Reasonable Anomaly Detection Based on Long-term Sequence Modeling

Yalong Jiang, *Member*, *IEEE*, Changkang Li, Wenrui Ding, Jinzhi Xiang and Zheru Chi, *Member*, *IEEE*

***Abstract*—Video anomaly detection is a challenging task due to the unpredictable nature of abnormal actions, sophisticated semantics and a lack in training data. The visual representations of most existing approaches are limited by short-term sequences which cannot provide necessary clues for achieving reasonable detections. In this paper, we propose to comprehensively represent the motion patterns in human actions by learning from long-term sequences. Firstly, a Stacked State Machine (SSM) model with distinctive basis functions is proposed to represent the temporal dependencies which are consistent across long-term observations. Secondly, the dependencies are leveraged in filtering out problematic motion estimations which are influenced by short-term observation noises, plausible motion parameters are obtained in this way. Finally, SSM model predicts future states based on past ones, the divergence between the predictions with inherent normal patterns and observed ones determines anomalies which violate normal motion patterns. To address the challenges in drone-based surveillance, a dataset which is more diversified than existing ones is built. Extensive experiments are carried out to evaluate the proposed approach on the dataset and existing ones. Improvements over state-of-the-art methods can be observed. The proposed dataset will be made publicly available. Code is available at https://github.com/AllenYLJiang/Anomaly-Detection-in-Sequences.**

***Index Terms*—Video Anomaly Detection, Long-term Sequences, Plausible Anomaly Detection, Drone-based Dataset.**

I. INTRODUCTION

DUE to the large variety of abnormal events and inaccessibility of task-specific data, anomaly detection is a quite challenging research problem in surveillance. Typical anomalies include robbing, burglary and so on. However, abnormal actions rarely occur and do not conform to any fixed pattern, it is difficult to obtain annotations on anomalies. As a result, unsupervised approaches are required to distinguish the events that do not match regular patterns [1] [2].

Existing unsupervised or weakly supervised approaches leverage either the unpredictability of human behaviors [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] or the divergence in deep features [13] [14] [15] [16] [17] [18] [19] [20] between normal and abnormal events. To address the lack in data, some approaches [21] such as [22] [23] [24] have also been proposed to explore the distinctiveness of motion patterns. Furthermore, [25] [26] [27] proposed to leverage the prior knowledge about anomalies in identifying irregular behaviors.

Although impressive results have been achieved, the above approaches have not considered whether anomaly detections are reasonable. For instance, short-term sequences of observations cannot well describe the properties about actions [14] [4]. Some short-term motion patterns are shared by regular and irregular behaviors, it's difficult to characterize actions with short sequences. Typical unreasonable anomalies include the normal samples which are determined to be anomalies because short-term observation noises [28] lead to counterfactual pose estimations, the detections violate the long-term consistent patterns governing subjects' motion. As a result, representing subjects' behaviors with long-term sequence modeling is necessary in achieving reasonable anomaly detections. In one way, the effective encoding of motion patterns underlying long sequences provides comprehensive clues for distinguishing actions, exhibiting robustness to short-term random variations. On the other way, short-term implausible motion estimations can be corrected with long-term consistent patterns. Long-term temporal analysis improves representational capacities in a way that is complementary to existing approaches such as [29].

Furthermore, some of existing methods can only conduct frame-level anomaly detection without localizing abnormal regions [30] [7], the interpretability of them is not satisfactory because the underlying factors leading to anomalies may come from background clutters. As a result, foreground-centered method [31] isolated foreground dynamics from backgrounds. The introduction of inherent embeddings [3] augmented the interpretations of anomalies. Our approach moves one way forward by building representations with invariant basis functions, empowering the investigation to the intrinsic and invariant properties governing long-term dynamics.

Existing datasets [32] [33] [34] [35] [36] [37] [38] [39] covering public scenarios such as campus, subway stations and human-related abnormal events have fueled the fast development of algorithms [40] [32] [41] [42] [43] [44] [45]. However, the datasets are captured with static cameras and only involve limited scenarios with ideal lighting conditions. Models trained on the finite spatio-temporal patterns struggle in adapting to novel motion. More importantly, in applications such as crowd monitoring, long-time inspection, searching and rescue, flexible view-points are required to circumvent obstacles. As a result, drone-based anomaly detection has enjoyed an increasing research and commercial interest.

Targeted at gaining complete and reasonable understandings

Manuscript received September 12, 2023, revised December 31, 2023, accepted June 16, 2024. This work is supported by Beijing Natural Science Foundation (4234085).

Yalong Jiang (corresponding author), Changkang Li and Wenrui Ding are with the Department of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: allenyljiang@buaa.edu.cn).

Jinzhi Xiang is with Beijing Institute of Technology.

Zheru Chi was with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong.

Fig. 1. Demonstration of different categories of actions in the proposed dataset, including typical types of anomalies. Besides stationary platforms, we also collect videos with moving drones. Different lighting conditions are also shown in each row.

about objects' movements, an approach is proposed to effectively learn long-term motion patterns. Besides, a large-scale dataset is built with DJI UAV [46]. Sufficient types of anomalies and diverse scenes including playgrounds, plots, gardens, streets are included. More importantly, the scenarios with drones following people are included, the observations include continuously changing view-points. As is shown by Fig. 1, the horizontal axes denote the variations in scenes and vertical axes show different types of actions.

Targeted at addressing the challenges in human action-related anomaly detection, the contributions of this paper are summarized as follows: Firstly, SSM models based on state machines are proposed for completely representing the motion patterns underlying long-term sequences of actions. The representations are leveraged in distinguishing anomalies through predicting future states. Secondly, the proposed SSM1 model restricts the estimated motion parameters to a plausible subspace, contributing to plausible anomaly detections with SSM2 model. Furthermore, a comprehensive dataset is built for evaluations under diverse scenes and time-variant motion patterns. Extensive experiments are conducted to demonstrate the superiority of the proposed method on both existing datasets and the proposed one. The rest of this article is organized as follows. Section II presents related work. Section III introduces the proposed method. Section IV presents experimental results, including the procedures for building the dataset. Finally, Section V concludes the article.

II. Related Work

A. *Existing Methods based on Reconstruction or Prediction*

Due to the imbalance of surveillance videos, unpredictable nature of anomalies and inaccessibility of annotations, existing approaches are mostly unsupervised. Reconstruction or prediction-based approaches produce larger error on irregular motion patterns than on normal ones. For instance, [13] proposed an encoder-decoder structure for reconstruction with inherent probabilistic modeling on latent feature encodings. [14] augmented encoders with memory modules, improving the sensitivity of reconstruction error to anomalies. [15] and [16] combined LSTM structures with encoder-decoder networks for fall detection. [17] proposed to integrate appearances with motion features. [18] compressed each video into a single frame before augmenting reconstruction loss with sparsity constraints. Other reconstruction-based methods include autoencoders [47] and adversarially learned models [48]. Prediction-based methods evaluate the divergence between the motion patterns in past and future frames. Typical ones include skeleton prediction [4] [5], LSTM-based prediction [6], GAN-based prediction [7], variational autoencoder-based prediction [8] and spatio-temporal two stream models [9]. [10] [11] [19] [12] proposed to combine prediction with reconstruction and built a pool of features for encoding normal dynamics, adapting hyperparameters to new scenes with only a few frames. Although remarkable improvements have been achieved [4], most of the approaches are based on short-term temporal variations which cannot completely describe objects' behaviors. Besides, short-term observations are easily influenced by noises such as occlusions, producing implausible results. As a result, we move toward comprehensive understandings about objects' behaviors and plausible anomaly detections by effectively encoding long-term motion patterns using state machines. Different from [20] which utilizes learned memory items in anomaly detection, our approach introduces invariant temporal basis functions in encoding temporal variations.

B. *Distance-based Methods*

Aside from the above-mentioned approaches, distance-based techniques built similarity metrics between video instances. For instance, clustering-based approaches measured the similarity in subjects' spatio-temporal embeddings [22] [23] [49] [50], the representations are obtained with 3D convolutions [24]. To improve the measurement of similarity,

[42] [49] enlarged the distances between normal and abnormal events while ensuring inner-class compactness, however, it required the labels on anomalies. [43] proposed to model anomaly detection as a one-versus-rest classification task. To tackle complex scenarios, [21] proposed a probabilistic framework for categorizing actions. [44] built a graph connecting different objects in each frame and cluster graphs. [27] was targeted at crowded scenes and integrated collective properties for multi-stage clustering. Differently, we propose to build complete and plausible representations to ensure valid distance computations and reasonable outlier detections.

C. Invariant and Generalizable Feature Representations

To generalize to novel circumstances, meta learning-based methods such as [11] [12] introduced adjustable feature representations which can adapt to new domains. Attention-based methods such as [36] attended to critical and domain-invariant features while reducing the influence of backgrounds. [32] conducted sparse encoding to focus on invariant motion features. To further improve generalization by combining complementary clues, [41] integrated multiple sub-tasks, including moving direction prediction, appearance consistency evaluation and object classification to better align with anomaly detection. [26] introduced casual temporal relations to enhance Multi-instance Learning (MIL)-based approaches [51] and built compact and discriminative representations. [29] [52] [53] applied invariant and robust representations in addressing unexpected feature patterns. Differently, our proposed approach approximates the dynamics of objects with the coefficients of invariant basis functions.

