Long Short-Term Dynamic Prototype Alignment Learning for Video Anomaly Detection

Chao Huang[1], **Jie Wen**[2*], **Chengliang Liu**[2] and **Yabo Liu**[2]
[1]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
huangch253@mail.sysu.edu.cn, wenjie@hit.edu.cn, liucl1996@163.com, yaboliu.ug@gmail.com

Abstract

Video anomaly detection (VAD) is the core problem of intelligent video surveillance. Previous methods commonly adopt the unsupervised paradigm of frame reconstruction or prediction. However, the lack of mining of temporal dependent relationships and diversified event patterns within videos limit the performance of existing methods. To tackle these problems, we propose a novel prototype-guided and dynamic-aware long-distance frame prediction paradigm for VAD. Specifically, we develop a prototype-guided dynamics matching network (PDM-Net) to enhance the discriminant and robustness of anomaly detector. To explore the temporal contexts, we equip PDM-Net with a long short-term dynamic prototype alignment learning mechanism, which stores long-term dynamic prototypes into memory bank and learns how to recall long-term dynamic prototypes with short-term dynamics. As a result, the short input sequences can recall long-term dynamic prototypes stored in the memory bank to achieve the task of long-distance frame prediction. Besides, a feature discrimination module is adopted to extract the representative dynamic features of various normal events meanwhile preserving the diversity of normal patterns. Experimental results on four datasets demonstrate the superiority of our method.

1 Introduction

With the ever-growing volumes of surveillance cameras, it is highly demanded how to effectively recognize the abnormal events that may seriously threaten public security in surveillance videos, such as explosion, fighting, crimes, and traffic accidents. It is impractical to promptly detect the abnormal events by watching all surveillance videos, because vast amounts of videos are captured every second. Therefore, we need to develop intelligent video surveillance system to automatically detect abnormal events, where video anomaly detection (VAD) [Huang *et al.*, 2022d; Huang *et al.*, 2022c; Huang *et al.*, 2021] is the core technology.

Generally, current methods can be grouped into weakly-supervised and unsupervised methods according to the manners of model training. Weakly-supervised VAD [Sultani *et al.*, 2018; Lv *et al.*, 2021b; Huang *et al.*, 2022a; Wu *et al.*, 2023b; Zhang *et al.*, 2022] requires both normal and anomalous data to train model. Sultani *et al.* [Sultani *et al.*, 2018] established a weakly-supervised dataset UCF-Crime, which activates this direction. Although remarkable gain has been achieved in prior works [Lv *et al.*, 2021b], weakly-supervised VAD still suffers from two drawbacks: 1) it can only recognize abnormal events included in the training set and cannot detect unseen anomalies; and 2) it still needs to collect amounts of abnormal samples, although time-intensive video temporal annotation is not required. Whereas, unsupervised methods generally formulate VAD as an outlier detection problem. The anomaly detector is trained with only normal data for learning the normal patterns. During the test phase, the learned model discriminates behaviors that outside of the learned normal patterns as abnormal events. Early works [Li *et al.*, 2013] mainly focus on manually designing appropriate features to represent videos. However, such approaches are difficult to transfer among different scenarios. With the recent advances in deep learning, deep neural networks (DNNs) based schemes have become the mainstream solutions to VAD. Generally, DNNs-based approaches follow two frameworks: frame reconstruction and prediction. Frame reconstruction-based VAD learns to reconstruct the normal events and detects abnormal events by the larger reconstruction errors. Prediction-based VAD takes prior frames as the input of model to predict current frame and detects frames with poor prediction as anomalies.

Despite the remarkable performance gain, DNNs-based VAD still suffers from three issues: *1) Previous DNNs-based approaches allow to reconstruct or predict anomalies well.* As the reconstruction goal is the original input, reconstruction-based models usually reduce the loss by simply memorizing the pixel-level details of inputs, which results in the abnormal frames can be also well reconstructed. Although prediction-based VAD can avoid this problem to some extent, it cannot guarantee larger prediction errors for abnormal behaviors because of the powerful generalization ability of DNNs. *2) Previous prediction-based methods lack sufficient abilities to exploit the long-term temporal contexts of video events.* Specifically, existing prediction-based methods

*Corresponding author.

only predict a short-distance frame (*e.g.,* the next frame) with tiny differences from the input sequence, which cannot explicitly exploits the temporal contexts. In other words, vanilla frame prediction with small information gap cannot explicitly exploit the temporal contexts. *3) The diversity of normal patterns has been overlooked.* For instance, it is normal that no one or large amounts of pedestrians pass by the avenue, but these two cases have completely different dynamic patterns.

In this work, we present a Prototype-guided Dynamics Matching Network (PDM-Net) for VAD to enhance the discriminant and robustness of anomaly detector. Specifically, PDM-Net adopts a novel prototype-guided and dynamic-aware video prediction framework for VAD. To tackle the *issue 1)*, we adopt a prototype module which learns representative motion prototypes from normal videos. During the test phase, the motion features of input sequence are reconstructed with normal prototypes, and then reconstructed features are used to predict the target frames. As a result, the model can lessen the generalization ability of model towards anomalies, because abnormal target frames are also predicted with the normal prototypes. To address the *issue 2)*, we propose a Prototype-guided and Dynamic-aware Long-Distance Frame Prediction (PDLP) paradigm for VAD. Our model first stores the normal long-term motion prototypes learned from normal long sequences into prototype module, and then it uses motion features extracted from short sequences to recall the stored normal long-term motion prototypes. Finally, the recalled normal long-term motion prototypes are exploited to help compensate the missing motion information between the short input sequence and long-distance target frame. To mitigate the *issue 3)*, we adopt a feature discrimination module to preserve the diversity of normal patterns.

The main contributions of this work are listed as follows:

1) We develop a novel prototype-guided and dynamic-aware long-distance frame prediction framework for VAD, which can jointly help the anomaly detector explicitly exploit the long-term temporal contexts and lessen the generalization ability of model on anomalies.
2) A long-term dynamic prototypical network is designed to learn the long-term dynamic prototypes of normal patterns, which facilitates the long-distance prediction of normal frames and suppresses those of abnormal frames.
3) Experimental results on four public datasets demonstrate the superiority of our method.

