Covariance-Aware Multi-Backbone Feature Fusion with Bidirectional Temporal Modelling for Frame-Level Video Anomaly Detection

Umar Shafiu Haruna[1], Abdulrauf Garba Sharifai[2], Muktar Danlami[3], Aminu Abdulkadir Mahmoud[4], Lawan Garba[5]
Department of Cybersecurity, Northwest University, Kano State, Nigeria[1, 3, 4]
Department of Computer Science, Northwest University, Kano State, Nigeria[2]
Department of Software Engineering, Northwest University, Kano State, Nigeria[5]

***Abstract*—Video anomaly detection remains challenging because abnormal events are rare, diverse, temporally dependent, and strongly influenced by scene context. This study evaluates a modular covariance-aware framework combining five frozen ImageNet-pretrained CNN backbones (VGG19, ResNet50, EfficientNetB0, DenseNet121, and InceptionV3), feature standardisation, principal component analysis (PCA), Mahalanobis normality scoring, early and late fusion, and bidirectional temporal convolutional modelling. Evaluation is conducted under documented dataset-specific protocols on CUHK Avenue, UCSD Ped2, and ShanghaiTech Campus. BiTCN late fusion attains AUC-ROC/AUC-PR values of 0.9984/0.9981 on Avenue and 0.8534/0.8481 on ShanghaiTech, while EfficientNetB0 with PCA-Mahalanobis scoring gives the strongest held-out Ped2 result of 0.9730/0.9938. On Avenue, the PCA-Mahalanobis ResNet50 configuration increases AUC-ROC from the 0.8906 centroid/Euclidean reference to 0.9980, an absolute change of 0.1074; because PCA and score geometry change together, this difference is attributed to the combined PCA-Mahalanobis configuration rather than Mahalanobis distance alone. The results further show that fusion and temporal refinement are dataset-dependent: the best single backbone outperforms all evaluated fusion rules on Ped2, and the bidirectional temporal model is interpreted as an offline refinement mechanism. Cross-paper values are used for contextual positioning when evaluation protocols differ.**

***Keywords*—Video anomaly detection; Mahalanobis distance; principal component analysis; CNN feature fusion; BiTCN; temporal modelling; UCSD Ped2; avenue; ShanghaiTech**

I. Introduction

Video Anomaly Detection (VAD) concerns the identification of events or behaviours that deviate from the patterns regarded as normal within a video scene. It has become an important component of intelligent surveillance because events of interest are usually infrequent, difficult to anticipate, and often costly to annotate exhaustively. More importantly, abnormality is inherently contextual: an action that is routine in one environment may be anomalous in another, and visually similar activities may carry different meanings depending on where and when they occur. Y. Liu et al. [1] describe this contextual dependence as one of the central difficulties in generalised VAD, while Abdalla et al. [2] identify the rarity, diversity, and scene dependence of anomalous events as persistent challenges across the field. Consequently, many VAD systems are required to learn predominantly from normal observations and identify deviations that may not have been explicitly represented during training.

Recent research has approached this problem through increasingly expressive models of normality and anomaly. Memory-based methods retain representative normal patterns, masked autoencoders learn regular visual structure through reconstruction, and transformer-based models capture broader spatial and temporal relationships. Ahn et al. [3], for example, use memory components to preserve different forms of normality, whereas Ristea et al. [4] employ self-distilled masked autoencoding for efficient anomaly detection. Other studies have moved toward semantic and open-world reasoning through large pretrained vision-language models [5], [6]. These developments have considerably expanded the range of information available to VAD systems. Nevertheless, recent reviews continue to identify representation quality, temporal modelling, generalisation across scenes, computational requirements, and reliable anomaly scoring as important research concerns [1], [2], [7]. There is therefore still value in examining how much can be achieved through a comparatively transparent pipeline in which the contribution of each modelling stage can be observed separately.

An important part of this problem is the relationship between visual representation and anomaly scoring. Deep pretrained CNNs provide strong frame-level representations, but their architectural biases differ [8]–[12]. Heterogeneous backbones can therefore provide alternative descriptions of the same frame, while the usefulness of combining them depends on the scene and scoring stage. The framework explicitly evaluates both individual and fused representations so that complementarity is established empirically rather than assumed.

The scoring stage is equally important. Euclidean distance measures geometric deviation from a reference point but does not account for covariance among feature dimensions, whereas Mahalanobis distance measures deviation relative to both the mean and covariance structure [13]. In the present framework, PCA provides dimensionality control before covariance-aware scoring, and the resulting PCA-Mahalanobis path is evaluated as a combined statistical normality model. This framing permits the observed gains to be interpreted directly in terms of the implemented pipeline without over-assigning them to one component in isolation.

A further challenge is temporal behaviour. Frame-level scores can fluctuate because individual observations may be blurred, occluded, or visually ambiguous. The BiTCN stage therefore examines score patterns over 16-frame windows using bidirectional dilated convolutions. Because it uses context from both temporal directions, the implemented BiTCN is positioned as an offline or bounded-delay refinement model, where neighbouring observations are available for analysis.

These considerations motivate the framework proposed in this study. Rather than relying on a single visual encoder or treating feature extraction, statistical scoring, fusion, and temporal modelling as isolated alternatives, the framework evaluates them within a common modular pipeline. Five frozen ImageNet-pretrained CNN backbones, namely VGG19, ResNet50, EfficientNetB0, DenseNet121, and InceptionV3, provide heterogeneous frame-level representations. PCA is used for dimensionality control where required, while Mahalanobis distance provides covariance-aware modelling of normal behaviour. The framework subsequently examines early feature fusion, conventional late score fusion, and bidirectional temporal modelling using a BiTCN. This arrangement allows the contribution of the representation, scoring geometry, fusion strategy, and temporal component to be examined separately instead of embedding them in a single end-to-end model.

The contribution of this study is therefore the controlled integration and component-wise examination of established mechanisms rather than novelty claims for CNN feature extraction, PCA, Mahalanobis distance, fusion, or temporal convolution individually. Particular attention is given to dataset-specific evaluation protocols, component-level trade-offs, and the limits these place on cross-paper numerical comparison.

The main contributions of this study are as follows:

1) Modular multi-backbone normality modelling: Five heterogeneous frozen CNN backbones are evaluated under common feature-extraction and scoring components, allowing representation diversity to be tested without assuming that fusion is universally beneficial.
2) PCA-stabilised covariance-aware anomaly scoring: PCA is used to reduce redundancy and stabilise covariance estimation before Mahalanobis scoring. The contribution is evaluated as an integrated PCA-Mahalanobis normality model, allowing dimensionality control and covariance-aware deviation scoring to operate within a single statistically coherent path.
3) Controlled comparison of fusion and temporal strategies: Individual scoring, early fusion, conventional late fusion, Early-Fusion BiTCN, and BiTCN late fusion are evaluated within the same modular framework. The analysis identifies dataset conditions under which individual representations, fusion, or temporal refinement provide the strongest anomaly separation, with bidirectional modelling positioned for offline or bounded-delay analysis.
4) Cross-dataset empirical analysis with protocol disclosure: CUHK Avenue, UCSD Ped2, and ShanghaiTech are analysed using the dataset-specific procedures documented in Section IV. Published results are used for contextual positioning rather than direct ranking when evaluation protocols differ.

The remainder of this study is organised as follows. Section II critically reviews recent work on VAD representation learning, fusion, covariance-aware scoring, and temporal modelling. Section III presents the proposed framework. Section IV documents the datasets, preprocessing, evaluation protocols, training procedure, metrics, and computational complexity. Section V reports the experimental results. Section VI provides contextual comparison with recent VAD studies while accounting for protocol differences. Section VII discusses dataset-specific behaviour, scoring, fusion, temporal modelling, and practical implications. Section VIII concludes the study and outlines directions for further validation and deployment-oriented extension.

II. Related Work

A. Recent Representation Learning and Semantic VAD

Recent VAD methods increasingly exploit strong pretrained or self-supervised representations. VideoPatchCore [3] demonstrates that pretrained features can be effective when coupled with a memory of normality, reducing the need for task-specific retraining, although memory construction and retrieval remain additional design choices. Ristea et al. [4] use self-distilled masked autoencoding to obtain efficient anomaly representations, but reconstruction-oriented objectives may still favour regular appearance over explicit modelling of feature-space covariance. State-space approaches such as VADMamba [14] improve temporal efficiency and long-range modelling, while introducing a different sequence-modelling assumption from the statistical score refinement studied here.

Semantic and open-world approaches broaden VAD beyond purely visual normality models. Wu et al. [5] support open-vocabulary anomaly recognition, while Zanella et al. [6] and Yang et al. [15] exploit large language or vision-language models for training-free or rule-based reasoning. Their strength is semantic flexibility, particularly for previously unseen anomaly categories; their limitation is dependence on large pretrained models, prompts, or external semantic knowledge. These methods therefore address a different deployment trade-off from compact statistical normality modelling over frozen visual features. Recent weakly supervised methods also incorporate prompt-enhanced context and textual clues to improve anomaly discrimination [16], [17].

