CHAPTER 2

LITERATURE SURVEY

Several approaches have been proposed for preprocessing, feature extraction and classification of anomalies utilizing deep learning techniques. Various Video Anomaly Detection (VAD) and Transfer Learning algorithms have been developed and refined over time. This chapter provides a comprehensive discussion of these techniques.

2.1 VIDEO ANOMALY DETECTION TECHNIQUES

Bouindour et al. (2019) proposed an unsupervised anomaly detection method using a modified 3D residual convolutional network for extracting spatiotemporal features and a robust classifier for detecting abnormal events in video surveillance. The methodology relied on training with normal events, minimizing redundancy and enabling adaptation to new normal events during testing. The model achieved strong performance, with Area Under the Curve (AUC) values of 86% in standard surveillance scenarios and 84% in more dynamic or cluttered environments, demonstrating its effectiveness in both frame-level and pixel-level anomaly detection tasks. However, its robustness in highly complex scenarios remained limited, highlighting potential challenges for deployment in real-world applications.

Xu et al. (2019) proposed an anomaly detection system for crowded scenes using a variational auto-encoder with convolutional kernels, which extracts appearance and motion features directly from raw video frames. Evaluated on the University of California, San Diego (UCSD) Pedestrian dataset, the model achieved 95.7% AUC on Pedestrian1 and 92.3% on Pedestrian2, demonstrating competitive performance compared to state-of-the-art methods. However, the reliance on Gaussian-based anomaly scoring limited adaptability to highly dynamic environments, potentially reducing detection robustness in complex real-world scenarios.

Alshammari & Rawat (2019) developed an intelligent multi-camera surveillance system that integrated multiple viewpoints to reduce blind spots and enhance anomaly detection accuracy. The system effectively handled overlapping fields of view, providing comprehensive coverage for urban surveillance. However, high computational demands and

variations in camera quality restricted scalability in large-scale environments. This work demonstrated the importance of multi-camera setups for improving detection reliability.

Nawaratne et al. (2020) proposed an Incremental Spatio-Temporal Learner that combined unsupervised deep learning, active learning and fuzzy aggregation to enable real-time Anomaly Detection (AD). The model demonstrated robust performance and computational efficiency on benchmark datasets, achieving 95.3% Accuracy with low false positive rates, making it suitable for diverse industrial applications. However, it faced challenges with sparse or noisy data, which limited adaptability in certain scenarios. The study underscored the need for advanced preprocessing to handle data inconsistencies effectively.

Direkoglu (2020) proposed an abnormal crowd behavior detection method using Motion Information Images (MIIs) and a Convolutional Neural Network (CNN). The approach achieved high Accuracy, with 99.08% on the University of Minnesota (UMN) dataset and 98.39% on Performance Evaluation of Tracking and Surveillance 2009 (PETS2009) dataset, demonstrating superior performance. Despite its effectiveness, the model relies on accurate optical flow computation and faces challenges with occlusions or varying crowd densities.

Al-Dhamari et al. (2020) applied Transfer Learning with VGGNet-19 for feature extraction and employed a Support Vector Machine (SVM) classifier for anomaly detection, accomplishing a high Accuracy of 97.44% and AUC of 0.9795. The model demonstrated the effectiveness of leveraging pre-trained networks for rapid deployment. However, its reliance on large labeled datasets limited its applicability in data-scarce environments. The study emphasized the need for semi-supervised techniques to overcome labeling challenges.

Singh et al. (2020) developed a neural network-based approach for real-time anomaly recognition using CCTV data. The proposed methodology achieved a 5.8% improvement in recognition accuracy compared to traditional methods, highlighting its effectiveness in detecting anomalies in surveillance videos. However, the model's high computational complexity and significant processing power requirements present limitations for deployment in resource-constrained environments.

Nasaruddin et al. (2020) proposed a visual attention mechanism for deep anomaly detection in surveillance videos. The model employed a robust background subtraction technique to extract motion and identify attention regions, which were then fed into a three-dimensional Convolutional Neural Network (CNN). By leveraging Convolution 3-dimensional to exploit spatiotemporal relations fully, the model effectively distinguished between normal and anomalous events, achieving enhanced anomaly localization with an Accuracy of 99.25%. However, the approach was prone to misclassification in crowded scenes, presented challenges in densely populated environments.

Ullah et al. (2021) proposed a Bi-Directional Long Short-Term Memory (BD-LSTM) framework for spatiotemporal anomaly detection, integrating pre-trained CNNs to extract deep features from video frames, which were then processed to capture both forward and backward temporal dependencies. The methodology achieved notable performance improvements, with a 3.41% Accuracy increase on University of Central Florida-Crime (UCF-Crime) dataset and an 8.09% improvement on UCFCrime2Local dataset compared to state-of-the-art methods, demonstrating its effectiveness in complex surveillance scenarios. However, the bidirectional processing introduced high memory consumption, limiting the framework's scalability for large-scale applications. Despite this limitation, the study showcased the potential of bidirectional architectures for accurate anomaly detection in surveillance networks.

Rendón-Segador et al. (2021) introduced ViolenceNet, a deep learning framework integrating DenseNet-3D, multi-head self-attention and BD-LSTM to encode spatiotemporal features for automatic violence detection in video surveillance. The model achieved 95.6% Accuracy on the most Violence Situations dataset (most complex) and 100% on the Movies Fights dataset (simplest), reducing network parameters to 4.5 million and achieving inference times under 0.3 seconds, demonstrating its efficiency. Yet, cross-dataset analysis revealed a drop in Accuracy to 70.08%–81.51%, highlighting challenges in generalization and the need for improved anomaly detection in diverse datasets.