D. State Machine Models for Time-Series Forecasting

Methods such as [54] [55] [56] [57] have exploited the effectiveness of state machines in modeling time series. As is indicated by [58], state machines with certain state transition matrices [54] outperform specialized transformers [59] in forecasting long sequences. According to [54], our approach models temporal dynamics using a sparse set of coefficients in terms of intrinsic and fixed basis functions. The proposed model memorizes long-term normal patterns effectively and generalizes well, exhibiting robustness to short-term domain-shifts and observation noises in anomaly detection.

E. Methods based on Unsupervised Data Augmentation

Some methods have been proposed to generate anomaly labels. Models such as [60] were trained on normal videos and leveraged to generate pseudo-labels on abnormal snippets. [45] employed a generator which was not fully trained to create abnormal samples as supervision. [40] assigned anomaly scores to objects, identifying the outliers as positive samples. [61] focused on infrequent normal samples during generation, harnessing novel sampling strategies.

However, most of the above-mentioned methods are based on frame-level analysis [30] and cannot distinguish abnormal pedestrians, human-level approaches [62] [43] [4] [5] brought improvements but are influenced by occlusions or unexpected spatio-temporal patterns when representing subjects' attributes. Visual features are sensitive to noises. Even if invariant physical rules have been introduced to augment visual features [63], the approaches were limited to simple rigid objects. Differently, the proposed approach encodes long-term temporal dependencies using invariant basis functions, exhibiting robustness to domain shifts and partial observations. The proposed method applies to complex dynamics.

F. Datasets for Human-related Video Anomaly Detection

Existing benchmarks for video anomaly detection are divided into large-scale ones and small-scale ones.

Large-scale Datasets: Typical large-scale benchmarks include ShanghaiTech [32] with 437 videos (856 × 480 resolution) covering running, fighting and other anomalies in 13 scenes, NWPU [64] including scene-dependent anomalies and the anomalies which can be anticipated in 43 scenes, and Ubnormal [65] which includes 29 virtual scenes and detailed annotations on a rich variety of anomalies. Differently, our dataset is captured with drones, covering flexible view-points and rich lighting conditions. Ours is complementary to the above ones. Besides, UCF Crime [36] contains normal and crime-related videos collected from real-world cameras, each video describes a different event, such as robbery, shooting and so on. Since this dataset does not come with frame-level annotations in the training set which includes both normal and abnormal behaviors, we cannot use it for learning the distribution of events. Besides, the resolution of videos is 320 ×240 which is difficult to observe detailed motions. As a result, high-resolution datasets with diversified scenes are beneficial. [64] built a dataset with 4,677 videos, focusing on traffic anomalies. [65] built an aerial video dataset with 87,488 frames, including abnormal geographical conditions. Street Scene [66] contains the anomalies being related to humans, vehicles and animals. The proposed dataset is complementary to Street Scene, the former focuses on detailed human motion while the latter mainly consists of abnormal backgrounds.

Small-scale Datasets: CUHK Avenue [33] involves 16 training and 21 test videos (640 × 360 resolution), respectively. The actions include throwing objects, loitering and running. UCSD [34] is divided into Pedestrian1 (Ped1) and Pedestrian2 (Ped2). Ped1 includes 34 training videos and 36 test videos with 40 irregular events. Ped2 contains 16 training videos and 12 test videos with 12 abnormal events. The Subway dataset [35] is an indoor dataset with unusual events including walking in wrong directions and loitering, the dataset is two hours long. UR Fall [37] contains 70 videos collected in a nursing home with people falling. UMN [38] consists of three crowded scenes with 1,453 frames, 4,144 frames and 2,144 frames, respectively (240×320 resolution). Scattered flee is included. IITB-Corridor [39] has 301,999 training frames and 181,567 test frames (1920 × 1080 resolution), abnormal events include protest, chasing, fighting and so on. However, the above-mentioned datasets are obtained with stationary cameras and contain only limited scenes. To conclude, a comprehensive dataset with practical variations is required to boost the research in this field.

III. Methodology For Anomaly Detection

In this section, we propose SSM (Stacked State Machine) model which encodes the relations between objects' states and the variations in states across long periods. The contributions

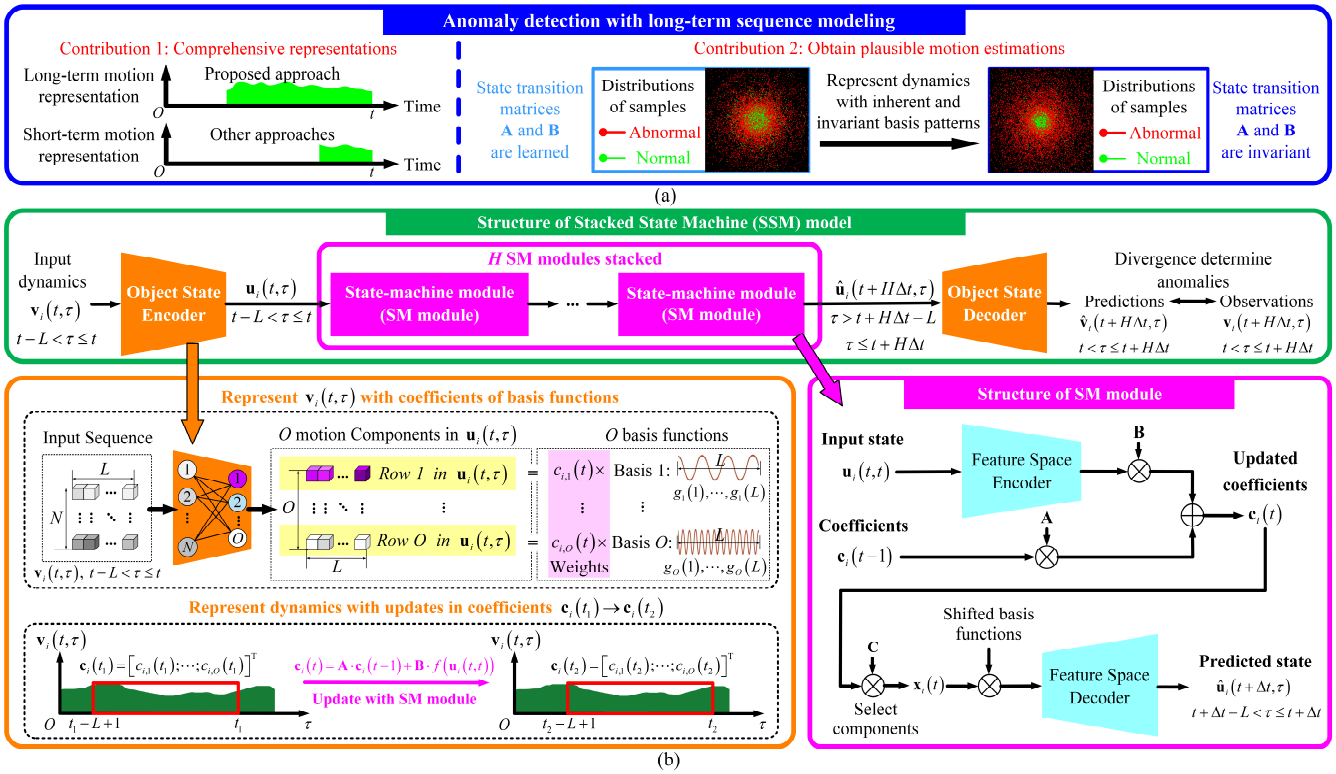


Fig. 2. Illustration of the proposed framework for detecting anomalies. (a) The representations of long-term sequences contribute in two ways. Firstly, more motion patterns can be derived from longer sequences. Secondly, plausible motion estimations are achieved with the representations which are based on fixed basis functions. The red points denote anomalies while green ones show normal actions. Restrictions from invariant basis functions and state transition matrices $\mathbf{A},\mathbf{B}$ contribute to more identifiable anomalies. (b) The structure of SSM model (same for SSM1 and SSM2) in predicting future states (poses and locations) based on past long-term observations. SSM model includes H SM modules each of which makes predictions at Δt in the future. Objects' states $\mathbf{v}_i(t,\tau)$ are encoded with the coefficients with respect to fixed basis functions. The dynamics in states are represented with SM modules in which the variations in coefficients are described by $\mathbf{A}$ and $\mathbf{B}$. SM modules memorize the weights of basis functions in normal dynamics. Finally, the coefficients of are multiplied with temporally shifted basis functions to produce predictions. Anomaly scores are determined by the divergence between predictions and future observations.

are in two folds. Firstly, SSM model extracts the motion patterns which are consistent across long sequences. The patterns are used for predicting future states, the divergence between predictions and future observations is harnessed in determining anomalies, as will be detailed in Section III-A. The learned motion patterns function in filtering out short-term noises and obtaining plausible motion estimations, as will be addressed in Section III-B, Section III-C and Fig. 2.

A. Modeling Long-Range Temporal Dependencies in Actions with State Machines

Different from short-range temporal dependencies which vary across the phases of actions or view-point changes, long-range dependencies provide more comprehensive clues in describing actions. In this method, state machines encode the intrinsic motion patterns governing long-range dynamics.

The proposed SSM model represents dynamics using the coefficients in terms of basis functions. The updates in objects' states are encoded with the updates in coefficients which compose state vectors. The state transition matrices $\mathbf{A}$ and $\mathbf{B}$ are determined by basis functions, $\mathbf{A}$ and $\mathbf{B}$ are invariant for fixed basis functions. The comparison between learned and invariant $\mathbf{A},\mathbf{B}$ is shown by Fig. 2(a). Due to the fact that the divergence between SSM model's predictions and future observations determines anomalies, we visualize the norms and angles of divergence vectors. The red points denote anomalies while green ones show normal actions. The larger the distance between each point and the center is, the more abnormal the point is determined to be. It can be seen that normal and abnormal samples are more separable with fixed basis functions. In learned settings, the lack in restrictions from fixed basis functions leads to unreasonable motion estimations where green and red points are confused.