B. Multi-Stream, Multi-Backbone, and Fusion Approaches

Multi-stream systems show that complementary evidence can improve anomaly detection when the streams encode genuinely different information. Pang et al. [18] combine appearance and skeleton cues, TDS-Net [19] integrates spatial and motion information, and M2VAD [20] exploits multiview and multimodal signals. These approaches can capture anomaly cues that a single stream may miss, but they require additional modalities or derived representations and are commonly evaluated under architecture-specific supervision and preprocessing protocols. Rahimpour et al. [21] likewise combine transformer representations with an ensemble of convolutional

auto-encoders, illustrating another route for aggregating complementary deep evidence.

Multi-backbone fusion is related but distinct: several encoders process the same visual input, potentially diversifying texture, residual, multiscale, and efficiency-oriented inductive biases. Early fusion can preserve cross-backbone interactions but enlarges the feature space and can increase redundancy and covariance-estimation cost. Late fusion preserves detector-specific score behaviour and is computationally compact after feature extraction, but it cannot recover interactions discarded by the independent scoring stages. Neither strategy is inherently superior; this motivates evaluating single-backbone, early-fusion, and late-fusion configurations rather than assuming complementarity.

C. Temporal and Bidirectional Modelling

Temporal context can distinguish persistent abnormal behaviour from isolated score fluctuations. Bai et al. [22] show that dilated temporal convolutions capture dependencies across multiple temporal ranges; Shen et al. [23] demonstrate the value of bidirectional context in VAD; Ismail and Aboshady [24] combine recurrent and transformer mechanisms; and VADMamba [14] uses state-space modelling for efficient temporal processing. These approaches strengthen sequence reasoning, but they differ in causality, receptive field, supervision, and computational cost.

Bidirectional processing defines a specific deployment scope: a decision at time t can use both preceding and succeeding observations. This is well suited to offline analysis and bounded-delay surveillance, where local future context is available. Strict online deployment can use the same score-modelling concept with a causal temporal variant restricted to current and preceding observations.

D. Covariance-Aware Scoring and Dimensionality Reduction

A Euclidean anomaly score measures distance from a reference point under an effectively spherical geometry. Mahalanobis distance [13] accounts for covariance and can weight deviations according to the normal variation of each direction. PaDiM [25], Sharif et al. [26], and Behrendt et al. [27] illustrate the value of covariance-aware scoring in deep anomaly-detection settings. The principal computational challenge is that covariance estimation becomes increasingly demanding as feature dimensionality approaches or exceeds the number of normal observations, which motivates dimensionality control before covariance modelling.

PCA provides a principled means of reducing this burden while retaining dominant variance [28], [29]. It can improve conditioning and reduce covariance-storage and inversion cost, but dimensionality reduction can also alter anomaly separability by itself. For that reason, an experiment that changes from raw Euclidean scoring to PCA-Mahalanobis scoring changes two factors simultaneously. The present study therefore interprets PCA and Mahalanobis as a combined configuration unless an identical PCA representation is used for both distance measures.

E. Critical Synthesis and Research Gap

The recent literature reveals a consistent set of trade-offs. Memory and semantic methods offer flexible representations but introduce memory or large-model dependencies; multi-stream systems exploit richer evidence but require additional modalities; temporal models capture sequence structure but may sacrifice causality; and covariance-aware scoring is transparent but sensitive to dimensionality and covariance estimation. Direct numerical comparison across these families is further complicated by differences in supervision, preprocessing, localisation, and AUC aggregation [1], [2], [7].

A specific gap therefore remains in understanding how heterogeneous frozen CNN representations, dimensionality control, covariance-aware scoring, feature- versus score-level fusion, and temporal refinement behave when examined progressively within one modular pipeline. The present framework addresses that intersection and emphasises component-wise interpretation rather than assuming that additional backbones, fusion, or temporal modelling must always improve performance.

III. Proposed Method

This section presents the proposed covariance-aware multi-backbone feature fusion framework for frame-level video anomaly detection. The framework is designed around the principle that video anomaly detection is primarily a normality-modelling problem. Instead of requiring large numbers of labelled abnormal examples, the model learns the statistical structure of normal video frames and assigns higher anomaly scores to frames whose deep representations deviate from that learned normal distribution. The architecture combines five frozen CNN backbones, PCA-based dimensionality reduction, Mahalanobis normality scoring, early and late fusion strategies, and BiTCN-based temporal score modelling.

A. Proposed Framework Architecture

Fig. 1 shows the complete processing flow of the proposed framework. The input video is first converted into a sequence of sampled frames. Each frame is resized to 224 × 224 and represented as a three-channel image compatible with ImageNet-pretrained CNN backbones. The model then extracts backbone-specific deep features using VGG19, ResNet50, EfficientNetB0, DenseNet121, and InceptionV3. These backbones are used as frozen feature extractors so that the study focuses on representation diversity, statistical scoring, fusion behaviour, and temporal score refinement rather than on backbone fine-tuning.

The theoretical motivation for using multiple frozen backbones is that each architecture encodes a different visual bias. VGG-style networks emphasise texture-rich appearance, residual networks support deeper semantic abstraction, EfficientNet provides a compact scaling strategy, DenseNet encourages feature reuse, and Inception captures multiscale responses. These differences justify testing whether their representations provide useful diversity, but the framework does not assume that combining them will necessarily outperform the best single backbone.

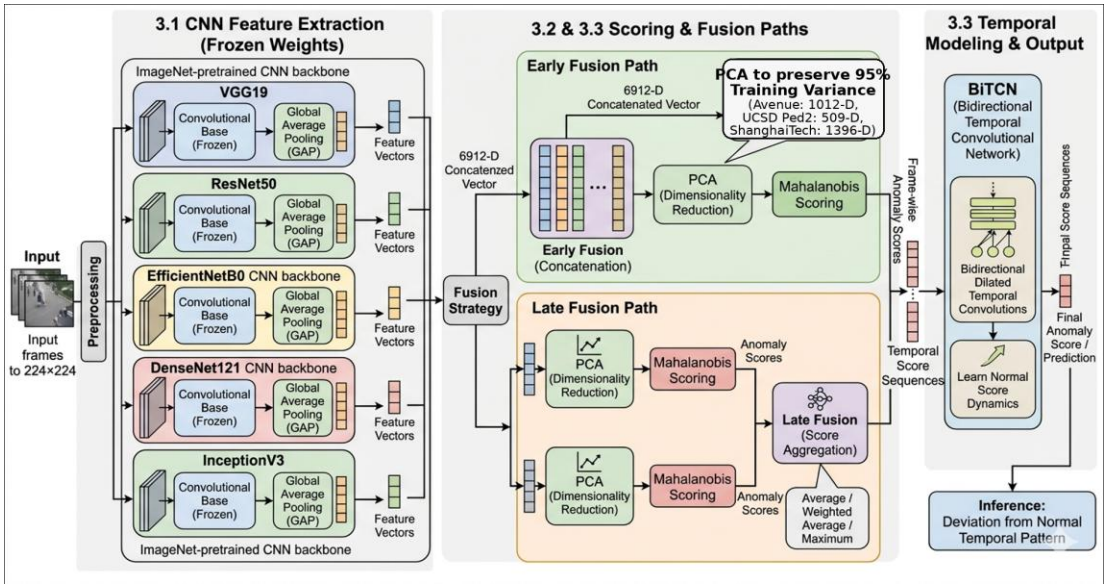

Fig. 1. Proposed covariance-aware multi-backbone feature fusion and BiTCN temporal modelling framework.

The architecture separates the pipeline into two main paths. In early fusion, the five backbone feature vectors are concatenated before PCA reduction and Mahalanobis scoring. In late fusion, each backbone is standardised, reduced, and scored independently before the resulting score streams are combined. The BiTCN component is placed after spatial scoring so that it models temporal score dynamics rather than raw pixels. Table I summarises the role of each processing stage and makes clear which components are statistical, fusion-based, or learnable.

TABLE I. MAIN COMPONENTS OF THE PROPOSED VAD PIPELINE

Step	Component	Role
1	Input preprocessing	Frames are sampled with stride 2, resized to 224 × 224, and converted to a three-channel representation suitable for ImageNet-pretrained CNNs.
2	Frozen CNN feature extraction	VGG19, ResNet50, EfficientNetB0, DenseNet121, and InceptionV3 provide alternative pooled descriptors; complementarity is evaluated rather than assumed.
3	Standardisation	Scaling parameters are learned from normal training features and then applied to validation/test features to avoid leakage.
4	PCA reduction	PCA retains 95% of training variance and stabilises covariance estimation in high-dimensional deep-feature space.
5	Mahalanobis scoring	Normal training features define a covariance-aware distribution; test frames are scored by deviation from this normal model.
6	Fusion	Early fusion combines features before scoring; late fusion combines score streams after independent backbone scoring.
7	BiTCN temporal modelling	Bidirectional dilated convolutions process 16-frame score sequences for offline temporal refinement.
8	Inference	Large deviations from normal spatial or temporal score patterns are treated as anomalous frames.