Li et al. (2021) proposed a Two-Stream Deep Spatial-Temporal Auto-Encoder for VAD, utilizing a spatial stream for appearance features and a temporal stream for motion patterns, integrating them through joint reconstruction error. The model demonstrated superior accuracy on the UCSD, Avenue and UMN datasets, leveraging optical flow to

improve continuity extraction and inter-frame motion understanding. However, the reliance on the optical flow increases the computational complexity, which limits the real-time processing efficiency in large-scale surveillance applications.

Cherian & Poovammal (2021) proposed an anomaly detection method using multiple instances learning and Iterative Dichotomiser 3 for feature extraction, followed by a deep neural network for classification. On a dataset with 128 hours of video, including 10% anomaly videos, the approach achieved an AUC of 81%, demonstrating its effectiveness but with the need for improvement in handling diverse anomalies.

Kim et al. (2022) proposed a Cross UNet (U-Shaped Network) framework incorporating two interconnected UNet subnetworks and a Cascade Sliding Window method for real-time anomaly scoring. The methodology was designed to enhance both detection accuracy and speed, with outputs from the contracting path of one subnetwork combined with corresponding outputs from the other to optimize spatial and temporal feature handling. The framework achieved competitive AUC scores of 97.0% on UCSD Pedestrian2, 90.8% on Avenue and 72.5% on ShanghaiTech datasets. However, risks of overfitting were identified due to the limited availability of training data, which restricted the model's generalizability in complex real-world applications.

Zhang, Z. (2022) proposed an ensemble GAN (Generative Adversarial Network) for VAD, training multiple generators and discriminators for improved normal data modeling. The method outperformed single GANs, achieving AUCs of 90.6% for CUHK Avenue and 75.1% for ShanghaiTech datasets. While enhancing detection accuracy, its computational complexity limits real-time applications.

Liu, T. et al. (2022) suggested a novel method that combined Hybrid Dilated Convolution (HDC) and Deeper Bidirectional Convolutional Long Short-Term Memory (DB-ConvLSTM) to enhance AD in surveillance videos. The methodology balanced the advantages of reconstruction-based and future frame prediction models by utilizing adaptive receptive fields to extract comprehensive spatial features while leveraging DB-ConvLSTM to capture temporal continuity across frames. The approach achieved competitive AUC scores of 96.6% on UCSD Pedestrian2 (Peds2), 85.1% on UCSD Pedestrian1(Peds1) and 86.5% on Avenue datasets, demonstrating its effectiveness across diverse video scenes.

However, the high computational requirements for processing extended sequences had limited its feasibility for real-time applications, highlighting the need for optimization in advanced architectures.

Wu et al. (2022) proposed a Self-Supervised Sparse Representation framework for VAD, integrating sparse coding for feature learning, temporal consistency constraints and a memory-augmented autoencoder to improve anomaly detection at the feature level. The model achieved AUC scores of 89.3% on Peds2 and 86.7% on the Avenue dataset while demonstrating efficient training with limited data. The method's reliance on pseudo anomaly generation introduced noise and its performance depends on learned sparse representations, potentially limiting real-world generalization.

Ganokratanaa et al. (2022) proposed the Deep Residual Spatiotemporal Translation Network (DR-STN), which utilized conditional GANs with Online Hard Negative Mining for anomaly detection in surveillance videos. The framework achieved an average AUC of 96.73% on UCSD, UMN and Chinese University of Hong Kong (CUHK) Avenue datasets, outperforming advanced methods by 7.6%. The method performed well in both frame-level and pixel-level evaluations; however, the dense optical flow computations added significant processing overhead, limiting its scalability for large-scale applications.

Hao et al. (2022) proposed a spatiotemporal consistency-enhanced network using a 3D CNN encoder, 2D CNN decoder and adversarial training for video anomaly detection. A 3D CNN discriminator ensured input-frame consistency, achieving advanced performance on ShanghaiTech, CUHK Avenue and Peds2 datasets. However, high computational overhead and training complexity limited its real-time feasibility.

Kotkar & Sucharita (2023) suggested a framework for VAD combining lightweight keyframe extraction, Modified Spatio-Temporal features and a Recurrent Neural Network (RNN) - Long Short-Term Memory (LSTM) classifier. Keyframe extraction reduced computational overhead by 50%, while robust feature extraction with Discrete Wavelet Transform and Principal Component Analysis (PCA) enhanced detection accuracy. The model achieved a 4.5% improvement in Accuracy and a 21% reduction in processing time compared to state-of-the-art methods. Though, its reliance on advanced preprocessing posed challenges for highly dynamic scenarios.

Yu et al. (2023) proposed an attention-guided residual frame learning approach using a Convolutional Long Short-Term Memory (ConvLSTM) model for VAD. It employs self-attention and residual learning to enhance anomaly focus and training stability. The model outperformed on the CUHK Avenue and Subway Exit datasets but struggle in highly dynamic scenes.

Pelvan et al. (2023) introduced a Context Tree-Based Anomaly Detection, a hierarchical ensemble approach for VAD that balances the bias-variance trade-off. It partitions the image plane into complex models, transitioning from simple to complex as more data becomes available. This method reduces training samples while improving locational anomaly detection, achieving AUC values of 0.93 for UCSD, 0.77 for Avenue, 0.77 for Shanghai and 0.48 for Street Scene datasets. But, its reliance on partitioning and weighted model combinations introduced computational overhead.

Monakhov et al. (2023) introduced Grid-Based Hierarchical Temporal Memory (GridHTM). This unsupervised anomaly detection framework that enhances Hierarchical Temporal Memory to improve noise tolerance and online learning in video surveillance. The model was tested on the Video and Image Retrieval and Analysis Tool (VIRAT) Dataset. It demonstrated real-time anomaly detection capabilities and its AS output effectively identified anomalous segments, as visualized in the GridHTM AS plot. However, despite its advantages, GridHTM's performance in highly heterogeneous environments and its ability to handle extreme class imbalances remain areas for further improvement.

