Pattern Recognition 167 (2025) 111759

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Unsupervised video anomaly detection by memory network with autoencoders in euclidean and non-euclidean spaces

Jinmyeong Kim, Sung-Bae Cho *

Dept. of Computer Science, Yonsei University, Seoul 03722, South Korea

ARTICLE INFO

Keywords:
Video anomaly detection
Autoencoder
Non-euclidean space
Constant curvature manifold

ABSTRACT

Conventional video anomaly detection methods, which predominantly rely on models trained in Euclidean space, are inherently limited in their ability to capture the complex, nonlinear dynamics in video data. This paper proposes a method employing two-stream autoencoder to distinguish and learn salient features in both Euclidean and constant curvature manifold (CCM) spaces. One autoencoder encodes and reconstructs latent vectors in Euclidean space to capture spatial correlations effectively and store these representations in a memory network. The other decoder processes latent vectors in non-Euclidean space of CCM to focus on learning temporally coherent features with semantic hierarchical relationships, which are also stored in the memory network. The proposed model leverages two-stream autoencoder to independently learn spatial and temporal features, integrating the reconstruction error maps from the autoencoders to enhance its ability to distinguish between normal and abnormal patterns. Experimental results demonstrate the model's superior detection performance against the conventional methods, achieving AUC scores of 99.3 %, 92.8 %, and 80.5 % on the UCSD Ped2, CUHK Avenue, and ShanghaiTech Campus anomaly detection datasets, respectively.

1. Introduction

Detecting anomalies is essential for a wide range of applications, such as public safety [1–3], cyber security [4] and industrial automation [5–7]. In particular, video anomaly detection (VAD) has become increasingly important as it allows systems to monitor and analyze real-time video content to distinguish between normal behaviors and abnormal occurrences. This capability is critical in dynamic environments where immediate identification of irregular events can prevent potential risks or failures. However, it faces significant challenges to implement robust video anomaly detection systems. One primary issue is the imbalance between normal and abnormal data with rare anomalous events, making it difficult to obtain a substantial amount of abnormal samples [8,9]. This scarcity limits the viability of supervised learning approaches. Moreover, unlike static images, video data exhibit temporal characteristics with complex inter-frame dependencies, such as motion patterns, object interactions and contextual shifts [10,11]. These nonlinear features complicate the detection, as models must grasp both spatial details within individual frames and dynamic changes across sequences.

To tackle these challenges, unsupervised learning approaches have been extensively pursued [9,10]. They allow models to learn inherent features from unlabeled data and identify whether new observations deviate from learned norms. Despite their potential, many current unsupervised learning techniques are confined to Euclidean space, which inadequately captures the intricate and nonlinear properties of video data. Addressing this limitation requires methods capable of effectively capturing the nonlinear complexities in video sequences [11]. Since anomalies often exhibit nonlinear characteristics, learning in a constant curvature manifold (CCM) effectively captures the intrinsic geometric properties of data [12,13]. By mapping high-dimensional complex relationships onto a lower-dimensional curved surface, CCM preserves the underlying nonlinear patterns that are difficult to represent in Euclidean space [14–16]. This approach naturally models the data distribution through curvature, enabling a concise and accurate representation of the dynamic and nonlinear features inherent in video sequences [11,17]. As a result, it significantly enhances the detection of anomalies that display such complex behaviors. Thus, leveraging learning in CCM is crucial for a deeper understanding of these properties and for improving the accuracy of anomaly detection.

In this paper, we propose a two-stream autoencoder-based method that integrates learning in CCM with learning in traditional Euclidean

* Corresponding author.
E-mail addresses: jmkim_@yonsei.ac.kr (J. Kim), sbcho@yonsei.ac.kr (S.-B. Cho).

https://doi.org/10.1016/j.patcog.2025.111759
Received 30 October 2024; Received in revised form 21 April 2025; Accepted 22 April 2025
Available online 23 April 2025

Table 1
Related studies in unsupervised video anomaly detection.

Task	Model	Architecture	Key techniques	Key benefits	Challenges
Reconstruction	ITAE [30]	Two-path Autoencoder	Normalizing flow -based density estimation	Captures both static and dynamic features	Increased complexity compared to single-path AE
	MAAM [23]	Autoencoder, Memory network	Memory module, patch-based convolution, margint loss	Enhances feature separation, reduces computation costs	Patch size selection for multi-scene datasets
	MemAE [1]	Autoencoder, Memory network	Memory-based normal pattern storage	Efficient detection of rare anomalies	Memory tuning needed
	StackRNN [8]	Stacked RNN	Long-range temporal modeling	Captures extended temporal patterns	High computational cost, vanishing gradient issue
	SSL [31]	Autoencoder, Jigsaw puzzle module	Decoupled spatial-temporal jigsaw puzzles	Learns features without labeled data	Potential noise in complex tasks
Frame Prediction	AMMC [26]	Autoencoder, Memory network, FlowNet	Integrate FlowNet for motion-based anomaly detection	High sensitivity to motion anomalies	Limited effectiveness for static anomalies
	AnoPCN [9]	Autoencoder, ConvLSTM	Spatiotemporal integration	Handles complex patterns	High computational cost, challenging optimization
	BDPN [32]	Autoencoder, Multimodal discriminator, FlowNet	Multimodal discriminators for robust detection	Enhanced detection robustness	High complexity, needs careful tuning
	CDAE [3]	Autoencoder, Appearance/Motion cluster	Clustering appearance and motion features	Captures complex anomaly types	Sensitive to initial cluster configuration
	FFP [24]	Autoencoder, FlowNet, Mask RCNN, Discriminator	Combines multiple strong models for high accuracy	High detection accuracy using diverse features	High computational complexity, difficult to train and tune
	MNAD [2]	Autoencoder, Memory network	Memory network for normal patterns	Reduces false positives in stable environments	Need careful tuning for memory size and update rules
	MSN—Net [33]	Autoencoder, Memory network	Memory network to efficiently detect deviations	Improved anomaly detection efficiency	Performance depends on effective memory management
	SAFA [34]	Autoencoder, Self-attention	Uses self-attention to focus on relevant data parts	Improves focus on critical features	Computationally expensive, limited performance on distributed anomalies
	Unmasking [21]	Unmasking module, Appearance/Motion module	Focuses on motion changes	Sensitive to dynamic anomalies	Limited detection for static anomalies
	VEC [35]	Autoencoder, Cascade R-CNN	Improves object localization using cascade R-CNN	Accurate localization in complex scenes	Relies on detection, misses subtleties

space to enhance VAD performance. The autoencoder in Euclidean space represents local spatial features using a memory network that stores normal patterns and spatially similar representations. The CCM part captures the nonlinear temporal dynamics of video data, forming a hierarchical representation of temporal features. The CCM memory network further exploits the exponentially increasing distance measure from the origin in the CCM space, effectively representing hierarchical nonlinear temporal patterns [18,19]. By leveraging the two geometrical perspectives, the proposed method overcomes the limitations of conventional methods and detects complex and subtle anomalies in video sequence effectively [11,20]. It is verified with the popular benchmark datasets such as UCSD Ped2, CUHK Avenue, and ShanghaiTech Campus. A comparative study is also conducted with autoencoder-based methods, which are widely used for unsupervised VAD [1,2,9].

Our main contributions are as follows:

- We propose a novel model for VAD that leverages the Euclidean and CCM spaces to amplify the reconstruction error of anomaly.
- Evaluation on the three benchmark datasets demonstrates that the proposed method outperforms over 15 baseline methods, validating its effectiveness in improving VAD performance.

The organization of this paper is as follows: Section 2 provides a review of relevant research on VAD, and Section 3 presents the architecture and details of the proposed method. The experimental results in Section 4 demonstrate the effectiveness of the proposed method, and Section 5 concludes with a summary of findings and potential directions for future research.

2. Related works

Table 1 summarizes the relevant works on VAD. In the early stages of deep learning, VAD methods primarily relied on hand-crafted features, such as histograms of oriented gradients (HOG) and histograms of optical flows (HOF), to extract spatial and motion insights. Techniques like AE-Conv2d and Unmasking [21] employed these features to train classifiers for distinguishing normal and abnormal events. However, such methods suffered from computational complexity and limited scalability, particularly when processing large datasets.

Recent advancements in deep neural networks have significantly improved VAD by leveraging reconstruction- and prediction-based paradigms [2,22,23]. Reconstruction-based methods, such as MNAD [3] and MAAM [23], utilize memory networks to learn normal patterns and amplify reconstruction errors for anomalies. Prediction-based approaches focus on spatiotemporal representation learning to anticipate future frames, assuming anomalies will exhibit higher prediction errors. For example, Frame-Pred [24] uses U-Net-based architectures combined with optical flow to ensure temporal consistency, while FastAno++ [22] improves anomaly detection by enhancing feature separation with anomaly distance learning and adaptively controlling skip connections through context-aware skip connection, facilitating real-time inference.

On the other hand, the methods incorporating skip connections were proposed to minimize reconstruction errors caused by the diversity of normal patterns, maintaining high sensitivity to abnormal patterns [9]. Furthermore, some methods integrate pre-trained FlowNet models to utilize optical flow for extracting more precise temporal feature to improve detection performance [9,23]. These methods contribute to more accurate anomaly detection by capturing temporal consistency in video sequences [25]. Meanwhile, appearance-motion memory networks capture the coherence between appearance and motion in video sequences, effectively storing normal spatiotemporal patterns and accurately detecting abnormal motions [26].

Liu et al. [27] categorized VAD methods into generalized frameworks, highlighting the progression from unsupervised to more comprehensive paradigms, such as weakly-supervised (WAED) and fully-unsupervised (FVAD). These approaches aim to generalize across

complex scenarios by combining local and global representations. Building on this, Zhao et al. [28] introduced a local-global normality framework that balances short-term spatiotemporal patterns (local normality) with long-term prototype patterns (global normality), enhancing adaptability to both simple and complex scenes.

Liu et al. [29] proposed a causality-inspired method to address the limitations of traditional VAD methods in capturing spatiotemporal dependencies. By enforcing representation consistency between appearance and motion, this method leverages causal relationships to improve normality learning and robustness to diverse anomalies.