B. Input Preprocessing and Multi-Backbone Feature Extraction

Let a surveillance video be represented as an ordered sequence of frames. This notation makes it possible to define every later operation at frame level, which is consistent with the frame-level AUC-ROC and AUC-PR evaluation used in the experiments.

$$V = \{I_t\}_{t=1}^{T} \tag{1}$$

where, *it* denotes the frame at time t and T is the total number of frames in the video. Each frame is preprocessed before feature extraction. The preprocessing operation standardises the frame size and channel format so that all CNN backbones receive inputs with the same spatial and colour structure.

$$x_t = \mathcal{P}(I_t), \qquad x_t \in \mathbb{R}^{224\times224\times3} \tag{2}$$

where, $\mathcal{P}(.)$ denotes the preprocessing function. In the implemented pipeline, this includes frame sampling, resizing to 224 × 224, and conversion to a three-channel representation. The proposed model uses five frozen CNN backbones. The backbone set is denoted as:

$$\mathcal{B} = \{B_1, B_2, B_3, B_4, B_5\} \tag{3}$$

The individual members of the backbone set are defined as:

$$\begin{aligned} B_1 &= \text{VGG19}, \quad B_2 = \text{ResNet50} \\ B_3 &= \text{EfficientNetB0} \\ B_4 &= \text{DenseNet121}, \quad B_5 = \text{InceptionV3} \end{aligned} \tag{4}$$

For a preprocessed frame, each backbone extracts a feature map, which is converted into a compact descriptor using global

average pooling. Global average pooling reduces each feature map to a fixed-length descriptor, allowing all subsequent statistical scoring operations to be performed at frame level.

$$f_t^{(k)} = \mathrm{GAP}\big(B_k(x_t)\big), \qquad k = 1,2,3,4,5 \tag{5}$$

where, $f_t^{(k)}$ is the feature vector produced by the k-th backbone for frame t. Since the backbones are not fine-tuned, their parameters remain unchanged during training:

$$\Delta\theta_{B_k} = 0, \qquad k = 1,2,3,4,5 \tag{6}$$

Freezing the backbone parameters reduces training cost, limits overfitting on normal-only training sets, and keeps the experimental focus on the contribution of feature diversity, PCA reduction, covariance-aware scoring, fusion strategy, and temporal score modelling.

C. Feature Standardisation and PCA Reduction

The feature vectors produced by different CNN backbones may have different numerical scales. If these raw features are used directly in a distance-based anomaly detector, dimensions with larger values may dominate the score even when they are not more informative. Therefore, standardisation is applied using statistics estimated only from the normal training features.

$$z_t^{(k)} = \frac{f_t^{(k)} - \mu_f^{(k)}}{\sigma_f^{(k)} + \epsilon} \tag{7}$$

where, $\mu_f^{(k)}$ and $\sigma_f^{(k)}$ are the mean and standard deviation of normal training features from the k-th backbone, and ε is a small constant used for numerical stability. Fitting the scaler only on normal training data prevents abnormal validation or test samples from influencing the normality model.

After standardisation, PCA is applied to reduce feature dimensionality while retaining the dominant variance structure of normal behaviour. PCA also makes covariance estimation more stable by projecting high-dimensional features into a lower-dimensional orthogonal space.

$$y_t^{(k)} = \mathrm{PCA}_k\big(z_t^{(k)}\big) \tag{8}$$

where, $y_t^{(k)}$ is the PCA-reduced representation obtained from the k-th backbone. The retained PCA dimension is selected to preserve at least 95% of the variance observed in the normal training features:

$$\frac{\sum_{i=1}^{r_k} \lambda_i^{(k)}}{\sum_{i=1}^{d_k} \lambda_i^{(k)}} \geq 0.95 \tag{9}$$

where, $\lambda_i^{(k)}$ is the i-th PCA eigenvalue for the k-th backbone, d_k is the original feature dimension, and r_k is the number of retained principal components. The value of r_k is selected such that the retained components preserve at least 95% of the total variance in the normal training features. This allows the model to reduce dimensionality while retaining the dominant structure of normal behaviour, thereby improving computational efficiency and stabilising covariance estimation before Mahalanobis scoring.

This step is important because direct covariance estimation in the original high-dimensional feature space can be numerically unstable, particularly when the number of normal training samples is limited. PCA therefore supports both computational efficiency and reliable Mahalanobis scoring.

D. Covariance-Aware Mahalanobis Scoring

The anomaly-scoring stage models the distribution of normal PCA-reduced features for each backbone. Instead of measuring simple distance from a centroid, Mahalanobis distance measures deviation from the normal distribution while considering the covariance among feature dimensions. This is suitable for deep CNN features because their dimensions are often correlated rather than independent.

$$\mu_k = \mathrm{mean}\big(y^{(k)}\big), \qquad \Sigma_k = \mathrm{cov}\big(y^{(k)}\big) \tag{10}$$

where, μk is the mean vector, and Σk is the covariance matrix of the PCA-reduced normal training features for backbone k. To make covariance inversion stable, a small diagonal regularisation term is added:

$$\tilde{\Sigma}_k = \Sigma_k + \lambda I \tag{11}$$

where, λ is a small positive regularisation coefficient, and I is the identity matrix. The regularised covariance matrix prevents unstable inversion when feature dimensions are highly correlated or when some components have very small variance.

The Mahalanobis anomaly score for frame *t* under the *k-th* backbone is then computed as:

$$s_t^{(k)} = \sqrt{\big(y_t^{(k)} - \mu_k\big)^T \tilde{\Sigma}_k^{-1} \big(y_t^{(k)} - \mu_k\big)} \tag{12}$$

A larger value of $s_t^{(k)}$ indicates that the frame representation lies farther from the learned normal distribution. The score is covariance-aware because deviations along low-variance normal directions are penalized more strongly than deviations along high-variance directions. This makes the scoring stage more informative than a simple Euclidean-style distance when the normal feature cloud is elongated or anisotropic.

E. Early and Late Fusion Strategies

The framework evaluates both early and late fusion because the point at which information is combined can affect anomaly separation. Early fusion combines representations before scoring, allowing the model to form one joint feature vector across all five backbones. This can capture cross-backbone feature relationships, but it also increases dimensionality and therefore requires PCA before covariance estimation.

$$f_t^E = \big[f_t^{(1)}; f_t^{(2)}; f_t^{(3)}; f_t^{(4)}; f_t^{(5)}\big] \tag{13}$$

The early-fused feature vector is then reduced using PCA:

$$y_t^E = \mathrm{PCA}_E(f_t^E) \tag{14}$$

The early-fusion Mahalanobis score is computed from the PCA-reduced fused feature distribution:

$$s_t^E = \sqrt{(y_t^E - \mu_E)^T \tilde{\Sigma}_E^{-1} (y_t^E - \mu_E)} \tag{15}$$

where, μ_E and $\tilde{\Sigma}_E$ are the mean vector and regularised covariance matrix of the early-fused normal training features after PCA reduction.

In late fusion, each backbone first produces an independent Mahalanobis score. These scores are then combined after scoring rather than before scoring. This preserves the individual behaviour of each backbone and allows the fusion stage to aggregate complementary anomaly evidence.

$$s_t = \left[s_t^{(1)}, s_t^{(2)}, s_t^{(3)}, s_t^{(4)}, s_t^{(5)}\right]^T \quad (16)$$

The average late-fusion score treats all five backbones equally:

$$s_t^{avg} = \frac{1}{5}\sum_{k=1}^{5} s_t^{(k)} \quad (17)$$

The weighted late-fusion score assigns different levels of importance to the five backbone scores:

$$s_t^w = \sum_{k=1}^{5} w_k\, s_t^{(k)}, \quad \sum_{k=1}^{5} w_k = 1, \quad w_k \geq 0 \quad (18)$$

where, w_k is the fixed coefficient assigned to the k-th backbone. In the validation-aware weighted-fusion experiment, the coefficients are derived from the individual backbone validation AUC-ROC values and normalised to sum to one. The resulting coefficients are fixed before final test evaluation; test labels and test-set AUC values are not used for weight selection. Weighted fusion is therefore reported only for the validation-aware Ped2 configuration.