Encoding states using coefficients of basis functions. Denote $\mathbf{v}_i(t,\tau)\big|_{\tau\in[1,t]} @\left[\mathbf{v}_i(1,1);...;\mathbf{v}_i(t,t)\right]\in ?^{N\times t},\forall i$ the N-D states of Object i from 1 to t, $\mathbf{v}_i(\tau,\tau)\in ¡^{N\times 1},\tau\in[1,t]$ denotes the state at τ. $\mathbf{v}_i(t,\tau)$ is cropped to $\mathbf{v}_i(t,\tau)\big|_{\tau\in[t-L,t]}=\left[\mathbf{v}_i(t-L+1,t-L+1);...;\mathbf{v}_i(t,t)\right]\in ¡^{N\times L}$ by ignoring the early moments with weak correlation to t. $N=34$ for humans' 17 2-D key-points and $N=8$ for the 4 key-points of low-resolution humans and other objects, as will be shown in Section III-B. The Object State Encoder in Fig. 2(b) is a fully-connected layer with input size N and output size O, it decomposes $\mathbf{v}_i(t,\tau)\big|_{\tau\in[t-L,t]}$ to O motion components in $\mathbf{u}_i(t,\tau)\big|_{\tau\in[t-L,t]}=\left[\mathbf{u}_i(t-L+1,t-L+1);...;\mathbf{u}_i(t,t)\right]\in {}^{O\times L}$, $\mathbf{u}_i(\tau,\tau)\in {}^{O\times 1}$.

TABLE I
NOTATIONS IN THE PROPOSED APPROACH

Name	Definition
$\mathbf{v}_i(t,\tau)\big\|_{\tau\in(t-L,t]} \in ¡^{N\times L}$	Sequence of Object i 's key-point coordinates from $t-L+1$ to t
$\mathbf{v}_i(\tau,\tau)\in ¡^{N\times 1}, \tau\in(t-L,t]$	Object i 's N coordinates at τ
$\mathbf{u}_i(t,\tau)\big\|_{\tau\in(t-L,t]} \in ¡^{O\times L}$	Sequence of Object i 's spatial features from $t-L+1$ to t
$\mathbf{u}_i(\tau,\tau)\in ¡^{O\times 1}, \tau\in(t-L,t]$	Object i 's O spatial features at τ
N	Number of Object i 's key-point coordinates
O	Number of fixed basis functions
L	Duration of past observations for making predictions in future K moments
$\left[g_o(1);...;g_o(L)\right],\ o\in[1,O]$	O Length-L distinctive basis functions
$c_{i,1}(t),...,c_{i,O}(t)$	Coefficients of basis functions
$\mathbf{A}$,$\mathbf{B}$	Matrices for updating the coefficients in $\mathbf{c}_i(t)$ based on new observations
$\mathbf{C}$	Matrix for assigning higher weights to normal motion patterns
$\mathbf{x}_i(t)$	Coefficients of selected basis functions for representing normal actions

$\mathbf{u}_i(t,\tau)\big|_{\tau\in(t-L,t]}$ and $\mathbf{v}_i(t,\tau)\big|_{\tau\in(t-L,t]}$ are simplified as $\mathbf{u}_i(t,\tau)$ and $\mathbf{v}_i(t,\tau)$, respectively. The notations are defined in Table I.

As is shown in Fig. 2(b), the $o-th$ row of $\mathbf{u}_i(t,\tau)$ is proportional to the $o-th$ basis function $\left[g_o(1);...;g_o(L)\right]\in ¡^{1\times L}$ which is an invariant Legendre polynomial [67], $o=1,...,O$:

$$\left[u_{i,o}(t-L+1);...;u_{i,o}(t)\right]=c_{i,o}(t)\cdot\left[g_o(1);...;g_o(L)\right] \tag{1}$$

where $u_{i,o}(\tau)$ is the $(o,\tau)-th$ entry in $\mathbf{u}_i(t,\tau)$. The rows in $\mathbf{v}_i(t,\tau)$ which are associated by Object State Encoder to the same row in $\mathbf{u}_i(t,\tau)$ share the same motion pattern. For instance, eyes and nose move in coordination. The coefficients $\mathbf{c}_i(t)=\left[c_{i,1}(t),...,c_{i,O}(t)\right]^{\mathrm{T}}\in ¡^{O\times 1}$ encode $\mathbf{v}_i(t,\tau)$.

Representing dynamics using state machines. At each moment t, the new observation $\mathbf{u}_i(t,t)\in ¡^{O\times 1}$ updates $\mathbf{c}_i(t)$:

$$d\mathbf{c}_i(t)/dt=\mathbf{A}_{HiPPO}\cdot\mathbf{c}_i(t)+\mathbf{B}_{HiPPO}\cdot\sum_{o=1}^{O}u_{i,o}(t) \tag{2}$$

$\mathbf{A}_{HiPPO}\in ¡^{O\times O}$ and $\mathbf{B}_{HiPPO}\in ¡^{O\times 1}$ update state vectors in the O-D state space, they are determined by Legendre basis functions [67] [54]. The way of obtaining $\mathbf{A}_{HiPPO}$, $\mathbf{B}_{HiPPO}$ and the comparisons between different basis functions are shown by Eq. (S24), Eq. (S25) and Table S-I in supplementary materials. $\sum_{o=1}^{O}u_{i,o}(t)$ denotes the summation of all rows in $\mathbf{u}_i(t,t)$. Then we discretize Eq. (2) with $t\in ¢$ being an integer:

$$\mathbf{c}_i(t)=\mathbf{A}\cdot\mathbf{c}_i(t-1)+\mathbf{B}\cdot\sum_{o=1}^{O}u_{i,o}(t) \tag{3}$$

$\mathbf{A}=\left(\mathbf{I}-0.5\mathbf{A}_{HiPPO}\right)^{-1}\left(\mathbf{I}+0.5\mathbf{A}_{HiPPO}\right)$, $\mathbf{B}=\left(\mathbf{I}-0.5\mathbf{A}_{HiPPO}\right)^{-1}\mathbf{B}_{HiPPO}$, the details for obtaining them are provided in Eq. (S22) in appendices. As is shown in Fig. 2(b), the transition from $\mathbf{v}_i(t_1,\tau)$ to $\mathbf{v}_i(t_2,\tau)$ updates coefficients from $\mathbf{c}_i(t_1)$ to $\mathbf{c}_i(t_2)$.

Making prediction with SSM model. State machines are used in modeling the temporal dependencies between past and future states. SSM model is shown by Fig. 2(b), it stacks H SM modules each of which makes predictions in a forthcoming period of Δt. As a result, SSM model predicts $\hat{\mathbf{u}}_i(t+H\Delta t,\tau)$ based on $\mathbf{u}_i(t,\tau)$.

A state machine resides in each SM module, a state vector $\mathbf{c}_i(t)$ with $O=64$ coefficients in Eq. (3) encodes Object i's time-variant states from $t-L+1$ to t. As O grows, more diversified bases enable the generalization to more actions, more normal and abnormal motion patterns in $\mathbf{v}_i(t,\tau)$ can be represented by $\mathbf{c}_i(t)$. At each moment t, the new observation $\mathbf{u}_i(t,t)$ updates $\mathbf{c}_i(t-1)$ to $\mathbf{c}_i(t)$, as is shown in Eq. (3). The transition indicates the dynamics in objects' poses.

SM module learns to make predictions by selecting the basis functions which account for normal temporal variations. For this purpose, $\mathbf{C}\in ¡^{O\times O}$ is proposed to select from $\mathbf{c}_i(t)$ the components that account for the motion patterns in normal actions. $\mathbf{C}$ is learned by training SM module to predict $\hat{\mathbf{u}}_i(\tau,\tau),\tau=t+1,...,t+K$ based on $\mathbf{u}_i(\tau,\tau),\tau=t-L+1,...,t$. The prediction error on abnormal sequences grows because the selected motion patterns cannot approximate abnormal actions.

$$\mathbf{x}_i(t)=\mathbf{C}\cdot\mathbf{c}_i(t) \tag{4}$$

where $\mathbf{C}$ increases the coefficients corresponding to normal motion components, producing $\mathbf{x}_i(t)\in ¡^{O\times 1}$. By combining Eq. (3) and Eq. (4), the relation between $\mathbf{x}_i(t)$ and $\mathbf{u}_i(\tau,\tau)$, $\tau=t-L+1,...,t$ can be obtained as Eq. (5):

$$\begin{aligned}\mathbf{x}_i(t)=\mathbf{C}\cdot\mathbf{A}^{L-1}\cdot\mathbf{B}\cdot\sum_{o=1}^{O}u_{i,o}(t-L+1)+\mathrm{L}\ +\\ \mathbf{C}\cdot\mathbf{A}\cdot\mathbf{B}\cdot\sum_{o=1}^{O}u_{i,o}(t-1)+\mathbf{C}\cdot\mathbf{B}\cdot\sum_{o=1}^{O}u_{i,o}(t)\end{aligned} \tag{5}$$

Denote $\mathbf{K}_L=\left[\mathbf{C}\mathbf{A}^{L-1}\mathbf{B},...,\mathbf{C}\mathbf{A}\mathbf{B},\mathbf{C}\mathbf{B}\right]\in ¡^{O\times L}$, $\mathbf{K}_{L,o}$, $u_{i,o}(\tau)$ and $x_{i,o}(t)$ are the $o-th$ rows of $\mathbf{K}_L$, $\mathbf{u}_i(\tau,\tau)$ and $\mathbf{x}_i(t)$, respectively. $x_{i,o}(t)$ can be obtained with:

$$x_{i,o}(t)=\left\langle\mathbf{K}_{L,o},\left[\sum_{o=1}^{O}u_{i,o}(t-L+1),...,\sum_{o=1}^{O}u_{i,o}(t)\right]\right\rangle,\forall o \tag{6}$$

where $x_{i,o}(t)\in ¡$ is the dot product of two vectors.