However, the memory-based methods alone may not sufficiently maximize reconstruction errors for anomalies and often require additional data for training auxiliary networks [1,2]. They also face limitations in effectively capturing the complex temporal dynamics of video data [3,26]. In order to address these challenges, it is crucial to employ methods capable of capturing nonlinear features effectively, such as those leveraging non-Euclidean spaces [21,23,25]. In this paper, we propose a method that learns spatial features through training in Euclidean space and captures temporal relationships through training in non-Euclidean space of CCM.

3. The proposed method

Fig. 1 shows the overall architecture of the proposed method for detecting anomalies in video sequence. The method receives consecutive frames $x_{t-n:t-1}$ into a shared encoder E_s, which learns features in both Euclidean and CCM spaces, converting them into feature vectors [21,23,36]. These feature vectors are then separately processed: Learning in Euclidean space is conducted through a memory network M_e and a decoder D_e, while learning in CCM space is facilitated by a memory network M_c and a decoder D_c [2,23]. The feature vectors produced by the shared encoder E_s are projected onto a negative curvature manifold using exponential mapping *expmap* [15,37].

Within the memory networks, similarity measures in each space capture normal spatial and temporal patterns, respectively [3,21]. The decoder outputs in Euclidean and CCM spaces, $\hat{x}_t^e$ and $\hat{x}_t^c$ are compared against the ground truth x_t, and their element-wise multiplication amplifies the reconstruction errors related to anomalies while reducing those for normal data. This amplification occurs because normal data produce consistent reconstructions in both spaces, resulting in minimal discrepancies when multiplied. In contrast, anomalies exhibit significant deviations in each space, and their discrepancies, when multiplied, are further magnified, effectively highlighting abnormal behaviors and enhancing detection sensitivity [1,2].

3.1. Learning in euclidean space

For learning in Euclidean space, a 2D convolutional structure comprising E_s, D_e, and a memory network M_e of size $m \times d$ is employed [27]. The shared encoder E_s takes consecutive frames $x_{t-n:t-1}$ from the video as input and extracts a latent vector f_s. The size of the latent vector f_s is w, h, and d, representing the width, height, and dimension of the feature map. Subsequently, the latent vector f_S is normalized to form f_e to facilitate the calculation of cosine similarity [2]. The normalized vector f_e is then input into the memory network M_e to measure its similarity to the stored memory items using cosine similarity [2,23].

The similarity measurement involves reshaping f_e into a matrix of size $(w \times h,\ d)$, allowing for the computation of similarities between each local feature in the feature map and M_e. Matrix multiplication of the reshaped f_e and M_e yields a similarity matrix S_e of size $(w \times h,\ m)$. This similarity matrix S_e is used to compute a weighted sum with M_e through matrix multiplication, resulting in f_{em}, a feature map of the same dimension as f_e [1,2]. Cosine similarity is employed as the similarity metric, as shown in Eq. (1), and the operational details of the memory network are depicted in Fig. 2.

$$S_e = matmul(f_e,\ M_e). \tag{1}$$

After obtaining f_{em} from the memory network, f_{em} is concatenated with f_e and passed into the decoder D_e, producing $\hat{x}_t^e$, the prediction for x_t. The mean squared error (MSE) between x_t and $\hat{x}_t^e$ is then computed to quantify the discrepancy between the prediction and the ground truth. Eq. (2) defines the loss function utilized for the learning process in

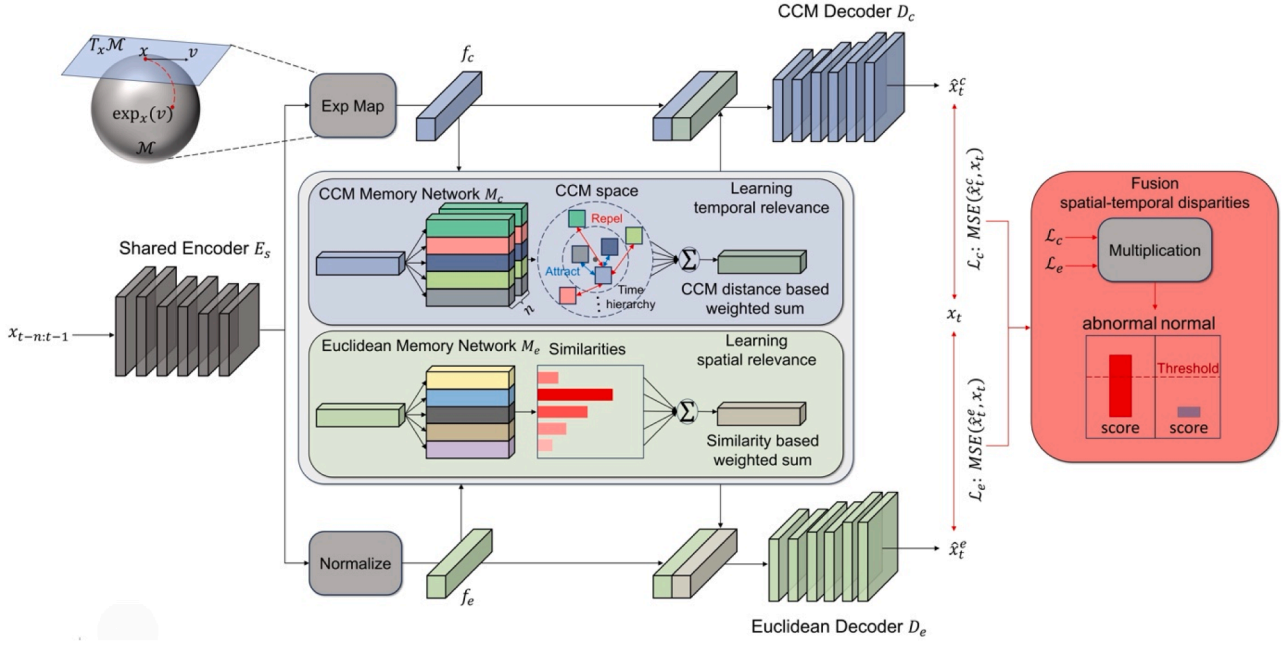

Fig. 1. Overview of the proposed method. The input sequence $\boldsymbol{x_{t-n:t-1}}$ is encoded by the shared encoder $\boldsymbol{E_s}$ to generate the latent features $\boldsymbol{f_c}$ (for CCM space) and $\boldsymbol{f_e}$ (for Euclidean space). The CCM Memory Network ($\boldsymbol{M_c}$) learns temporal relationships by projecting features into the CCM space using the exponential map and computing distances for weighted sum operations. The Euclidean Memory Network ($\boldsymbol{M_e}$) learns spatial relationships by calculating similarity-based weighted sums. The decoders $\boldsymbol{D_c}$ and $\boldsymbol{D_e}$ reconstruct the inputs as $\hat{\boldsymbol{x}}_{\boldsymbol{c}}^{\boldsymbol{t}}$ and $\hat{\boldsymbol{x}}_{\boldsymbol{e}}^{\boldsymbol{t}}$, respectively.

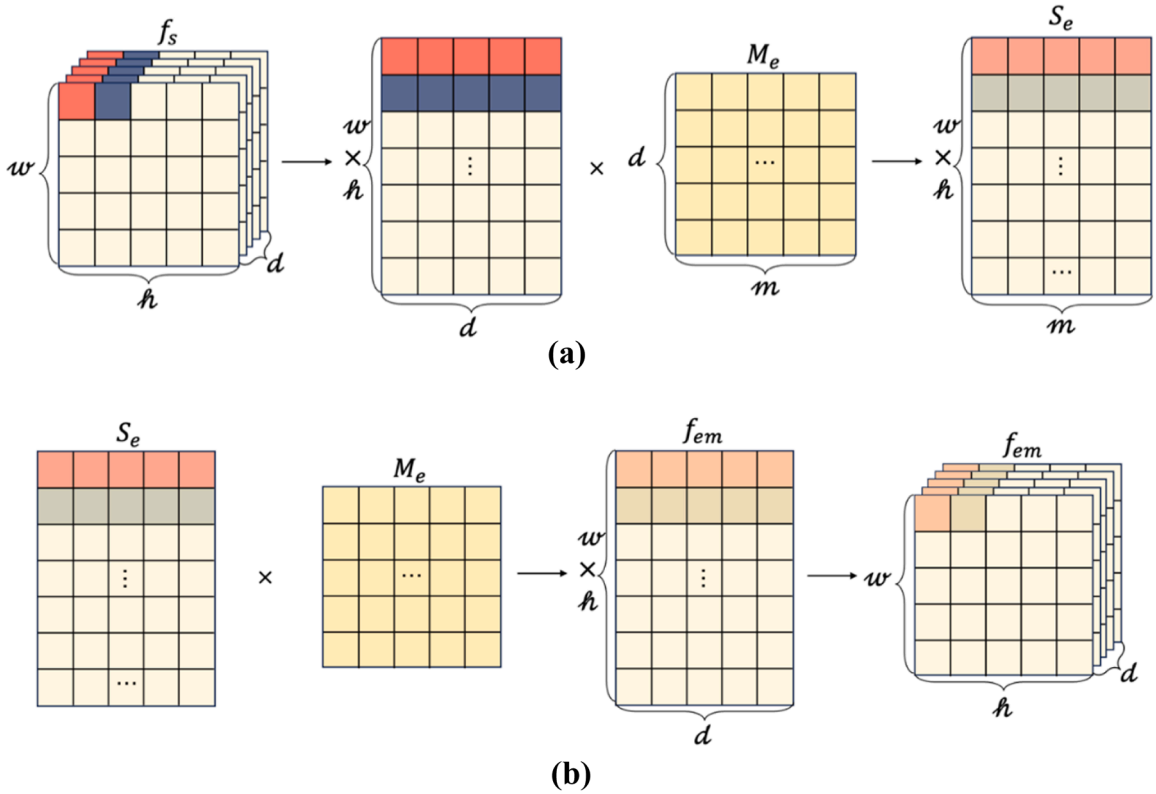

Fig. 2. The process of memory network in Euclidean space; (a) The process of calculating similarity matrix S_e, and (b) The process of calculating f_{em}.

Euclidean space.

$$\mathscr{L}_e = MSE(D_e(concat(f_s, matmul(S_e,\ M_e))),\ x_t). \tag{2}$$

The Euclidean memory network M_e learns and stores the spatial features of the normal data provided as input. This memory network captures common patterns of objects or entities appearing in localized regions with a fixed background, positioning similar features close to each other in the latent space [2,23]. For instance, given N different video sequence samples, each sample may contain different objects or entities in the same localized region, leading to distinct features. As illustrated in the left side of Fig. 3, f_{em} comprises d feature maps of size $w \times h$. The features in the first column of the first row tend to cluster in similar locations, and the features across the columns of the first row exhibit similar patterns. This method allows the model to effectively learn spatially similar features, enhancing its ability to detect consistent patterns in the video sequence [1,2,23].

