Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

IRA-C2NN: Interactive Routing Algorithm-Optimized Channel-Wise Deep CNN with Hybrid Energy-Variation Keyframe Selection for Anomaly Detection in Surveillance Video

Mangesh Nikose
Symbiosis Skills and Professional University
Anni U. Gupta
MET'S IOE, Nashik , India
Megha Patil
Symbiosis Skills and Professional University
Dhanashree Barbole
dhanashree.barbole@sitpune.edu.in

Symbiosis Institute of Technology, Symbiosis International (Deemed University)

Research Article

Keywords: Video anomaly detection, surveillance, keyframe selection, YOLOv5, channel-wise normalization, hybrid activation, metaheuristic optimization, convolutional neural network

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11534185/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

IRA-C2NN: Interactive Routing Algorithm-Optimized Channel-Wise Deep CNN with Hybrid Energy-Variation Keyframe Selection for Anomaly Detection in Surveillance Video

Dr. Mangesh Nikose[1], Dr. Anni U. Gupta[2], Dr. Megha Patil[3], Dr. Dhanashree Barbole[4]
[1,3]School of Mechatronics Engineering, Symbiosis Skills and Professional University, Pune, India
[2]Assistant Professor, Department of Electronics and Telecommunication, MET'S IOE, Nashik,India
[4]Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
mangesh.nikose@sspu.ac.in[1], anniugupta@gmail.com[2], megha.patil@sspu.ac.in[3], dhanashree.barbole@sitpune.edu.in[4]
*Corresponding Author: Dhanashree Barbole (dhanashree.barbole@sitpune.edu.in)

Abstract

Automatic anomaly detection in surveillance video is essential for public safety, but existing deep models suffer from limited generalisation, high computational cost and manual hyper-parameter tuning. This paper proposes IRA-C2NN, an interactive-routing-algorithm-optimised channel-wise deep convolutional neural network. Redundant frames are first removed by a hybrid energy-variation keyframe selector that fuses DWT, biorthogonal-wavelet and GLCM energies; objects are detected and tracked with YOLOv5; and complementary LPQ-ensemble spatial-pyramid HOG, spatiotemporal activity-pattern (STAP) and statistical features are extracted. A channel-wise CNN with channel-wise response layer normalisation and a hybrid ReLU/sigmoid activation classifies the features, while a hybrid pigeon-inspired/swarm optimiser, the interactive routing algorithm (IRA), tunes its hyper-parameters. On the UCSD and ShanghaiTech datasets, IRA-C2NN achieves 96.03% and 96.21% accuracy with 80% training data, and 95.00% and 96.60% under 10-fold cross-validation, outperforming nine baselines on all evaluated metrics by 0.18–1.97 percentage points over the strongest competitor. Ablation, statistical, runtime and ROC analyses confirm the complementary role of the feature families and the stability of the model.

Keywords : Video anomaly detection, surveillance, keyframe selection, YOLOv5, channel-wise normalization, hybrid activation, metaheuristic optimization, convolutional neural network.

I. INTRODUCTION

Video surveillance has become a core component of smart cities, transport hubs, campuses and retail environments, where camera networks continuously record public spaces to deter crime and accidents [1], [2]. Because abnormal events such as fights, thefts, accidents or the appearance of non-pedestrian objects on a walkway are rare, manual inspection of the footage is monotonous, error-prone and does not scale, while the large share of redundant footage inflates storage and processing costs [3], [4]. Automatic video anomaly detection (VAD), which learns what is normal and flags what is not, is therefore an active research area [2], [5].

Deep learning has substantially advanced VAD through U-Net and autoencoder reconstruction or prediction models [6]–[8], convolutional and recurrent classifiers [9]–[11], generative adversarial networks [12], [13] and, more recently, transformers [14], [15]. Nevertheless, three limitations persist. First, many models are evaluated on a single data source and generalise poorly to other camera views, lighting conditions and crowd densities. Second, per-frame processing of every frame is computationally heavy and ignores the fact that consecutive frames are highly redundant. Third, hand-tuned hyper-parameters and ordinary batch-

dependent normalisation limit accuracy and training stability, and few frameworks combine object-level, texture-level and attention-based spatiotemporal cues in one feature vector.

This paper presents IRA-C2NN, an interactive-routing-algorithm-optimised channel-wise deep convolutional neural network for surveillance-video anomaly detection. The main contributions are:

- A *hybrid energy-variation keyframe selector* that fuses discrete-wavelet, biorthogonal-wavelet and grey-level co-occurrence-matrix energies and keeps only frames whose energy exceeds the video mean, removing redundant frames before any heavy processing.
- A *multi-cue feature representation* that combines YOLOv5 object evidence with LPQ-ensemble spatial-pyramid HOG texture descriptors, spatiotemporal activity patterns (STAP) from joint spatial/temporal attention, and statistical moments.
- A *channel-wise deep CNN (C2NN)* with channel-wise response layer normalisation (CRLN) and a two-branch hybrid ReLU/sigmoid activation that stabilises training without batch dependency.
- An *Interactive Routing Algorithm (IRA)*, a hybrid of pigeon-inspired routing and swarm interaction, that tunes the C2NN hyper-parameters with fast convergence.
- A comprehensive evaluation on the UCSD and ShanghaiTech datasets under both training-percentage and 10-fold protocols against nine baselines, with ablation, statistical, runtime and ROC analyses.

The remainder of the paper is organised as follows. Section II reviews related work, Section III formulates the problem and describes the datasets, Section IV details IRA-C2NN, Section V presents the experimental setup, Section VI reports and discusses the results, and Section VII concludes.

II. RELATED WORK

A. Reconstruction- and Prediction-Based Models

U-Net based detectors improve efficiency and localisation. Kim *et al.* combined a Cross U-Net with a cascade sliding window to detect anomalies quickly and accurately [6], and Taghinezhad and Yazdi proposed a multi-scale multi-path U-Net for unsupervised frame prediction [16]. Autoencoders learn normal appearance and flag poorly reconstructed frames: spatio-temporal dissociation [7], masked autoencoders with multi-memory modules [17], attention-based residual autoencoders [8] and double-flow ConvLSTM variational autoencoders [18] are representative designs, and design trade-offs are examined in [19]. These models generalise reasonably but degrade on complex, cluttered scenes and often miss long-range, multi-scale context.

B. Classical Machine Learning and Deep Classifiers

Machine-learning pipelines pair hand-crafted motion or appearance features with classifiers such as XGBoost or pixel/frame-level learners [20], [21], which are accurate but sensitive to feature design and hardware. Deep classifiers range from two-stream fusion with YOLO object cues [22], ResNet–recurrent hybrids [11], keyframe-driven spatiotemporal pipelines [23] and 3-D CNNs [9] to CNN–BiLSTM [10] and attention-based CNN–LSTM models [24]. Deep one-class and contrastive formulations have also been proposed [25], [26], and AnomalyNet [27] is a representative surveillance-oriented deep detector. Their scalability across datasets remains limited.

C. Generative and Transformer Models

GAN-based detectors offer robustness through residual spatiotemporal translation [12] and multi-channel generative frameworks [13], but they are expensive to train. Transformers capture global dependencies, e.g., bi-directional ViT–ConvLSTM hybrids [14], TransAnomaly [15] and meta-learning for multi-scenario adaptation [28], yet real-time deployment is challenging. Context-tree ensembles with YOLOv5 object cues [29] and continual-learning detectors [30] complete the picture.

