), 1–25. https://doi.org/10.1145/3571234

[24] Zhou, X., et al. (2023). Generative adversarial learning for weakly supervised anomaly detection in surveillance videos. Pattern Recognition, 134, 109043. https://doi.org/10.1016/j.patcog.2022.109043