3.2. Learning in CCM space

The proposed method is inspired by existing studies [11,19], which emphasize that non-Euclidean spaces with negative curvature are highly effective for learning nonlinear data structures. While Euclidean space provides strong visual feature representations, it has limitations in encoding feature dependencies, such as temporal relationships. In contrast, CCM space effectively characterizes feature interactions but lacks rich visual encoding. By leveraging both spaces, we fuse spatial and temporal representations to reconstruct frames, enhancing anomaly detection [11,16].

Building on this foundation, we utilize CCM spaces to model nonlinear temporal patterns and represent hierarchical relationships inherent in video data. Video sequences inherently exhibit nonlinear temporal variations, and CCM spaces provide a suitable structure to capture such complexities. Specifically, we calculate the relationships between consecutive frames using the distance metric in CCM spaces,

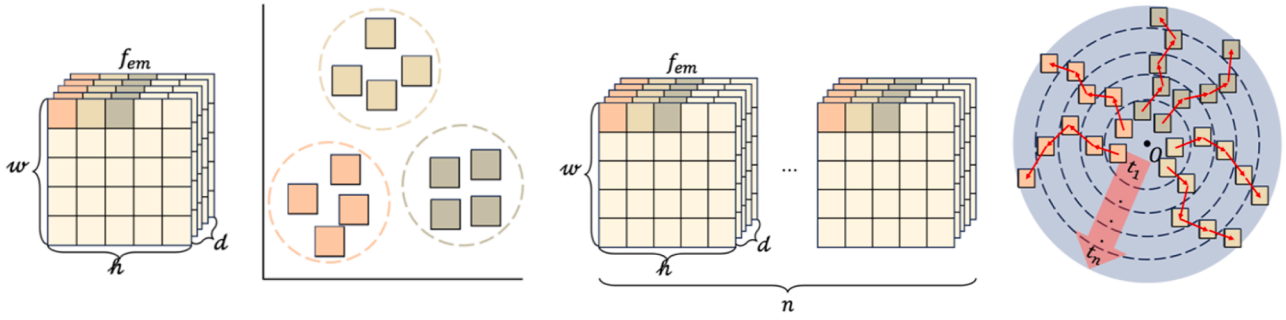

Fig. 3. Visualization of memory network operations in different spaces of the proposed method. The first two subfigures show spatial feature maps (f_{em}) extracted from video frames and their distinction using a Euclidean autoencoder, which clusters features based on spatial regions. The last two subfigures illustrate temporal feature maps across n frames and their organization in the CCM space, where concentric circles represent hierarchical temporal layers ($t_1, \ldots, t_n$), with features near the origin (t_1) capturing minor variations and those farther away (t_n) reflecting larger temporal changes.

ensuring that similar frames are positioned closer together while dissimilar frames are placed farther apart. This method aligns with the intuitive fact that the current frame is more similar to the first subsequent frame than to the second. Repeated learning in CCM spaces takes advantage of their exponentially increasing distances, enabling more discriminative representations compared to Euclidean spaces [13,17]. Furthermore, the unique properties of CCM naturally divide hierarchical structures, allowing the model to learn frame-to-frame relationships and enhance the distinction between normal and abnormal patterns.

For learning in CCM space, the architecture utilizes E_s, D_c, and a memory network M_c with dimension $m \times d$. The encoder E_s processes consecutive video frames $x_{t-n:t-1}$ to generate a latent vector f_s, which is then mapped into the CCM using exponential mapping, leveraging CCM's capability to handle non-linear data distributions and capture complex temporal relationships [11,20]. In the CCM, which has negative curvature, points are restricted to a unit ball defined by:

$$\mathbb{P}^n = \{x \in \mathbb{R}^n : \ \| x\|^2 < 1\}, \tag{3}$$

where $\| x\|^2$ is the squared Euclidean norm, limiting points to lie within a unit ball. The proposed method leverages the CCM memory network to learn hierarchical temporal features, utilizing Eq. (4) for CCM distance computation. In this context, x and y represent two distinct feature vectors within the CCM, with x typically corresponding to a latent feature derived from $x_{t-n:t-1}$, and y representing another feature vector in the memory network. Eq. (4) computes the geodesic distance between x and y, a critical metric for capturing deviations that may indicate anomalies:

$$d_{\mathbb{P}^n}(x,y) = \cosh^{-1}\left(1 + \frac{2\| x-y\|^2}{(1-\| x\|^2)(1-\| y\|^2)}\right). \tag{4}$$

By exploiting the CCM's negative curvature, distances increase exponentially as points move further from the origin [16,19,37]. This property enhances the model's sensitivity to abnormal patterns, as it amplifies the separation between normal and anomalous features. For example, temporal changes in $x_{t-n:t-1}$ that align with normal patterns result in smaller distances, whereas deviations from these patterns yield significantly larger distances.

Additionally, the hierarchical temporal structure arises naturally in the CCM due to the Poincaré disk's geometric properties. As time progresses, feature vectors derived from consecutive frames ($x_{t-n:t-1}$) move further from the origin, representing accumulated temporal changes. The exponential growth of distances in the CCM allows for small temporal variations in earlier frames to be represented near the origin, while more significant temporal deviations in later frames are projected farther outward. This property enables the CCM to effectively capture the hierarchical progression of temporal features, as normal patterns remain within concentric layers near the origin, and anomalous patterns deviate further outward.

The Möbius gyrovector space within the CCM enables operations like Möbius addition in Eq. (5), preserving the data's geometric properties when projecting features onto the CCM:

$$x \oplus y := \frac{(1 + 2\langle x,y\rangle + \| y\|^2)x + (1-\| x\|^2)y}{1 + 2\langle x,y\rangle + \| x\|^2\| y\|^2}. \tag{5}$$

The exponential map $\exp_x(v)$ projects the latent vector f_s onto the manifold, embedding it in the CCM space while preserving its curvature properties:

$$\exp_x(v) = x \oplus \left(\frac{\tanh\left(\frac{\| v\|}{2}\right)v}{\| v\|}\right). \tag{6}$$

Conversely, the logarithmic map $\log_x(y)$ maps point back to the tangent space, allowing the model to extract geodesic distance-related features effectively. The logarithmic map is defined as:

$$\log_x(y) = 2\tanh^{-1}(\| u\|)\frac{u}{\| u\|}, \tag{7}$$

where u is a tangent vector representing the difference between x and y in the tangent space.

The hierarchical structure is further strengthened by the concentric organization of temporal features within the CCM. For instance, in Fig. 3, concentric circles represent temporal layers corresponding to different time step $(t-n,\ t-n+1,\ \ldots,\ t-1)$. Features closer to the origin capture earlier frames with minor temporal variations, while features further outward capture more significant changes in later frames. In Fig. 3, n represents the time dimension. The first row's first column, second column, and third column represent different spatial regions. In the proposed method, the distance computation is performed on the same spatial region across n time frames (e.g., the first column of the first row at $n=0,\ n=1,\ \ldots,\ n=n$). This is because the frames at $n=0$ and $n=1$ are visually more similar, and the difference between these consecutive frames is smaller compared to the frame at $n=n$. As a result, $n=0$ and $n=1$ are positioned closer in the space, while $n=n$ is positioned further away. This structure enables the CCM to model hierarchical temporal dynamics effectively, ensuring that normal patterns cluster within specific layers while anomalous patterns deviate significantly across the manifold.

By leveraging the exponential growth of distances in CCM, the memory network M_c stores the semantic temporal features of normal data, ensuring that similar features are positioned close together while those representing different temporal variations are spaced further apart. This spatial arrangement within the CCM facilitates the accurate modeling of semantic temporal hierarchical features across consecutive frames, as illustrated in Fig. 3, where concentric circles represent different temporal layers within the manifold. This organization allows the model to capture not only the spatial but also the hierarchical temporal relationships of objects in video sequences, enhancing the ability to identify anomalies in complex scenes [2,14].

Leveraging the unique properties of CCM along with the memory network enables the extraction of semantic temporal hierarchical features from consecutive video frames [11,16,19]. The latent vector f_C, which is mapped onto the CCM, is fed into the memory network M_c to calculate the distance between f_c and M_c in the CCM [16,36]. This distance is measured by reshaping f_c into dimension $\left(m \times h \times\ n,\ \frac{d}{n}\right)$, capturing the temporal features across n consecutive frames within the entire feature map. The reshaped f_c is then multiplied by M_c using matrix multiplication, resulting in a similarity matrix S_c of size $(mxh,\ n,\ m)$. Following this, another matrix multiplication between the similarity matrix S_c and the memory network M_c is conducted to compute a weighted sum, producing f_{cm}, which retains the same dimension as f_c [15,20]. The distance measuring method employed aligns with the distance metrics defined for CCM, as denoted in Eq. (4). This method effectively captures both the temporal and spatial characteristics of the video sequence, facilitating the accurate modeling of semantic temporal hierarchical features across consecutive frames.

Subsequently, the feature vector f_{cm} derived from the memory network is concatenated with f_c and fed into the decoder D_c. This process yields the prediction $\hat{x}_t^c$ for the input x_t. The mean squared error is then calculated between x_t and $\hat{x}_t^c$ to measure the difference between the predicted output and the actual target. Eq. (8) specifies the loss function used for training in the CCM space.

$$\mathscr{L}_c = MSE(D_c(concat(f_c, matmul(S_c,\ M_c))),\ x_t). \tag{8}$$

The CCM memory network M_c learns and stores the semantic temporal features of the normal data provided as input. This memory network captures the temporal common patterns of objects or entities appearing in localized regions, ensuring that similar features are positioned close to each other in the latent space [2,14]. For instance, given

N different video sequence samples, each sample captures the temporal changes of objects or entities within the same localized region across n consecutive frames, thereby learning the temporal characteristics. The fundamental operation of the CCM memory network is similar to that of the Euclidean memory network, except that it measures distances within the CCM instead of using cosine distance [19,20]. Fig. 3 illustrates an example of learning temporal features. The right side of the figure depicts the latent vectors positioned within the CCM. The concentric circles represent different temporal layers, illustrating how features from various time steps $(n_1, n_2, \ldots, n_5)$ occupy specific regions within the manifold. Similar temporal features are located closely together within the CCM, while features from different time steps, representing temporal variations, are spaced further apart [20,31]. This spatial distribution within the CCM effectively captures the temporal dynamics of the video data and represents semantic hierarchical relationships.