To make predictions, $x_{i,1}(t),...,x_{i,O}(t)$ are multiplied with the shifted versions of Legendre basis functions, producing $\hat{\mathbf{u}}_i(t+\Delta t,\tau)$ where each row is the multiplication between $x_{i,o}(t)$ and one shifted basis function. The Feature Space Encoder and Feature Space Decoder in SM module learn the temporal dependencies in $\mathbf{u}_i(t,\tau)$. The output tensor from the SSM model is truncated to produce N-D predictions in future K moments $\hat{\mathbf{v}}_i(t+1,t+1),...,\hat{\mathbf{v}}_i(t+K,t+K)$, $K=H\Delta t$. Detailed formula derivations are in supplementary materials.

Different from transformers and LSTMs [68] [19] where the temporal relations between past and futures states are learned from data, the basis functions in Eq. (1) and corresponding $\mathbf{A}$ and $\mathbf{B}$ for updating states in Eq. (3) are invariant. As is illustrated in [58], appropriate choices of invariant state

transition matrices contribute to the handling of long-range temporal dependencies. In the proposed approach, long-range

sequence modeling exhibits robustness to short-term variations and occlusions. The advantages of invariant $\mathbf{A},\mathbf{B}$ over learned ones will be demonstrated in ablation studies.

The reason for stacking H SM modules instead of one in SSM model is to increase the order in modeling temporal variations. Specifically, the matrices $\mathbf{C}$ in different SM modules are different. The temporal variations from $\mathbf{u}_i(t,\tau)$ to $\mathbf{u}_i(t+\Delta t,\tau)$ and those from $\mathbf{u}_i(t+\Delta t,\tau)$ to $\mathbf{u}_i(t+2\Delta t,\tau)$ are encoded with different basis functions. H is 3 in our implementations. The hyper-parameters of the encoders and decoders in Fig. 2(b) and (c) are shown by Table II:

TABLE II
HYPER-PARAMETERS OF ENCODERS AND DECODERS IN SSM MODEL, FC: FULLY-CONNECTED LAYER

Module name	Layer type	Number of input channels	Number of output channels
Object State Encoder	FC	$N=34$	$O=64$
Object State Decoder	FC	$O=64$	$N=34$
Feature Space Encoder	FC	$L=30$	$L=30$
Feature Space Decoder	FC	$L=30$	$L=30$

B. SSM1 Model for Achieving Plausible Motion Estimations

As is shown in Fig. 2(a), **A** and **B** which are learned from training data lead to implausible results, high anomaly scores are produced on normal samples. To address this issue, SSM1 model with the structure in Fig. 2(b) leverages invariant **A** and **B** in achieving plausible motion representations.

The method for achieving plausible motion estimations is divided into 3 steps. Firstly, tracking is performed to build the trajectory of each Object $i,\forall i$. Secondly, SSM1 model is only trained on training data. As is shown by Fig. 2(b), the model has a matrix $\mathbf{C}$ for selecting normal motion components. We freeze $\mathbf{C}$ to be an identity matrix in SSM1 model to facilitate the representations of potential abnormal motion components. Thirdly, SSM1 model's predictions are averaged with observed states $\mathbf{v}_i(t,\tau)$ for smoothing before feeding into SSM2 during training and testing. Object i 's trajectory $\mathbf{u}_i(t,\tau)$ includes poses and locations. Human poses represent key-point coordinates, including eyes, ears, nose, shoulders, elbows, wrists, hips, knees and ankles. For objects with other categories and low-resolution humans in data [34], the central points of the top, bottom, left and right sides of a bounding box are defined as key-points.

Step1: Tracking. Appearance similarity is leveraged in obtaining objects' trajectories. To measure the similarities between pairs of objects from different frames, each bounding box is divided into 8×8 patches before applying optimal transport [69] in associating the patches from two bounding boxes. The similarity between a pair of boxes is measured by the sum of matching error between associated patches.

Suppose $p_{t_1,h,i}, h=1,...,P$ the $h-th$ block of object i at instant t_1, $p_{t_2,q,j}, q=1,...,Q$ denotes the $q-th$ block of object j at t_2. $p_{t_1,h,i}$ and $p_{t_2,q,j}$ are cropped from the two bounding boxes which surround the key-points in $\mathbf{v}_i(t_1,t_1)$ and those in $\mathbf{v}_j(t_2,t_2)$, respectively. $\mathbf{v}_i(\tau,\tau)\in ¡^{N\times 1}, \tau\in(t-L,t]$ is shown in Table I. $i\in[1,M_1]$ and $j\in[1,M_2]$ where M_1 and M_2 show the numbers of detections at t_1 and t_2, respectively. $\mathbf{W}(i,j,t_1,t_2)\in ¡^{P\times Q}$ denotes the pairing relations between the blocks of Object i and those of Object j, each entry in $\mathbf{W}(i,j,t_1,t_2)$ takes value from $\{0,1\}$. $\mathbf{C}(i,j,t_1,t_2)\in ¡^{P\times Q}$ denotes the matching error between blocks. The overall matching error is minimized with respect to $\mathbf{W}$ to obtain the optimal pairing relations between blocks.

$$\mathbf{W}_*(i,j,t_1,t_2)=\underset{\mathbf{W}(i,j,t_1,t_2)}{\arg\min}\sum_{h=1}^{P}\sum_{q=1}^{Q}\mathbf{W}(i,j,t_1,t_2,h,q)\cdot\mathbf{C}(i,j,t_1,t_2,h,q) \quad (7)$$

$\mathbf{W}(i,j,t_1,t_2,h,q)$ and $\mathbf{C}(i,j,t_1,t_2,h,q)$ are the $(h,q)-th$ entries in $\mathbf{W}(i,j,t_1,t_2)$ and $\mathbf{C}(i,j,t_1,t_2)$, respectively. Define $\mathbf{U}(t_1,t_2)\in ¡^{M_1\times M_2}$ the associations between the objects at t_1 and those at t_2. The $(i,j)-th$ entry $\mathbf{U}(t_1,t_2,i,j)$ takes 1 if Object i and Object j share the same identity and 0 otherwise.

$$\mathbf{U}_*(t_1,t_2)=\underset{\mathbf{U}(t_1,t_2)}{\arg\min}\sum_{i=1}^{M_1}\sum_{j=1}^{M_2}\mathbf{U}(t_1,t_2,i,j)\sum_{h=1}^{P}\sum_{q=1}^{Q}\begin{pmatrix}\mathbf{W}_*(i,j,t_1,t_2,h,q)\cdot\\ \mathbf{C}(i,j,t_1,t_2,h,q)\end{pmatrix} \quad (8)$$

The $\mathbf{W}_*(i,j,t_1,t_2)$ and $\mathbf{U}_*(t_1,t_2)$ are achieved with optimal transport [69]. To improve the performance under crowded cases with overlapping objects, we match Object i 's earlier E observations at $t_1-E+1,...,t_1$ with Object j at t_2:

$$\mathbf{U}_*(t_1,t_2)=\underset{\mathbf{U}(t_1,t_2)}{\arg\min}\sum_{i=1}^{M_1}\sum_{j=1}^{M_2}\mathbf{U}(t_1,t_2,i,j)\sum_{e=0}^{E-1}\sum_{h=1}^{P}\sum_{q=1}^{Q}\begin{pmatrix}\mathbf{W}_*(i,j,t_1-e,t_2,h,q)\cdot\\ \mathbf{C}(i,j,t_1-e,t_2,h,q)\end{pmatrix} \quad (9)$$

j is matched to i if it has the highest similarity to i 's appearances across $E=5$ moments.

Step 2: Pre-training SSM1. Denote SSM1's function as:

$$SSM_1(\mathbf{v}_i(t,\tau))=\hat{\mathbf{v}}_i(t+1,t+1),...,\hat{\mathbf{v}}_i(t+K,t+K) \quad (10)$$

In Object i 's trajectory, SSM1 model predicts the poses at future K moments. SSM1 model shares the same structure as SSM2 model. Different from SSM2 whose matrix $\mathbf{C}$ is learned to select normal motion components, SSM1 model freezes $\mathbf{C}$ to be identity matrix to represent more diversified motion components. During training and testing, SSM1 model pre-processes the input to SSM2 model through averaging.