The maximum late-fusion score selects the strongest anomaly response among the five backbones:

$$s_t^{max} = \max\left(s_t^{(1)}, s_t^{(2)}, s_t^{(3)}, s_t^{(4)}, s_t^{(5)}\right) \quad (19)$$

Average fusion is useful for smoothing detector behaviour, weighted fusion allows dataset-specific emphasis, and maximum fusion is sensitive to cases where only one backbone strongly detects an abnormal pattern. These alternatives provide an ablation-friendly way of studying how representation diversity contributes to video anomaly detection.

F. BiTCN Temporal Score Modelling

Frame-level anomaly scores may fluctuate because individual frames can be blurred, occluded, visually ambiguous, or similar to normal behaviour. Temporal modelling is therefore introduced to examine whether local score sequences follow normal temporal dynamics or form temporally coherent abnormal patterns.

The BiTCN component is placed after Mahalanobis scoring. This design choice is theoretically important because the temporal model does not operate as a black-box pixel-level video classifier. Instead, it receives score sequences that already express the degree of deviation from normal feature distributions. The temporal component therefore refines interpretable anomaly evidence rather than replacing the covariance-aware scoring stage.

For late-fusion BiTCN, the temporal input is formed by collecting score vectors over a fixed-length window ending at frame t:

$$Q_t = \{s_{t-L+1}, s_{t-L+2}, \ldots, s_t\}, \qquad L = 16 \quad (20)$$

where, L is the temporal window length. In this study, L = 16 frames. The BiTCN uses dilated temporal convolutions to capture short- and medium-range score dependencies efficiently. The dilation rates are:

$$\delta \in \{1,2,4,8\} \quad (21)$$

The BiTCN maps the score window into a temporal representation and then converts that representation into a final temporal anomaly score:

$$h_t = \mathrm{BiTCN}(Q_t), \qquad s_t^T = g(h_t) \quad (22)$$

where, h_t is the learned temporal representation and $g(.)$ is the output mapping that converts the BiTCN representation into a temporal anomaly score. The bidirectional structure allows the model to use both forward and backward score context during offline benchmark evaluation, while the use of score sequences keeps the temporal model compact and interpretable.

G. Final Anomaly Score and Inference

The final anomaly score depends on the selected configuration. The framework is modular, so the score may come from early-fusion Mahalanobis scoring, standard late-fusion scoring, maximum fusion, or BiTCN temporal score modelling.

$$A_t = \begin{cases} s_t^E, & \text{early fusion} \\ s_t^{avg}, & \text{average late fusion} \\ s_t^w, & \text{weighted late fusion} \\ s_t^{max}, & \text{maximum late fusion} \\ s_t^T, & \text{BiTCN temporal modelling} \end{cases} \quad (23)$$

The final score A_t is a continuous frame-level anomaly score. For binary inference, a frame is classified as anomalous when its score exceeds a selected threshold τ:

$$\hat{y}_t = \begin{cases} 1, & A_t > \tau \\ 0, & A_t \leq \tau \end{cases} \quad (24)$$

where, $\hat{y}_t = 1$ denotes an anomalous frame and $\hat{y}_t = 0$ denotes a normal frame. In the experimental evaluation, continuous anomaly scores are used to compute AUC-ROC and AUC-PR rather than relying only on a single fixed threshold.

The complete proposed framework can therefore be summarised as the following frame-level anomaly detection pipeline:

$$\begin{aligned} &I_t \rightarrow \mathcal{P}(I_t) \rightarrow \{B_k\}_{k=1}^{5} \rightarrow \mathrm{PCA} \\ &\rightarrow \text{Mahalanobis Score} \rightarrow \text{Fusion} \\ &\rightarrow \mathrm{BiTCN} \rightarrow A_t \end{aligned} \quad (25)$$

This formulation shows that the framework produces anomaly evidence through a staged and interpretable process. First, backbone-specific CNN features are extracted from each frame. Second, the features are standardised and compressed using PCA. Third, Mahalanobis distance measures covariance-aware deviation from normal behaviour. Fourth, fusion combines evidence across the five backbone streams. Finally, BiTCN temporal modelling transforms the anomaly-score sequence to capture temporal structure.

IV. Experimental Setup

The experimental study was structured as a modular evaluation rather than as a single end-to-end configuration. Four groups were considered: 1) frozen multi-backbone features with centroid/Euclidean scoring as a reference baseline; 2) PCA-stabilised Mahalanobis normality modelling; 3) early and late fusion; and 4) bidirectional temporal refinement using BiTCN. The centroid/Euclidean and PCA-Mahalanobis paths therefore represent two complete scoring configurations, and their difference is interpreted as the effect of the combined dimensionality-control and covariance-aware normality-modelling path.

Five ImageNet-pretrained CNNs—VGG19, ResNet50, EfficientNetB0, DenseNet121, and InceptionV3—were used throughout the study. Their parameters were frozen so that the experiments focused on the behaviour of the extracted representations and the subsequent normality-modelling stages rather than on backbone fine-tuning. The evaluation was performed on CUHK Avenue, UCSD Ped2, and ShanghaiTech Campus using frame-level anomaly scores.

A. Datasets

Three established VAD datasets were used to expose the framework to different surveillance conditions. The datasets themselves are standard benchmarks, but the evaluation protocol is not identical across the three experiments. The dataset-specific construction is therefore documented explicitly so that each result is interpreted under the population on which it was computed.

CUHK Avenue contains 16 training videos and 21 testing videos recorded in a relatively constrained surveillance scene, with anomalies including running, throwing objects, and other scene-inconsistent actions [30]. In this study, stride-2 sampled training-video frames form the normal reference population and sampled testing-video frames form the comparison population used for the Avenue component analysis. The resulting metric is therefore interpreted under the documented study-specific construction rather than assumed to be identical to the canonical per-frame benchmark protocol.

UCSD Ped2 represents a controlled pedestrian walkway in which normal activity is dominated by pedestrians and anomalies include bicycles, vehicles, and other non-pedestrian motion [31]. For the validation-aware experiment, the processed optical-flow data were partitioned at sequence level into 11 normal training entries, 2 normal validation entries, 3 normal test entries, 9 abnormal validation entries, and 15 abnormal test entries. After stride-2 sampling, this yielded 870 normal training frames, 150 normal validation frames, 255 normal test frames, 780 abnormal validation frames, and 1,230 abnormal test frames. This is a study-specific held-out protocol and is not presented as the canonical Ped2 benchmark split.

ShanghaiTech Campus provides a substantially more heterogeneous setting with multiple camera views, scene layouts, illumination conditions, crowd densities, and anomaly categories [32]. The data organisation used in this study contained 238 training and 199 testing sequence entries. Because this organisation differs from the split commonly quoted in some ShanghaiTech studies, the resulting values are interpreted under the dataset organisation described here rather than assumed to be protocol-identical to every published result. Table II summarises the three dataset protocols and their evaluation roles.

TABLE II. Summary of Datasets and Evaluation Protocols

Dataset	Study-specific scale/protocol	Input representation	Evaluation role/caveat
CUHK Avenue	16 training videos; 21 testing videos; stride-2 sampling; metric contrasts sampled training-normal frames with sampled testing frames.	RGB frames	Avenue component comparison under the stated normal-versus-test metric construction.
UCSD Ped2	Study-specific split: 11 normal train, 2 normal validation, 3 normal test; 9 abnormal-validation and 15 abnormal-test sequence entries; 870/150/255 normal and 780/1,230 abnormal validation/test frames after stride sampling.	Three-channel optical-flow representation	Leakage-aware held-out protocol for model selection and final evaluation; reported as a study-specific Ped2 split.
ShanghaiTech	238 training and 199 testing sequence entries; stride-2 processing.	RGB frames	Multi-scene comparison under the stated data organisation; metric population combines normal-training scores with frame-labelled test scores.

B. Data Splitting and Preprocessing

Frames were sampled at a stride of two to reduce near-duplicate observations and computational cost. RGB frames were converted to the required three-channel format and resized to 224 × 224 pixels before backbone-specific preprocessing. Each CNN received the corresponding ImageNet preprocessing function supplied with its Keras implementation. A batch size of 64 was used for feature extraction, and a fixed random seed of 42 was applied to the stochastic components of the experiments.

Avenue and ShanghaiTech used their dataset-specific training and testing directories or sequence lists, whereas Ped2 used the study-specific video-level partition described above. For learnable temporal configurations, normal training sequences were reserved for early stopping where required. Feature scalers, PCA transformations, covariance models, and BiTCN parameters were fitted without using final evaluation labels. The construction of the metric population nevertheless differs by dataset and is documented explicitly in Section IV-E.

The Ped2 experiments used precomputed optical-flow arrays. Two-channel flow maps were converted to three-channel representations by appending the flow magnitude as the third channel; single-channel arrays were repeated across three channels. Values falling outside the [0, 1] range were min–max normalised before resizing. This conversion allowed the optical-flow representation to be processed by the same pretrained CNN family used for the RGB-based Avenue and ShanghaiTech experiments.