By employing the CCM memory network, the model learns and stores both semantic hierarchical structures and temporal patterns efficiently. This capability enhances the model's ability to understand and process complex video sequences, making it highly effective for tasks such as VAD and recognition [2,15,20]. By capturing the underlying semantics and hierarchical relationships in the data, this method provides a robust method for deep learning models to understand the intricate temporal and spatial characteristics inherent in video sequence [11,19,20].

3.3. Detecting anomaly with anomality score

To assess the degree of normality or abnormality in a video frame during the testing phase, we operate under the premise that queries extracted from a normal video frame will closely match the memory items, which encapsulate typical patterns of normal behavior [1,2]. We measure the L2 distance between each query and its nearest memory item using the following formula:

$$\mathscr{D}_e(f_e, matmul(S_e,\ M_e)) = \frac{1}{K}\sum_{k}^{K} \|f_e^k - (matmul(S_e,\ M_e))^k\|^2. \tag{9}$$

$$\mathscr{D}_c(f_c, matmul(S_c,\ M_c)) = \frac{1}{K}\sum_{k}^{K} d\left(f_c^k,\ (matmul(S_c,\ M_c))^k\right). \tag{10}$$

We utilize the memory items to implicitly evaluate the abnormality score by examining how well the video frame is reconstructed using these memory items. The assumption here is that abnormal patterns in a video frame will not be adequately reconstructed by the memory items. In line with the method detailed in [3], we compute the peak signal-to-noise ratio (PSNR) between the original video frame and its reconstructed counterpart:

$$P(x_t, \hat{x}_t) = 10\log_{10}\frac{\max(\hat{x}_t)}{\|x_t - \hat{x}_t\|_2^2/N}, \tag{11}$$

where N is the number of pixels in the video frame. A lower PSNR value suggests that the frame x_t is abnormal, whereas a higher PSNR indicates normality [2,33]. To incorporate temporal context, we compute a temporally smoothed PSNR using a weighted sliding window. For a given frame x_t, the temporally weighted PSNR is defined as:

$$P_t^{weighted} = \frac{\sum_{i=-w}^{w} w_i \cdot P(x_{t+i}, \hat{x}_{t+i})}{\sum_{i=-w}^{w} w_i}, \tag{12}$$

where w is the window size, $w_i = \exp\left(-\frac{i^2}{2\sigma^2}\right)$ is the Gaussian weight for temporal smoothing, and $P(x_{t+i},\ \hat{x}_{t+i})$ represents the PSNR value of the frame x_{t+i}.

We also compute the discrepancy between the outputs from both spaces and the ground truth separately. The resulting discrepancies are then combined using element-wise multiplication. This method amplifies smaller errors to become even smaller and larger errors to become more definite, thereby enabling more robust anomaly detection [2,33]. Consistent with the methodology outlined in [2], each error metric in Eqs. (9), (10) and (11) is normalized to a [0, 1] range using min-max normalization [23]. The final abnormality score SA for each video frame is determined by combining these two metrics:

$$SA = P_t^{weighted}(x_t, \hat{x}_t^e) \otimes P_t^{weighted}(x_t, \hat{x}_t^c) + \mathscr{D}_e(f_e, matmul(S_e,\ M_e)) + \mathscr{D}_c(f_c, matmul(S_c,\ M_c)). \tag{13}$$

To sum up, Fig. 4 shows the algorithm of the proposed method. This method exploits both Euclidean and CCM-based features to enhance the detection of anomalies by effectively capturing both spatial and temporal patterns in video sequence [11,26].

3.4. Synergistic integration of euclidean and CCM streams

The rationale for adopting a dual-stream architecture—spatial learning in Euclidean space and temporal learning in CCM space—stems from the observation that anomalies in video data rarely manifest in a single domain. Euclidean space excels at capturing local appearance and structural consistency within frames, making it well-suited for detecting spatial anomalies such as unusual objects or unexpected locations. Conversely, the CCM stream models nonlinear temporal variations across consecutive frames, allowing the system to identify behavioral anomalies or deviations in motion trajectories. While these components perform independently, their integration yields a synergistic effect in which spatial and temporal inconsistencies reinforce each other during anomaly scoring.

To illustrate this synergy, consider a scenario in which a person walking in a typical surveillance scene suddenly appears riding a bicycle. Visually, the presence of a bicycle constitutes a spatial anomaly that the Euclidean stream can detect. However, the abrupt change in motion pattern is better captured in the CCM stream, which tracks hierarchical temporal inconsistencies. By fusing the reconstruction errors from both streams via element-wise multiplication, the model amplifies the combined deviation, resulting in highly sensitive and discriminative anomaly detection. This fusion enables the detection of subtle or compound anomalies that may go unnoticed when relying on a single modality.

Moreover, as shown in our ablation study (Table 5 in Experiments), the combined model consistently outperforms its single-stream counterparts across all benchmark datasets. Notably, the dual-stream approach achieves a 3∼5 % improvement in AUC over Euclidean-only and CCM-only models, particularly excelling in the complex ShanghaiTech Campus dataset. These empirical results highlight that the proposed architecture is not merely beneficial but necessary. In real-world scenarios where anomalies exhibit multifaceted irregularities, a uni-modal detection scheme—whether spatial or temporal—is fundamentally limited. The proposed dual-stream structure reflects an architectural imperative to comprehensively model the complex interplay of appearance and dynamics inherent in video data.

4. Experiments

4.1. Dataset

We evaluate the proposed method on three pedestrian video datasets in Table 2: UCSD Ped2 [38], CUHK Avenue [39], and Shanghaitech Campus [8]. Following the standard preprocessing, all images are resized to 256 × 256 and normalized to values between −1 and 1. The history length n is set to 4, and the memory size m is set to 10, in line with the frame-pred strategy. The training process spans 60 epochs for UCSD Ped2 and CUHK Avenue datasets, while for the Shanghaitech Campus dataset, training is conducted over 10 epochs. All experiments utilize the Adam optimizer with a learning rate of 2×10^{-4}, which is decayed following the CosineAnnealingLR strategy.

Algorithm 1. Video anomaly detection using two-stream autoencoders and memory networks

Input: Consecutive video frames $x_{t-n:t-1}$
Parameters for Euclidean and CCM-based autoencoders
Output: Anomaly score S_t for each video frame x_t
Step 1: Feature extraction
Input $x_{t-n:t-1}$ into shared encoder E_s
Obtain latent vector f_s
Step 2: Learning in Euclidean space
Normalize f_s to form f_e
Input f_e into Euclidean memory network M_e
Measure cosine similarity between f_e and M_e
Compute weighted sum to get f_{em}
Concatenate f_{em} with f_e
Input into decoder D_e to reconstruct x_t^e
Step 3: Learning in CCM space
Map f_s to CCM using exponential mapping to get f_c
Input f_c into CCM memory network M_c
Measure CCM-specific distance between f_c and M_c
Compute weighted sum to get f_{cm}
Concatenate f_{cm} with f_c
Input into decoder D_c to reconstruct x_t^c
Step 4: Anomaly detection
Compute PSNR between x_t and x_t^e
Compute PSNR between x_t and x_t^c
Compute disparity between f_e and $matmul(S_e, M_e)$
Compute disparity between f_c and $matmul(S_c, M_c)$
Compute anomaly score SA

Fig. 4. The proposed video anomaly detection algorithm.

Table 2
The details of video anomaly detection benchmark datasets.

Dataset	Total (frames/videos)	Train (frames/videos)	Test (frames/videos)	Anomalies
Ped2	4560/28	2550/16	2010/12	skater, biker, cart
Avenue	30,652/37	15,328/16	15,324/21	object throwing, loitering, running
Shanghaitech Campus	317,398/437	274,515/330	42,883/107	chasing, fighting, loitering, running

Table 3
The comparison of AUC (%) scores with unsupervised VAD methods.

Task	Method	Dataset		
		Ped2	Avenue	Shanghaitech Campus
Reconstruction	MAAM [23]	97.7	90.9	71.3
	MemAE [1]	94.1	83.3	71.2
	ITAE [30]	99.2	88.0	76.3
	StackRNN [8]	92.2	81.7	68.0
	VEC [35]	97.3	90.2	74.8
Frame Prediction	AMMC [26]	96.6	85.6	70.3
	AnoPCN [9]	96.8	86.2	73.6
	CDAE [3]	96.5	86.8	73.3
	CRC [29]	98.7	92.5	78.3
	Dual GroupGAN [10]	96.6	85.5	73.1
	FastAno++ [22]	98.1	87.8	75.2
	FFP [24]	95.4	88.5	72.8
	HF^2-VAD [40]	**99.3**	91.1	76.2
	LGC—Net [28]	97.1	89.3	73.0
	MNAD [2]	97.0	88.5	70.5
	MSN—Net [33]	97.6	89.4	73.4
	SAF3D [41]	96.7	84.7	74.8
	SAFA [34]	96.8	87.3	76.4
	TransMem [42]	98.1	88.5	72.5
	Unmasking [21]	82.2	80.6	68.3
	VAD-CL [43]	92.2	86.2	73.8
	Ours	**99.3**	**92.8**	**80.5**

Specifically, the UCSD Ped2 dataset consists of 16 video sequences for training and 12 sequences for testing, capturing various anomaly events [38]. The CUHK Avenue dataset comprises 16 training videos and 21 testing videos that include different types of anomalous activities [39]. The Shanghaitech Campus dataset, which covers a broader set of scenarios, includes 330 videos for training and 107 videos for testing [8]. Evaluation of the proposed method is conducted using AUC-ROC and TPR-FPR metrics, adhering to the experimental protocols outlined in [1,2,29].