D. Research Gaps

The surveyed literature reveals recurring gaps: (i) long-range and multi-scale spatiotemporal context is incompletely captured; (ii) accuracy falls on complex data and across datasets; (iii) high inference time and hardware demand hinder real-time use; and (iv) tuning is manual or computationally heavy. IRA-C2NN targets these gaps through redundancy-aware keyframe selection, multi-cue features, normalisation without batch dependency and metaheuristic hyper-parameter search.

III. PROBLEM FORMULATION AND DATASETS

A. Formulation

Let the collected data be

$$\mathcal{D} = \{(V_i, y_i)\}_{i=1}^{N} \tag{1}$$

where N is the number of surveillance clips, y_i is the ground-truth label (normal or anomalous) of clip V_i, and each clip is a frame sequence

$$V_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,T}\} \tag{2}$$

with T the number of frames in the clip. The goal is to learn a mapping from keyframe-level features to the label that is robust to shadows, occlusion, background clutter and viewpoint change. Training minimises the categorical cross-entropy

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c} y_{i,c}\log p_{i,c} \tag{3}$$

where $p_{i,c}$ is the predicted probability of class c for clip i.

B. Datasets

Two public benchmarks are used (Table I). The UCSD dataset [31] was recorded by a stationary, elevated camera overlooking pedestrian walkways; videos are 8-bit grey-scale at 10 frames/s, crowd density ranges from sparse to dense, and anomalies include bicycles, carts, skaters and pedestrians on the grass or off the walkway. It comprises two subsets (Ped1 and Ped2). The ShanghaiTech Campus dataset [32] contains about 270 000 training frames, over 130 abnormal events, multiple camera angles and uneven illumination, with pixel- and frame-level annotations.

TABLE I

SUMMARY OF THE DATASETS

Property	**UCSD**	**ShanghaiTech**
Camera	Static, elevated	Multiple angles
Subsets / scenes	Ped1, Ped2	13 scenes
Format	8-bit grey, 10 fps	—
Training frames	—	≈270 000
Anomalies	Bikes, carts, skaters, off-path walking	Over 130 abnormal events
Illumination	Fixed outdoor view	Uneven lighting
Annotation	Frame-level (pixel-level on a subset)	Frame- and pixel-level

C. Evaluation Metrics

With true/false positives and negatives (TP, TN, FP, FN) the five metrics are

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad F1 = \frac{2\,\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{6}$$

IV. PROPOSED IRA-C2NN FRAMEWORK

A. Overview

Fig. 1 shows the pipeline. Input clips are reduced to keyframes by hybrid energy-variation selection, the keyframes are processed by YOLOv5 for object detection and tracking, three complementary features are extracted, and the C2NN classifier, whose hyper-parameters are tuned by the IRA, outputs the anomaly decision.

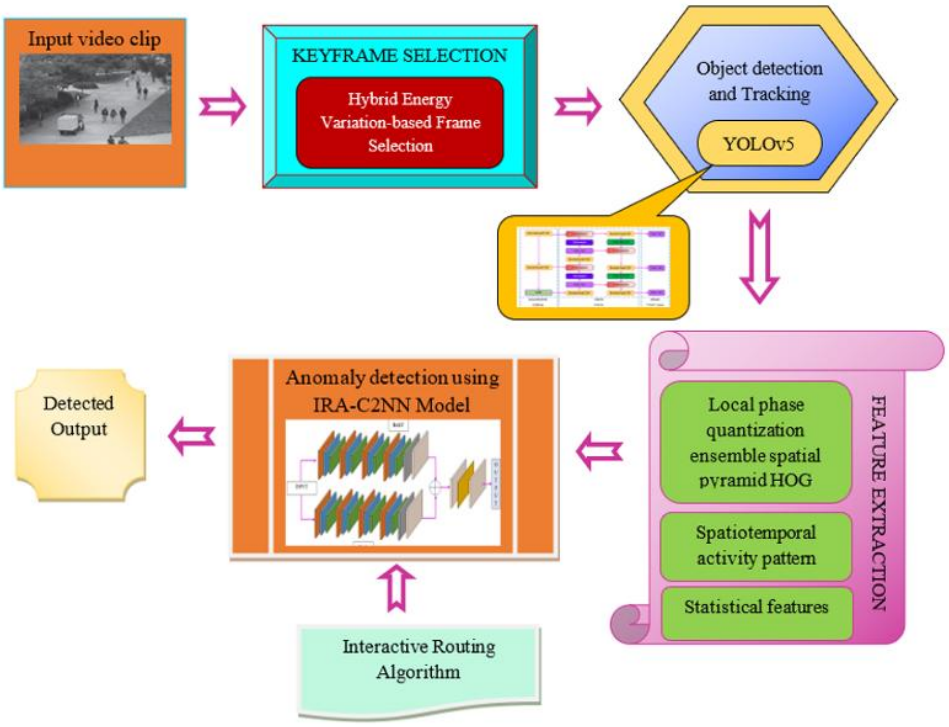


Fig. 1. Overall workflow of the proposed IRA-C2NN anomaly detection framework.

B. Hybrid Energy-Variation Keyframe Selection

Consecutive frames in surveillance video are highly redundant. For every frame f_t we compute three energies. The DWT energy [33] is the sum of squared coefficients of the LL, LH, HL and HH sub-bands, the biorthogonal-wavelet energy [34] is computed analogously with symmetric linear-phase filters, and the GLCM energy [35] is the angular second moment of the normalised co-occurrence matrix P_t:

$$E_{t,\mathrm{DWT}} = \sum_{b,u,v} [W_t^b(u,v)]^2 \tag{7}$$

$$E_{t,\mathrm{GLCM}} = \sum_{i,j} P_t(i,j)^2 \tag{8}$$

The hybrid energy of a frame is the average of the three, and a frame is selected as keyframe when it exceeds the mean energy of the video:

$$E_t = \tfrac{1}{3}(E_{t,\mathrm{DWT}} + E_{t,\mathrm{BWT}} + E_{t,\mathrm{GLCM}}) \tag{9}$$

$$\mathcal{K} = \{t : E_t > \bar{E}\}, \quad \bar{E} = \tfrac{1}{T}\sum_{t=1}^{T} E_t \tag{10}$$

C. Object Detection and Tracking with YOLOv5

Keyframes are passed to YOLOv5 [36], whose CSPDarknet53 backbone [37] extracts features, spatial-pyramid pooling preserves fine-grained localisation, a PANet neck [38] aggregates multi-

level features, and three convolutional head maps predict objects at three scales. Detected boxes are associated between adjacent keyframes to track people and vehicles and to expose motion that departs from the scene norm.