Feature standardisation was performed separately for each backbone using statistics estimated from normal training features. The fitted scalers were then applied unchanged to validation and test data. In the early-fusion path, the five standardised feature vectors were concatenated before dimensionality reduction. PCA was fitted on the normal training representation only, with the number of components selected automatically to preserve at least 95% of the training variance; the same PCA transformation was subsequently applied to validation and test representations. Temporal windows were constructed independently within each video so that no BiTCN sequence crossed a video boundary. Where padding was required at the beginning of a sequence, the first available observation was repeated to preserve one output score per sampled frame.

C. Implementation Details

The experiments were implemented in Python using TensorFlow/Keras, NumPy, OpenCV, and scikit-learn and executed in GPU-enabled environments. The modular runs used TensorFlow 2.19–2.20 and NVIDIA Tesla T4/P100 accelerators. The CNN classification heads were removed (include_top=False), all pretrained convolutional parameters were frozen, and global average pooling was applied to obtain one descriptor per frame. The resulting feature dimensions were 512 for VGG19, 2,048 for ResNet50, 1,280 for EfficientNetB0, 1,024 for DenseNet121, and 2,048 for InceptionV3. Their concatenation therefore produced a 6,912-dimensional early-fusion representation before PCA.

The reference experiments used centroid/Euclidean distance, while the covariance-aware configuration used PCA where required, followed by regularised Mahalanobis scoring with $\lambda = 10^{-4}$. Since these are complete alternative scoring paths, the reported comparison is attributed to the integrated PCA-Mahalanobis configuration. A PCA-Euclidean variant on the same reduced representation is reserved as a finer-grained future ablation for separating dimensionality-reduction and scoring-geometry effects.

Three late-fusion rules were examined after the individual backbone scores had been placed on comparable scales: arithmetic mean, weighted mean, and maximum fusion. Mean fusion provides a consensus score across detectors, whereas maximum fusion preserves the strongest individual anomaly response. For the validation-aware Ped2 experiments, the weighted-fusion coefficients were obtained from the individual validation AUC-ROC values and normalised to sum to one before being fixed for final-test evaluation.

The temporal module used a 16-frame BiTCN window with kernel size 3 and dilation factors {1, 2, 4, 8}. Each forward and backward branch consisted of residual dilated Conv1D blocks with ReLU activation, batch normalisation, spatial dropout of 0.2, and L2 regularisation of 10^{-4}. The backward branch operated on a reversed temporal sequence and was returned to the original temporal order before fusion with the forward branch. The two branches were concatenated, followed by global temporal average pooling and dense layers of 64 and 16 units. A dropout rate of 0.3 was applied between the dense layers. The early-fusion temporal configuration used 128 filters, while the compact late-fusion score model used 64 filters.

Two BiTCN configurations were evaluated. In Early-Fusion BiTCN, the standardised backbone features were concatenated, PCA-reduced, arranged into temporal windows, and supplied to the temporal network. In BiTCN late fusion, the input at each time step was the five-dimensional vector of backbone-specific anomaly scores. The latter therefore models the temporal agreement and disagreement of the individual detectors without reintroducing the full high-dimensional feature representation.

D. Training Protocol

The study follows a normality-modelling protocol. The pretrained CNN backbones were not fine-tuned; they served only as fixed feature extractors. The centroid reference, feature-scaling parameters, PCA projection, normal mean, and regularised covariance matrix were estimated from normal training data. For Ped2, validation data were used for model control and fusion calibration before evaluation on the held-out custom test subset. Avenue and ShanghaiTech compare several configurations on their fixed evaluation populations; their highest scores are therefore treated as descriptive configuration comparisons rather than as estimates from a single validation-selected final model.

Early-Fusion BiTCN was trained only on normal temporal sequences using centre-frame reconstruction. For a 16-frame input window, the network reconstructed the feature vector at the centre of the sequence, and the mean squared reconstruction error provided the temporal anomaly evidence. This objective allows the temporal model to learn the regular evolution of normal representations without requiring abnormal examples for gradient-based training.

BiTCN late fusion was trained on sequences of normalised backbone anomaly scores. Its target represents the temporal variability of each normal score window, computed from the standard deviation of the input sequence and scaled to a unit interval. This provides a compact normal-only learning objective that maps local score dynamics to a temporal anomaly signal without requiring abnormal training examples. Future work can extend the temporal study with causal variants and lightweight smoothing controls to characterise deployment trade-offs across different surveillance settings.

Both BiTCN variants were optimised using Adam with an initial learning rate of 10^{-3} and mean squared error loss. Training was limited to 50 epochs with a batch size of 64. Early stopping monitored validation loss with a patience of 10 epochs and restored the best model weights. ReduceLROnPlateau reduced the learning rate by a factor of 0.5 after five epochs without improvement, with a minimum learning rate of 10^{-6}. Mean absolute error was retained as an auxiliary optimisation metric.

E. Evaluation Metrics

Performance was summarised using Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and average precision (AP). AUC-ROC measures the ranking of positive observations above negative observations across thresholds. AP summarises the precision-recall trade-off as a weighted mean of precision values over increases in recall. For consistency with the experimental result tables, AP is denoted AUC-PR throughout the remainder of the study; it should not be interpreted as trapezoidal integration of the precision-recall curve.

The metric populations were constructed as follows. For Avenue, sampled training-video frames form the negative class and sampled testing-video frames form the positive class, yielding a normal-training versus test-video separation measure under the study protocol. For ShanghaiTech, normal-training scores are concatenated with frame-labelled test scores before AUC-ROC and AP are computed. Ped2 uses the study-specific validation/test construction described in Section IV-A, with validation results used for model selection and fusion weighting and the held-out test subset used for final reporting. These dataset-specific constructions are stated explicitly so that Tables III–V can be interpreted within the population on which each metric was computed.

Taken together, the experiments support a component-wise assessment within each documented protocol. Where design changes occur jointly, most notably PCA reduction with covariance-aware scoring, the corresponding results are interpreted at configuration level. The temporal experiments similarly quantify the behaviour of the implemented BiTCN path, while causal and lightweight smoothing variants remain useful extensions for a more deployment-oriented temporal comparison.

F. Computational Complexity and Reproducibility

Let N denote the number of processed frames, K = 5 the number of backbones, d_k the descriptor dimension of backbone k, and r_k its PCA-reduced dimension. The frozen feature-extraction stage scales as $\mathcal{O}(N\sum_{k=1}^{K} C_k)$, where C_k is the forward-inference cost of backbone k. The five descriptor dimensions are 512, 2,048, 1,280, 1,024, and 2,048, giving an early-fusion dimension D = 6,912 before PCA. A dense covariance matrix in that unreduced space contains D^2 = 47,775,744 elements, which is approximately 382 MB in float64 before intermediate storage; this quantifies the numerical and memory motivation for PCA.

For a feature matrix with N observations and d dimensions, covariance-based PCA requires approximately $\mathcal{O}(Nd^2 + d^3)$ for covariance formation and eigendecomposition, while transformation to r retained components costs $\mathcal{O}(Nd_k r_k)$. Covariance estimation in the reduced space costs $\mathcal{O}(Nr_k^2)$, dense covariance inversion costs $\mathcal{O}(r_k^3)$ during fitting, and Mahalanobis scoring costs $\mathcal{O}(r_k^2)$ per frame, compared with $\mathcal{O}(r_k)$ for Euclidean distance. Average, weighted, and maximum score fusion each require $\mathcal{O}(K)$ operations per frame after individual scores have been produced.

For a temporal convolution with sequence length L = 16, kernel width q = 3, and channel dimensions c_{in} and c_{out}, the dominant convolutional cost is $\mathcal{O}(Lq\, c_{\text{in}} c_{\text{out}})$ per layer and is approximately doubled because the sequence is processed in both forward and backward directions. The five classification-head-free CNN backbones contain approximately 76.5 million frozen parameters in total: 20.0M for VGG19, 23.6M for ResNet50, 4.05M for EfficientNetB0, 7.04M for DenseNet121, and 21.8M for InceptionV3. Assuming 32-bit floating-point storage, th,ese parameters require approximately $76.5 \times 10^6 \times 4 \approx 306 \times 10^6$ bytes ≈ 306 MB of memory, excluding intermediate activations and other runtime memory requirements. The instantiated BiTCN configurations contain approximately 1.24-1.75 million parameters, of which about 1.23-1.75 million are trainable; thus, the full parameter inventory is approximately 77.7-78.3 million parameters while the CNN feature extractors remain frozen. A fixed seed of 42 is used for reproducibility across learnable runs. The parameter inventory and asymptotic analysis provide an implementation-independent efficiency characterisation, while repeated-run stability and platform-specific latency, throughput, and peak-memory profiling are natural extensions for deployment-focused evaluation.