4.2. Quantitative results

Table 3 and Fig. 5 highlight the superior performance of the proposed method in unsupervised VAD, achieving the highest AUC scores across the Ped2 (99.3 %), Avenue (92.8 %), and Shanghaitech Campus (80.5 %) datasets. This consistent high performance underscores the method's ability to effectively capture and learn temporal patterns, which are crucial for identifying anomalies in dynamic video scenes. The proposed method outperforms other competitive methods such as AMMC [26], CRC [29], FastAno++ [22], and HF^2-VAD [40]. which explicitly use optical flow to model temporal features. This implies that the proposed method extracts spatial and temporal features more effectively, resulting in better anomaly detection. Its ability to generalize well across different datasets indicates that it not only captures the

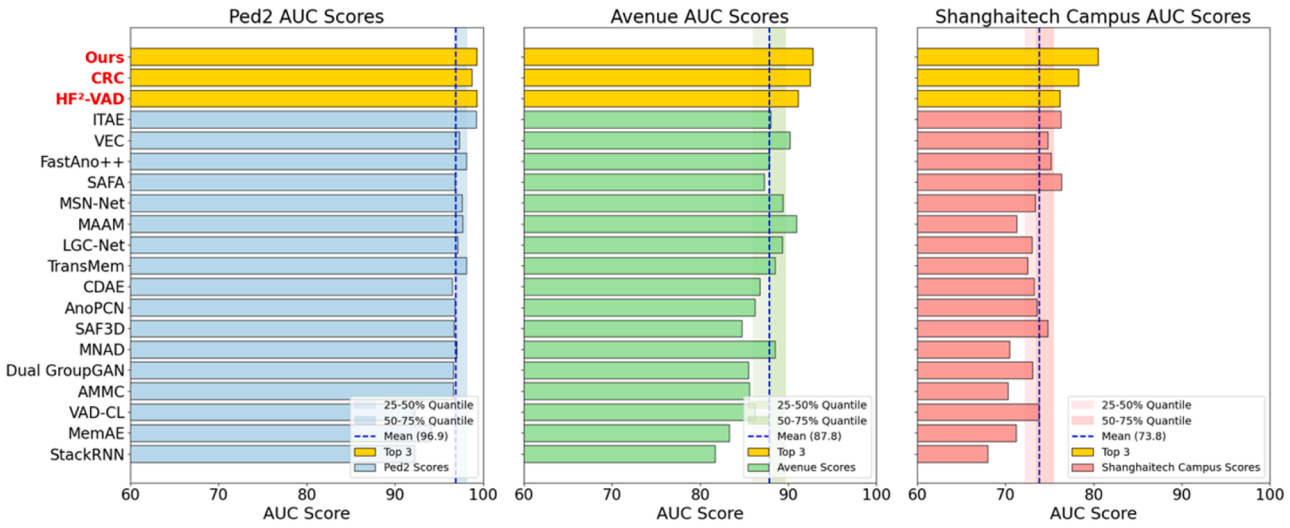

Fig. 5. AUC scores of various unsupervised video anomaly detection methods across different datasets.

immediate spatial anomalies but also understands complex temporal dynamics, enabling robust and accurate anomaly detection across diverse video environments, as confirmed by the results in Table 3 and Fig. 5 [2,3].

Fig. 6 and 7 provide a detailed comparative analysis of the anomaly detection capabilities of four models: the proposed model, AMMC [26], HF^2-VAD [40], and MNAD [2], using test videos from two benchmark datasets, UCSD Ped2 and CUHK Avenue. The performance is measured using normality scores plotted against the frame sequence, with shaded red regions indicating the ground truth anomaly intervals.

Fig. 6 focuses on the UCSD Ped2 dataset, specifically test videos #2 and #3. The proposed model demonstrates robust anomaly detection, characterized by a sharp and immediate decline in the normality score at the onset of the anomaly in test video #2, maintaining a low score throughout the anomaly interval. This behavior signifies high sensitivity and specificity, indicating effective temporal modeling and feature extraction capabilities. In test video #3, the model exhibits a significant increase in normality score toward the end of the sequence, showcasing its ability to adapt to varying anomaly types and providing strong temporal consistency. These clear patterns of response demonstrate the model's reliability in distinguishing between normal and abnormal events. In contrast, the AMMC, HF^2, and MNAD models display more variability in their normality scores during the anomaly intervals. AMMC shows moderate fluctuation, indicating less stable anomaly detection and potential susceptibility to noise, which could lead to higher false positive rates. HF^2 and MNAD exhibit even greater fluctuations, showing difficulties in maintaining consistent detection accuracy. These fluctuations imply that HF^2 and

MNAD may have limitations in effectively capturing complex temporal dependencies or differentiating subtle anomalies from normal behavior, potentially compromising their performance in real-world scenarios.

Fig. 7 extends the analysis to the CUHK Avenue dataset, examining test videos #4 and #12. The proposed method outperforms the others again, demonstrating precise anomaly detection. In test video #4, the model maintains high normality scores during normal frames and shows a sharp decline during the anomaly interval, quickly returning to high scores post-anomaly. This consistent behavior underscores the model's robustness in anomaly detection and quick recovery, demonstrating effective handling of spatial and temporal features. In test video #12, characterized by multiple anomalies, the proposed method accurately identifies each anomaly, with distinct drops in normality score, reflecting its ability to handle complex anomaly patterns and maintain detection reliability under varied conditions.

AMMC, HF^2, and MNAD, however, exhibit less consistent performance in Fig. 7. AMMC displays irregular dips in normality score even outside designated anomaly intervals, pointing to potential false positives. HF^2 and MNAD show significant variability, particularly in test video #12, where their normality scores fluctuate considerably, indicating instability and less effective anomaly detection. This variability may result from inadequate temporal modeling, feature extraction, and generalization capabilities, which are crucial for handling diverse anomaly scenarios [1,2,40].

4.3. Qualitative results

Fig. 8 and 9 present the anomaly detection capabilities of the proposed method for the ShanghaiTech Campus [8] and CUHK Avenue [39] datasets, respectively. These figures demonstrate how the method processes video frames to detect anomalies by showing the progression from raw input frames through various stages of feature extraction to the final anomaly-highlighted output. The first column of each figure displays raw video frames, establishing a baseline for normal behavior. The subsequent columns show: (1) reconstruction results from the Euclidean autoencoder, (2) the reconstruction error between the ground truth frames and the Euclidean autoencoder's output, (3) reconstruction results from the CCM autoencoder, (4) the reconstruction error between the ground truth frames and the CCM autoencoder's output, and (5) the element-wise multiplication of the Euclidean branch's reconstruction error (3rd column) and the CCM branchr's reconstruction error (5th column). This method allows the model to effectively filter out irrelevant information and concentrate on potential anomalies. The final column highlights the areas of high activation, indicating detected anomalies. The model successfully identifies significant deviations, such as unusual movements or behaviors, emphasizing its capability to detect both subtle and overt anomalies. This indicates that the model's feature extraction and attention mechanisms are well-tuned for distinguishing abnormal activities [23,26]. The consistent performance across both datasets underscores the robustness and adaptability of the proposed model. By effectively isolating and emphasizing anomalous features, the model demonstrates strong potential for real-world applications in surveillance and monitoring systems.

4.4. Ablation study

The ablation study examines the effect of memory size on the performance of the proposed method across three datasets. The results as shown in Table 4 indicate that the model's performance is relatively

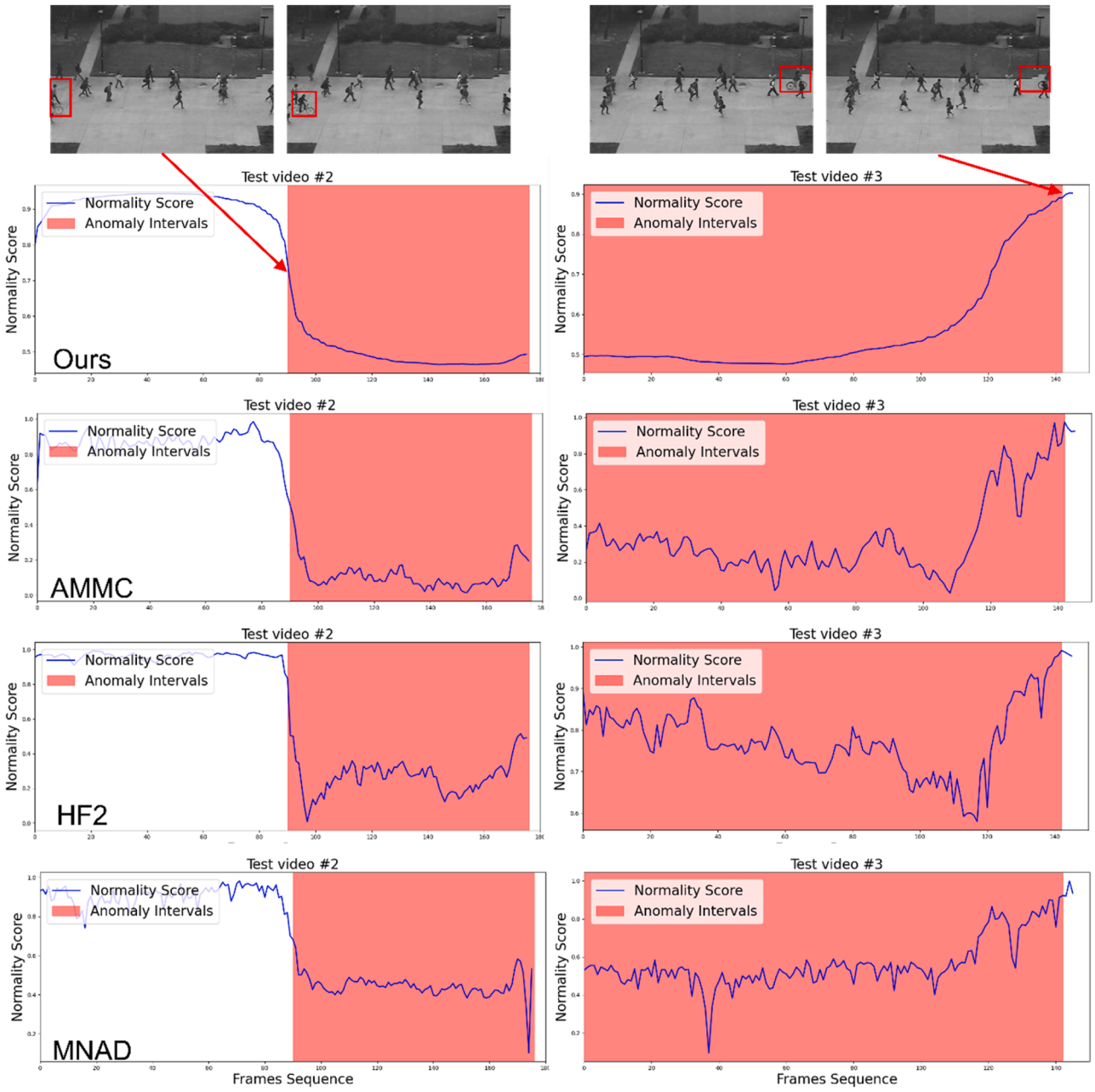

Fig. 6. Comparative analysis of anomaly detection models on UCSD Ped2.

robust to the changes in memory size, particularly for the UCSD Ped2 and CUHK Avenue datasets. For the UCSD Ped2 dataset, performance remains consistently high, around 99.1 % to 99.3 %, across all tested memory sizes (10, 20, 30, and 50). This stability implies that the model is highly effective at anomaly detection, irrespective of the memory size. The CUHK Avenue dataset shows slight variations in performance, with accuracy ranging from 92.1 % to 92.8 %. While there is a minor dip in performance at a memory size of 20, the model quickly recovers, indicating that memory size has only a limited impact on its detection capability.