D. Feature Extraction

1) LPQ-ensemble SP-HOG. Local phase quantisation (LPQ) [39] computes the short-term Fourier transform in a local window at low frequencies, decorrelates the coefficients and quantises the signs of the real and imaginary parts into an 8-bit code (0–255), which yields a blur-insensitive texture map. With F_t the local STFT, $s_k(x)=1$ if the k-th real/imaginary component $g_k(x)\geq 0$ and 0 otherwise:

$$F_t(u,x)=\sum_y f(x-y)\,e^{-j2\pi u^T y} \tag{11}$$

$$\mathrm{LPQ}(x)=\sum_{k=1}^{8} s_k(x)\,2^{k-1} \tag{12}$$

The LPQ map is fed to a spatial-pyramid HOG (SP-HOG): HOG descriptors [40] are computed on grids at several pyramid levels and concatenated, capturing both local gradients and global layout and removing the fixed-grid limitation of plain HOG.

2) STAP. Spatiotemporal activity patterns combine a spatial attention module, which weights locations through softmax-gated convolutions to highlight anomalous regions, with a temporal attention module that re-weights time steps by their relevance; the two outputs are summed with the input via residual connections.

3) Statistical features. Mean, standard deviation, variance, skewness, kurtosis and entropy of each region are concatenated. The final vector concatenates the LPQ-ensemble SP-HOG (F_1), STAP (F_2) and statistical (F_3) features:

$$\mathbf{F}=[\mathbf{F}_1 \,\|\, \mathbf{F}_2 \,\|\, \mathbf{F}_3] \tag{13}$$

E. Channel-Wise Deep CNN with CRLN and Hybrid Activation

The C2NN (Fig. 2) processes the feature vector with two parallel branches, one using ReLU and one using sigmoid activation. Each branch stacks four convolution blocks (convolution, layer normalisation, max pooling; 8, 16, 32 and 128 channels), followed by flatten, dropout and a dense layer. The two 2-D branch outputs are concatenated, passed through a 128-unit dense layer with normalisation, and a final 2-way softmax layer. ReLU keeps positive activations unchanged and mitigates vanishing gradients, whereas the sigmoid branch models the positive/negative decision smoothly; fusing them gives the hybrid activation, where $\sigma(x)=1/(1+e^{-x})$ is the sigmoid:

$$\mathbf{z}=[\mathrm{ReLU}(\mathbf{x}) \,\|\, \sigma(\mathbf{x})] \tag{14}$$

CRLN normalises every activation channel of each sample independently of the other samples in the batch [41], reducing internal covariate shift and removing batch-size dependence [42]. With the per-sample, per-channel mean $\mu_{n,c}$ and variance $s_{n,c}^2$, learnable scale γ_c and shift β_c:

$$\hat{x} = \gamma_c \frac{x - \mu_{n,c}}{\sqrt{s_{n,c}^2 + e}} + \beta_c \tag{15}$$

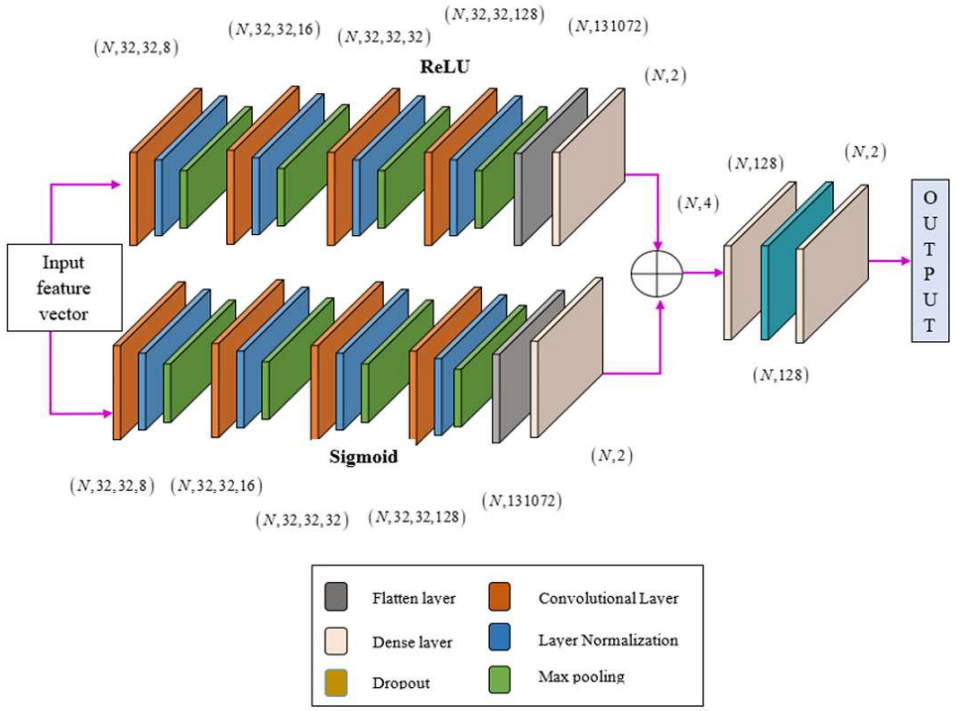


Fig. 2. Architecture of the IRA-C2NN classifier: parallel ReLU and sigmoid branches of convolution (orange), layer-normalisation (blue) and max-pooling (green) blocks, followed by dense fusion layers.

F. Interactive Routing Algorithm (IRA)

The IRA is a hybrid metaheuristic that merges the homing and escape routing behaviour of the pigeon *Columba livia* [43] with swarm interaction [44]. It addresses slow convergence and local optima of the standard algorithms [45], [46]. A population of candidate hyper-parameter vectors is initialised uniformly in the search bounds,

$$X_i = lb + r \odot (ub - lb) \tag{16}$$

$$f(X_i) = \mathrm{Acc}_{\mathrm{val}}(\mathrm{C2NN}(X_i)) \tag{17}$$

where $r \in [0,1]^d$ is random and the objective f is the validation accuracy of the C2NN configured by X_i. At each iteration an early-warning value is compared with a safety threshold. In the *safety phase* no predator is present and a Lévy flight $L(\lambda)$ explores the space broadly:

$$X_i^{t+1} = X_i^t + \eta L(\lambda) \odot X_i^t \tag{18}$$

In the *danger phase* solutions risk converging to local optima and are guided by leaders. In the *leading case* a solution is attracted to the global best X_g using memory M, distance to the threat D and the sound intensity I of the leader; in the *following case* it follows the local leader X_l through a velocity update with inertia weight w and acceleration factors c_1, c_2:

$$X_i^{t+1} = X_g^t + M_i^t \odot I_g^t D_i^t, \quad D_i^t = \left| X_g^t - X_i^t \right| \tag{19}$$

$$v_i^{t+1} = w v_i^t + c_1 r_1 (X_l^t - X_i^t) + c_2 r_2 (X_g^t - X_i^t) \tag{20}$$

$$X_i^{t+1} = X_i^t + v_i^{t+1} \tag{21}$$

Fitness is re-evaluated after each update and the best solution is retained until the maximum number of iterations. Algorithm 1 summarises the training procedure.

1) Select keyframes with the hybrid energy rule (Section IV-B).
2) Detect and track objects with YOLOv5; extract F (Section IV-D).
3) Initialise the IRA population; evaluate $f(X_i)$ by training C2NN on the training split.
4) For each iteration, update every solution by the safety, leading or following rule, re-evaluate and keep the global best.
5) Train the final C2NN with the best hyper-parameters; classify test keyframes.

Algorithm 1. Training procedure of IRA-C2NN.