2 Related Work

2.1 Video Anomaly Detection

Unsupervised video anomaly detection (VAD) [Huang *et al.*, 2022b; Wu *et al.*, 2023a; Huang *et al.*, 2024; Huang *et al.*, 2020] is typically formulated as an out-of-detection task in previous works, where a normal model is trained on only normal videos to recognize events outside of the learned model as anomalies. Some early works [Li *et al.*, 2013] adopt object detection and tracking methods to extract high-level features. However, these conventional approaches require the prior knowledge to design appropriate features. Recently, DNNs-based VAD methods have demonstrated superior performance over hand-crafted features-based methods. Many approaches adopt deep Auto-Encoder (AE) to learn the normal patterns and quantify the extent of abnormalities using reconstruction errors [Hasan *et al.*, 2016]. As the reconstruction target is the input frame, these deep AE-based models overlook temporal information of videos and typically reduce their loss by memorizing the pixel details of input frame. Subsequently, several variants of AE are developed to capture the temporal patterns of normal videos. For example, Luo *et al.* [Luo *et al.*, 2021] introduced recurrent neural networks into AE. Liu *et al.* [Liu *et al.*, 2018] utilized adversarial training to help AE perform frame prediction task. To lessen the generalization ability of DNNs, Gong *et al.* and Park *et al.* [Gong *et al.*, 2019; Park *et al.*, 2020] introduced a memory module into deep AE. Lv *et al.* [Lv *et al.*, 2021a] dynamically learned the prototypes of normal patterns via an attention mechanism for quantifying the normalities of each pixels. These memory-guided methods directly learn normal patterns from the short-distance frame prediction, i.e., utilizing several previous frames to predict current frame. Thus, they lack sufficient abilities to explicitly exploit the long-term temporal contexts of video events. To explicitly exploit the temporal contests, our proposed prototype-guided dynamic matching network (PDM-Net) learns how to match the learned normal long-term prototypes with the short-term inputs.

2.2 Prototypical Networks

Prototypical learning has proven its effectiveness in various pattern recognition tasks. The earliest prototypical learning can be traced back to the k-nearest neighbor which represents data with k nearest neighbors [Wang *et al.*, 2022]. Generally, conventional prototypical learning is mainly built on hand-crafted features. Recently, prototypical networks combining prototypical learning and DNN have demonstrated their superiority in various tasks [Wang *et al.*, 2021d; Jiang *et al.*, 2022; Wang *et al.*, 2021c; Liu *et al.*, 2023b; Wang *et al.*, 2023; Wang *et al.*, 2021b; Wang *et al.*, 2021a]. Prototypical networks can directly extract deep features by neural networks and incorporate prototypical learning to optimize the neural networks. For instance, Liu *et al.* [Liu *et al.*, 2023a] presented a region prototypical network to improve the performance of weak supervised image segmentation, which learns region prototypes to locate the inactivated objects more accurately. Lee *et al.* [Lee *et al.*, 2021] proposed a memory alignment method for improving the performance of video prediction, which learns the prototypes across training samples and uses the learned prototypes to facilitate video prediction at test time. Instead of directly learning prototypes from the inputs, PDM-Net learns long-term motion prototypes from normal video clips and then recalls them using the short inputs for predicting the long-distance frames.

3 Methodology

3.1 Problem Formulation

Prediction-based VAD typically leverages frame prediction to train a normal model. Then anomalies are detected by poor prediction based on the following assumption: the

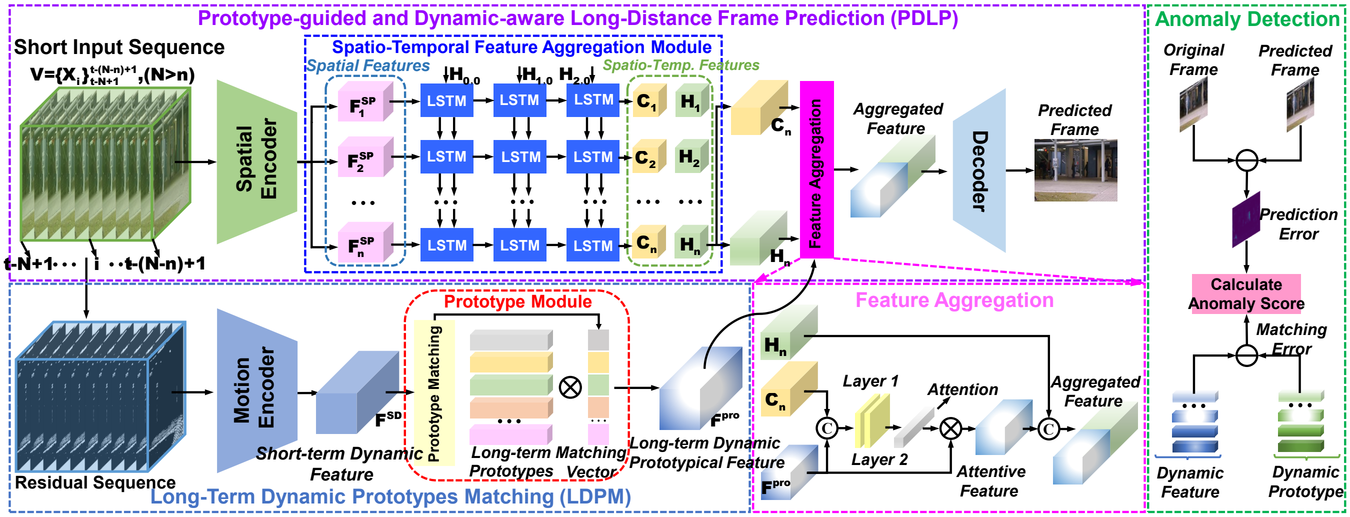


Figure 1: Architecture of our PDM-Net at inference phase. The lower branch is the long-term dynamic prototypes matching, which uses the motion features of short sequence to match the long-term motion prototypes of normal patterns. The upper branch is PDLP, which predict the target frame with the help of matched long-term motion prototypical features. Finally, the prediction error between original frame X_{t+1} and predicted frame $\widehat{X}_{t+1}$ as well as the matching error between feature $\mathbf{F}^{SD}$ and its prototype $\mathbf{\Gamma}^*$ are used to calculated the anomaly score.

learned normal model can predict normal future frames well while poorly predict abnormal frames. Assume $\mathbf{V}_{t-N+1:t} = \{X_i\}_{i=t-N+1}^{t}$ indicates a video sequence containing N consecutive frames, and X_t is the t-th frame. Given previous N frames $\mathbf{V}_{t-N+1:t}$, the goal is to train a predictor $\mathcal{P}$ that minimizes the difference between predicted frame $\widehat{X}_{t+1} = \mathcal{P}(\mathbf{V}_{t-N+1:t})$ and actual frame X_{t+1}.