V. Results and Analysis

A. CUHK Avenue

Under the Avenue evaluation construction described in Section IV-E, BiTCN late fusion produces the highest separation score. As shown in Table III, the early-fusion centroid/Euclidean reference gives AUC-ROC = 0.8906 and AUC-PR = 0.8938, while ResNet50 + PCA-Mahalanobis reaches 0.9980/0.9979 and BiTCN late fusion reaches 0.9984/0.9981. The increase from 0.8906 to 0.9980 is +0.1074 AUC-ROC and is attributed to the integrated PCA-Mahalanobis configuration because dimensionality treatment and score geometry change together. BiTCN late fusion provides a further +0.0004 AUC-ROC refinement over the strongest individual configuration. These values characterise normal-training versus test-video separation under the documented Avenue protocol and are interpreted on that basis.

TABLE III. Avenue Results Under the Documented Evaluation Protocol

Method	AUC-ROC	AUC-PR	Interpretation within Avenue protocol
Baseline early fusion	0.8906	0.8938	Centroid/Euclidean reference configuration.
ResNet50 + PCA-Mahalanobis	0.9980	0.9979	Best individual PCA-Mahalanobis result; gain interpreted for the integrated PCA-Mahalanobis path.
Ensemble max + PCA-Mahalanobis	0.9970	0.9969	Strong late-fusion result under the same Avenue protocol.
BiTCN late fusion	0.9984	0.9981	Highest Avenue separation score; bidirectional temporal refinement.

B. UCSD Ped2

Under the study-specific Ped2 held-out split, EfficientNetB0 + PCA-Mahalanobis is the strongest configuration, with AUC-ROC = 0.9730 and AUC-PR = 0.9938. Table IV shows that weighted and average late fusion reach 0.9402 and 0.9399 AUC-ROC, respectively, while BiTCN late fusion reaches 0.9192. EfficientNetB0 therefore exceeds the best conventional fusion by 0.0328 AUC-ROC and BiTCN late fusion by 0.0538. Fig. 2 and 3 provide the corresponding metric and ROC visualisations. This result demonstrates dataset-dependent complementarity: for the controlled Ped2 setting, the individual EfficientNetB0 representation is already the most discriminative option within the evaluated framework.

TABLE IV. UCSD PED2 FINAL-TEST RESULTS, INCLUDING EARLY-FUSION BITCN AND BITCN LATE FUSION

Method	AUC-ROC	AUC-PR	Type	Interpretation within Ped2 held-out split
EfficientNetB0 + PCA-Mahalanobis	0.9730	0.9938	Individual	Best validation-selected held-out result in the study-specific split.
Ensemble weighted	0.9402	0.9856	Late fusion	Best conventional fusion result, but below EfficientNetB0.
Ensemble average	0.9399	0.9856	Late fusion	Very close to weighted fusion; below EfficientNetB0.
VGG19 + PCA-Mahalanobis	0.9317	0.9839	Individual	Strong texture-oriented backbone.
ResNet50 + PCA-Mahalanobis	0.9304	0.9831	Individual	Stable residual representation.
Early fusion + PCA-Mahalanobis	0.9298	0.9834	Feature concatenation	Concatenation does not outperform EfficientNetB0.
BiTCN late fusion	0.9192	0.9818	Late Fusion + BiTCN	Bidirectional temporal refinement of late-fusion score sequences.
Early-Fusion BiTCN	0.9169	0.9807	Early Fusion + BiTCN	Bidirectional temporal modelling over the early-fused representation.
InceptionV3 + PCA-Mahalanobis	0.9128	0.9804	Individual	Moderate final-test performance.
Ensemble max	0.9127	0.9794	Late fusion	Less stable on Ped2.
DenseNet121 + PCA-Mahalanobis	0.9024	0.9774	Individual	Lowest final-test result among listed configurations.

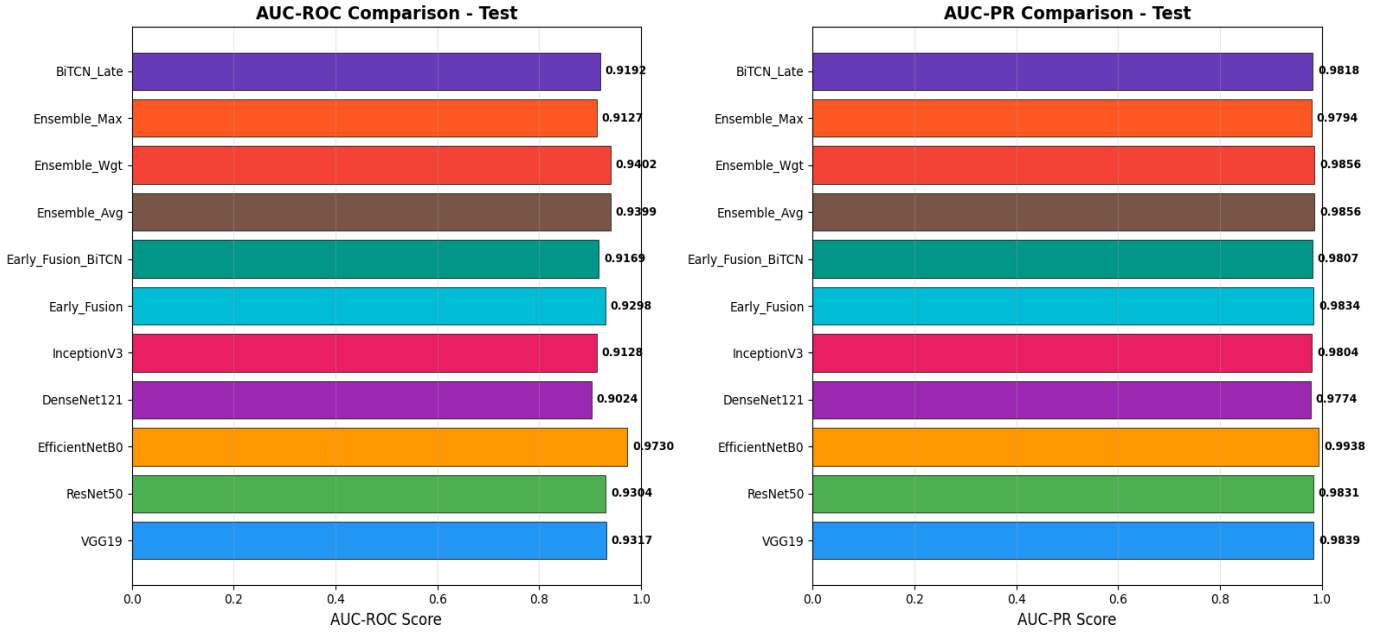

Fig. 2. UCSD Ped2 final-test AUC-ROC and AUC-PR comparison, including early-fusion BiTCN and BiTCN late fusion.