V. Experimental Setup

Protocols. Two validation schemes are used on both datasets: (i) training percentage (TP) of 40–80 % with the remainder for testing, with performance tracked for 20–100 epochs, and (ii) 10-fold cross-validation. Metrics are those of Section III-C. *Baselines.* IRA-C2NN is compared with nine methods: CNN [47], CNN-LSTM [24], deep one-class NN (DeepOC-NN) [25], RNN [48], 3-D ConvNets [9], temporal-aware contrastive network (TAC-Net) [26], the hybrid-activation DCNN (HAFDCNN) [49], and its pigeon-inspired (PIO-HAFDCNN) [45] and particle-swarm (PSO-HAFDCNN) [46] optimised variants.

VI. RESULTS AND DISCUSSION

A. Effect of Training Percentage and Epochs

Fig. 3 shows that IRA-C2NN improves steadily with both training data and training epochs (with small fluctuations on UCSD). On UCSD, accuracy rises from 90.1–91.2 % (TP = 40 %) to 92.9–93.8 % (TP = 60 %) and 95.4–96.03 % (TP = 80 %) over 20–100 epochs. On the more diverse ShanghaiTech data the gain is larger, from 83.1 % (TP = 40 %, epoch 20) to 96.2 % (TP = 80 %, epoch 100), showing that the model benefits from additional varied footage.

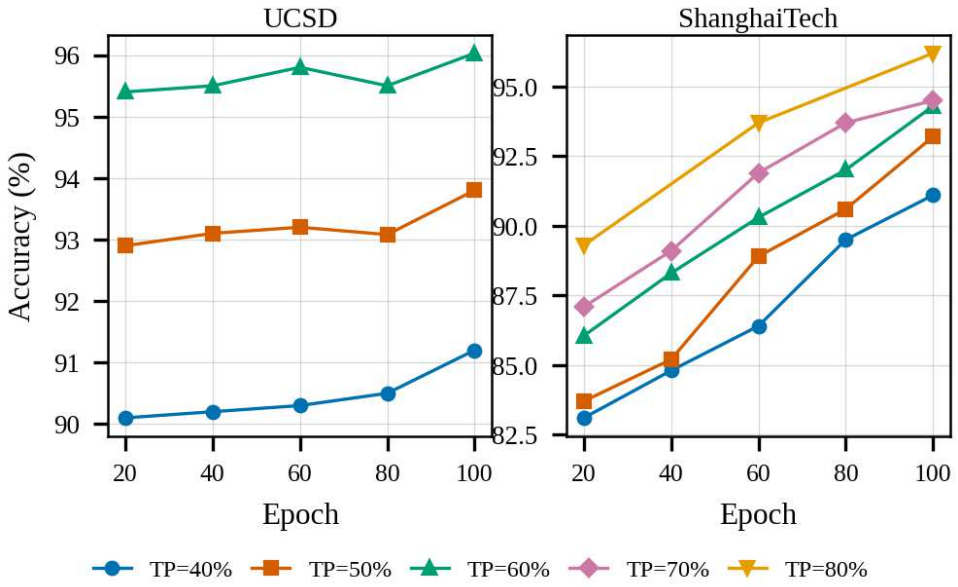


Fig. 3. Accuracy of IRA-C2NN versus epochs for different training percentages (TP) on UCSD (left) and ShanghaiTech (right).

B. Comparison at 80 % Training

Tables II and III and Fig. 4 compare all methods at TP = 80 %. On UCSD, IRA-C2NN reaches 96.03 % accuracy, 96.33 % F1, 97.00 % sensitivity, 94.32 % specificity and 95.66 % precision, exceeding the strongest baseline, PSO-HAFDCNN, on every metric by 0.38–0.39 pp and the mean of the nine baselines by 1.6–3.1 pp. On ShanghaiTech the margins are larger (accuracy 96.21 % versus 94.49 %, +1.72 pp; sensitivity +1.97 pp; specificity +1.58 pp), and the classical CNN, RNN, 3-D ConvNet and TAC-Net baselines trail by 6.5–10.3 pp in accuracy. The gap between sensitivity (97.00 %) and specificity (94.32 %) on UCSD indicates that residual errors are mostly false alarms, whereas the two are balanced on ShanghaiTech (95.38 % / 96.62 %).

TABLE II
PERFORMANCE (%) ON UCSD, TP = 80 %

Method	Acc.	F1	Sens.	Spec.	Prec.
DeepOC-NN	94.49	93.50	95.45	87.81	91.63
CNN	94.64	94.17	95.60	89.96	92.78
RNN	94.64	93.91	95.60	88.96	92.28
3D ConvNets	93.31	93.71	93.73	92.46	92.90
TAC-Net	92.48	90.79	93.19	91.07	93.21
CNN-LSTM	94.88	94.66	95.84	91.19	93.51
HAFDCNN	95.04	95.08	96.00	92.35	94.17
PIO-HAFDCNN	95.07	95.36	96.03	93.38	94.70
PSO-HAFDCNN	95.65	95.94	96.61	93.94	95.28
IRA-C2NN	**96.03**	**96.33**	**97.00**	**94.32**	**95.66**

TABLE III
PERFORMANCE (%) ON SHANGHAITECH, TP = 80 %

Method	Acc.	F1	Sens.	Spec.	Prec.
DeepOC-NN	85.55	87.02	87.89	84.38	92.79
CNN	85.93	93.74	88.27	84.75	93.94
RNN	89.75	85.75	89.32	89.96	93.44
3D ConvNets	87.91	89.96	89.44	87.14	94.05
TAC-Net	88.89	92.10	89.53	88.58	94.36
CNN-LSTM	87.13	86.59	88.99	86.19	94.67
HAFDCNN	91.52	93.86	90.61	91.98	95.33
PIO-HAFDCNN	92.39	94.82	91.71	92.73	95.86
PSO-HAFDCNN	94.49	95.14	93.41	95.04	96.43
IRA-C2NN	**96.21**	**96.77**	**95.38**	**96.62**	**96.81**