To exploit the temporal information, we develop a novel prototype-guided and dynamic-aware long-distance prediction (PDLP) paradigm, as shown in Figure 1. Specifically, PDLP utilizes the short input video sequence $\mathbf{V}_{(t-N+1):(t-(N-n)+1)}$ to predict its long-distance target frame X_{t+1}. Theoretically, the predictor $\mathcal{P}$ needs to bridge the larger information gap by explicitly exploiting the long-term temporal contexts of videos, so as to accurately predict the target frame X_{t+1}. Actually, only simply increasing the distance between input sequence and target frame cannot enable our model to obtain better performance (detailed in Ablation Study). Thus, we introduce the prototype module to tackle this problem. The dynamic features $\mathbf{F}^{SD}$ of short inputs are used as queries to match the normal long-term dynamic prototypes learned from the normal long sequences. Then, the recalled prototype features $\mathbf{F}^{pro}$ are aggregated as $\mathbf{F}_k^{Agg}$ to predict the target frame X_{t+1}

$$\{\mathcal{P}^*, \mathcal{F}^*\} = \underset{\mathcal{P},\mathcal{F}}{\arg\min}\ \|X_{t+1} - \mathcal{P}([\mathbf{V}; \mathcal{F}(\mathbf{V})])\|, \quad (1)$$

where $\mathbf{V}$ is the short sequence $\mathbf{V}_{(t-N+1):(t-(N-n)+1)}$, and $\mathcal{F}$ indicates the prototypical module. Finally, the prediction error and the matching error are used to calculated the anomaly score for detecting anomalies.

3.2 PDLP-Enabled Video Anomaly Detection

Figure 1 shows the architecture of our prototype-guided dynamics matching network (PDM-Net) at the inference phase. The short input sequence goes through two branches to predict the long-distance target frame X_{t+1}. One (lower branch of Figure 2) is long-term dynamic prototypes matching (LDPM), which uses the dynamic features of short sequence to match the long-term dynamic prototypes of normal patterns stored in the prototype module. The other one (upper branch of Figure 2) is the proposed PDLP module, which utilizes the spatio-temporal features to predict the target frame with the assistant of matched long-term dynamic prototypical features.

As for the branch of LDPM, the residual sequence $\mathbf{R}$ is taken as the input of dynamic encoder $\mathcal{E}_{SDM}$. The dynamic feature $\mathbf{F}^{SD} = \mathcal{E}_{SDM}(\mathbf{R})$ is extracted for matching the long-term dynamic prototypes $\mathbf{\Gamma}$ from the prototype module. Specifically, a matching vector $\mathbf{M}$ is calculated according to the prototype matching strategy. $\mathbf{F}^{SD}$ is represented as the weighted combination of prototypes $\mathbf{\Gamma}$, *i.e.*, $\mathbf{F}^{pro} = \mathbf{\Gamma} \otimes \mathbf{M}$. In this way, $\mathbf{F}^{pro}$ of the short input can be considered to contain the long-term temporal dynamic contexts. Then, $\mathbf{F}^{pro}$ is embedded into the upper branch PDLP to predict the long-distance target frame X_{t+1}.

With regards to the branch of PDLP, each frame of the short sequence is independently fed into the spatial encoder $\mathcal{E}_{SP}$ for extracting the appearance features $\mathbf{F}_k^{SP} = \mathcal{E}_{SP}(X_k)$. Further, $\mathbf{F}_k^{SP}$ is fed into a spatio-temporal aggregation module composed of a stack of convolutional LSTMs (ConvLSTM) [Shi *et al.*, 2015] in time step orders to capture the temporal relations. Each cell in ConvLSTM outputs a cell memory C_k and a hidden state H_k. Then, we design a feature aggregation module to aggregate the spatio-temporal features and prototypical feature, as shown in Figure 1. Specifically, $\mathbf{F}^{pro}$ and C_k are concatenated and fed into the multi-layer perception to generate the channel-wise attention $\mathbf{A}_k^{pro} = MLP(\mathbf{F}^{pro} © C_k)$. © indicates concatenate operator. Further, the channel-wise refined prototypical feature $\widehat{\mathbf{F}}_k^{pro} =$

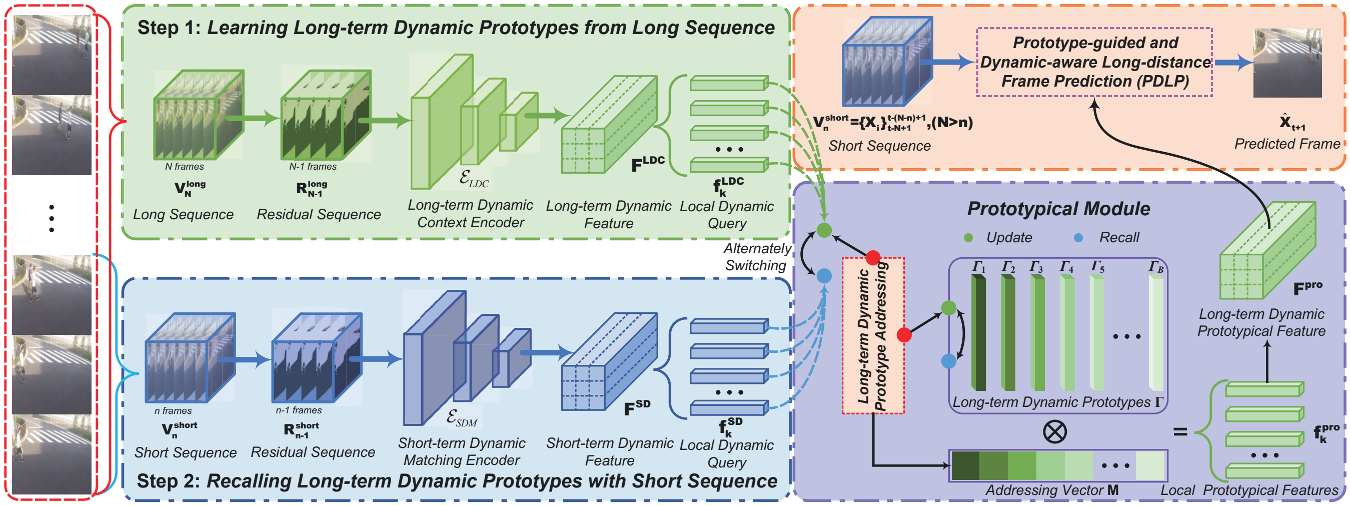


Figure 2: Detailed design of our prototype module with long-term and short-term dynamic contexts matching. To match the long-term and short-term dynamic in the prototype module, our model is trained with two steps: (i) learning long-term dynamic context prototypes from long sequence, (ii) recalling long-term dynamic prototypes with short sequence.