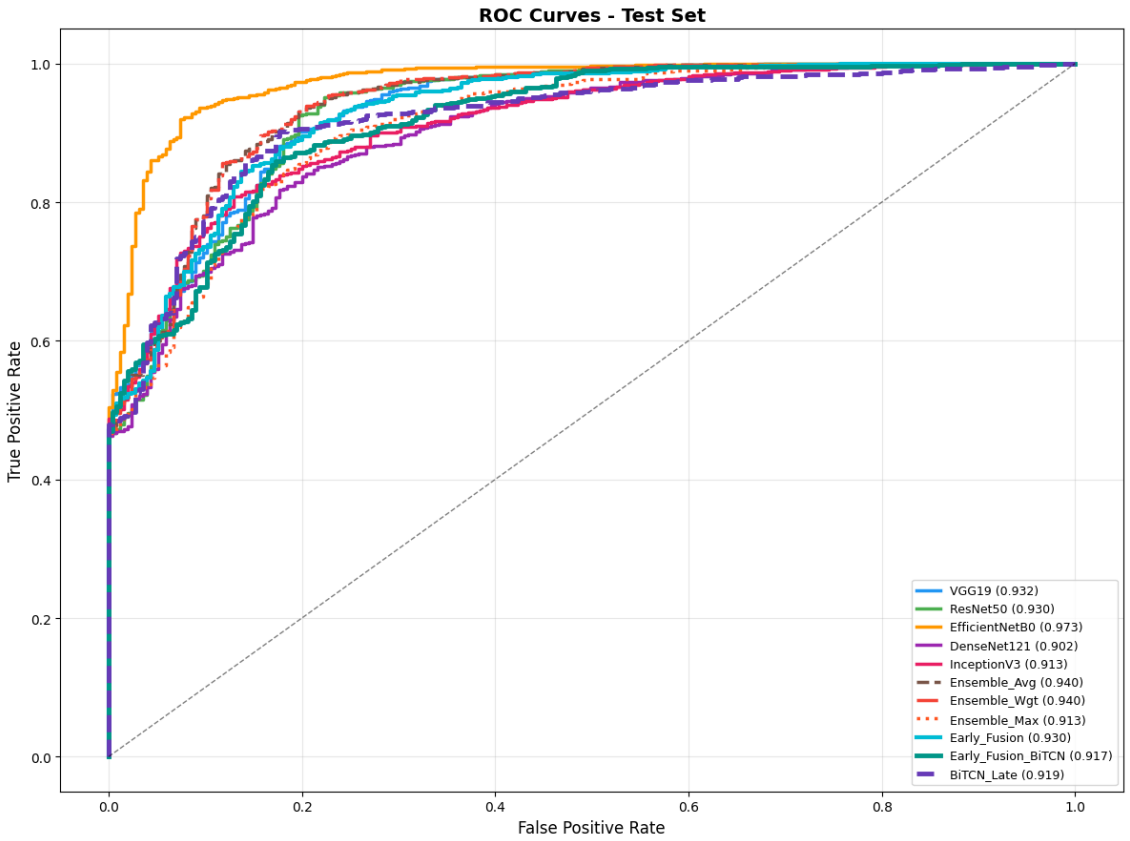

Fig. 3. UCSD Ped2 final-test ROC curves, including early-fusion BiTCN and BiTCN late fusion.

C. ShanghaiTech

ShanghaiTech provides the most heterogeneous scene collection in the study. Table V shows that the centroid reference gives AUC-ROC = 0.7050 and AUC-PR = 0.1150, EfficientNetB0 + PCA-Mahalanobis gives 0.7612/0.1641, Early-Fusion BiTCN gives 0.8237/0.8154, and BiTCN late fusion gives 0.8534/0.8481. Relative to the centroid reference, BiTCN late fusion improves AUC-ROC by +0.1484 and AP by +0.7331. AP can change more sharply than ROC AUC because it is sensitive to positive prevalence and to the ranking of positives in high-precision regions. The result is therefore interpreted within the metric population defined in Section IV-E, where normal-training scores are combined with frame-labelled test scores.

TABLE V. SHANGHAITECH RESULTS UNDER THE DOCUMENTED EVALUATION PROTOCOL

Method	AUC-ROC	AUC-PR	Interpretation within ShanghaiTech protocol
Baseline early fusion	0.7050	0.1150	Centroid reference under the stated train-normal + test-score metric construction.
EfficientNetB0 + PCA-Mahalanobis	0.7612	0.1641	Integrated PCA-Mahalanobis configuration.
Early fusion + BiTCN	0.8237	0.8154	Substantial AP increase under the documented metric population.
Ensemble max late fusion	0.8370	0.8269	Strong non-BiTCN score-fusion result under the same metric construction.
BiTCN late fusion	0.8534	0.8481	Highest ShanghaiTech score under the documented protocol; bidirectional temporal refinement.

VI. CONTEXTUAL COMPARISON WITH RECENT VAD METHODS

Table VI summarises recent traceable results reported by their original authors together with the best dataset-specific result obtained by the proposed framework. Because supervision level, input representation, object localisation, score construction, and AUC aggregation vary across studies, the comparison is used as protocol-aware literature context rather than as a controlled head-to-head ranking. Tables III-V provide the detailed within-study results and the evaluation setting for each proposed configuration.

TABLE VI. RECENT LITERATURE CONTEXT AND BEST PROPOSED RESULTS ACROSS THE THREE DATASETS; VALUES ARE REPORTED UNDER THE RESPECTIVE STUDY PROTOCOLS

Method	Year/venue	Modelling strategy	UCSD Ped2 AUC (%)	Avenue AUC (%)	ShanghaiTech AUC (%)	Note
AnomalyRuler	Yang et al. [15] / ECCV 2024	Few-normal-shot LLM/VLM rule reasoning	97.9	89.7	85.2	Rule-based reasoning; not directly identical to this training protocol.
Self-distilled masked AE	Ristea et al. [4] / CVPR 2024	Efficient self-distilled masked autoencoder	95.4 micro / 98.4 macro	91.3 micro / 90.9 macro	79.1 micro / 84.7 macro	Reports micro and macro AUC.
VideoPatchCore	Ahn et al. [3] / ACCV 2024	Training-free memory method	NR	92.8	85.1	Reports Avenue, ShanghaiTech, and Corridor.
VADMamba	Lyu et al. [14] / ICME 2025	State-space prediction and optical-flow reconstruction	98.5	91.5	77.0	Fast state-space VAD method.
UniVAD	Lee et al. [33] / WACV 2026	Unified anomaly-type model with multiple streams	NR	NR	89.5 micro / 91.6 macro	Strong multi-stream ShanghaiTech comparator.
Proposed framework (best dataset-specific configuration)	This work	PCA-stabilised covariance-aware scoring with dataset-dependent fusion/temporal refinement	97.30	99.84	85.34	Ped2: EfficientNetB0 + PCA-Mahalanobis; Avenue and ShanghaiTech: BiTCN late fusion. Values are reported under the study-specific protocols described in Section IV.

VII. DISCUSSION

A. Interpretation of Dataset-Specific Behaviour

The three evaluation settings reveal complementary aspects of the framework. Avenue provides the strongest anomaly separation, Ped2 shows that a compact single-backbone covariance-aware configuration can be sufficient, and ShanghaiTech tests robustness under substantial cross-scene variability. Taken together, the results support a dataset-adaptive interpretation in which representation, fusion, and temporal refinement are selected according to scene characteristics rather than assuming one universally optimal configuration.

On Ped2, EfficientNetB0 + PCA-Mahalanobis reaches 0.9730 AUC-ROC, compared with 0.9402 for the best conventional fusion and 0.9192 for BiTCN late fusion. This result is informative because it shows that, in a relatively controlled pedestrian-walkway scene, a strong individual representation coupled with covariance-aware scoring can capture the dominant abnormal patterns without requiring the full multi-backbone temporal path. The modular design therefore supports selecting the least complex configuration that preserves strong discrimination for a given scene.

ShanghaiTech provides the strongest cross-scene stress test because of its multiple scenes, viewpoints, illumination conditions, and anomaly types. Under the evaluation organisation described in Section IV, the progression from centroid scoring to PCA-Mahalanobis and temporal fusion produces substantial improvement in anomaly separation, indicating that dimensionality control, covariance-aware scoring, and temporal context can be valuable in heterogeneous surveillance settings. Because published methods use different evaluation organisations and reporting conventions, the cross-paper comparison is used for contextual positioning rather than direct numerical ranking.

B. Effect of PCA-Mahalanobis Scoring

The experiments demonstrate a substantial change in anomaly separation when moving from the centroid/Euclidean reference to the PCA-Mahalanobis path. Because PCA reduction and covariance-aware scoring are introduced together in this configuration, the observed change is attributed to the combined PCA-stabilised covariance model rather than to either component in isolation. This conservative attribution keeps the interpretation aligned with the implemented ablation. A PCA-Euclidean control on the same reduced representation is a useful future refinement for decomposing the individual contribution of scoring geometry more finely.

PCA remains important for numerical tractability. The five pooled descriptors concatenate to 6,912 dimensions, for which a dense covariance matrix contains 47,775,744 elements. Reducing the representation before covariance estimation lowers storage and inversion cost and mitigates ill-conditioning. This is a computational justification for PCA, distinct from any claim that PCA alone improves AUC.

The use of Mahalanobis distance is consistent with prior work demonstrating the value of covariance-aware anomaly scoring [25], [26], [27]. In the present framework, covariance-aware scoring is applied after dimensionality control, providing a statistically structured alternative to centroid distance while keeping covariance estimation tractable. The results therefore support the practical value of the combined PCA-stabilised covariance path within a heterogeneous feature pipeline.

C. Fusion and Temporal Score Modelling

The fusion results illustrate why a modular architecture is useful. Early fusion combines all backbone features before scoring and can capture cross-backbone interactions, but it also creates a larger and more heterogeneous representation. On Ped2, EfficientNetB0 alone outperforms early fusion, showing that a single well-matched representation may already be

sufficiently discriminative for a controlled scene. Rather than forcing fusion in every case, the framework enables empirical selection between individual, early-fusion, and late-fusion configurations.

Late fusion preserves the statistical behaviour of each backbone score stream and provides a low-dimensional mechanism for combining independent anomaly evidence. Its benefit is dataset-dependent: on Ped2, EfficientNetB0 alone reaches 0.9730 AUC-ROC compared with 0.9402 for the best standard fusion result, whereas on Avenue BiTCN late fusion reaches 0.9984 compared with 0.9980 for ResNet50. These results show that heterogeneous fusion is most useful when the constituent detectors provide complementary evidence, while the modular framework can retain a strong individual backbone when fusion offers no advantage.