The inputs of SSM1 firstly undergo normalization to ignore task-irrelevant parameters. In body joint coordinates $\mathbf{v}_i(\tau,\tau)\in ¡^{N\times 1}, \forall\,\tau\in(t-L,t+K]$. The 2-D average of all key-points' coordinates which denotes the center of Object i at τ is duplicated for $(N/2)$ times, producing $\mathbf{l}_i(\tau)\in ¡^{N\times 1}$. To avoid the influences from objects' scales and absolute locations which are irrelevant to anomaly detection, we firstly subtract the earliest offset from each sequence:

$$\mathbf{v}_i(\tau,\tau)\leftarrow\frac{\mathbf{v}_i(\tau,\tau)-\mathbf{l}_i(t-L+1)}{D_i(t-L+1)}, \tau=t-L+1,...,t+K \quad (11)$$

where $\mathbf{l}_i(t-L+1)$ denotes the center coordinate of Object i at the earliest moment, it is subtracted from the coordinates of all

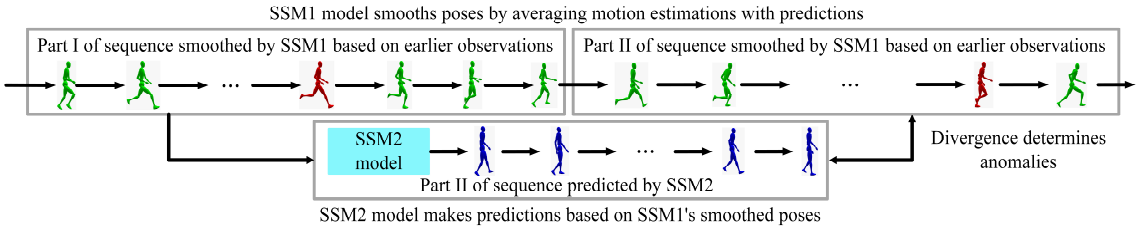


Fig. 3. Illustration of SSM2 model's function in making predictions based on SSM1's smoothed sequences for anomaly detection.

body joints throughout the period, with the relative positions between different moments kept. The coordinates are normalized by $D_i(t-L+1)$ which is the length of the diagonal of Object i's bounding box at $t-L+1$.

Eq. (11) facilitates the joint modeling of displacements and pose variations, the joint modeling outperforms the solely consideration of poses, as will be shown in ablation studies.

Step 3: Smoothing poses. Observation noises, such as occlusions, special viewpoints, lead to implausible pose estimations. SSM1 model's predictions are averaged with pose estimations and locations to filter out noises:

$$\hat{\mathbf{p}}_i(\tau,\tau)=c_i(\tau)\mathbf{v}_i(\tau,\tau)+(1-c_i(\tau))\hat{\mathbf{v}}_i(\tau,\tau),\ \tau=t+1,...,t+K \quad (12)$$

where $\mathbf{v}_i(\tau,\tau)$ denotes the initially estimated pose based on Frame τ, $c_i(\tau)$ is the confidence of its bounding box. When poses cannot be confidently estimated, SSM1 model's predicted $\hat{\mathbf{v}}_i(\tau,\tau)$ has a higher weight. Eq. (12) is performed in a sliding window fashion with window length $L+K$. As is shown by the green sequence in Fig. 3, SSM1 pre-processes the estimated poses before anomaly detection. The red objects show the moments when the object is detected with low confidences (lower than 0.5), SSM1 model corrects the estimated motion parameters before feeding them to SSM2.

C. SSM2 Model for Anomaly Detection

SSM2 model's inputs are pre-processed by SSM1 during training and testing. We model SSM2's function in Eq. (13), $\hat{\mathbf{p}}_i(t,\tau)\in\mathbb{R}^{O\times L}$ denotes the poses smoothed by SSM1 model:

$$SSM_2\left(\hat{\mathbf{p}}_i(t,\tau)\right)=\hat{\mathbf{v}}_i(t+1,t+1),...,\hat{\mathbf{v}}_i(t+K,t+K) \quad (13)$$

$\hat{\mathbf{p}}_i(\tau,\tau),\tau\in(t-L,t+K]$ is the smoothed pose at τ in Object i's trajectory. The anomaly score of i at t is the sum of divergence between $\hat{\mathbf{v}}_i(t+k,t+k)$ and $\hat{\mathbf{p}}_i(t+k,t+k)$ for all k:

$$s_i(t)=\sum_{k=1}^{K}MSE\left(\hat{\mathbf{v}}_i(t+k,t+k),\hat{\mathbf{p}}_i(t+k,t+k)\right) \quad (14)$$

where $MSE(\ )$ denotes mean squared error. In the training phase, SSM1 model learns to make predictions on original data while SSM2 model learns to make predictions on the data smoothed by SSM1 model. In the inference phase, SSM1 model pre-processes input sequences according to Eq. (12). SSM2 model only memorizes the normal motion patterns, the state machines in SSM2 cannot generate realistic pose sequences when observing abnormal input motion patterns, leading to higher prediction error on anomalies. The reason for employing SSM1 for pre-processing is that it memorizes more diversified motion patterns with identity matrix $\mathbf{C}$, producing low prediction error on normal and abnormal sequences. The benefits of pre-processing will be shown in Table VI.

Each score $s_i(\tau),\forall\tau$ is normalized with respect to the number of key-points $N/2$. In each frame, the anomaly score is the maximum one among all objects:

$$s(t)=\max_i s_i(t) \quad (15)$$

IV. Experimental Results

This section presents the procedures for building the dataset as well as the comparisons with existing baseline approaches.

A. Introduction to the Proposed Dataset and Existing Ones

The dataset includes 476 videos collected with DJI Mini 2 UAV [46]. The videos cover multifarious scenes including gardens, squares, streets, grasslands, snowfield, basketball courts and so on. Some videos involve multiple action classes. The resolution is 3840×2160 which facilitates detailed motion representations and region-level analysis [66]. Frame rate is 30 per second. The dataset contains 630,833 frames in total. The durations of video clips in the dataset sum up to 350 minutes with an average clip length of 44 seconds. For each class of action, there are a minimum number of 2 clips, a maximum number of 128 clips and an average number of 36 clips.

Fig. 4 shows the statistics of the dataset. Fig. 4(a) visualizes scene complexity. As is introduced in [70], the number of scenes in a video determines its scene complexity. Specifically, we leverage CLIP [71] for categorizing frames into the 365 classes defined in Places-365 dataset [72], this is achieved by measuring the similarities between frames and scene-related words. Each bar shows the amount of videos with a certain scene complexity. Horizontal axes show different scene complexities. The largest complexity is 29 in our dataset, most videos involve 1 to 10 scenes. When drones chase after people, walking, running and fighting are observed.

Fig. 4(b) shows the proportions of abnormal frames in the videos. The proportions are quantized into bins, each bar shows the percentage of videos with a specific proportion of abnormal frames. It can be seen that a sufficient number of videos are untrimmed with normal and abnormal actions. As a result, the approach applies to long-range analysis where different actions appear in the same sequence.

Fig. 4(c) illustrates the distributions of drones' temporal variations. When drones follow and chase after people, we compute the optical flow vectors [73] on background regions

outside bounding boxes, the average orientation and amplitude (in pixels) of optical flow vectors between each pair of frames represent the drone's direction and velocity at that moment.

Fig. 4(d) visualizes the temporal variations of objects. According to [74], we quantize the optical flow vectors in each bounding box into 8 bins for representing an object. The 8-D representations are visualized using t-SNE [75]. It can be seen that the dataset involves more diversified dynamics and is complementary to ShanghaiTech [32]. Fig. 4(e) visualizes the distributions of anomalies by listing the numbers of frames in each class of action. It can be inferred that the duration of the dataset surpasses existing datasets in most actions.

The dataset is split into a training set, a validation set and a test set which consist of 255,509, 35,623 and 339,701 frames, respectively. The training set involves only normal behaviors, including walking, waving, having picnic. The abnormal behaviors in validation and test sets include running, fighting, falling, robbing, arresting, cycling, playing basketball, playing football and scattered flee. Playing football and basketball are categorized into abnormal actions because they involve adversarial actions which are similar to fighting or running. Frame-level anomaly labels are provided. For instance, the frames with abnormal actions are labeled '1' while those with only normal actions are labeled '0'. For fair comparison, we do not use any annotations on anomalies during training, the validation and test sets are merged into one in our experiments. The validation set can be used by other methods which require anomaly annotations during training.

Two working modes of drones are adopted: hovering and following. In the former mode, the platforms of cameras are stationary while in the latter mode, drones chase after people with view-points changing. Different from existing datasets, remarkable variations in people's scales can be observed. Specifically, we discretize the view-points into three bins: $(30°,45°]\cup[135°,150°)$, $(45°,60°]\cup[120°,135°)$ and $(60°,120°)$, Fig. 4(f) shows the statistics on view-point variations in different actions. To further enrich the dataset, crowded scenes are included with occluded objects. Moreover, the videos are captured at morning, noon and afternoon, different angles between the optical axes of cameras and sunlight are included.

Besides, the proposed approach is also evaluated on existing benchmarks: the CUHK Avenue dataset [33], ShanghaiTech [32], UCSD Ped datasets [34], UR Fall dataset [37] and Street Scene [66] which have been introduced in Section II-E. There are two versions of ShanghaiTech dataset [32] [76], the latter contains the annotations of anomalies in training set. As our approach is targeted at practical applications where no anomaly labels are provided, we use the first version. The anomalies in UCSD data are related to bicycles and cars. The UR Fall dataset includes 40 depth videos with only normal behaviors for training and 30 for test. We convert 1-channel grayscale depth images into 3-channel ones by simply copying.

ROC (Receiver Operation Characteristic) Curve (AUC) is leveraged for evaluation [7] [17]. The metric is computed by continuously changing the threshold of anomaly scores. Then AUC is obtained through cumulating under the ROC curve. A higher AUC value indicates a better performance. We evaluate the performance with two clearly defined frame-level AUCs as metrics: Macro-averaged AUC (macro-AUC) first computes the AUC for each video, then averages the resulting AUCs of videos, and micro-averaged AUC (micro-AUC) first concatenates the scores of all videos and then computes AUC.