$\mathbf{A}_k^{pro} \otimes \mathbf{F}^{pro}$ and the hidden feature H_k from the ConvLSTMs are concatenated to embed long-term dynamic contexts to the aggregated feature $\mathbf{F}_k^{Agg} = \widehat{\mathbf{F}}_k^{pro} \copyright H_k$. Since $\mathbf{F}_k^{Agg}$ can provide the prior of long-term dynamic contexts to the short input sequence, it is fed into the frame decoder $\mathcal{D}$ to generate the predicted frame $\widehat{X}_{t+1} = \mathcal{D}(\widehat{\mathbf{F}}_k^{pro} \copyright H_k)$. Finally, the prediction error between the original frame X_{t+1} and predicted frame $\widehat{X}_{t+1}$ as well as the matching error between dynamic feature $\mathbf{F}^{SD}$ and its prototype $\mathbf{\Gamma}^*$ are used to calculated the anomaly score (detailed in Section 3.5). At the training phase, we employ a prediction loss to constrain the PDLP, which is defined as

$$\mathcal{L}_{pre} = \|X_{t+1} - \widehat{X}_{t+1}\|_2^2 + \|X_{t+1} - \widehat{X}_{t+1}\|_1. \tag{2}$$

3.3 Dynamic Prototype Matching Learning

Inspired by [Lee *et al.*, 2021], we introduce a long short-term dynamic prototype matching scheme to accurately match the long-term dynamic with the short sequence. Figure 2 illustrates the overall training procedure of our prototype module. The prototype module is trained alternately with two steps: 1) *learning long-term dynamic prototypes from normal long sequence*; and 2) *recalling the corresponding normal long-term dynamic prototypes with the short sequence*. Notably, the prototype module is updated only at the first step.

Long-term Dynamic Prototypes Learning

As for the first step, the normal long sequence $\mathbf{V}_N^{long}$ with N consecutive frames is taken as the input. Then, the corresponding residual sequence $\mathbf{R}_{N-1}^{long}$ is fed into the long-term dynamic context encoder $\mathcal{E}_{LDC}$ to extract the long-term dynamics $\mathbf{f}^{LDC}$. $\mathbf{f}^{LDC} = \{\mathbf{f}_k^{LDC}\}_{k=1}^{K}$ ($K = h \times w$) is divided to exploit local dynamic contexts. Notably, the local dynamic features $\mathbf{f}_k^{LDC} \in \mathbb{R}^c$ are regarded as the query items.

The prototype module $\mathbf{\Gamma} = \{\boldsymbol{\Gamma}_i\}_{i=1}^{B}$ contains B prototype items $\boldsymbol{\Gamma}_i$ with c channels. A matching matrix $\mathbf{M} = \{\mathbf{m}_k\}_{k=1}^{K}$ is created to address the location of prototype module, and each vector $\mathbf{m}_k$ is used to match a query feature $\mathbf{f}_k^{LDC}$ with all prototype items. Each element value $m_{k,i}$ of $\mathbf{m}_k$ can be used as the matching probability, which is calculated as

$$m_{k,i} = \frac{\exp(d(\mathbf{f}_k^{LDC}, \boldsymbol{\Gamma}_i))}{\sum_j^B \exp(d(\mathbf{f}_k^{LDC}, \boldsymbol{\Gamma}_j))}, \tag{3}$$

where $d(\cdot,\cdot)$ is the cosine similarity. The prototype module outputs the local prototypical features $\mathbf{f}_k^{pro} = \sum_{i=1}^{B} m_{k,i} \cdot \boldsymbol{\Gamma}_i$. The prototypical feature $\mathbf{F}^{pro} = \{\mathbf{f}_k^{pro}\}_{k=1}^{K}$ is derived as an ensemble of K local feature $\mathbf{f}_k^{pro}$. Finally, $\mathbf{F}^{pro}$ is integrated into PDLP to help the short sequence $\mathbf{V}_n^{short}$ predict the long-term target frame X_{t+1}, as described in Section 3.2.

As for the parameters updating of prototype module, we update the prototype item $\boldsymbol{\Gamma}_i$ with all query features whose nearest prototype item is $\boldsymbol{\Gamma}_i$. Let $\boldsymbol{\mathcal{Z}}_i$ indicate the set of the corresponding query features for $\boldsymbol{\Gamma}_i$. Similar to Eq. 3), the attention weight is calculated by

$$\omega_{k,i} = \frac{\exp(d(\mathbf{f}_k^{LDC}, \boldsymbol{\Gamma}_i))}{\sum_j^K \exp(d(\mathbf{f}_j^{LDC}, \boldsymbol{\Gamma}_i))}, \tag{4}$$

and renormalized with the query features in $\boldsymbol{\mathcal{Z}}_i$ as $\widehat{\omega}_{k,i} = \frac{\omega_{k,i}}{\max_{j \in \boldsymbol{\mathcal{Z}}_i} \omega_{j,i}}$. The parameter updating is represented as

$$\boldsymbol{\Gamma}_i \leftarrow \mathcal{N}(\boldsymbol{\Gamma}_i + \sum\nolimits_{k \in \boldsymbol{\mathcal{Z}}_i} \widehat{\omega}_{k,i} \cdot \mathbf{f}_k^{LDC}), \tag{5}$$

where $\mathcal{N}(\cdot)$ denotes ℓ_2 norm.

Prototypes Matching Learning

At the second step, the short sequence $\mathbf{V}_n^{short}$ with n ($n < N$) consecutive frames is taken as the input of model. The goal is to train our model to learn how to recall the long-term dynamic contexts stored in the prototype module using the short inputs. Similar to the first step, the residual sequence $\mathbf{R}_{n-1}^{short}$ is used to obtain the corresponding dynamics. The short-term dynamic feature $\mathbf{f}^{SD}$ is extracted by a short-term dynamic matching encoder $\mathcal{E}_{SDM}$. The local dynamic feature $\mathbf{f}_k^{SD}$ of $\mathbf{f}^{SDM} = \{\mathbf{f}_k^{SD}\}_{k=1}^{K}$ ($K = h \times w$) is adopted as the query item. The prototype matching progress is same as the first step. Then the recalled prototypical feature is embedded into LPLP for video prediction. At this step, the parameters of prototype module are fixed. Except for the prototype module, all network parameters are optimized.