Zhang, Q. et al. (2023) proposed a VAD model based on an auto-encoder with an attention mechanism to enhance feature representation and reduce unnecessary background information. By incorporating deep separable convolution to lower model complexity, the method achieved AUC scores of 80.5% on Peds1, 97.9% on Peds2 and 85.9% on CUHK Avenue datasets, improving detection Accuracy by 1.9%, 1.4% and 6.6%, respectively, over the baseline model. But its efficiency decline under limited computing resources and background noise can still impact detection accuracy in complex real-world scenarios.

Zhang, L. et al. (2023) suggested Cascaded Memory-Augmented Autoencoder, for VAD, integrating memory-enhancing modules, Squeeze-and-Excitation attention mechanisms and skip connections to improve feature learning and reconstruction quality.

The model achieved an AUC of 99.2% on Peds2 and 89.4% on CUHK Avenue datasets, demonstrating superior performance in detecting anomalies. Yet, reliance on memory modules introduces scalability challenges and computational complexity could delay real-time deployment.

Liu et al. (2023) proposed a Prompt-based Feature Mapping Framework to address the anomaly and scene gap in VAD by generating unseen anomalies with unbounded types and adapting feature representations across different surveillance environments. The method achieves state-of-the-art performance on three benchmark datasets, demonstrating improved generalization capabilities over previous approaches. However, the framework relies on synthetic anomaly generation, which introduce domain adaptation challenges and its effectiveness depends on the quality of the anomaly prompts and mapping network.

Gupta et al. (2024) introduced a VAD model using a modified Deep Neural Network (DNN) integrated with the Histo Sigmoid of Orientation and Enthalpy with Fast Accelerated Segment Test (HSOE-FAST) method for feature extraction, where input frames are pre-processed using a Gaussian filter before classification. The extracted features are then classified using the modified Deep Neural Network (DNN) approach, which enhances anomaly detection accuracy in surveillance footage. The model achieved approximately 99% Accuracy, outperforming existing techniques by effectively identifying anomalies, but its reliance on a specific feature extraction technique HSOE-FAST and Gaussian filtering limit generalization across diverse datasets with varying noise levels and complex background dynamics.

AlMarri et al. (2024) proposed a multi-head architecture for weakly supervised VAD, using margin loss and stochastic segment shuffling to handle noisy labels and data imbalance. The model achieved AUC scores of 85.47% on UCF-Crime dataset and 91.53% on Extensive and Diverse-Violence (XD-Violence) dataset, reducing annotation reliance while maintaining accuracy. Though, variability in performance across datasets highlighted challenges in generalization.

Wang, J. et al. (2024) suggested a Diverse Motion-conditioned Adversarial Predictive Network for VAD, integrating conditional variational generation and adversarial learning. A motion-guided generator leverages optical flow, while diversity regularization

preserves normal patterns. A video discriminator improves anomaly detection, but reliance on synthetic anomalies limits real-world adaptability.

Tran et al. (2024) proposed a transformer-based Spatio-temporal unsupervised anomaly detection framework that predicts future frames in aerial traffic videos, identifying anomalies based on high reconstruction errors. The model outperformed on the University of Information Technology - Anomalous Drone (UIT-ADrone) dataset. It significantly outperformed state-of-the-art methods on the Drone-Anomaly dataset, demonstrating its effectiveness in handling small, multiscale objects and complex backgrounds. But challenges remain due to high object similarity, limited anomalous event occurrences and the need for further optimization in remote sensing applications.

Mei et al. (2024) introduced a graph-based domain adversarial framework for VAD, improving generalization across datasets. It integrates graph convolution and domain-invariant learning with a gradient reversal layer for robust feature mapping. The model achieved AUC gains of 12.47% on Avenue and up to 8.64% on UCF-Crime, enhancing cross-domain detection. However, it depends on labeled source data, limiting fully unsupervised adaptability.

Wang, Z. et al. (2024) proposed a Spatio-Temporal Generalization model for VAD, addressing domain generalization challenges. The method classifies anomalies into Normal Object and Abnormal Behavior, Abnormal Object and Normal Behavior (AONB) and Abnormal Object and Abnormal Behavior (AOAB). Integrating contrastive learning and adaptive data augmentation improves AONB and AOAB detection without auxiliary datasets. While achieving strong generalization, its effectiveness declines in highly diverse anomaly scenarios.

Dhevanandhini & Yamuna (2024) suggested an intelligent video surveillance system using Modified Barnacles Mating Optimization for segmentation and Chaotic Hummingbird Optimization for feature selection. A Hybrid CNN-Supreme Gradient Boost classifier enhances object detection accuracy, outperforming state-of-the-art models on Penn-Fudan, Daimler and Inria datasets. While improving precision, its computational complexity limits real-time scalability.

Lv & Sun (2024) introduced a Long-Term Context (LTC) module within vision-language learning models to enhance contextual understanding for VAD. The methodology improved the modeling of long-range dependencies, eliminating the need for thresholding and enabling textual explanations for detected anomalies. The model had achieved top performance on UCF-Crime and TAD benchmarks datasets, with AUC improvements of +3.86% and +4.96%, respectively and additional gains in weakly supervised settings (+0.88% and +2.44% on UCF-Crime). However, scalability to larger datasets and real-time applications had remained a limitation, underscoring challenges in adapting such methods to diverse scenarios.

Li & Chen (2024) proposed a Dynamic Multiple-Instance Learning (Dy-MIL) framework for weakly supervised VAD. It uses a dynamic ranking method with k-max-selection to improve anomaly separation using only video-level labels. The model outperformed others on ShanghaiTech, UCF Crime and Nanjing University of Technology (NUT) datasets, but weak supervision limits its ability to detect novel anomalies.

Ganagavalli & Santhi (2024) introduced an improved YOLO-based deep learning framework for automated crime detection and tracking in video surveillance, fine-tuning the model with optimized hyperparameters. The system achieved an AUC of 0.91 for detecting vandalism behavior and 0.8299 across 14 crime activity classes, outperforming existing methods in precision, training loss, testing loss and F1 Score. Still, challenges remain in real-time processing, scalability to diverse environments and handling occlusions in crowded scenes, which impact detection accuracy.