On the other hand, the results on ShanghaiTech Campus dataset demonstrate more noticeable sensitivity to the changes in memory size, with performance varying between 79.7 % and 80.5 %. The lower overall performance compared to the other two datasets implies that the ShanghaiTech Campus presents more challenging anomaly detection scenarios, potentially due to more complex or subtle anomalies. For this dataset the performance dip at a memory size of 20 indicates that memory configuration might play a more significant role in detection performance of the model [1,2,26].

Overall, this paper reveals that while the proposed method is generally robust to variations in memory size, selecting an optimal memory size such as 50, could help maintain or slightly enhance the performance. This flexibility and robustness make the model suitable for real-world applications where memory constraints may vary, ensuring consistently high anomaly detection performance.

For UCSD Ped2, the combined method demonstrates a significant advantage over both Euclidean-only and CCM-only methods, as shown in Table 5. While the CCM-only model is more effective at capturing temporal irregularities, the fusion of spatial and temporal features provides a substantial performance boost. This integrated method increases sensitivity to subtle anomalies, particularly benefiting from the simpler anomaly types present in this dataset. In the case of CUHK Avenue, the CCM-only method outperforms its Euclidean counterpart, further confirming the importance of modeling temporal dynamics in

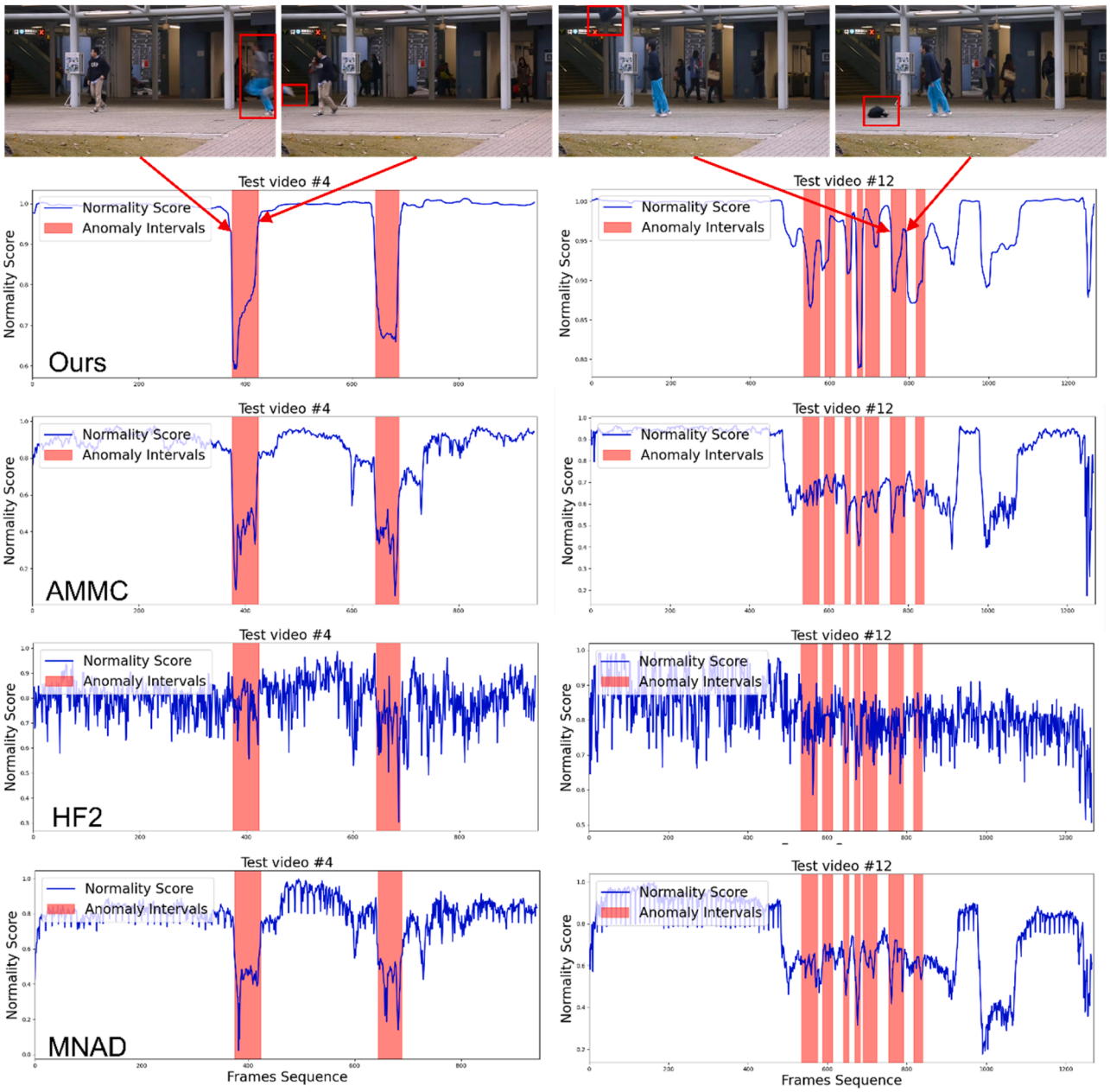

Fig. 7. Comparative analysis of anomaly detection models on CUHK Avenue.

datasets where motion-based anomalies dominate, as reflected in Table 5. Nevertheless, the combined method, utilizing both spatial and temporal features, delivers the best results. By leveraging complementary information from both domains, the fusion leads to a more robust and accurate anomaly detection performance. For the more challenging ShanghaiTech Campus dataset, which involves complex and diverse anomaly types, the CCM-only method again surpasses the Euclidean-only model, underscoring the necessity of temporal modeling. However, as demonstrated in Table 5, the highest performance is achieved through the combined use of both spatial and temporal features. This dual-geometrical method efficiently handles the intricate spatiotemporal patterns in the dataset, resulting in superior detection accuracy across a wide range of scenarios.

4.5. Inference time analysis

Table 6 presents a comparative analysis of frame-per-second (FPS) performance with state-of-the-art methods for video anomaly detection. The methods are categorized into reconstruction-based and frame prediction-based approaches. In general, frame prediction methods demonstrate faster inference speeds compared to their reconstruction-based counterparts.

Reconstruction-based methods are relatively slower, with MAAM achieving 51 FPS, MemAE operating at 38 FPS, and StackRNN at just 10 FPS. In contrast, frame prediction methods exhibit substantially improved efficiency—MSN—Net and SAFA reach 78 FPS and 76 FPS respectively, while TransMem and MAND achieve 72 FPS and 65 FPS. FastAno++ and Dual GroupGAN also report faster performance with 60 FPS and 57 FPS respectively.

The proposed method attains 64 FPS, positioning it as a competitive solution among fast-performing frame prediction methods. Its inference speed is comparable to MAND (65 FPS) and TransMem (72 FPS), and only slightly behind the fastest models like MSN—Net and SAFA. Furthermore, it significantly outperforms slower baselines such as