3.4 Feature Discrimination

The feature discrimination contains two components, i.e., prototype dispersion and feature compactness. They are respectively embedded into the first and second training step for preserving the diversity of prototypes and reducing the intraclass variances of normal patterns. Specifically, the prototype dispersion enables the prototype module to learn representative prototypes of normal patterns by enforcing prototypes far away from each other at the first training step. Inspired by [Lai *et al.*, 2021], we evaluate the dispersibility between prototypes from the perspective of interclass scatter matrix by

$$\mathbf{D} = \sum\nolimits_{i=1}^{B} \frac{k_i}{K} (\boldsymbol{\Gamma}_i - \bar{\boldsymbol{\Gamma}})^\top (\boldsymbol{\Gamma}_i - \bar{\boldsymbol{\Gamma}})), \tag{6}$$

where k_i denotes the number of query features that $\boldsymbol{\Gamma}_i$ is their nearest prototype, and $\bar{\boldsymbol{\Gamma}}$ indicates the mean vector of all prototypes in the prototype module. The trace $Tr(\mathbf{D})$ of interclass scatter matrix can be used to measure the dispersibility of prototypes. Thus, we encourage the prototype module to learn various long-term dynamic prototypes by maximizing $Tr(\mathbf{D})$, and the prototype dispersion loss can be represented as $\mathcal{L}_{dis} = -Tr(\mathbf{D})$. Notably, $\mathcal{L}_{dis}$ only optimizes the prototype module at the first training step.

Moreover, the feature compactness enforces the query features close to their nearest prototypes in the latent space for reducing the intraclass variations. In this way, query features with the same dynamic patterns can be accurately matched to the same prototype. Here, we adopt a feature compactness loss [Park *et al.*, 2020] which measures the average distance between query feature and their prototypes

$$\mathcal{L}_{com} = \frac{1}{K} \sum\nolimits_{k=1}^{K} \|\mathbf{f}_k^{SD} - \boldsymbol{\Gamma}_k^*\|_2, \tag{7}$$

where $\boldsymbol{\Gamma}_k^*$ is the most-relevant prototype of $\mathbf{f}_k^{SD}$.

3.5 Training and Inference

Training. As mentioned above, the optimization of PDM-Net contains two steps. At the first step, the network weights of long-term dynamic context encoder $\mathcal{E}_{LDC}$ (θ), prototype model $\mathcal{F}$ ($\mathbf{\Gamma}$) and PDLP networks $\mathcal{P}$ (ϕ) are optimized by $\mathcal{L}_{1st} = \mathcal{L}_{pre} + \lambda(\mathcal{L}_{dis} + \mathcal{L}_{com})$. The prototype model $\mathbf{\Gamma}$ is updated according to Eq. 5. At the second step, the weights of short-term dynamic matching encoder $\mathcal{E}_{SDM}$ (γ) and PDLP networks $\mathcal{P}$ (ϕ) are optimized by $\mathcal{L}_{2nd} = \mathcal{L}_{pre} + \lambda\mathcal{L}_{com}$. In this step, the network weights of prototype module are fixed. The optimization of model is described as Algorithm 1.

Algorithm 1: Optimization of our PDM-Net

Input: $\mathbf{V}_n^{short}$, $\mathbf{V}_N^{long}$, Target frame X_{t+1}
Output: The network parameters $\{\theta, \phi, \gamma, \mathbf{\Gamma}\}$
Initialize the network parameters $\{\theta, \phi, \gamma, \mathbf{\Gamma}\}$;
for *each iteration* **do**
  Step 1: Learn Long-term Dynamic Prototype
  Calculate residual sequence: $\mathbf{R}_{N-1}^{long}$;
  Long-term dynamic: $\mathbf{f}^{LDC} = \mathcal{E}_{LDC}(\mathbf{R}_{N-1}^{long})$;
  Extract prototypical feature: $\mathbf{F}^{pro} = \mathcal{F}(\mathbf{f}^{LDC})$;
  Predict target frame: $\widehat{X}_{t+1} = \mathcal{P}([\mathbf{V}_n^{short}; \mathbf{F}^{pro}])$;
  Calculate loss: $\mathcal{L}_{1st} = \mathcal{L}_{pre} + \lambda(\mathcal{L}_{dis} + \mathcal{L}_{com})$;
  Update $\{\theta, \phi\}$ using Stochastic Gradient:
    $\theta \leftarrow \theta - \nabla_\theta(\mathcal{L}_{pre} + \lambda(\mathcal{L}_{dis} + \mathcal{L}_{com}))$;
    $\phi \leftarrow \phi - \nabla_\phi(\mathcal{L}_{pre} + \lambda(\mathcal{L}_{dis} + \mathcal{L}_{com}))$;
  Update the prototype model $\mathcal{F}$:
    $\boldsymbol{\Gamma}_i \leftarrow \mathcal{N}(\boldsymbol{\Gamma}_i + \sum_{k \in \boldsymbol{\mathcal{Z}}_i} \widehat{\omega}_{k,i} \cdot \mathbf{f}_k^{LDC})$;
  Step 2: Prototypes Matching Learning
  Calculate residual sequence: $\mathbf{R}_{n-1}^{short}$;
  Short-term dynamic: $\mathbf{f}^{SD} = \mathcal{E}_{SDM}(\mathbf{R}_{n-1}^{short})$;
  Recall long-term prototypes: $\mathbf{F}^{pro} = \mathcal{F}(\mathbf{f}^{SD})$;
  Predict target frame: $\widehat{X}_{t+1} = \mathcal{P}([\mathbf{V}_n^{short}; \mathbf{F}^{pro}])$;
  Calculate loss: $\mathcal{L}_{2nd} = \mathcal{L}_{pre} + \lambda\mathcal{L}_{com}$;
  Update $\{\gamma, \phi\}$ using Stochastic Gradient:
    $\gamma \leftarrow \gamma - \nabla_\gamma(\mathcal{L}_{pre} + \lambda\mathcal{L}_{com})$;
    $\phi \leftarrow \phi - \nabla_\phi(\mathcal{L}_{pre} + \lambda\mathcal{L}_{com})$;
end

Inference. As shown in Figure 1, our PDM-Net only takes short sequence as the input at the testing phase. The LDPM branch extracts the short-term dynamic features to match the long-term dynamic prototypes from the prototype module for assisting the PDLP branch to generate the predicted frame $\widehat{X}_{t+1}$. To quantify the abnormal extent of a video frame, the prediction error between the predicted frame $\widehat{X}_{t+1}$ and actual frame X_{t+1} is used to calculate the anomaly score. Besides, we also evaluate the matching error between the query feature $\mathbf{f}_k^{SD}$ and its prototype $\boldsymbol{\Gamma}_k^*$. Here, we apply the peak signal to noise ratio to measure the prediction error, namely $\mathcal{S}_{pre}$. The lower $\mathcal{S}_{pre}$ indicates poorer prediction and X_{t+1} tends to be abnormal. We calculate the matching error as

$$\mathcal{S}_{mat} = \frac{1}{K} \sum\nolimits_{k=1}^{K} \|\mathbf{f}_k^{SD} - \boldsymbol{\Gamma}_k^*\|_2. \tag{8}$$

The query feature $\mathbf{f}_k^{SD}$ is similar to the corresponding normal prototype $\boldsymbol{\Gamma}_k^*$ indicates X_{t+1} is normal. Following the previous methods, we adopt the same normalization function [Liu *et al.*, 2018] to normalize $\mathcal{S}_{pre}$ and $\mathcal{S}_{mat}$ to $[0, 1]$. The final anomaly score is calculated as

$$\mathcal{S} = \beta(1 - \mathcal{M}(\mathcal{S}_{pre})) + (1 - \beta)\mathcal{M}(\mathcal{S}_{mat}), \tag{9}$$

where β is the balance weight, and $\mathcal{M}(\cdot)$ is the normalization function over the whole video frames.