The BiTCN experiments show that temporal score modelling can provide useful refinement when anomaly evidence is temporally coherent, most notably in the Avenue and ShanghaiTech settings. By operating on compact backbone-specific score sequences, the temporal stage adds sequence context without reintroducing the full feature dimensionality. The bidirectional configuration uses both preceding and succeeding observations and is therefore designed for offline or bounded-delay analysis, where local future context is available. Causal temporal modelling and lightweight smoothing baselines provide natural deployment-oriented extensions for future work.

D. Benchmark Positioning and Practical Implications

Published VAD results are presented as contextual benchmark positioning. Contemporary studies vary in supervision level, input representation, preprocessing, anomaly definition, and reporting convention; Table VI therefore shows how the obtained values relate to recent literature while keeping those protocol differences explicit. This provides a useful external performance context without overinterpreting non-identical experimental settings as a controlled head-to-head ranking.

From an engineering perspective, the framework is modular and inspectable: feature extraction, PCA, covariance estimation, score fusion, and temporal modelling can be diagnosed independently. The complete five-backbone path has an approximate static parameter inventory of 76.5M, compared with about 4.05M for EfficientNetB0 alone and 23.6M for ResNet50 alone. The full extractor therefore contains about 18.9 times the parameters of EfficientNetB0 and 3.24 times those of ResNet50 before the temporal module is added. Together with the analytical complexity results, these figures provide a reproducible, implementation-independent characterisation of computational scale. Deployment latency, throughput, and peak memory remain hardware- and software-dependent and can be profiled for the intended operational platform.

VIII. Conclusion and Future Work

This study presented a modular frame-level VAD framework integrating heterogeneous frozen CNN representations, PCA-based dimensionality reduction, covariance-aware Mahalanobis scoring, feature- and score-level fusion, and bidirectional temporal score modelling. Across the three documented evaluation settings, the results demonstrate that anomaly separation depends not only on the visual encoder but also on feature-space geometry, fusion strategy, and temporal context. BiTCN late fusion achieves AUC-ROC/AUC-PR values of 0.9984/0.9981 on Avenue and 0.8534/0.8481 on ShanghaiTech, while EfficientNetB0 with PCA-Mahalanobis scoring provides the strongest Ped2 result at 0.9730/0.9938. The findings also demonstrate the practical value of modularity: a simpler single-backbone configuration can be preferred when it is already highly discriminative, whereas temporal fusion can be introduced when score dynamics provide additional information. Analytical complexity and parameter-count analysis further characterise the accuracy–complexity trade-off among the evaluated configurations. Overall, the study supports PCA-stabilised covariance-aware modelling as an interpretable basis for deep-feature VAD and shows that heterogeneous representations and temporal refinement are most effective when applied selectively according to dataset characteristics. Future work will investigate causal temporal variants, repeated-run stability, deployment-specific efficiency profiling, object- and scene-aware representations, and unified cross-dataset evaluation protocols.

References

[1] Y. Liu et al., "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," ACM Computing Surveys, vol. 56, no. 7, Art. no. 189, pp. 1-38, 2024, doi: 10.1145/3645101.

[2] M. Abdalla, S. Javed, M. Al Radi, A. Ulhaq, and N. Werghi, "Video anomaly detection in 10 years: A survey and outlook," Neural Computing and Applications, vol. 37, no. 32, pp. 26321-26364, 2025, doi: 10.1007/s00521-025-11659-8.

[3] S. Ahn, Y. Jo, K. Lee, and S. Park, "VideoPatchCore: An effective method to memorize normality for video anomaly detection," in Computer Vision - ACCV 2024, M. Cho, I. Laptev, D. Tran, A. Yao, and H. Zha, Eds., Lecture Notes in Computer Science, vol. 15474. Springer, 2025, pp. 312-328, doi: 10.1007/978-981-96-0908-6_18.

[4] N.-C. Ristea et al., "Self-distilled masked auto-encoders are efficient video anomaly detectors," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 15984-15995, doi: 10.1109/CVPR52733.2024.01513.

[5] P. Wu et al., "Open-vocabulary video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18297-18307. doi: 10.1109/CVPR52733.2024.01732.

[6] L. Zanella, W. Menapace, M. Mancini, Y. Wang, and E. Ricci, "Harnessing large language models for training-free video anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 18527-18536, doi: 10.1109/CVPR52733.2024.01753.

[7] E. Dilek and M. Dener, "An overview of transformers for video anomaly detection," Neural Computing and Applications, vol. 37, pp. 17825-17857, 2025, doi: 10.1007/s00521-025-11218-1.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700-4708.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine

Learning (ICML), Proceedings of Machine Learning Research, vol. 97, 2019, pp. 6105-6114.

[13] P. C. Mahalanobis, "On the generalised distance in statistics," Proceedings of the National Institute of Sciences of India, vol. 2, pp. 49-55, 1936.

[14] J. Lyu et al., "VADMamba: Exploring state space models for fast video anomaly detection," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2025, pp. 1-6.

[15] Y. Yang, K. Lee, B. Dariush, Y. Cao, and S.-Y. Lo, "Follow the rules: Reasoning for video anomaly detection with large language models," in Computer Vision - ECCV 2024, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Lecture Notes in Computer Science, vol. 15139. Springer, 2025, pp. 304-322, doi: 10.1007/978-3-031-73004-7_18.

[16] Y. Pu, X. Wu, L. Yang, and S. Wang, "Learning prompt-enhanced context features for weakly-supervised video anomaly detection," IEEE Transactions on Image Processing, vol. 33, pp. 4923-4936, 2024, doi: 10.1109/TIP.2024.3451935.

[17] T. Liu, K.-M. Lam, and B.-K. Bao, "Injecting text clues for improving anomalous event detection from weakly labeled videos," IEEE Transactions on Image Processing, vol. 33, pp. 5907-5920, 2024, doi: 10.1109/TIP.2024.3477351.

[18] W. Pang, Q. He, Y. Li, and N. Ahmed, "Detecting video anomalies by jointly utilizing appearance and skeleton information," Expert Systems with Applications, vol. 246, Art. no. 123135, 2024, doi: 10.1016/j.eswa.2023.123135.

[19] A. Hussain et al., "TDS-Net: Transformer-enhanced dual-stream network for video anomaly detection," Expert Systems with Applications, vol. 256, Art. no. 124846, 2024, doi: 10.1016/j.eswa.2024.124846.

[20] S. Paulraj and S. Vairavasundaram, "M2VAD: Multiview multimodality transformer-based weakly supervised video anomaly detection," Image and Vision Computing, vol. 149, Art. no. 105139, 2024, doi: 10.1016/j.imavis.2024.105139.

[21] S. M. Rahimpour, M. Kazemi, P. Moallem, and M. Safayani, "Video anomaly detection using transformers and ensemble of convolutional auto-encoders," Computers & Electrical Engineering, vol. 120, Art. no. 109879, 2024, doi: 10.1016/j.compeleceng.2024.109879.

[22] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.

[23] 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 Transactions on Image Processing, vol. 33, pp. 6865-6880, 2024, doi: 10.1109/TIP.2024.3512369.

[24] A. Ismail and D. Aboshady, "A novel lightweight dual-stream recurrent transformer model for anomaly detection in driving dashcam videos," Neural Computing and Applications, vol. 37, pp. 18527-18548, 2025, doi: 10.1007/s00521-025-11380-6.

[25] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in Pattern Recognition: ICPR International Workshops and Challenges. Springer, 2021, pp. 475-489, doi: 10.1007/978-3-030-68799-1_35.

[26] M. H. Sharif, L. Jiao, and C. W. Omlin, "CNN-ViT supported weakly-supervised video segment level anomaly detection," Sensors, vol. 23, no. 18, Art. no. 7734, 2023, doi: 10.3390/s23187734.

[27] F. Behrendt et al., "Leveraging the Mahalanobis distance to enhance unsupervised brain MRI anomaly detection," in Medical Image Computing and Computer Assisted Intervention - MICCAI 2024. Springer, 2024, pp. 394-404, doi: 10.1007/978-3-031-72120-5_37.

[28] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, Art. no. 20150202, 2016, doi: 10.1098/rsta.2015.0202.

[29] M. Altin and A. Cakir, "Exploring the influence of dimensionality reduction on anomaly detection performance in multivariate time series," IEEE Access, vol. 12, pp. 85783-85794, 2024, doi: 10.1109/ACCESS.2024.3415088.

[30] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2013, pp. 2720-2727.

[31] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (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. Computer Vision (ICCV), 2017, pp. 341-349.

[33] K. Lee, Y. Jo, S. Ahn, and S. Park, "Unified video anomaly detection model for detecting different anomaly types," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2026, pp. 784-794, doi: 10.1109/WACV61042.2026.00083.