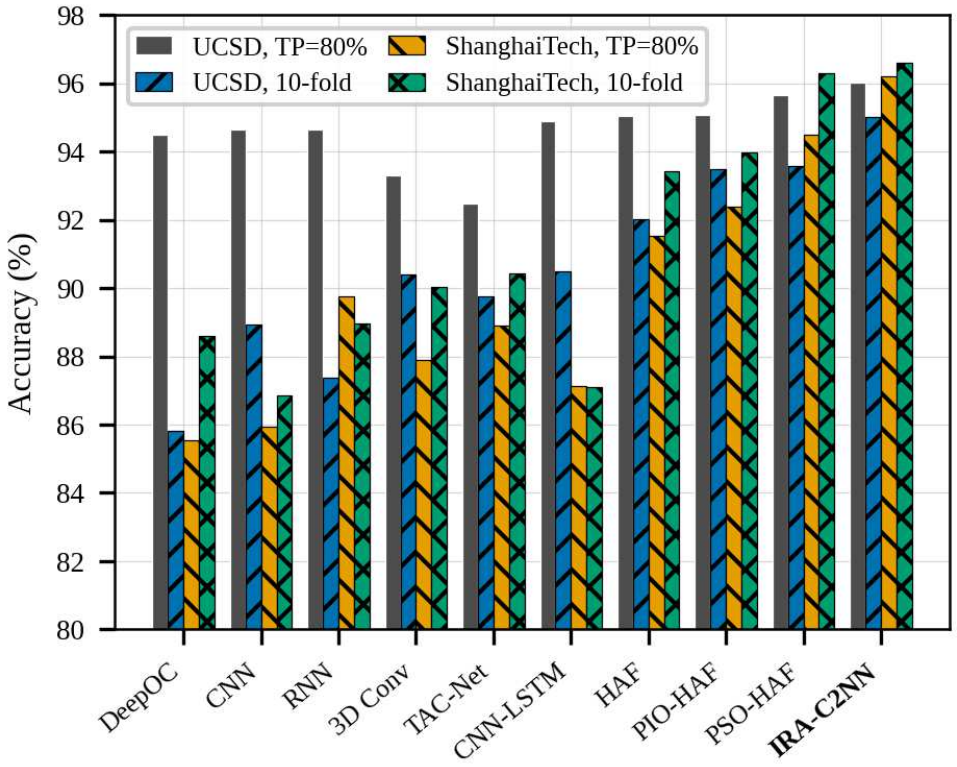


Fig. 4. Accuracy of all methods on UCSD and ShanghaiTech under TP = 80 % and 10-fold validation.

C. Comparison under 10-Fold Cross-Validation

Under 10-fold validation (Tables IV and V) IRA-C2NN again ranks first on all ten dataset–metric pairs. On UCSD it attains 95.00 % accuracy (+1.42 pp over the best baseline) with the largest margin in specificity (93.86 % versus 92.08 %), and on ShanghaiTech 96.60 % accuracy and 96.35 % precision (+1.51 pp over PIO-HAFDCNN). Fig. 5 contrasts the five-metric profile of IRA-C2NN with that of PSO-HAFDCNN in all four settings. Over all 20 dataset–protocol–metric combinations the margin over the best baseline is

0.18–1.97 pp (mean 0.97 pp), so the improvement is consistent but, against the strongest optimised baselines, modest in size.

TABLE IV
PERFORMANCE (%) ON UCSD, 10-FOLD

Method	Acc.	F1	Sens.	Spec.	Prec.
DeepOC-NN	85.82	84.51	87.69	76.41	81.55
CNN	88.93	87.80	90.16	81.97	85.56
RNN	87.38	86.16	88.93	79.19	83.56
3D ConvNets	90.40	90.06	92.57	86.06	85.56
TAC-Net	89.76	88.95	92.00	85.28	86.57
CNN-LSTM	90.48	89.44	91.40	84.74	87.57
HAFDCNN	92.03	91.08	92.63	87.52	89.58
PIO-HAFDCNN	93.50	94.09	95.11	92.08	93.09
PSO-HAFDCNN	93.58	92.71	93.87	90.30	91.59
IRA-C2NN	**95.00**	**95.46**	**96.34**	**93.86**	**94.60**

TABLE V
PERFORMANCE (%) ON SHANGHAITECH, 10-FOLD

Method	Acc.	F1	Sens.	Spec.	Prec.
DeepOC-NN	88.59	89.03	90.44	87.66	83.30
CNN	86.86	86.80	86.64	86.96	87.31
RNN	88.98	88.85	88.46	89.24	85.30
3D ConvNets	90.04	89.95	89.69	90.21	87.31
TAC-Net	90.44	89.91	88.42	91.45	88.32
CNN-LSTM	87.10	87.03	86.79	87.26	89.32
HAFDCNN	93.44	93.37	93.17	93.57	91.33
PIO-HAFDCNN	93.98	93.81	93.31	94.32	94.84
PSO-HAFDCNN	96.30	96.21	95.96	96.47	93.33
IRA-C2NN	**96.60**	**96.58**	**96.50**	**96.65**	**96.35**

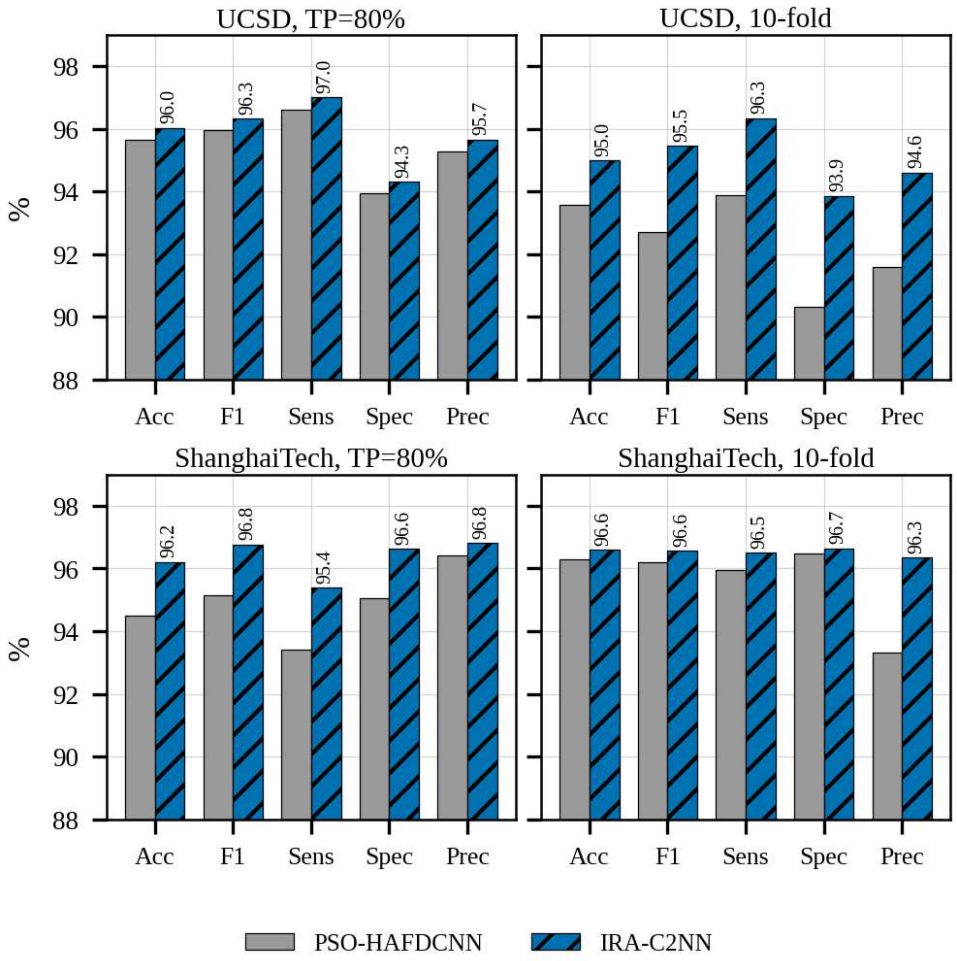


Fig. 5. Five-metric profile of IRA-C2NN versus the strongest baseline, PSO-HAFDCNN, in four evaluation settings.