Fu et al. (2024) proposed Spatiotemporal Masked Autoencoder with Multi-memory modules and Skip connections (SMAMS), to enhance VAD by capturing high-level semantic and temporal contextual features. The model achieved AUC scores of 99.9% on Peds2, 94.8% on CUHK Avenue and 78.9% on ShanghaiTech datasets, outperforming state-of-the-art methods in anomaly detection accuracy. But, its reliance on masked video patches introduced information loss and its computational complexity could pose challenges for real-time surveillance applications.

2.2 TRANSFER LEARNING TECHNIQUES

Perera & Patel (2019) introduced a Transfer Learning (TL) framework for one-class classification, leveraging CNNs to detect anomalies as deviations from normal data

representations. The methodology used compactness and descriptiveness loss functions alongside a parallel CNN architecture to learn low intra-class variance features, reducing dependence on labeled anomaly data. The model demonstrated significant improvements over state-of-the-art methods in anomaly detection and novelty detection tasks. However, its applicability was limited to one-class settings, restricting its use in multi-class anomaly detection scenarios.

Liu et al. (2020) proposed a TL framework leveraging action recognition models to enhance anomaly detection in surveillance videos. The approach utilized techniques such as training on small-scale datasets, temporal modules and separate networks for distinguishing hard examples from normal activities, improving efficiency and reducing training time. The framework demonstrated strong performance on the CitySCENE benchmark dataset, achieving category accuracies ranging from 25.0% to 100%, with notable results for categories like explosion by 100% and normal activities by 97.5%. However, the model's performance relied heavily on dataset compatibility, limiting its generalizability across diverse scenarios.

Song et al. (2021) introduced a tri-directional TL approach for predicting gastric cancer morbidity across different regions and disease types, integrating univariate regression, multivariate Gaussian modeling and mapping-based deep transfer learning to adapt predictive models. The proposed method achieved over 80% Accuracy in the source region and up to 78% in target regions, demonstrating its effectiveness in leveraging pollutant-related data for disease prediction with limited labeled samples. Still, its performance be constrained by regional data variability, requiring further adaptation to diverse environmental and demographic conditions.

Khaire & Kumar (2022) proposed a semi-supervised CNN-Bidirectional Long Short-Term Memory (BiLSTM) framework leveraging RGB-Depth (RGB-D) data and TL to detect anomalies in critical surveillance environments like Bank-ATMs. The model, trained on weakly labeled normal samples, achieved competitive results on a custom RGB-D ATM dataset, Avenue and UCFCrime2Local datasets. While effective, its reliance on RGB-D data limited its generalization to non-multi-modal scenarios.

Liu, T. et al. (2022) suggested the Spatio-Temporal Prediction and Reconstruction Network (STPR-net) using TL with hybrid dilated convolutions and bidirectional

ConvLSTM to enhance VAD. The model achieved AUC scores of 85.1%, 96.6% and 86.5% on Peds1, Peds2 and CUHK Avenue datasets, respectively, outperforming state-of-the-art methods. Yet, its reliance on complete normal training data limited scalability in inconsistent environments.

Solanki et al. (2023) conducted a comparative analysis of TL models for VAD, evaluating deep learning architectures such as Visual Geometry Group (VGG16), (VGG19), Residual Network (ResNet50) and Densely Connected Convolutional Network 121 (DenseNet121) on subsets of the UCF Crime dataset. Their findings show that DenseNet121 performs best with a Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) score of 0.85, demonstrating its effectiveness in detecting anomalies in real-world surveillance footage. However, the study focuses on pre-trained models without extensive fine-tuning and not generalize well to unseen datasets or complex anomaly patterns beyond the training data.

Hafeez et al. (2023) proposed a Generative Transfer Learning (GTL) algorithm for VAD that utilizes an unsupervised Transformer-based framework comprising a feature extractor, generator transformer and discriminator transformer to identify anomalies through high reconstruction errors. Tested on the UCF-Crime and ShanghaiTech datasets, this approach demonstrates its potential in handling imbalanced and unlabeled data but face computational challenges and sensitivity to adversarial anomalies.

Kommanduri & Ghorai (2023) introduced a supervised VAD method for TL using Mobile Network Version 2 (MobileNetV2). Fine-tuned on annotated data, the model achieved high accuracy with low computational cost. While effective, its reliance on labeled data limits adaptability to unseen anomalies.

Dilek & Dener (2024) proposed a frame-level VAD method leveraging TL and Fine-Tuning (FT) with 20 pre-trained CNN-based deep learning models, including VGG, ResNet, MobileNet and EfficientNet. The approach demonstrated state-of-the-art performance, achieving AUC scores of 100% on Peds1 and Peds2 datasets; and 98.41% on CUHK Avenue datasets, surpassing existing Accuracy, Precision, Recall and F1 Score. Despite its efficiency and accuracy, reliance on pre-trained models and high resource demands limits its use in resource-constrained environments, emphasizing the need for lighter models.

Nie et al. (2024) proposed an Unsupervised VAD framework that integrates One-Class Classification (OCC) and Weakly-Supervised (WS) learning through an interleaving training module, introducing Weighted OCC (WOCC) for improved label consistency and an adaptive thresholding strategy for refining pseudo-labels. The method achieves superior performance, with an AUC of 88.18% for WS and 82.57% for Spatio-Temporal Graph Normalizing Flow (STG-NF) on the ShanghaiTech dataset, along with a notable improvement from 70.48% to 74.76% on Unsupervised Behavior Anomaly Detection (UBnormal) dataset. Despite its advantages, the approach requires user intervention for threshold refinement and pseudo-label randomness affects performance, necessitating multiple training cycles for stability.