Method	**SHTech**	**Ped1**	**Ped2**	**Avenue**
Conv-AE[Hasan *et al.*, 2016]	60.9	75.0	90.0	70.2
CLSTM-AE[Luo *et al.*, 2017]	55.0	75.5	88.1	77.0
FP[Liu *et al.*, 2018]	72.8	83.1	95.4	85.1
MemAE[Gong *et al.*, 2019]	71.2	-	94.1	83.3
PCM [Ye *et al.*, 2019]	73.6	-	96.8	86.2
DeepOC[Wu *et al.*, 2019]	-	75.5	96.9	86.6
MNAD [Park *et al.*, 2020]	70.5	-	97.0	88.5
IPR[Tang *et al.*, 2020]	73.0	-	96.3	85.1
ClusterAE [Chang *et al.*, 2020]	73.3	-	96.5	86.0
ISTL[Nawaratne *et al.*, 2020]	-	75.2	91.1	76.8
SSPCA [Ristea *et al.*, 2022]	69.8	-	-	84.8
SSMC [Madan *et al.*, 2022]	70.6	-	-	84.6
ITAE [Cho *et al.*, 2021]	71.8	-	96.8	85.5
FastAno [Park *et al.*, 2022]	72.2	-	96.3	85.3
ITAEGM [Cho *et al.*, 2021]	73.0	-	97.3	86.0
MESDnet [Fang *et al.*, 2021]	73.2	-	95.6	86.3
Multispace [Zhang *et al.*, 2021]	73.6	-	95.4	86.8
F^2PN[Luo *et al.*, 2022]	73.0	84.3	96.2	85.7
AMMCN [Cai *et al.*, 2021]	73.7	-	96.6	86.6
PDM-Net	**74.2**	**85.2**	**97.7**	**88.1**

Table 1: Comparison with other methods. [AUC (%)]

Dataset	Index	Pred.	LD-Pred.	Prot.	FD	AUC
Ped1	A	✓				82.5%
	B		✓			82.0%
	C	✓		✓		83.8%
	D		✓	✓		84.5%
	E		✓	✓	✓	**85.2%**
Ped2	A	✓				94.1%
	B		✓			93.7%
	C	✓		✓		95.8%
	D		✓	✓		96.9%
	E		✓	✓	✓	**97.7%**
Avenue	A	✓				84.8%
	B		✓			84.3%
	C	✓		✓		86.1%
	D		✓	✓		87.0%
	E		✓	✓	✓	**88.1%**

Table 2: Results of the ablation studies. [AUC (%)]

4 Experiments

4.1 Experimental Setup

We conduct experiments on four unsupervised VAD benchmarks, UCSD Ped1 [Li *et al.*, 2013], UCSD Ped2 [Li *et al.*, 2013], CUHK Avenue [Lu *et al.*, 2013], and ShanghaiTech [Luo *et al.*, 2021] to evaluate performance. Under the unsupervised setting, the training set only consists of normal videos. Following prior works [Liu *et al.*, 2018], Area Under ROC (AUC) is adopted as the evaluation metric. We resize each input frames to intensity of [0, 1] and resolution of 256×256. Adam with the initial learning rate of 0.0002 is adopted to optimize our PDM-Net. The number B of prototypes in the prototype module is set as 100. The lengths of long and short sequences are set as to 13 and 9, respectively. λ and β are set as 0.1 and 0.6.

4.2 Comparison with Other Methods

We compare our PDM-Net with state-of-the-art VAD approaches. The detailed results are listed in Tabel 1. On ShanghaiTech, PDM-Net is superior to all the methods involved in the comparison with an average AUC of 74.2%, which is higher 0.5% points than that of AMMCN. Our PDM-Net outperforms the vanilla prediction framework-based methods FP and PCM, which validates the effectiveness of our PDLP paradigm. On UCSD Ped1, PDM-Net can obtain a new state-of-the-art performance with an average AUC of 85.2%, which is higher 0.9% than the previous best result of 84.3% AUC reported by F^2PN. On UCSD Ped2, PDM-Net is superior to all the baselines involved in the comparison with an AUC of 97.7%, which is higher 0.4% than that of previous state-of-the-art method ITAEGM. Compared with previous memory-guided methods MemAE and MNAD, PDM-Net can not only exploit the long-term temporal contexts of normal videos, but also lessen the generalization ability of model to anomalies. The improvement indicates the superiority of our prototype module with dynamics matching learning. On Avenue, PDM-Net obtains a competitive performance with the averaged AUC of 88.1%.

4.3 Ablation Study

We study the contribution of each component in our PDM-Net, and the experimental results are shown in Table 2. A basic model (module A) adopts vanilla prediction framework, which is trained with the prediction loss $\mathcal{L}_{pre}$ (Eq. 2). It obtains 82.5% AUC on UCSD Ped1, 94.1% AUC on UCSD Ped2, and 84.8% AUC on Avenue. Then, we adopt the long-distance frame prediction (module B), which only simply increases the distance between the input sequences and target frames. The performance of module B is degraded by 0.4% to 0.5% on all datasets compared with Module A. Without the assistance of long-term dynamic context prototypes in the prototype module, the model lacks sufficient ability to exploit the long-term temporal contexts for bridging the large information gap between input and target, resulting in poor prediction on both normal and abnormal frames.

Impact of Prototype Module

We embed the prototype module into the model A and B to construct the module C and D, respectively. Clearly, embedding our prototype module enables module C to obtain 1.3% to 1.7% improvements over the basic module A on all datasets. Comparing module B and module D, we can observe that embedding prototype module enables module D to obtain the performance improvements of 2.5%, 3.2% and 2.7% of AUC scores on three datasets. The significant performance improvements demonstrate the validity of our prototype module.