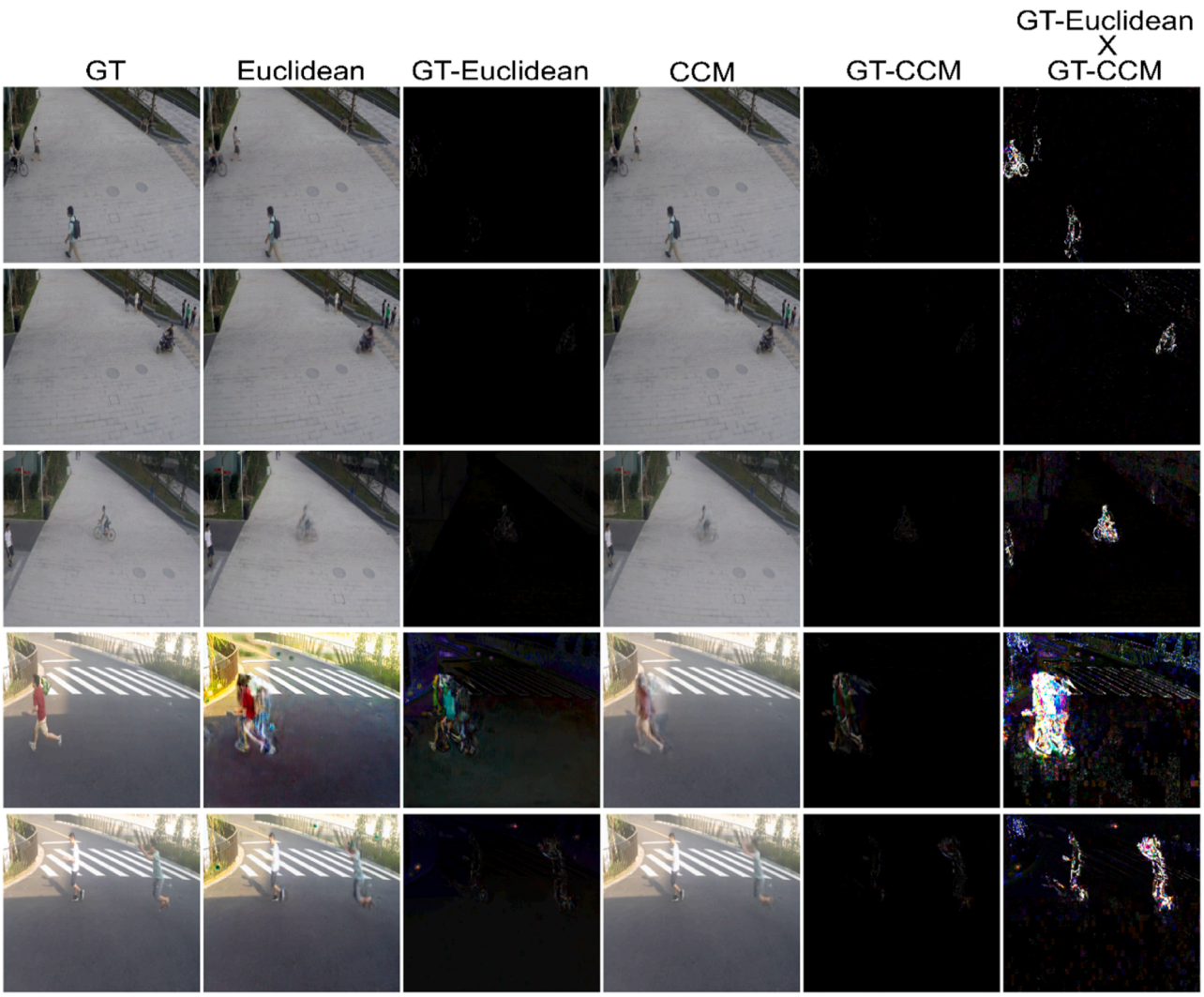

Fig. 8. Visualization of the ouputs on the Shanghaitech Campus dataset.

Unmasking (20 FPS) and HF^2-VAD (10 FPS).

In addition to runtime efficiency, the proposed method also achieves strong detection performance, reporting 92.8 % on the Avenue dataset and 80.5 % on the ShanghaiTech dataset. These results highlight its ability to strike a balance between speed and accuracy, making it a suitable candidate for real-time video anomaly detection applications.

4.6. Generalization ability on large-scale dataset

To evaluate the generalization capability of the proposed method beyond conventional benchmarks, we conduct additional experiments on the NWPU Campus dataset [44] which is the largest dataset for unsupervised video anomaly detection to date. This dataset contains 43 distinct scenes, 28 types of anomalous events, and approximately 1.47 million frames, making it over 4.6 times larger in frame count and over 3 times richer in scene diversity compared to the ShanghaiTech Campus dataset (317,398 frames, 13 scenes).

Notably, NWPU Campus is the only dataset that incorporates scene-dependent anomalies, where certain events are abnormal in specific contexts but not in others—for instance, cycling on a footpath is anomalous only in certain scenes. In contrast, datasets like ShanghaiTech treat object appearance (e.g., a bicycle) as anomalous regardless of context, making early prediction infeasible. NWPU also supports anomaly anticipation by ensuring observable build-up to anomalous events.

As shown in Table 7, our method achieves an AUC of 68.2 % on NWPU Campus, outperforming both reconstruction-based (e.g., MemAE: 61.9 %) and prediction-based methods (e.g., AMMC: 64.5 %). This demonstrates that our method not only generalizes well to large-scale, diverse environments, but is also robust to complex and scene-dependent anomalies. These results strongly support the model's scalability and real-world applicability.

5. Concluding remarks

In this paper, we propose a two-stream autoencoder model to address the limitations of single model trained exclusively in Euclidean space. Our method integrates learning in both Euclidean and CCM spaces,

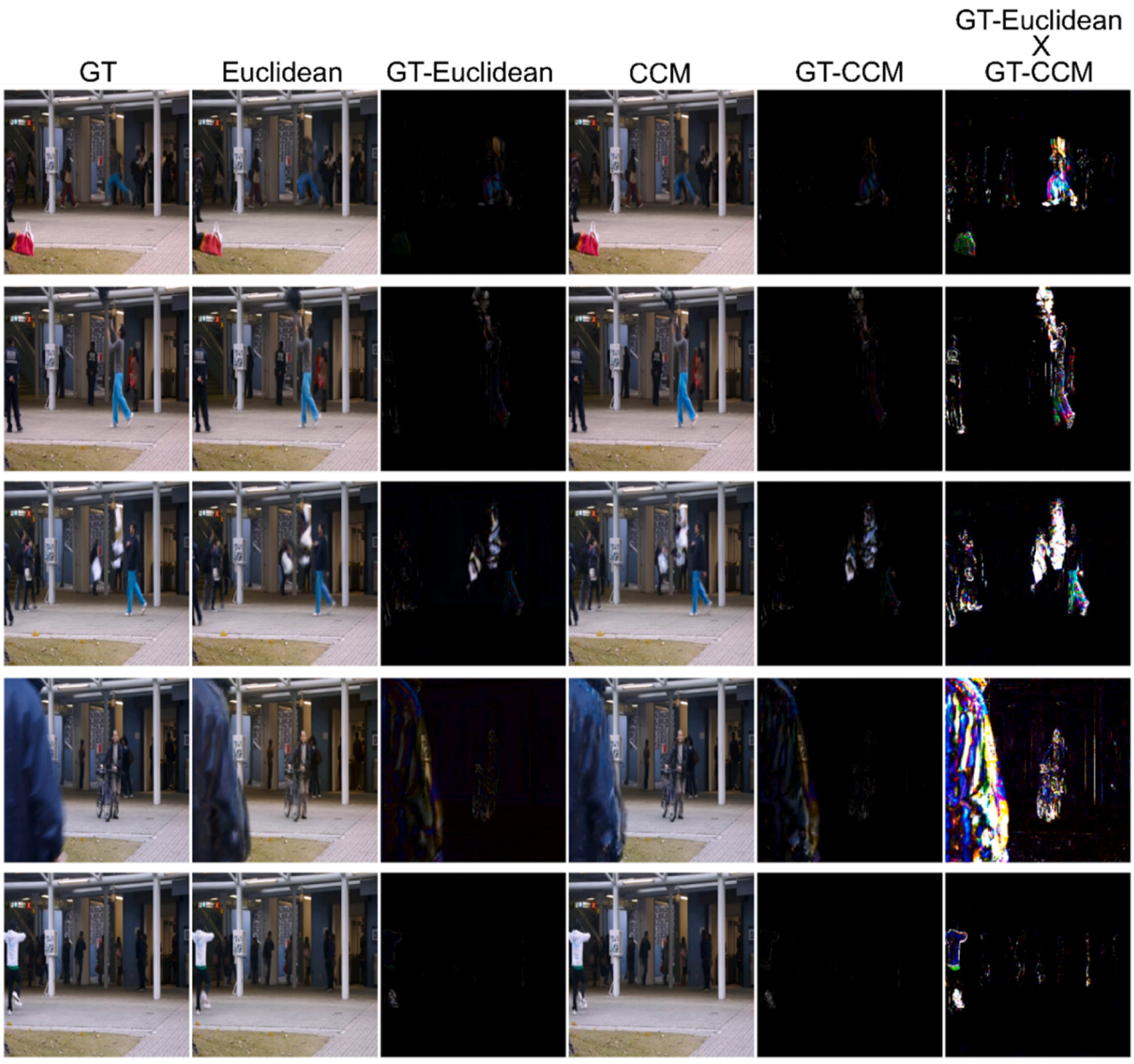

Fig. 9. Visualization of the outputs on the Avenue dataset.

Table 4
Performance impact of varying memory sizes.

Memory Size	UCSD Ped2	CUHK Avenue	ShanghaiTech Campus
10	99.3	92.8	80.5
20	99.3	92.1	79.7
30	99.1	92.4	80.1
50	99.2	92.2	80.2

Table 5
Ablation study comparing the performance (AUC %) of Euclidean-only, CCM-only, and the combined Euclidean-CCM methods across three benchmark datasets (UCSD Ped2, CUHK Avenue, ShanghaiTech Campus). The combined method consistently yields the highest performance, demonstrating the benefit of fusing spatial and temporal features.

Dataset	Euclidean-only	CCM-only	Combined
UCSD Ped2	95.2	96.1	99.3
CUHK Avenue	87.5	89.0	92.8
ShanghaiTech Campus	75.2	77.5	80.5

which leverages CCM space learning effectively captures nonlinear temporal features in video data, extracting the temporal structures between consecutive frames. This results in higher discrepancy of original and reconstructed images for anomalies, while the Euclidean autoencoder minimizes the discrepancy for normal data. By amplifying the discrepancy associated with anomalies with distance measurement in CCM space, our method distinctly separates normal instances from abnormal ones. Comparative study with several unsupervised learning-based autoencoder models confirms the superior performance of the proposed method, and deeper analysis shows that the CCM autoencoder captures nonlinear temporal features effectively and amplifies the discrepancy significantly. Its ability to reliably detect and highlight anomalies makes it a valuable tool for environments where accurate anomaly detection is essential.

However, the performance of the proposed method on large-scale and diverse datasets, such as ShanghaiTech Campus and NWPU Campus, shows some limitations. Specifically, the proposed method struggles with context-dependent anomalies, the events that are considered abnormal only in specific scenarios. This issue highlights the need for more sophisticated scene-aware modeling and the

Table 6
Comparison of frame per second (FPS) with the state of the art methods.

Task	Method	FPS (s)	Task	Method	FPS (s)
Reconstruction	MAAM [23]	51	Frame prediction	FFP [24]	25
	MemAE [1]	38		HF^2-VAD [40]	10
	ITAE [30]	19		LGC—Net [28]	19
	StackRNN [8]	10		MAND [2]	65
Frame prediction	AMMC [26]	18		MSN—Net [33]	78
	AnoPCN [9]	10		SAF3D [41]	47
	CDAE [3]	32		SAFA [34]	76
	CRC [29]	32		TransMem [42]	72
	Dual GroupGAN [10]	57		Unmasking [21]	20
	FastAno++ [22]	60		Ours	64

Table 7
AUC performance comparison on the NWPU Campus dataset.

Task	Method	AUC (%)
Reconstruction	MemAE [1]	61.9
Frame prediction	MNAD [2]	62.5
	HF^2-VAD [40]	63.7
	AMMC [26]	64.5
	Ours	68.2

incorporation of contextual understanding to improve the model's generalization capabilities. Therefore, equipping the model with the ability to distinguish between scene-specific normal and abnormal behaviors becomes essential for deployment in complex real-world environments, where the nature of anomalies can vary across different contexts significantly.