Majhi et al. (2024) introduced a Human-Scene Network optimized with self-rectifying loss, using TL to learn subtle and strong spatiotemporal cues for WS VAD. The model achieved state-of-the-art performance on UCF-Crime, ShanghaiTech and Indian Institute of Technology Bombay - Corridor datasets, excelling in five of six scenarios. Despite its success, the high computational requirements limited its feasibility for real-time deployment, highlighting the need for more resource-efficient approaches.

2.3 PREPROCESSING TECHNIQUES

Jayachandra & Kamal (2019) proposed a real-time marine video image preprocessing technique using filters and the Point Spread Function to enhance underwater video frames. The method achieved high-quality results with PSNR values above 60dB, demonstrating effective noise reduction. However, the computational demands of the preprocessing steps posed challenges for real-time applications in resource-limited environments.

Dos Santos et al., (2020) introduced a preprocessing technique such as Contrast Limited Adaptive Histogram Equalization for contrast enhancement and the Wiener filter for noise reduction to improve image and video quality. These methods achieved 95.05% Accuracy, 75.64% sensitivity and 96.96% specificity on the Digital Retinal Images for Vessel Extraction dataset for blood vessel segmentation, demonstrating their effectiveness in medical imaging. Yet, they introduce computational overhead, limiting real-time applicability and requiring extensive tuning for diverse datasets.

Ahlawat et al. (2020) emphasized the significance of preprocessing in improving handwritten digit recognition accuracy using CNNs. The study employed techniques such as scaling, noise reduction, centering, slant correction and skew estimation to standardize input images before feeding them into the model, ensuring better segmentation and feature extraction. By optimizing CNN design parameters like layer depth, stride size and kernel configurations, the proposed approach achieved a remarkable 99.87% Accuracy on the Modified National Institute of Standards and Technology dataset or MNIST dataset, surpassing ensemble-based architectures while reducing computational complexity. Still, the reliance on dataset-specific preprocessing limits the model's adaptability to more diverse handwriting styles and real-world variations.

Liu et al. (2021) proposed a CNN-Enhanced Graph Convolutional Network for hyperspectral image classification, combining CNNs for pixel-level and Graph Convolutional Networks (GCN) for pixel-level feature fusion. A graph encoder-decoder facilitated collaboration between Euclidean and non-Euclidean data, eliminating preprocessing-based graph encoding. The model achieved state-of-the-art performance across three datasets but faced challenges with computational complexity for large-scale applications.

Zhang, G. et al. (2022) enhanced pedestrian detection in YOLO-V4 by applying preprocessing techniques like image resolution adjustment, multiscale feature fusion and anchor box refinement. These optimizations improved feature representation and pedestrian size detection, achieving 87.8% Average Precision (AP) on CrowdHuman and 95.5% AP on CityPersons. Yet, the increased computational complexity limited real-time feasibility on resource-constrained devices.

Mumcu et al. (2022) analyzed adversarial attacks on VAD using frame manipulation and resolution adjustments, simulating effects like slowing, freezing and fast-forwarding. Testing on CUHK Avenue and ShanghaiTech datasets exposed vulnerabilities in models like Future Frame Prediction Network, Memory-guided Normality for Anomaly Detection and Modular Online Video Anomaly Detector, with PSNR dropping from 40–50dB in original videos to 30–45dB under frame manipulation and further decreasing to 25–45dB in in low-resolution adversarial datasets. While the study highlighted model weaknesses, it lacked mitigation strategies for real-world applications.

Janaki & Lakshmi (2024) introduced the Hybrid Model-Based Esophageal Disorder Diagnosis (HMEDD) framework for early Esophageal disorder detection using gastroscopic images, leveraging six augmentation techniques to enhance dataset diversity. These preprocessing methods improved the Esophageal Convolutional Neural Network (EsoNet) model, achieving 92.15% Accuracy across six disorder categories. However, the framework's generalizability was limited by the small dataset size and lack of validation on diverse clinical data.

2.4 FEATURE EXTRACTION TECHNIQUES

Gong et al. (2019) proposed Memory-Augmented AutoEncoder (MemAE) for unsupervised VAD, using memory modules to store normal patterns. By querying relevant memory items, MemAE accurately reconstructed normal data while amplifying anomalies, improving detection in complex scenarios. It achieved high performance across datasets but faced scalability issues due to high memory and computational demands, necessitating more efficient implementations.

Direkoglu (2020) proposed a feature extraction method for detecting abnormal crowd behaviors using MIIs generated from optical flow vectors. MIIs, based on optical flow magnitude and angle differences, visually represented crowd motion, enabling a CNN to effectively classify behaviors such as panic and escape. The method achieved remarkable Accuracy, with 99.08% on the UMN dataset and 98.39% on the PETS2009 dataset. However, the computational demands of generating MIIs from optical flow limited its applicability for real-time or large-scale scenarios, emphasizing the need for more efficient solutions.

Balasundaram & Chellappan (2020) developed an Intelligent Video Analytics Model (IVAM) for anomaly detection, integrating motion and appearance features with an attention mechanism to prioritize salient regions in video frames. Motion vectors combined with CNN-based spatial features enabled accurate detection of abnormal activities in public spaces like airports and shopping malls, achieving high classification accuracy with a low error rate. The system demonstrated real-time applicability in structured environments but faced challenges in highly dynamic scenarios due to the limitations of a fixed attention mechanism. This highlights the need for adaptive attention mechanisms to handle complex and dynamic scenes effectively.

Doshi & Yilmaz (2020) proposed an online anomaly detection methodology combining TL and continual learning to address challenges in surveillance video analysis. The approach utilizes neural network-based models for feature extraction and statistical detection methods to enable continual learning, reducing training complexity and adapting to new data without catastrophic forgetting. This makes it particularly effective in resource-constrained and dynamic environments. However, while the method excelled in high-dimensional scenarios, it struggled with complex datasets requiring deeper contextual understanding, highlighting its limitations in handling more intricate anomaly patterns.