B. Implementation Details

We adopt Yolo [77] pretrained on COCO [78] for detecting objects. In computing anomaly scores, temporally sliding windows with stride 1 are adopted. For Object i, the length of sliding window for determining the anomaly score at moment t is $L+K$, as is shown in Eq. (13). Different values of L will be compared in ablation studies. The algorithm runs on an Intel Core i7 processor and NVIDIA RTX-3080 GPU.

For human boxes with longer sides less than 64 pixels, the number of key-points is determined to be 4 based on Section III-B, this is the case in UCSD [34] and Street Scene [66].

Different from other datasets, Street Scene [66] involves low-resolution subjects, anomalies are determined by backgrounds. The visual features of objects' neighboring regions are extracted for representing anomalies. Specifically, patches of two scales 192×192 and 384×384 are cropped from around the center of each detected object. The first scale applies to small-scale objects with longer side of bounding box shorter than 64 pixels, the second scale applies to larger objects with longer side no less than 64. To reduce computational complexity, Principal Component Analysis (PCA) is leveraged in reducing patch sizes from 192×192 and 384×384 to 150 principal components which explain 99% of the variances. Denote $\mathbf{a}_t^i \in ¡^{150}$ the average of Object i's tube of patches across L frame, encoding foregrounds and backgrounds. The principal components of PCA are offline-computed. We crop training images for training PCA module. The anomaly score is shown by Eq. (16):

$$s_i(t) = D\left(\mathbf{a}_t^i\right) - s_{\min}^{train} \Big/ s_{\max}^{train} - s_{\min}^{train} \tag{16}$$

where $D\left(\mathbf{a}_t^i\right)$ is the distance between $\mathbf{a}_t^i$ and its nearest neighbor in training set. $s_{\min}^{train}$ and $s_{\max}^{train}$ are the minimum and maximum distances between training samples and their nearest neighbors in training data, respectively.

The proposed SSM1 model and SSM2 model are trained on training data, each training sample is a Length- $L+K$ sequence with N coordinates at each moment. The learnable parameters of SSM1 reside in encoders and decoders while the learnable weights of SSM2 also reside in $\mathbf{C}$. An input tensor has shape $B\times(L+K)\times N$ where $B=256$ denotes batch size. Pre-training lasts for 15 epochs with Adam optimizer and learning rate 5e-5. The learning rate decay is 0.99, learning rate is multiplied by 0.99 after each epoch.

The inference costs are shown by Table III which includes the average values on the proposed dataset. The corresponding accuracies are in Table VI. Alpha Pose [81] is applied to independently detect the skeleton in each bounding box. The running time in Table III is obtained by implementing the methods on the proposed dataset where more objects reside in each frame than existing datasets. So the speed is slower than those claimed in the papers using the data with less objects. For [80], we run the method for 4 times on the 4 quarters (left-

top, right-top, left-bottom and right-bottom) of each high-resolution frame, else the accuracy collapses.

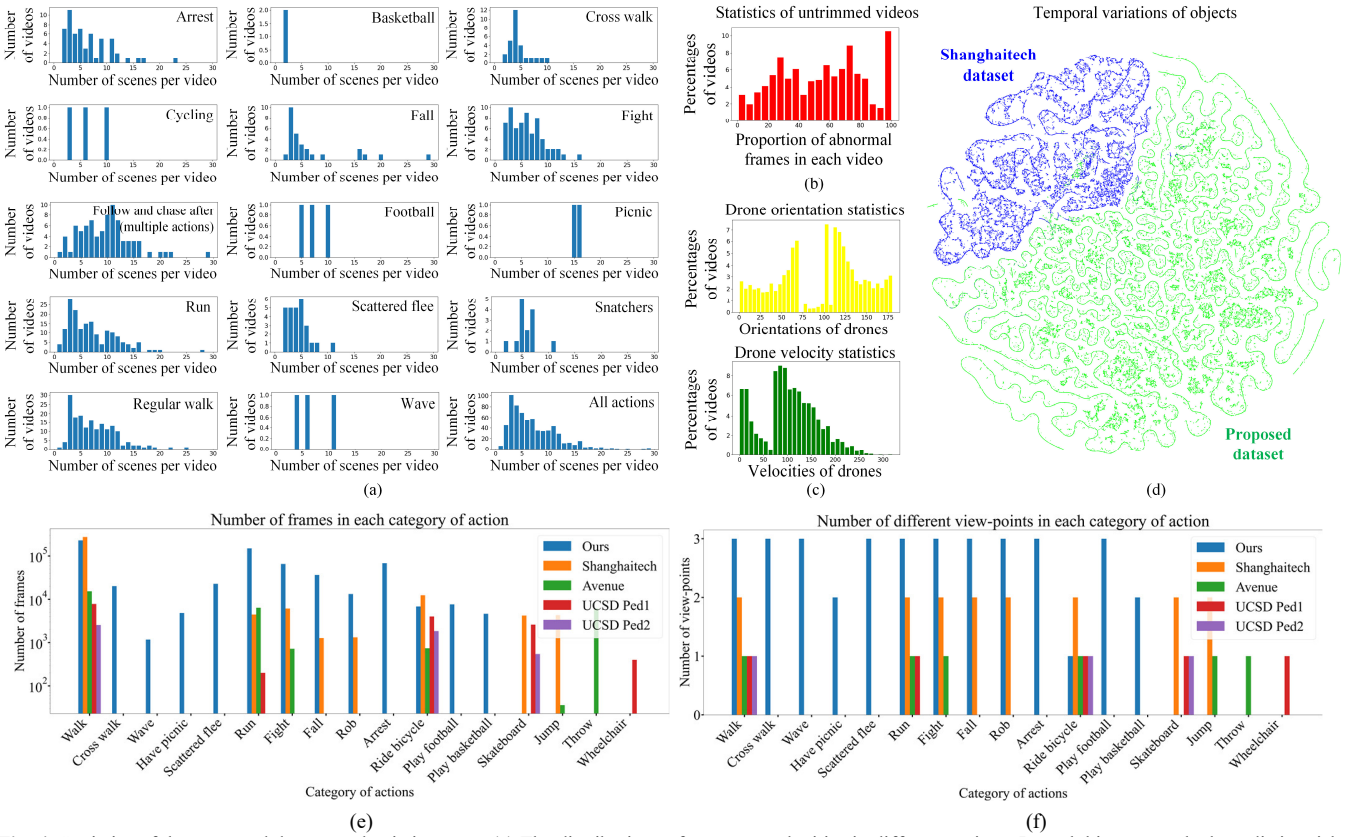


Fig. 4. Statistics of the proposed dataset and existing ones. (a) The distributions of scene complexities in different actions. In each histogram, the bars distinguish videos as to how many scenes are there in each video. (b) Statistics of untrimmed videos. The histogram shows the amounts of videos with different proportions of abnormal frames. (c) Statistics of drones' temporal variations. The two histograms show the proportions of videos captured by drones with different velocities and orientations. (d) Statistics of objects' temporal variations. The distributions are visualized by t-SNE [75]. (e) The numbers of frames in each category of action. (f) The numbers of view-points in each category of action.

TABLE III
PROCEDURES AND CONSUMPTIONS IN INFERENCE PHASE
(RUNNING THROUGHOUT THE PERIOD $[t-L+1, t+K]$)

	Name of operation / Approach	Time Consumption on average (per frame, ms)	GPU Memory Consumption (Gigabytes)
Ours	Detection, pose estimation and tracking	22.0824	2.603
	Pre-processing with SSM1	3.46	0.883
	Prediction with SSM2	3.46	0.883
Other	Solve puzzles [79]	45.14	2.861
	Restoration [80]	21.23	5.214
	Multi-Task [41]	56.49	3.163

C. Quantitative Results on Existing Benchmarks

The proposed approach is compared with other methods on existing benchmarks, as is shown in Table IV. The baselines include reconstruction-based or prediction-based approaches such as [4] [14], transfer-learning based ones [40], clustering-based ones [82], and multi-task approaches such as [41]. It can be observed that the proposed state machine-based approach outperforms existing ones. Specifically, the accuracy in Table IV is achieved with $L=30$. In the early moments when Object i just appears with trajectory shorter than 30, L equals to the length of trajectory. Efficiency decreases as L grows.

Besides, the comparison between the third from last row and the last row shows the advantage of coupling pose variations with the relative positions between observations of the same object at different moments, as is shown in Eq. (11). According to [83], velocities indicate anomalies, the coupling provides more comprehensive representations. The comparison between the last two rows shows the contribution of more basis functions with more diversified motion patterns.

Additional experiments are conducted on other datasets such as IITB-Corridor [39] which characterizes semi-outdoor scenes with complex lighting conditions. Table V shows the advantages over existing state-of-the-art methods. On default, the representations of low-resolution and non-human objects require the relative offsets between observations because pose-related features cannot be accurately and confidently derived.

D. Evaluation on the Proposed Dataset

As is shown by Table VI, the proposed approach outperforms eight other methods on the proposed dataset. The results are obtained with our re-implementation based on the settings addressed in the papers.