In the future, we intend to delve into the development and application of more sophisticated and advanced architectural methods, such as transformer-based models and their variants, to significantly enhance the model's proficiency in capturing intricate and long-range temporal dependencies in video data. By leveraging the attention mechanisms inherent in transformer architectures, we can achieve a more nuanced understanding of temporal patterns and contextual relationships, thereby improving the overall anomaly detection performance. Additionally, we plan to integrate semi-supervised learning techniques, which would allow the model to effectively leverage both labeled and unlabeled data. This method is expected to facilitate the model's adaptability to new and previously unseen types of anomalies, even when only a limited amount of labeled data is available. We anticipate that these improvements will not only boost the model's robustness and generalization capabilities but also enable it to operate more effectively in dynamic and evolving real-world environments where the nature of anomalies can vary over time.

CRediT authorship contribution statement

Jinmyeong Kim: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Sung-Bae Cho:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Sung-Bae Cho reports financial support was provided by Institute of Information & Communications Technology Planning & Evaluation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Yonsei Fellow Program funded by Lee Youn Jae, IITP grant funded by the Korea government (MSIT) (No. RS-2022-II220113, Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework), and Air Force Defense Research Sciences Program funded by Air Force Office of Scientific Research.

Data availability

Data will be made available on request.

References

[1] D. Gong, L. Liu, V. Le, B. Saha, M.R. Mansour, S. Venkatesh, A. van den Hengel, Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection, in: IEEE/CVF Int. Conf. on Computer Vision, 2019, pp. 1705–1714.
[2] H. Park, J. Noh, B. Ham, Learning memory-guided normality for anomaly detection, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020, pp. 14372–14381.
[3] Y. Chang, Z. Tu, W. Xie, J. Yuan, Clustering driven deep autoencoder for video anomaly detection, Eur. Conf. Comput. Vis. 16 (2020) 329–345.
[4] S.J. Bu, S.B. Cho, Deep character-level anomaly detection based on a convolutional autoencoder for zero-day phishing URL detection, Electron. 10 (12) (2021) 1492.
[5] G.B. Jang, S.B. Cho, Multi-instance attention network for anomaly detection from multivariate time series, Cybern. Syst. 55 (6) (2024) 1417–1440.
[6] P. Yan, A. Abdulkadir, P.P. Luley, M. Rosenthal, G.A. Schatte, B.F. Grewe, T. Stadelmann, A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: methods, applications, and directions, IEEE Access. 12 (2024) 3768–3789.
[7] G.B. Jang, S.B. Cho, Anomaly detection for health monitoring of heavy equipment using hierarchical prediction with correlative feature learning, in: Int. Conf. on Soft Computing Models in Industrial and Environmental Applications, 2022, pp. 598–608.
[8] W. Luo, W. Liu, S. Gao, A revisit of sparse coding based anomaly detection in stacked RNN framework, in: IEEE Int. Conf. on Computer Vision, 2017, pp. 341–349.
[9] M. Ye, X. Peng, W. Gan, W. Wu, Y. Qiao, Anopcn: video anomaly detection via deep predictive coding network, ACM Int. Conf. Multimed. (2019) 1805–1813.
[10] Z. Sun, P. Wang, W. Zheng, M. Zhang, Dual GroupGAN: an unsupervised four-competitor (2V2) approach for video anomaly detection, Pattern. Recognit. 153 (2024) 110500.
[11] J. Leng, Z. Wu, M. Tan, Y. Liu, J. Gan, H. Chen, and X. Gao, "Beyond Euclidean: dual-space representation learning for weakly supervised video violence detection," *Advances in Neural Information Processing Systems*, vol. 36.
[12] Y. Zhang, L. Luo, W. Xian, H. Huang, Learning better visual data similarities via new grouplet non-Euclidean embedding, in: IEEE/CVF Int. Conf. on Computer Vision, 2021, pp. 9918–9927.
[13] D. Grattarola, D. Zambon, L. Livi, C. Alippi, Change detection in graph streams by learning graph embeddings on constant-curvature manifolds, IEEE Trans. Neural Netw. Learn. Syst. 31 (6) (2019) 1856–1869.
[14] D. Zambon, L. Livi, C. Alippi, Anomaly and change detection in graph streams through constant-curvature manifold embeddings, in: Int. Joint Conf. on Neural Networks, 2018, pp. 1–7.
[15] H. Sun, L. Wang, L. Zhang, L. Gao, Hyperbolic space-based autoencoder for hyperspectral anomaly detection, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–15.
[16] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: going beyond euclidean data, IEEE Signal. Process. Mag. 34 (4) (2017) 18–42.
[17] V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, V. Lempitsky, Hyperbolic image embeddings, in: IEEE/CVF Int. Conf. on Computer Vision 153, 2020, pp. 6418–6428.

[18] R. Ma, P. Fang, T. Drummond, M. Harandi, Adaptive Poincaré point to set distance for few-shot classification, in: AAAI Conf. on Artificial Intelligence 36, 2022, pp. 1926–1934.
[19] M. Nickel, D. Kiela, Poincaré embeddings for learning hierarchical representations, Adv. Neural Inf. Process. Syst. 30 (2017) 6338–6347.
[20] C. Chadebec, S. Allassonnière, A geometric perspective on variational autoencoders, Adv. Neural Inf. Process. Syst. 35 (2022) 19618–19630.
[21] R.T. Ionescu, S. Smeureanu, B. Alexe, M. Popescu, Unmasking the abnormal events in video, in: IEEE Int. Conf. on Computer Vision, 2017, pp. 2895–2903.
[22] C. Park, D. Kim, M. Cho, M. Kim, M. Lee, S. Park, S. Lee, Fast video anomaly detection via context-aware shortcut exploration and abnormal feature distance learning, Pattern. Recognit. 157 (2025) 110877.
[23] L. Wang, J. Tian, S. Zhou, H. Shi, G. Hua, Memory-augmented appearance-motion network for video anomaly detection, Pattern. Recognit. 138 (2023) 109335.
[24] W. Luo, W. Liu, D. Lian, S. Gao, Future frame prediction network for video anomaly detection, IEEE Trans. Pattern. Anal. Mach. Intell. 44 (11) (2021) 7505–7520.
[25] M.Z. Zaheer, J.H. Lee, M. Astrid, S.I. Lee, Old is gold: redefining the adversarially learned one-class classifier training paradigm, in: IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020, pp. 14183–14193.
[26] R. Cai, H. Zhang, W. Liu, S. Gao, Z. Hao, Appearance-motion memory consistency network for video anomaly detection, AAAI Conf. Artif. Intell. 35 (2021) 938–946.
[27] Y. Liu, D. Yang, Y. Wang, J. Liu, J. Liu, A. Boukerche, L. Song, Generalized video anomaly event detection: systematic taxonomy and comparison of deep models, ACM. Comput. Surv. 56 (7) (2024) 1–38.
[28] M. Zhao, X. Zeng, Y. Liu, J. Liu, C. Pang, Rethinking prediction-based video anomaly detection from local–global normality perspective, Expert. Syst. Appl. 262 (2025) 125581.
[29] Y. Liu, Z. Xia, M. Zhao, D. Wei, Y. Wang, S. Liu, L. Song, Learning causality-inspired representation consistency for video anomaly detection, ACM Int. Conf. Multimed. (2023) 203–212.
[30] M. Cho, T. Kim, W.J. Kim, S. Cho, S. Lee, Unsupervised video anomaly detection via normalizing flows with implicit latent features, Pattern. Recognit. 129 (2022) 108703.
[31] G. Wang, Y. Wang, J. Qin, D. Zhang, X. Bao, D. Huang, Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles, in: European Conf. on Computer Vision, 2022, pp. 494–511.
[32] C. Chen, Y. Xie, S. Lin, A. Yao, G. Jiang, W. Zhang, Y. Qu, R. Qiao, B. Ren, L. Ma, Comprehensive regularization in a bi-directional predictive network for video anomaly detection, in: AAAI Conf. on Artificial Intelligence 36, 2022, pp. 230–238.
[33] Y. Liu, D. Li, W. Zhu, D. Yang, J. Liu, L. Song, MSN-net: multi-scale normality network for video anomaly detection, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2023, pp. 1–5.
[34] Z. Ye, Y. Li, Z. Cui, Y. Liu, L. Li, L. Wang, C. Zhang, Unsupervised video anomaly detection with self-attention based feature aggregating, IEEE 26th Int. Conf. Intell. Transp. Syst (2023) 3551–3556.
[35] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, M. Kloft, Cloze test helps: effective video anomaly detection via learning to complete video events, ACM Int. Conf. Multimed. (2020) 583–591.
[36] H. Shao, A. Kumar, P.T. Fletcher, The riemannian geometry of deep generative models, in: IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2018, pp. 315–323.
[37] S. Cho, J. Lee, D. Kim, Hyperbolic VAE via latent Gaussian distributions, Adv. Neural Inf. Process. Syst. 36 (2024).
[38] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern. Anal. Mach. Intell. 36 (1) (2013) 18–32.
[39] C. Lu, J. Shi, J. Jia, Abnormal event detection at 150 FPS in matlab, in: IEEE Int. Conf. on Computer Vision, 2013, pp. 2720–2727.
[40] Z. Liu, Y. Nie, C. Long, Q. Zhang, G.S. Li, A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction, in: IEEE/CVF Int. Conf. on Computer Vision, 2021, pp. 13588–13597.
[41] A. Niaz, S.U. Amin, S. Soomro, H. Zia, K.N. Choi, Spatially aware fusion in 3D convolutional autoencoders for video anomaly detection, IEEE Access. Vol 12 (2024) 104770–104784.
[42] Z. Wang, X. Gu, X. Gu, J. Hu, Enhancing video anomaly detection with learnable memory network: a new approach to memory-based auto-encoders, Comput. Vis. Image Underst. 241 (2024) 103946.
[43] S. Qiu, J. Ye, J. Zhao, L. He, L. Liu, E. Bicong, X. Huang, Video anomaly detection guided by clustering learning, Pattern. Recognit. 153 (2024) 110550.
[44] C. Cao, Y. Lu, P. Wang, Y. Zhang, A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation, in: IEEE Int. Conf. on Computer Vision, 2023, pp. 20392–20401.