Impact of PDLP Framework

We adopt our proposed prototype-guided and dynamic-aware long-distance frame prediction (PDLP) framework (module D), *i.e,* embedding the prototype module into the long-distance frame prediction-based module B. Compared with the vanilla frame prediction framework (Module A), PDLP framework-based module D can achieve the performance gains of 2.0%, 2.8%, and 2.2% of AUC scores on three datasets, which demonstrates the superiority of our PDLP framework. Compared with the module C which is embedded the same prototype module, our PDLP framework can

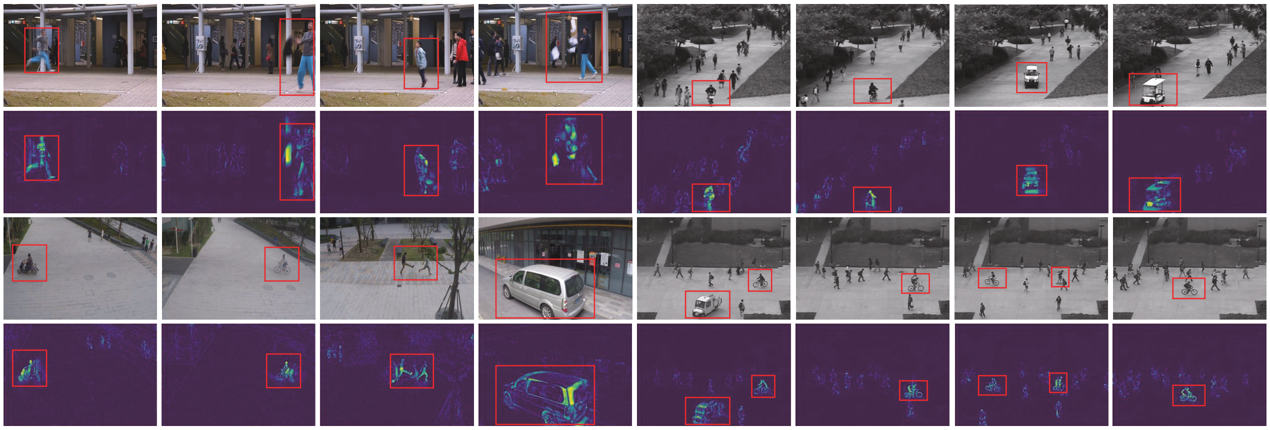
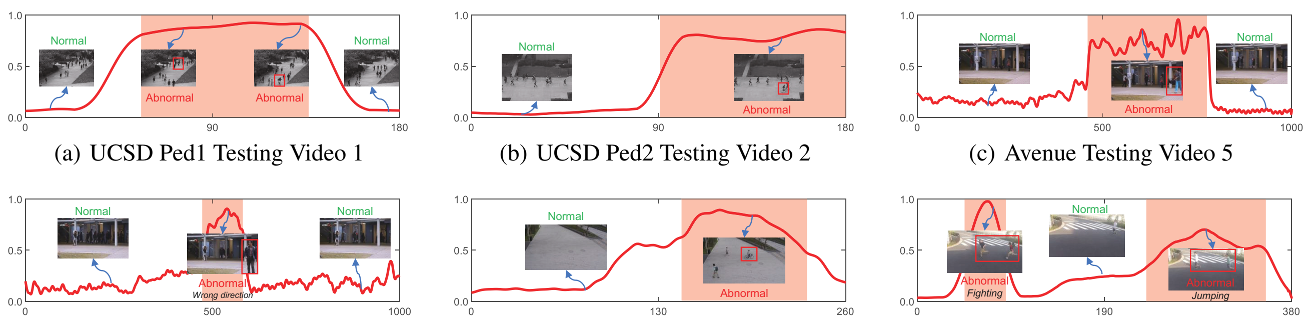


Figure 3: Visual demonstration of the frame prediction for detecting anomalies. The first and third rows are the actual frames. The second and fourth rows are residual frames between actual and predicted frames. Red boxes indicate the abnormal regions.



(a) UCSD Ped1 Testing Video 1
(b) UCSD Ped2 Testing Video 2
(c) Avenue Testing Video 5
(d) Avenue Testing Video 15
(e) ShanghaiTech Testing Video 01 0014
(f) ShanghaiTech Testing Video 03 0035

Figure 4: Examples of anomaly score curves and representative frames. Light red regions is the ground truth of abnormal events.

improves the performance by 0.7% on UCSD Ped1, by 1.1% on UCSD Ped2 and by 0.9% on Avenue, which indicates that the performance gain of our PDLP does not simply come from the superposition of modules, but from the more effective coupling between modules.

Impact of Feature Discrimination

Then, we embed the feature discrimination module into PDLP framework-based module D to construct our overall model (module E). The feature discrimination enables our model to obtain improvements of 0.7%, 0.8% and 1.1% on three datasets, respectively. The performance improvement demonstrates the effectiveness of feature discrimination.

Visualization of Long-Distance Prediction

We show the actual and residual frames to check the reliability of our model for VAD. In Figure 3, the second and fourth rows are residual frames between actual and predicted frames. Red boxes represent the abnormal regions. It is obvious that the prediction of anomalies by our PDM-Net is significantly worse than that of normal regions, which indicates that our method can achieve the goal of lessening the generalization ability of learned model to anomalies.

Visualization of Anomalous Event Detection

Figure 4 shows some instances of anomaly score curves from our PDM-Net. It is clear that our PDM-Net is able to correctly response to normal and abnormal events in time. Specifically, the curve rises sharply when an anomalous event suddenly appears and continuously remains at a quite high level when the anomalous event is in progress. When the objects resulting in the anomalies disappears, the curves drop to a relatively low level.

5 Conclusion

In this paper, we propose a novel prototype-guided dynamics matching framework that can jointly exploit the long-term temporal contexts and preserve the diversity of normal patterns, while lessening the generalization ability of model to anomalies. Specifically, a prototype-guided and dynamic-aware long-distance frame prediction (PDLP) framework is adopted to exploit the temporal contexts of normal events. Moreover, a prototype module with dynamic matching learning is designed to provide the short normal inputs with long-term dynamic prototypes of normal events, which helps the model to bridge the large information gap for achieving PDLP. Extensive experimental results on four datasets demonstrate the superiority of our method.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62301621) and Shenzhen Science and Technology Program (20231121172359002).

References

[Cai *et al.*, 2021] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. Appearance-motion memory consistency network for video anomaly detection. In *Proc. AAAI Conf. Artif. Intell.*, volume 35, pages 938–946, 2021.

[Chang *et al.*, 2020] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *Proc. Eur. Conf. Comput. Vis.*, pages 329–345. Springer, 2020.

[Cho *et al.*, 2021] MyeongAh Cho, Taeoh Kim, and Sangyoun Lee. Unsupervised video anomaly detection via flow-based generative modeling on appearance and motion latent features. In *Proc. AAAI Conf. Artif. Intell.*, 2021.