E. Ablation Study

Ablation studies are shown in Table VI. Firstly, the influences of L is illustrated. On the benchmarks as well as the proposed dataset, a period of time in video that spans 30 frames covers rich dynamics. At the early moments of Object i's appearances, trajectory length is less than L. From the moment when Object i's sequence of observations is longer

TABLE IV
FRAME-LEVEL AREA UNDER THE CURVE (%) COMPARISON ON EXISTING BENCHMARK DATASETS

	Algorithm	ShanghaiTech		Avenue		UCSD Ped2	
Reconstruction or prediction	Dynamics [10]	73.8	-	89.5	-	96.9	-
	Normal Graph [5]	74.1	-	87.3	-	-	-
	Memorizing [14]	72.8	-	84.9	-	95.4	-
	Frame Predict [7]	72.8	-	85.1	-	95.4	-
	Object-centric [43]	84.9	-	90.4	-	97.8	-
	Solve Puzzles [79]	84.3	-	92.2	-	99.0	-
	Regularity [4]	73.4	-	-	-	-	-
	Stacked RNN [32]	-	-	81.7	-	92.2	-
	Old is Gold [45]	-	-	-	-	98.1	-
	Hierarchical [84]	82.9	-	88.8	-	-	-
Transfer / Meta	Latent Space [13]	72.5	-	-	-	95.4	-
	Continual [40]	71.6	-	86.4	-	97.8	-
	Anomaly3D [24]	80.6	-	89.2	-	95.8	-
	Few-shot [11]	77.9	-	85.8	-	96.2	-
	Memory for Prediction [12]	70.5	-	88.5	-	97.0	-
Clustering	Clustering [17]	73.3	-	86.0	-	96.5	-
	Margin Learn [42]	76.8	-	89.2	-	-	-
	Scene-Aware [44]	74.7	-	89.6	-	-	-
	Graph-embed [82]	76.1	-	-	-	-	-
	Aggregation [85]	74.7	-	89.9	-	97.6	-
Multi-task	Multi-Task [41]	-	90.2	-	92.8	-	99.8
	Self-supervised [86]	-	89.5	-	92.9	-	-
	AI-VAD [74]	85.1	89.6	93.3	96.2	99.1	99.9
	Ubnormal [87]	-	90.5	-	93.2	-	-
	Agnostic [88]	-	89.3	-	92.3	-	-
	AnomalyNet [18]	-	-	-	-	94.9	-
	Semantic Contrast [62]	83.4	-	93.7	-	98.1	-
Ours	Ours without relative offsets in Eq. (11)	87.3	90.4	93.0	95.8	-	-
	Ours with $O=32$ Eq. (2)	88.3	91.2	92.6	95.6	98.8	99.3
	Ours	**89.1**	**92.1**	**93.8**	**96.7**	99.2	99.8

than L, we only keep the most-updated L observations. It can be seen that a larger L improves SSM model's capability in representing the motion patterns governing objects' actions. In comparison with short-term settings, long-term sequence modeling introduces longer basis functions. According to Section III-A, each motion sequence is represented with the weighted versions of O length-L basis functions as motion components. As L grows, the distinctiveness between components is enhanced. For instance, different basis functions can only be orthogonal to each other when $L \geq O$. Additionally, more distinctive basis functions facilitate more comprehensive representations about the diversified motion patterns in actions.

The comparison between state machines and transformers is also included in Table VI. Specifically, the transformer proposed in [68] is used for replacing SSM1 and SSM2 models in predicting objects' states at future K moments based on past L observations. It can be seen that state machines outperform transformers in representing the time-invariant patterns in sequences and detecting outliers that violate the patterns, especially under long-term observations. Besides, the **A** and **B** which are invariant facilitate state machines in learning the intrinsic temporal dependencies that generalize to different scenarios. Moreover, the comparison between different H shows the contribution of an appropriate number of SM modules in SSM models.

TABLE V
MICRO-AUC, TBDR AND RBDR (%) ON OTHER DATASETS

Data	Approach	AUC	
IITB-Corridor	Frame Prediction [7]	64.65	
	Learning Regularity [4]	64.27	
	Multi-timescale [39]	67.12	
	PoseCVAE [89]	67.34	
	Ours with $L=30$	72.83	
UCSD Ped1	Dynamics [10]	85.1	
	Frame Prediction [7]	83.1	
	Few-shot [11]	86.3	
	Ours with $L=30$	87.3	
UR Fall	Deepfall [16]	89	
	Few-shot [11]	90.6	
	Ours with $L=30$	92.4	
Street Scene	Approach	TBDR	RBDR
	Street Scene [66]	53	21
	Ours with $L=30$	55	24

In the fourth from last row, it can be seen that if **A** and **B** are learned from data, performance drops. In the third from last row of Table VI, it can be seen that SSM1 model functions in filtering out unreasonable pose estimations and improves performance with plausible motion estimations. In all but the last two rows in Table VI, $E=1$ in Eq. (9), the comparison between different E indicates that the consideration of appearances at multiple moments improves anomaly detection accuracy by enhancing the tracker. Besides, a more complex tracker [90] contributes to improvements.

F. Evaluation Under Unsupervised Settings

In this section, the pre-trained parts in the proposed method, including object detector and pose estimator, are replaced with unsupervised counterparts. Specifically, the reconstruction model [12] is introduced to represent each frame, as is shown by "Recon." in Table VII. The model encodes each of the $I=1024$ locations (32 horizontal and 32 vertical grids) in one frame with $\mathbf{q}_i(t,t) \in \mathbb{R}^{1024}, i \in [1,I]$ at t. $\mathbf{q}_i(t,t)$ is obtained by concatenating a 512-D input embedding with a 512-D vector which represents the memory acquired from training data [12]. The I vectors function in re-constructing each input frame.

Then we use SSM models in representing the dynamics in $\mathbf{q}_i(t,t), \forall i$ at each location i, this setting is shown by "Recon. + SSM" in Table VII. The input shape of Object State Encoder and output shape of Object State Decoder are reset to 1024. $\hat{\mathbf{q}}_i(\tau,\tau)$ and $\hat{\mathbf{q}}_i(\tau,\tau)$ denote the input and output of SSM2 model, in the same fashion as Eq. (14).

$$s(t) = \max_{i \in [1,I]} \left(\sum_{k=1}^{K} MSE\left(\hat{\mathbf{q}}_i(t+k,t+k), \mathbf{q}_i(t+k,t+k)\right) \right) \quad (17)$$

As is shown by Eq. (17), the error in predicting sequence $\mathbf{q}_i(t+1,t+1),...,\mathbf{q}_i(t+K,t+K)$ based on $\mathbf{q}_i(t-L+1,t-L+1),...,\mathbf{q}_i(t,t)$

TABLE VI
COMPARISONS OF AUC(%) ON THE PROPOSED DATASET, †: OUR RE-IMPLEMENTATIONS

Algorithm	Proposed Dataset		ShanghaiTech (micro and macro AUC, %)		Avenue	
Dynamics [10] †	61.8	65.1	73.8	-	89.5	-
Object-centric [43] †	64.0	67.4	84.9	-	90.4	-
Self-supervised Multi-Task [41] †	72.9	76.9	-	90.2	-	92.8
Restoration [80] †	63.7	66.7	73.8	-	89.9	-
Attribute-based [74] †	71.2	74.9	85.1	89.6	93.3	96.2
Solve puzzles [79] †	70.4	74.0	84.3	-	92.2	-
NWPU [91] †	68.8	72.4	79.2	-	86.8	-
Normalizing flows [92] †	67.2	70.3	76.3	-	88.0	-
Ours with $L=5$	70.5	73.2	84.3	87.2	91.5	93.3
Ours with $L=10$	74.3	77.2	87.6	90.6	93.0	95.9
Ours with $L=15$	74.8	77.6	88.0	90.9	93.2	96.2
Ours with $L=20$	75.1	77.9	88.2	91.2	93.4	96.3
Ours with $L=25$	75.9	78.8	88.7	91.6	93.5	96.5
Ours with $L=40$	76.2	79.2	89.1	92.0	93.6	96.5
Ours with SSM models being replaced by transformers, $L=10$	74.3	77.1	87.7	90.7	93.0	95.9
Ours with SSM models being replaced by transformers, $L=20$	75.2	77.9	88.1	91.1	93.3	96.2
Ours with SSM models being replaced by transformers, $L=30$	73.9	76.7	86.9	90.0	92.9	95.8
Number of SM modules in SM model $H=1$	73.9	76.8	86.7	89.6	92.6	95.5
Number of SM modules in SM model $H=2$	75.4	78.3	88.5	91.4	93.3	96.2
Ours with SSM1 model and SSM2 model, $L=30$	76.3	79.2	89.1	92.1	93.8	96.7
Ours with **A** and **B** being learned from training data	74.1	76.8	86.8	89.8	92.5	95.4
Ours without SSM1 model, $L=30$	75.8	78.7	88.7	91.7	93.5	96.5
Ours combining multiple moments for tracking, $E=5$ in Eq. (12)	76.5	79.4	89.2	92.2	93.8	96.7
Ours with FairMOT [90] for tracking	76.9	79.8	89.4	92.3	93.9	96.8

TABLE VII
MICRO-AUC (%) ON SHANGHAITECH WITH UNSUPERVISED SETTINGS

Settings	Recon. [12]	Recon. + SSM	Detect. + Track. + Recon. + SSM
AUC	69.83	79.26	84.86
Weights (Million: 10^6)	16.41	16.55	78.45
Complexity (BFLOPS: 10^9 FLOPS)	4.22	8.61	159.53
Inference time (ms / frame)	16.39	33.70	44.86

using SSM2 model determines anomaly scores. Additionally, we have also experimented with "Detect. + Track. + Recon. + SSM" in Table VII, the modules for detection and tracking are kept. The region inside each bounding box is split into $I=4$ parts (2 horizontal and 2 vertical grids) each of which is encoded by a vector $\mathbf{q}_i(t,t)\in ¡^{1024}, i\in[1,I]$ at moment t. The reconstruction model [12] and SSM models run on each object.