D. Statistical Analysis

Fig. 6 shows the mean ± standard deviation of accuracy across runs/folds. IRA-C2NN has the highest mean accuracy in all four settings: 93.76 % (TP) and 93.07 % (10-fold) on UCSD, and 93.91 % and 96.12 % on ShanghaiTech, against 92.79 %, 91.64 %, 91.83 % and 93.95 % for PSO-HAFDCNN. Its variability is also the lowest among all methods in ShanghaiTech 10-fold ($\sigma = 0.43$), indicating stable behaviour across folds.

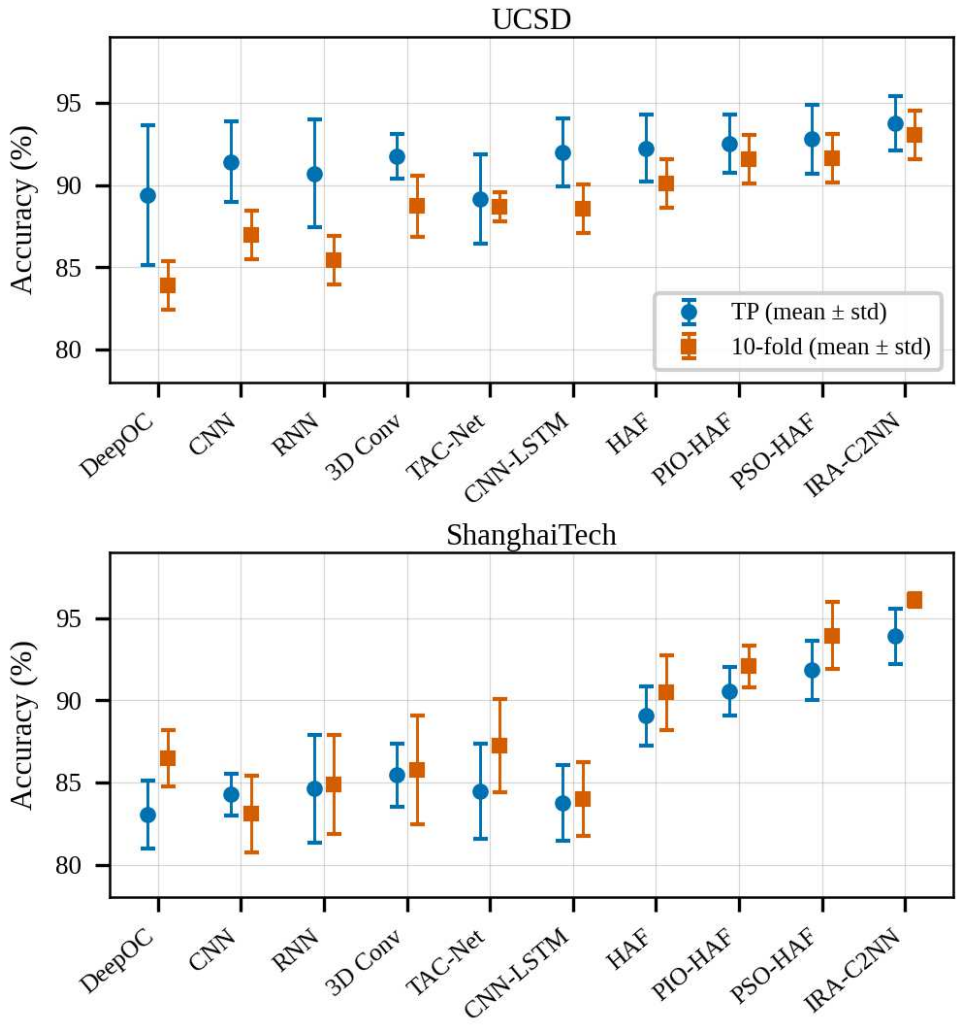


Fig. 6. Mean ± standard deviation of accuracy for every method under the TP and 10-fold protocols.

E. Ablation Study

Fig. 7(a) quantifies each feature family. Without extracted features, accuracy is 89.45 %; statistical, STAP and LPQ-ensemble SP-HOG features alone give 91.18 %, 90.53 % and 92.45 %, respectively, and their combination reaches 96.03 %, i.e., +6.6 pp over no features and +3.6 pp over the best single family, which confirms the complementarity of the cues. Fig. 7(b) compares the network variants: plain C2NN reaches 95.78 %, the hybrid-activation variant HAF-C2NN 94.86 % and the full IRA-C2NN 96.03 %. The IRA search therefore recovers and slightly exceeds the plain network (+0.25 pp) after hybrid activation alone lowered accuracy, which suggests that the benefit of hybrid activation depends on well-tuned hyper-parameters.

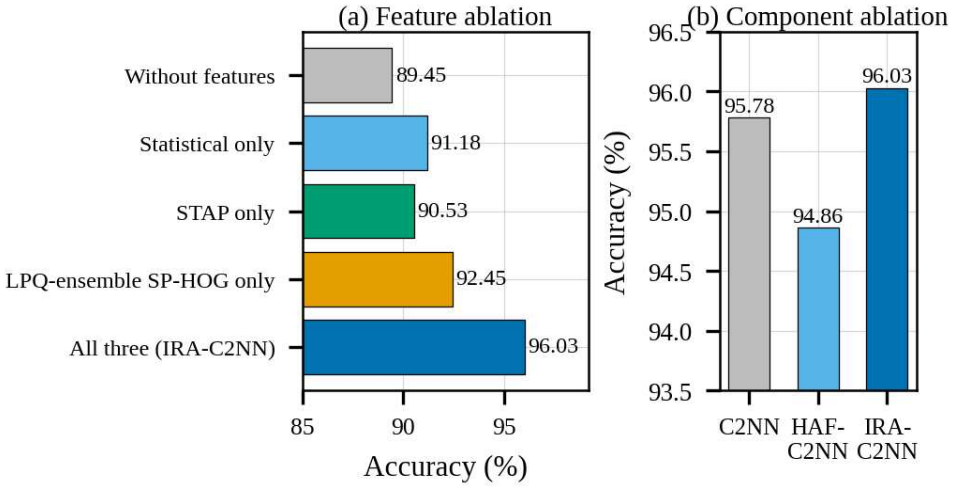


Fig. 7. Ablation study (UCSD, TP = 80 %): (a) contribution of feature families; (b) contribution of network components.

F. Computation Time and ROC Analysis

At iteration 100, IRA-C2NN needs 15.29 s, compared with 15.49 s for PSO-HAFDCNN and 15.60–15.81 s for the remaining methods (Fig. 8). The saving of 0.2–0.5 s (1.3–3.3 %) is small, and is attributed to keyframe selection, which reduces the number of frames to process. The ROC curves in Fig. 9 show IRA-C2NN dominating the baselines, with the highest true-positive rate at each operating point.