[Fang *et al.*, 2021] Zhiwen Fang, Joey Tianyi Zhou, Yang Xiao, Yanan Li, and Feng Yang. Multi-encoder towards effective anomaly detection in videos. *IEEE Trans. Multimedia*, 23:4106 – 4116, 2021.

[Gong *et al.*, 2019] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1705–1714, 2019.

[Hasan *et al.*, 2016] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 733–742, 2016.

[Huang *et al.*, 2020] Chao Huang, Zongju Peng, Yong Xu, Fen Chen, Qiuping Jiang, Yun Zhang, Gangyi Jiang, and Yo-Sung Ho. Online learning-based multi-stage complexity control for live video coding. *IEEE Transactions on Image Processing*, 30:641–656, 2020.

[Huang *et al.*, 2021] Chao Huang, Zehua Yang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, and Yaowei Wang. Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection. *IEEE Transactions on Cybernetics*, 52(12):13834–13847, 2021.

[Huang *et al.*, 2022a] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Transactions on Cybernetics*, 2022.

[Huang *et al.*, 2022b] Chao Huang, Yabo Liu, Zheng Zhang, Chengliang Liu, Jie Wen, Yong Xu, and Yaowei Wang. Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 307–315, 2022.

[Huang *et al.*, 2022c] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE transactions on neural networks and learning systems*, 2022.

[Huang *et al.*, 2022d] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Trans. Ind. Informat.*, 18(8):5171–5179, 2022.

[Huang *et al.*, 2024] Chao Huang, Yushu Shi, Bob Zhang, and Ke Lyu. Uncertainty-aware prototypical learning for anomaly detection in medical images. *Neural Networks*, page 106284, 2024.

[Jiang *et al.*, 2022] Qiuping Jiang, Yudong Mao, Runmin Cong, Wenqi Ren, Chao Huang, and Feng Shao. Unsupervised decomposition and correction network for low-light image enhancement. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19440–19455, 2022.

[Lai *et al.*, 2021] Yuandu Lai, Yahong Han, and Yaowei Wang. Anomaly detection with prototype-guided discriminative latent embeddings. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 300–309. IEEE, 2021.

[Lee *et al.*, 2021] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3054–3063, 2021.

[Li *et al.*, 2013] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):18–32, 2013.

[Liu *et al.*, 2018] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6536–6545, 2018.

[Liu *et al.*, 2023a] Weide Liu, Xiangfei Kong, Tzu-Yi Hung, and Guosheng Lin. Cross-image region mining with region prototypical network for weakly supervised segmentation. *IEEE Trans. Multimedia*, 25:1148–1160, 2023.

[Liu *et al.*, 2023b] Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, and Yong Xu. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23776–23786, 2023.

[Lu *et al.*, 2013] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2720–2727, 2013.

[Luo *et al.*, 2017] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *Proc. IEEE Int. Conf. Multimedia and Expo*, pages 439–444, 2017.

[Luo *et al.*, 2021] Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(3):1070–1084, 2021.

[Luo *et al.*, 2022] Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. Future frame prediction network for video anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7505–7520, 2022.

[Lv *et al.*, 2021a] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15425–15434, 2021.

[Lv *et al.*, 2021b] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE Trans. Image Process.*, 30:4505–4515, 2021.

[Madan *et al.*, 2022] Neelu Madan, Nicolae-Catalin Ristea, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised masked convolutional transformer block for anomaly detection. *arXiv preprint arXiv:2209.12148*, 2022.

[Nawaratne *et al.*, 2020] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Trans. Ind. Informat.*, 16(1):393–402, 2020.

[Park *et al.*, 2020] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14372–14381, 2020.

[Park *et al.*, 2022] Chaewon Park, MyeongAh Cho, Minhyeok Lee, and Sangyoun Lee. Fastano: Fast anomaly detection via spatio-temporal patch transformation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2249–2259, 2022.

[Ristea *et al.*, 2022] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13576–13586, 2022.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 802–810, 2015.

[Sultani *et al.*, 2018] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6479–6488, 2018.

[Tang *et al.*, 2020] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognit. Lett.*, 129:123–130, 2020.

[Wang *et al.*, 2021a] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30:1771–1783, 2021.

[Wang *et al.*, 2021b] Wei Wang, Shenglun Chen, Yuankai Xiang, Jing Sun, Haojie Li, Zhihui Wang, Fuming Sun, Zhengming Ding, and Baopu Li. Sparsely-labeled source assisted domain adaptation. *Pattern Recognition*, 112:107803, 2021.

[Wang *et al.*, 2021c] Wei Wang, Baopu Li, Mengzhu Wang, Feiping Nie, Zhihui Wang, and Haojie Li. Confidence regularized label propagation based domain adaptation. *IEEE TCSVT*, 32(6):3319–3333, 2021.

[Wang *et al.*, 2021d] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE TNNLS*, 34(1):264–277, 2021.

[Wang *et al.*, 2022] Rui-Qi Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Meta-prototypical learning for domain-agnostic few-shot recognition. *IEEE Trans. Neural Netw. Learn. Syst.*, 33(11):6990–6996, 2022.

[Wang *et al.*, 2023] Wei Wang, Mengzhu Wang, Xiao Dong, Long Lan, Quannan Zu, Xiang Zhang, and Cong Wang. Class-specific and self-learning local manifold structure for domain adaptation. *Pattern Recognition*, 142:109654, 2023.

[Wu *et al.*, 2019] Peng Wu, Jing Liu, and Fang Shen. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE Trans. Neural Netw. Learn. Syst.*, 31(7):2609–2622, 2019.

[Wu *et al.*, 2023a] Lian Wu, Chao Huang, Lunke Fei, Shuping Zhao, Jianchuan Zhao, Zhongwei Cui, and Yong Xu. Video-based fall detection using human pose and constrained generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[Wu *et al.*, 2023b] Lian Wu, Chao Huang, Shuping Zhao, Jinkai Li, Jianchuan Zhao, Zhongwei Cui, Zhen Yu, Yong Xu, and Min Zhang. Robust fall detection in video surveillance based on weakly supervised learning. *Neural networks*, 163:286–297, 2023.

[Ye *et al.*, 2019] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. Anopcn: Video anomaly detection via deep predictive coding network. In *Proc. ACM Int. Conf. Multimedia*, pages 1805–1813, 2019.

[Zhang *et al.*, 2021] Yu Zhang, Xiushan Nie, Rundong He, Meng Chen, and Yilong Yin. Normality learning in multispace for video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.*, 31(9):3694 – 3706, 2021.

[Zhang *et al.*, 2022] Dasheng Zhang, Chao Huang, Chengliang Liu, and Yong Xu. Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. *IEEE Signal Processing Letters*, 29:1197–1201, 2022.