Wu et al. (2020) proposed a VAD framework that leverages pre-trained CNNs for feature extraction and context mining, followed by a denoising autoencoder to achieve efficient anomaly detection with low model complexity, making it suitable for resource-constrained Internet of Things (IoT) edge devices. The approach achieves AUC scores of 84.1% on Peds1 and 92.4% on Peds2, demonstrating competitive performance compared to state-of-the-art methods while reducing computational costs. However, the reliance on pre-trained CNNs limit adaptability to unseen scenarios and the method primarily focuses on high-level object-based features, potentially missing fine-grained motion-based anomalies.

Feng et al. (2021) proposed a two-stream autoencoder-based architecture for anomaly detection in surveillance videos, focusing on separate spatial and temporal feature extraction. The spatial stream, comprising convolutional and deconvolutional layers, efficiently extracts low-level visual features, while the temporal stream integrates ConvLSTM layers to capture temporal dependencies across adjacent frames. Their combined spatial and temporal model achieved high frame-level AUC scores of 80.3 on Avenue, 84.5 on Peds2, 87.3 on Subway Entrance and 90.8 on Subway Exit, outperforming several existing methods. However, the approach is computationally intensive due to dual-stream processing and constrained by limited scalability for longer video sequences, highlighting areas for further optimization.

Huang, C. et al. (2022) suggested the Temporal-Aware Contrastive Network (TAC-Net), which utilizes deep contrastive self-supervised learning to improve feature discrimination for anomaly detection in intelligent video surveillance systems. By integrating multiple self-supervised tasks and incorporating temporal context, TAC-Net captures high-level semantic features and achieves superior anomaly detection performance

across three benchmark datasets. The method's unsupervised approach makes it scalable to real-world applications, but challenges with unbalanced datasets and high computational costs highlight areas for further improvement. TAC-Net demonstrates the effectiveness of contrastive learning in refining anomaly classification while addressing limitations of single-task-based approaches.

Liu, Y. et al. (2022) proposed a collaborative normality learning framework for WS VAD, integrating an unsupervised auto-encoder to learn prototypical Spatio-temporal patterns of normal videos and a regression module to refine anomaly detection. The model achieved state-of-the-art performance on three benchmark datasets, demonstrating improved distinguishability between normal and abnormal events through iterative fine-tuning. However, its reliance on video-level labels limit precise temporal anomaly localization and performance could be affected by noisy labels in weakly supervised settings.

Mandal et al. (2021) explored the use of Residual Network-50 (ResNet-50) for feature extraction in VAD, focusing on challenging scenarios like occlusions. The model demonstrated robust feature representation under difficult conditions, including low-quality video data, but its performance diminished with extremely degraded inputs. Leveraging ResNet-50's architecture, the study showcased its adaptability for anomaly detection tasks while achieving a Precision of 89.33%, Recall of 89.70% and F1 Score of 89.7% for unmasked data, with significantly lower metrics for masked data. Despite its effectiveness, the study highlighted limitations in handling occlusion and poor video quality, emphasizing the need for further optimization for real-world applications.

Baccouche (2021) proposed a YOLO-based fusion model for breast lesion detection and classification, integrating multiple YOLO architectures to simultaneously localize and classify abnormal mammogram lesions. The model preprocesses raw images, detects lesions and classifies them as masses or calcifications. Evaluated on three datasets (Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM), Portuguese Breast Cancer Mammography Dataset (INbreast) and a private dataset), it achieved detection accuracy rates of 95.7%, 98.1% and 98% for mass lesions and 74.4%, 71.8% and 73.2% for calcification lesions, respectively. Despite its strong performance in detecting small and overlapping lesions, the method was computationally intensive and dataset-dependent, underscoring the need for optimization for broader applications.

Taghinezhad & Yazdi (2023) introduced an unsupervised VAD framework based on a U-Net-like architecture focused on frame prediction. The method utilizes a time-distributed 2D CNN encoder-decoder and incorporates a memory module to store normal patterns during training, enabling poor predictions for anomalous inputs. The model employs an upstream multi-branch structure with dilated convolutions to enhance feature extraction, capturing multiscale contextual information. The method was evaluated on the Peds1 dataset, achieving an AUC of 83.8 and an EER of 22.2, on the Peds2 dataset with an AUC of 97.6 and an EER of 6.6 and on the CUHK Avenue dataset, where it obtained an AUC of 89.0 and an EER of 18.1, outperforming existing models and demonstrating superior anomaly detection capabilities. However, its reliance on memory networks and computationally intensive multi-path structures poses challenges for real-time applications.

Hao et al. (2022) proposed a spatiotemporal consistency-enhanced network for VAD, combining a 3D CNN-based encoder and a 2D CNN-based decoder to extract spatiotemporal features. Adversarial training with a 3D CNN-based discriminator improves prediction consistency, enabling robust anomaly detection. Tested on ShanghaiTech, CUHK Avenue and Peds2, the model achieved state-of-the-art performance but is computationally intensive, limiting its use in real-time applications.

Huang, H. et al. (2022) introduced a spatial-temporal ConvLSTM model for vehicle driving intention prediction, combining LSTM for temporal features and ConvLSTM for spatial interactions and temporal evolution of surrounding vehicles. Tested on a real road dataset, the model improved Accuracy by 3.1%, Precision by 10.5% and Recall by 6.1% over traditional LSTM. While effective, its computational complexity limits real-time applicability.

Zhou et al. (2022) proposed a context-aware anomaly detection framework for underwater exploration, extracting environmental and contextual features. It combines a patch-level autoencoder with a context-enhanced autoregressive network, sharing a common encoder for latent feature extraction. The autoregressive branch captures semantic dependencies, while anomaly detection is based on weighted reconstruction and feature similarity losses. Tested on Canadian Institute for Advanced Research-10 (CIFAR-10) and three underwater datasets, it achieved Area Under the Receiver Operating Characteristic

Curve (AUROC) gains of 6.36%, 32.45% and 40.17%, demonstrating effectiveness. However, its domain-specific design restricts broader applicability.