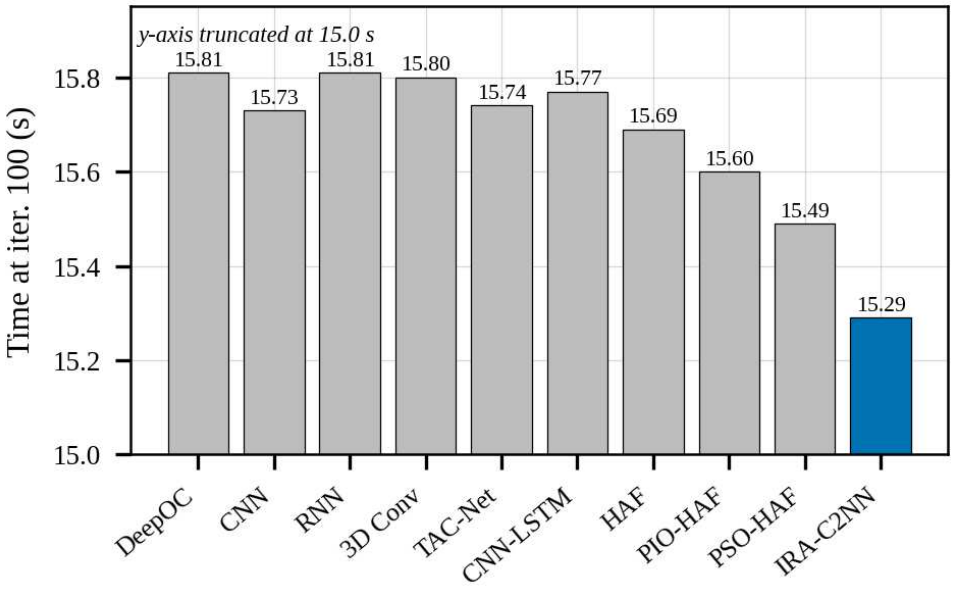


Fig. 8. Computation time of all methods at iteration 100 (y-axis truncated at 15.0 s).

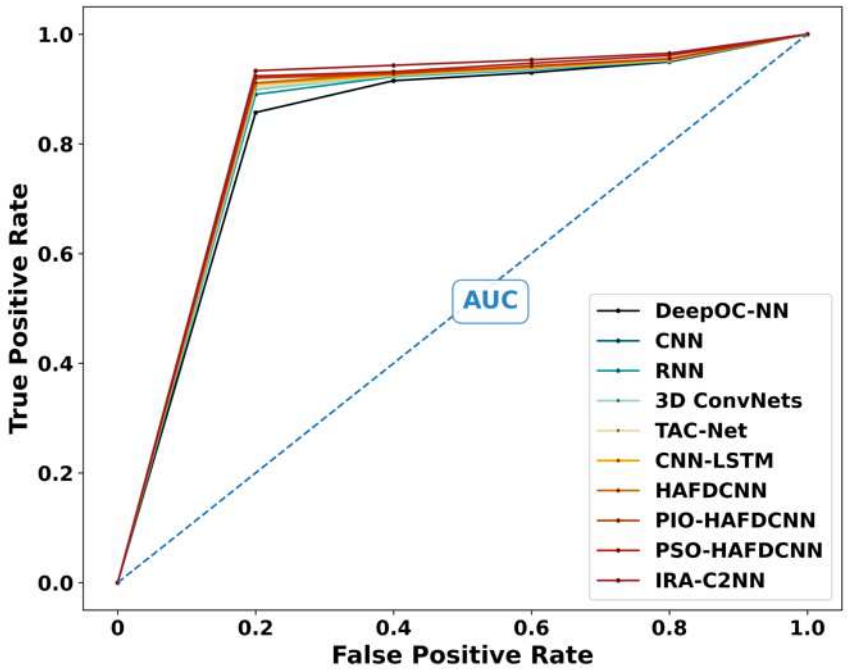


Fig. 9. ROC curves of IRA-C2NN and the baselines.

G. Limitations

The model degrades on long-distance objects and in rain, where objects are small or partially occluded. Evaluation uses random training/test percentages and threshold-based metrics rather than the frame-level AUC on the official train/test splits commonly reported for UCSD and ShanghaiTech; the ROC curves are therefore indicative only and direct comparison with recent state-of-the-art VAD methods remains to be done. The runtime difference is also small and was measured once per method. Future work will add hybrid attention and self-/weakly-supervised learning, evaluate AUC and EER on the official splits, and incorporate explainable-AI components for transparency.

VII. CONCLUSION

This paper proposed IRA-C2NN, a surveillance-video anomaly detector that couples hybrid energy-variation keyframe selection, YOLOv5 object evidence, LPQ-ensemble SP-HOG, STAP and statistical features with a channel-wise CNN using CRLN and a hybrid activation, tuned by the interactive routing algorithm. On UCSD and ShanghaiTech it reached 96.03 % and 96.21 % accuracy at 80 % training and 95.00 % and 96.60 % under 10-fold validation, ranking first in all evaluated settings and metrics against nine baselines, with the highest mean accuracy and complementary contributions from the three feature families. Evaluation on official splits with AUC, robustness to adverse weather and long-range objects, and explainability are the main directions for future work.

Declarations

The authors confirm that this manuscript is original and has not been published elsewhere nor is it under consideration by any other publication. All authors have read and approved the final manuscript and they have no conflicts of interest.

Data Availability Statement

The dataset is available from the corresponding author upon reasonable request.

Acknowledgments

Not applicable.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors contributed to the study conception and design. Author 2 and 3 performed conceptualization, methodology development, data collection, and drafted the original manuscript. Author 4 carried out software implementation, formal analysis, and contributed to manuscript review and editing. Author 1 provided supervision, project administration, and contributed to the critical revision of the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval

Not applicable. This study did not involve human participants, animals, or any procedures requiring ethical approval.

Consent to participate

Not applicable. This study did not involve human participants.

Consent to publish

Not applicable. No individual data, images or other information that requires consent for publication is included in this manuscript.

Code availability

The implementation and execution instructions are available from the corresponding author available on reasonable request.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] H. T. Duong, V. T. Le, and V. T. Hoang, "Deep learning-based anomaly detection in video surveillance: A survey," *Sensors*, vol. 23, no. 11, Art. no. 5024, 2023.

[2] K. U. Duja, I. A. Khan, and M. Alsuhaibani, "Video surveillance anomaly detection: A review on deep learning benchmarks," *IEEE Access*, vol. 12, pp. 164811–164842, 2024.

[3] N. Choudhry, J. Abawajy, S. Huda, and I. Rao, "A comprehensive survey of machine learning methods for surveillance videos anomaly detection," *IEEE Access*, vol. 11, pp. 114680–114713, 2023.

[4] E. Şengönül *et al.*, "An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey," *Appl. Sci.*, vol. 13, no. 8, Art. no. 4956, 2023.

[5] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *ACM Comput. Surv.*, vol. 53, no. 6, pp. 1–26, 2020.

[6] Y. Kim, J. Y. Yu, E. Lee, and Y. G. Kim, "Video anomaly detection using cross U-Net and cascade sliding window," *J. King Saud Univ.—Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3273–3284, 2022.

[7] Y. Chang *et al.*, "Video anomaly detection with spatio-temporal dissociation," *Pattern Recognit.*, vol. 122, Art. no. 108213, 2022.

[8] V. T. Le and Y. G. Kim, "Attention-based residual autoencoder for video anomaly detection," *Appl. Intell.*, vol. 53, no. 3, pp. 3240–3254, 2023.

[9] R. Maqsood, U. I. Bajwa, G. Saleem, R. H. Raza, and M. W. Anwar, "Anomaly recognition from surveillance videos using 3D convolution neural network," *Multimedia Tools Appl.*, vol. 80, no. 12, pp. 18693–18716, 2021.

[10] W. Ullah, A. Ullah, I. U. Haq, K. Muhammad, M. Sajjad, and S. W. Baik, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools Appl.*, vol. 80, no. 11, pp. 16979–16995, 2021.

[11] M. Qasim and E. Verdu, "Video anomaly detection system using deep convolutional and recurrent models," *Results Eng.*, vol. 18, Art. no. 101026, 2023.

[12] T. Ganokratanaa, S. Aramvith, and N. Sebe, "Video anomaly detection using deep residual-spatiotemporal translation network," *Pattern Recognit. Lett.*, vol. 155, pp. 143–150, 2022.

[13] T. H. Vu, J. Boonaert, S. Ambellouis, and A. Taleb-Ahmed, "Multi-channel generative framework and supervised learning for anomaly detection in surveillance videos," *Sensors*, vol. 21, no. 9, Art. no. 3179, 2021.

[14] G. Shen, Y. Ouyang, J. Lu, Y. Yang, and V. Sanchez, "Advancing video anomaly detection: A bi-directional hybrid framework for enhanced single- and multi-task approaches," *IEEE Trans. Image Process.*, vol. 33, pp. 6865–6880, 2024.

[15] H. Yuan, Z. Cai, H. Zhou, Y. Wang, and X. Chen, “TransAnomaly: Video anomaly detection using video vision transformer,” *IEEE Access*, vol. 9, pp. 123977–123986, 2021.

[16] N. Taghinezhad and M. Yazdi, “A new unsupervised video anomaly detection using multi-scale feature memorization and multipath temporal information prediction,” *IEEE Access*, vol. 11, pp. 9295–9310, 2023.

[17] Y. Fu, B. Yang, and O. Ye, “Spatiotemporal masked autoencoder with multi-memory and skip connections for video anomaly detection,” *Electronics*, vol. 13, no. 2, Art. no. 353, 2024.

[18] L. Wang, H. Tan, F. Zhou, W. Zuo, and P. Sun, “Unsupervised anomaly video detection via a double-flow ConvLSTM variational autoencoder,” *IEEE Access*, vol. 10, pp. 44278–44289, 2022.

[19] E. Cruz-Esquivel and Z. J. Guzman-Zavaleta, “An examination on autoencoder designs for anomaly detection in video surveillance,” *IEEE Access*, vol. 10, pp. 6208–6217, 2022.

[20] H. Tutar, A. Güneş, M. Zontul, and Z. Aslan, “A hybrid approach to improve the video anomaly detection performance of pixel- and frame-based techniques using machine learning algorithms,” *Computation*, vol. 12, no. 2, Art. no. 19, 2024.

[21] M. Alarfaj *et al.*, “Automatic anomaly monitoring in public surveillance areas,” *Intell. Autom. Soft Comput.*, vol. 35, no. 3, pp. 2655–2671, 2023.

[22] Y. Yang, Z. Fu, and S. M. Naqvi, “Abnormal event detection for video surveillance using an enhanced two-stream fusion method,” *Neurocomputing*, vol. 553, Art. no. 126561, 2023.

[23] V. A. Kotkar and V. Sucharita, “Fast anomaly detection in video surveillance system using robust spatiotemporal and deep learning methods,” *Multimedia Tools Appl.*, vol. 82, no. 22, pp. 34259–34286, 2023.

[24] S. Vosta and K. C. Yow, “KianNet: A violence detection model using an attention-based CNN-LSTM structure,” *IEEE Access*, vol. 12, pp. 2198–2209, 2024.

[25] P. Wu, J. Liu, and F. Shen, “A deep one-class neural network for anomalous event detection in complex scenes,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2609–2622, Jul. 2020.

[26] C. Huang *et al.*, “Abnormal event detection using deep contrastive learning for intelligent video surveillance system,” *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5171–5179, Aug. 2022.

[27] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, “AnomalyNet: An anomaly detection network for video surveillance,” *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[28] D. K. Singh, D. R. Pant, G. Gautam, and B. Shrestha, “Meta-learning approach for adaptive anomaly detection from multi-scenario video surveillance,” *Appl. Sci.*, vol. 15, no. 12, Art. no. 6687, 2025.

[29] S. Ö. Pelvan, B. Can, and H. Ozkan, “A hierarchical approach for improved anomaly detection in video surveillance,” *IEEE Access*, vol. 11, pp. 101644–101665, 2023.

[30] K. Doshi and Y. Yilmaz, “Continual learning for anomaly detection in surveillance videos,” in *Proc. IEEE/CVF CVPR Workshops*, 2020, pp. 254–255.

[31] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 1975–1981.

[32] W. Luo, W. Liu, and S. Gao, “A revisit of sparse coding based anomaly detection in stacked RNN framework,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 341–349.

[33] S. G. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[34] A. Cohen, I. Daubechies, and J.-C. Feauveau, “Biorthogonal bases of compactly supported wavelets,” *Commun. Pure Appl. Math.*, vol. 45, no. 5, pp. 485–560, 1992.

[35] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[36] G. Jocher *et al.*, “YOLOv5,” Ultralytics, 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[37] C.-Y. Wang *et al.*, “CSPNet: A new backbone that can enhance learning capability of CNN,” in *Proc. IEEE/CVF CVPR Workshops*, 2020, pp. 390–391.

[38] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8759–8768.

[39] V. Ojansivu and J. Heikkilä, “Blur insensitive texture classification using local phase quantization,” in *Proc. Int. Conf. Image Signal Process.*, 2008, pp. 236–243.

[40] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 886–893.

[41] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” 2016, *arXiv:1607.06450*.

[42] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, “Activation functions in deep learning: A comprehensive survey and benchmark,” *Neurocomputing*, vol. 503, pp. 92–108, 2022.

[43] S. Goel, “Pigeon optimization algorithm: A novel approach for solving optimization problems,” in *Proc. Int. Conf. Data Mining Intell. Comput. (ICDMIC)*, 2014, pp. 1–5.

[44] D. Wang, D. Tan, and L. Liu, “Particle swarm optimization algorithm: An overview,” *Soft Comput.*, vol. 22, no. 2, pp. 387–408, 2018.

[45] H. Alazzam, A. Sharieh, and K. E. Sabri, “A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer,” *Expert Syst. Appl.*, vol. 148, Art. no. 113249, 2020.

[46] Z. Wang, Y. Fu, C. Song, P. Zeng, and L. Qiao, “Power system anomaly detection based on OCSVM optimized by improved particle swarm optimization,” *IEEE Access*, vol. 7, pp. 181580–181588, 2019.

[47] C. Wu, S. Shao, C. Tunc, and S. Hariri, “Video anomaly detection using pre-trained deep convolutional neural nets and context mining,” in *Proc. IEEE/ACS 17th Int. Conf. Comput. Syst. Appl. (AICCSA)*, 2020, pp. 1–8.

[48] Y. Lu, K. M. Kumar, S. S. Nabavi, and Y. Wang, “Future frame prediction using convolutional VRNN for anomaly detection,” in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, 2019, pp. 1–8.

[49] C. Wu, S. Shao, C. Tunc, P. Satam, and S. Hariri, “An explainable and efficient deep learning framework for video anomaly detection,” *Cluster Comput.*, vol. 25, no. 4, pp. 2715–2737, 2022.