Asad et al. (2022) suggested a multi-stream two-stage architecture for VAD, with the first stage using a 3D Convolutional Autoencoder (3DCAE) to extract appearance and motion features. The second stage clusters latent features to eliminate noise and employs a Deep one-class Support Vector Data Description (SVDD) classifier for anomaly scoring. Tested on UCSD Pedestrian, Shanghai Tech and Avenue datasets, the model showed significant performance improvements but faces scalability challenges due to computational complexity and reliance on dynamic flow inputs.

Ragedhaksha et al., (2023) used a pre-trained CNN for spatial feature extraction in live CCTV feeds, enabling object detection, counting and anomaly detection with seamless integration via Application Programming Interface (API) calls. The approach is efficient and practical but lacks temporal feature analysis, limiting its effectiveness in dynamic scenarios requiring motion understanding, highlighting the need for improvements to address temporal dependencies in video processing.

Lv & Sun (2024) proposed a LTC module for Vision-Language Learning Models, enhancing long-range contextual understanding. While achieving notable AUC improvements, the model's scalability to larger datasets remained unexplored.

Liu et al. (2024) proposed a reactive deep learning-based model for quality assessment and anomaly detection in airport surveillance videos, integrating 3D CNNs (e.g., jitter, occlusion and malfunction) with 2D CNNs for image quality assessment. The model achieved 96.48% Accuracy in anomaly detection, surpassing existing methods by at least 3.39% and demonstrated an average correlation of 0.9014 in image quality assessment, highlighting its robustness. However, its performance be affected by varying lighting conditions and camera distortions, requiring further enhancements for real-world adaptability.

Shin et al. (2024) proposed a multi-stage deep learning model for weakly supervised video anomaly detection, combining two-stream feature extraction using a Vision Transformer based Contrastive Language-Image Pretraining module and a CNN-based Inflated 3D Convolutional Network module with Temporal Contextual Aggregation. The

features are processed by an Uncertainty-Regulated Dual Memory Unit, integrating GCN and Multi-Head Self-Attention for hierarchical feature extraction. Tested on ShanghaiTech, XD-Violence and UCF-Crime datasets, the model outperformed state-of-the-art methods but requires significant computational resources, limiting real-time applicability.

Amin et al. (2024) suggested a multi-attention-based deep learning system for anomaly detection in pandemic surveillance videos, integrating EfficientNet-B0 with the Convolutional Block Attention Module to enhance feature extraction and spatial information retention. The model demonstrated a substantial improvement in classification accuracy, increasing from 87% to 96%, effectively detecting anomalies such as incorrect mask usage, sneezing and spitting. However, potential limitations include reliance on predefined anomaly classes, which reduce the adaptability to novel or unseen anomalous behaviors.

Lee et al. (2024) introduced the Group-Based Lightweight Human Behavior Recognition Framework (GLBRF), which applies two-dimensional convolutional neural networks (2D CNNs) and location-based grouping to efficiently classify human interactions in video surveillance. The model achieved a 98% Accuracy with grouping, significantly outperforming the 68% Accuracy without it, demonstrating its effectiveness in behavior recognition while maintaining a low computational burden. However, its reliance on a relatively small dataset limit generalizability to more complex and diverse real-world scenarios.

Mishra & Jabin (2024) developed a VAD algorithm using deep autoencoders to extract spatiotemporal features and identify anomalies with a regularity score-based thresholding mechanism. The model achieved AUC scores of 86.4% on UCSD Pedestrian1 and 88.9% on Avenue datasets, showing competitive performance. However, its reliance on spatiotemporal features posed challenges in handling complex anomaly scenarios.

Guo et al. (2024) proposed a Two-Stream Spatial-Temporal Auto-Encoder network with adversarial training for VAD, focusing on robust feature extraction. The model comprises a spatial auto-encoder stream for extracting appearance features from detected objects and a temporal auto-encoder stream for encoding motion patterns using stacked optical flow maps. To improve the AD, an adversarial training branch is introduced, leveraging a pseudo-abnormal dataset to simulate abnormal events and increase reconstruction errors for

anomalies. Evaluated on benchmark datasets, the method achieved competitive performance, with an AUC of 86.7 and an EER of 19.5 on the Avenue dataset. However, the reliance on adversarial training increases computational demands, potentially limiting its scalability for real-time applications.

2.5 CLASSIFICATION TECHNIQUES

Fedorov et al. (2019) proposed a traffic flow estimation system using the Faster R-CNN two-stage detector combined with the Simple Online and Real-time Tracker and a region-based heuristic algorithm to classify vehicle movement direction. The model achieved a mean absolute percentage error of less than 10% in counting and classifying vehicles during weekday rush hours, enhanced through focal loss, adaptive feature pooling, an additional mask branch and anchor optimization. However, variations in environmental conditions, including lighting changes and occlusions, can impact system performance, potentially limiting its adaptability to diverse urban settings.

Alshammari & Rawat (2019) proposed a machine learning-based multi-camera surveillance system to enhance anomaly detection by reducing blind spots and improving coverage. The system extracts feature through target detection and tracking across video streams for threat identification. While effective for urban surveillance, computational challenges in synchronizing multi-camera data limit scalability in large-scale applications.

Fanta et al. (2020) introduced an optimized Gated Recurrent Unit (GRU) architecture, called Single-Tunnelled GRU, for anomaly detection in long-term sequences. The model eliminates the reset gate in standard GRUs to emphasize past content and replaces hyperbolic tangent activation with sigmoid activation for improved performance in deeper networks. Tested on the CUHK Avenue and UCSD datasets, the optimized GRU outperformed standard GRU and LSTM networks in detection accuracy and computational efficiency. However, while effective for structured surveillance environments with long-term dependencies, the model struggled with real-time dynamic anomaly scenarios, highlighting the need for further optimization in highly variable settings.

Ma (2021) proposed a partially supervised deep learning framework for anomaly detection, focusing on spatiotemporal feature extraction using a Variational Autoencoder

(VAE). The model constrains the hidden layer representation of normal samples to a Gaussian distribution, allowing anomalies to be identified based on their deviation from this distribution. Evaluated on the UCSD and Avenue datasets, the method achieved frame-level AUCs of 92.3% and 82.1%, respectively, with an impressive processing speed of 571 frames per second, demonstrating both effectiveness and efficiency. However, the high computational demands of the framework pose challenges for real-time applications, highlighting the need for more resource-efficient solutions.

Vu et al. (2022) suggested a robust anomaly detection framework that utilizes adversarial training and multilevel representations of intensity and motion data to improve accuracy in noisy environments. The method combines denoising Autoencoders for feature extraction, conditional GANs for level-wise representation generation and a consolidation process to detect anomalies across multiple representation levels. Tested on UCSD Peds1, UCSD Peds2 and Avenue datasets, the model showed significant improvements in pixel-level Equal Error Rate (EER), with gains of 11.35%, 12.32% and 4.31%, respectively. While effective, the approach incurs high computational costs, limiting its applicability in real-time settings and highlighting the need for more efficient solutions.

Sabeena et al. (2022) proposed a Fuzzy Convolution Bi-Directional Long Short-Term Memory classifier for Parkinson's disease detection, leveraging an Optimization-Based Ensemble Feature Selection approach that integrates Fitness-Memory-Based Optimization Algorithm, Levy Flight Cuckoo Search Algorithm and Adaptive Firefly Algorithm to refine feature subsets. The model achieved 98.77% Accuracy, surpassing SVM at 88.13%, CNN at 94.18% and Fully Connected Long Short-Term Memory-Convolutional Neural Network at 95.16%, using the University of California-Irvine (UCI) Parkinson's dataset, evaluated via Leave-One-Person-Out Cross Validation. Nevertheless, the model's reliance on complex ensemble feature selection and deep learning techniques increases computational demands, delaying real-time clinical deployment.

Ramadan et al. (2022) introduced a deep learning model for human baggage classification using the pre-trained DenseNet-161 architecture and a "fit-one-cycle policy" to optimize training efficiency. The model achieved 96%–98.75% Accuracy for binary classification and 96.67%–98.33% Accuracy for multi-class classification, outperforming existing methods while ensuring faster processing. However, its reliance on re-annotated

datasets and specific human stance attributes limits its adaptability to diverse real-world surveillance conditions with varying angles and occlusions.

Munteanu et al. (2022) proposed an AI-based system using deep learning for human activity recognition in video to assist Alzheimer's patients by monitoring their eating and hydration habits, providing reminders and allowing remote caregiver supervision. The model achieved 96% Accuracy in image classification, 74% Accuracy in object detection and 78% Accuracy in activity recognition, with training completed within 48 hours and a fast response time of two seconds on a portable development board. However, its performance can be constrained by standard computational hardware, potential inaccuracies in activity recognition and challenges in real-time adaptability across different environments.

Khan, S. et al. (2022) developed a CNN-based rolling prediction algorithm for traffic accident detection using video surveillance systems, achieving an Accuracy of 82%. The method leverages CNNs to extract features from video frames and detect anomalies, specifically traffic accidents, with a domain-specific Vehicle Accident Images dataset constructed for training. While effective for structured traffic systems, the algorithm's applicability is limited to traffic-related scenarios, reducing its versatility for broader surveillance tasks. This study highlights the potential of CNNs for domain-specific anomaly detection while emphasizing the need to adapt such approaches to diverse and complex environments.

Khan, A. et al. (2022) proposed a transfer learning approach using the pre-trained InceptionResNetV2 model to classify Smoking and Not Smoking images in a newly curated dataset, achieving 96.87% Accuracy, 97.32% Precision and 96.46% Recall, surpassing other CNN-based methods. While the model performs well across diverse environments, its reliance on static images impact real-time detection under challenging conditions like poor lighting or occlusions. Furthermore, dependence on a single pre-trained model and the lack of video-based analysis limit optimization and contextual understanding.

Ul Amin et al. (2022) deployed the Efficient Deep Learning Model for Anomaly Detection (EADN), a hybrid CNN-LSTM framework that captures spatiotemporal patterns for anomaly detection. The model segments video input using a shot boundary detection algorithm, with a time-distributed 2D CNN extracting features and LSTM cells learning

temporal dependencies. Achieving high AUC scores on the UCSD Peds1 dataset of 93.0, UCSD Peds2 of 97.0, CUHK Avenue of 97.0 and UCF-Crime of 98.0, EADN outperformed state-of-the-art methods. However, despite its effectiveness across these datasets, the model faces challenges in real-time applications due to its high computational demands.

Kumar & Patel (2024) proposed a real-time framework for detecting abnormal human activity in video surveillance using a deep learning architecture combining CNN, RNN and a temporal attention mechanism. This methodology effectively extracted spatiotemporal features from unprocessed video streams, enabling accurate classification of aberrant activities. The model achieved impressive accuracies of 96.94%, 98.95% and 62.04% on the UCF50, UCF110 and UCF-Crime datasets, respectively, outperforming state-of-the-art algorithms. However, the framework's performance on more complex datasets like UCF-Crime highlighted limitations in handling diverse and subtle anomalies.

2.6 SUMMARY

This chapter provides a detailed study of previous research on VAD, focusing on TL, preprocessing, feature extraction and classification techniques. This survey highlights the progress made and the challenges that shape future research in developing adaptable VAD systems.