Noticeably, the proposed SSM models with distinctive and fixed basis functions boost the performance of reconstruction model because the inherent properties of temporal variations in $\mathbf{q}_i(t,t)$ are well represented. Better performance can be achieved with better unsupervised feature extractors.

G. Subjective Results

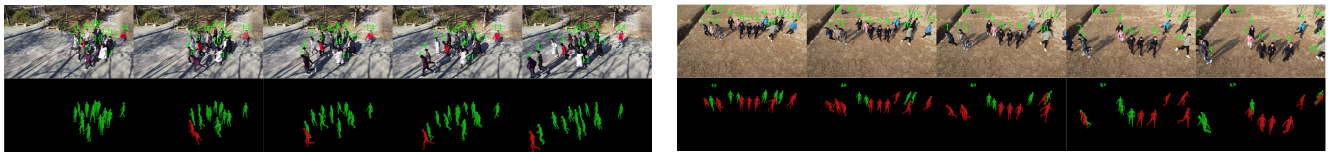
Some subjective results are shown in Fig. 5 where abnormal people are marked in red in the second rows of (a) and (b), the people with normal behaviors are green-marked, the numbers in the first rows denote identities. Abnormal behaviors in crowded scenes can be detected. Moreover, the quantitative results in Fig. 6 illustrate the contributions of different components. For instance, the combination of poses with displacements outperforms the solely modeling of poses. A person who has velocities with respect to grounds but no pose variations cannot be conducting normal actions such as standing or walking. Besides, the advantages of state machines and long sequences are shown. The involvement of SSM1 model for pre-processing also contributes to improvements.

H. Limitations

Since this approach relies on spatial features which are influenced by resolutions, remote objects, especially partially occluded ones, inevitably degenerate the performance, as is illustrated in Fig. 7. In the future, the approach will encode structural and semantic features which are resolution-invariant.

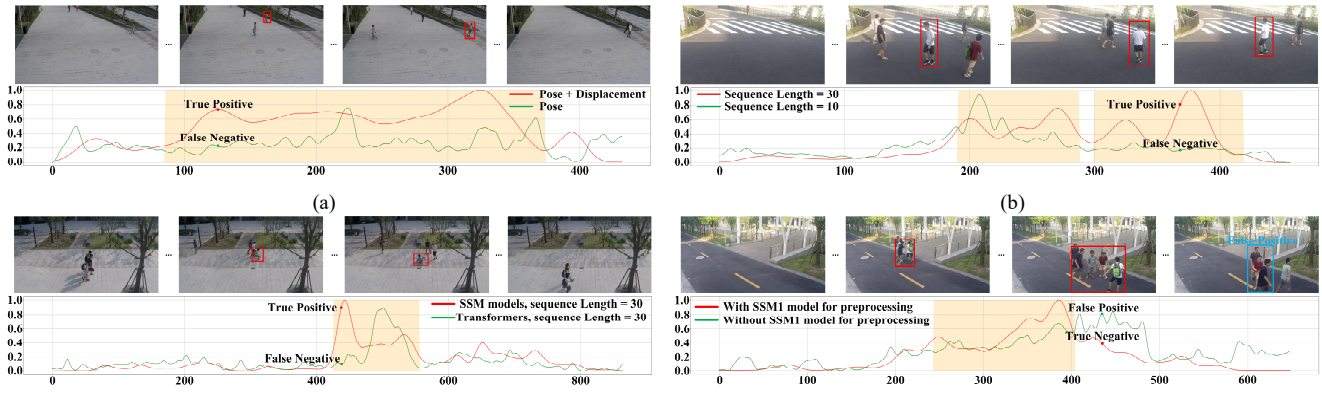
Fig. 7. Limitations of the approach on occluded low-resolution objects.

Besides, the proposed approach only focuses on the dynamics of foreground objects while ignoring backgrounds on the proposed dataset. For instance, adversarial actions such as playing basketball should be normal ones on basketball courts. As a result, our future work will move towards scene-dependent approaches [91] where anomaly detections are conditioned on scenes. Specifically, $\mathbf{v}_i(t,\tau)$ will be enriched to describe not only objects' actions but also scene-related contexts. The annotations of the proposed dataset will be refined to include scene-dependent anomalies.



(a) (b)

Fig. 5. Two examples demonstrating the results. The two rows in each example denote input images and predicted anomalies. In the second row of each example, humans in red are conducting abnormal behaviors. (a) Running. (b) Scattered flee.



(a) (b) (c) (d)

Fig. 6. Examples of anomaly detections. (a) The joint modeling of pose variations and displacements contributes to the accurate detection of skate-boarding which shows irregular relations between poses and displacements. (b) The modeling of long sequences enables SSM2 model to comprehensively describe human motion, loitering can be detected in this way. (c) State machines outperform transformers in extracting long-term motion patterns governing normal actions. Irregular sequences violating the patterns can be detected. (d) Mutual occlusions occur when humans are overlapped with each other, leading to implausible pose estimations and false alarms. SSM1 model filters out problematic pose estimations with learned motion patterns, contributing to plausible anomaly detections.

V. Conclusion

In this article, an approach is proposed to base anomaly detections on long-term sequence modeling. A framework built upon state machines is proposed for modeling the long-term dependencies which comprehensively describe actions. The motion patterns which are consistent across long periods function in smoothing motion estimations and obtain reasonable anomaly detections with prediction error. A large-scale drone-based dataset for anomaly detection is also built, it consists of self-collected 476 high-resolution videos with 630,833 frames. The proposed approach outperforms state-of-the-art methods on both existing benchmarks and the proposed dataset. The approach generalizes to non-human objects, including vehicles, bicycles and so on. In the future, more diversified basis functions will be explored to represent more types of anomalies.

References

[1] Lin, X., Chen, Y., Li, G. and Yu, Y., "A Causal Inference Look At Unsupervised Video Anomaly Detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022.

[2] Ergen, T. and Kozat, S. S. , "Unsupervised anomaly detection with LSTM neural networks," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 31, no. 8, pp. 3127-3141, 2019.

[3] Zhang, Y., Nie, X. , He, R., Chen, M. and Yin, Y., "Normality Learning in Multispace for Video Anomaly Detection," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 31, no. 9, pp. 3694-3706, 2020.

[4] Morais, R., Le, V., Tran, T., Saha, B., Mansour, M. and Venkatesh, S., "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11996-12004.

[5] Luo, W., Liu, W. and Gao, S., "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing,* vol. 444, pp. 332-337, 2021.

[6] Shi, X., Chen, Z., Wang, H., Yeung, D.Y. and Wong, W. C., "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *arXiv preprint* , 2015, arXiv:1506.04214.

[7] Liu, W., Luo, W., Lian, D. and Gao, S., "Future frame prediction for anomaly detection--a new baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6536--6545.

[8] Kingma, D. P. and Welling, M. , "Auto-encoding variational bayes," in *arXiv preprint* , 2013, arXiv:1312.6114.

[9] Nguyen, T. N. and Meunier, J., "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1273-1283.

[10] Lv, H., Chen, C., Cui, Z., Xu, C., Li, Y. and Yang, J., "Learning Normal Dynamics in Videos with Meta Prototype Network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15425-15434.

[11] Lu, Y., Yu, F., Reddy, M.K.K. and Wang, Y., "Few-Shot Scene-Adaptive Anomaly Detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 125-141.

[12] Park, H., Noh, J. and Ham, B., "Learning memory-guided normality for anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372-14381.

[13] Abati, D., Porrello, A., Calderara, S. and Cucchiara, R., "Latent space autoregression for novelty detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481-490.

[14] Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S. and Hengel, A. V. D., "Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705-1714.

[15] Nogas, J., Khan, S. S. and Mihailidis, A., "Fall detection from thermal camera using convolutional lstm autoencoder," in *Proceedings of the 2nd workshop on Aging, Rehabilitation and Independent Assisted Living, IJCAI Workshop*, Jul. 2018.

[16] Nogas, J., Khan, S. S. and Mihailidis, A., "Deepfall: non-invasive fall detection with deep spatio-temporal convolutional autoencoders," *J. Health. Inform.,* vol. 4, no. 1, pp. 50-70, 2020.

[17] Chang, Y., Tu, Z., Xie, W. and Yuan, J. , "Clustering Driven Deep Autoencoder for Video Anomaly Detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2020, pp. 329-345.

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

[19] Liu, Z., Nie, Y., Long, C., Zhang, Q. and Li, G., "A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13588-13597.

[20] Gao, J., Zhong, B. and Chen, Y., "Robust Tracking via Learning Model Update With Unsupervised Anomaly Detection Philosophy," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 33, no. 5, pp. 2330-2341, 2022.

[21] J. Li, Q. Huang, Y. Du, X. Zhen, S. Chen and L. Shao, "Variational Abnormal Behavior Detection With Motion Consistency," *IEEE Trans. Image Process.,* vol. 31, pp. 275-286, 2022.

[22] Liu, A. A., Su, Y. T., Nie, W. Z. and Kankanhalli, M., "Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 39, no. 1, pp. 102-114, 2017.

[23] Elkholy, A., Hussein, M. E., Gomaa, W., Damen, D. and Saba, E., "Efficient and robust skeleton-based quality assessment and abnormality detection in human action performance," *IEEE J. Biomed. Health Inform.,* vol. 24, no. 1, pp. 280-291, 2019.

[24] Asad, M., Yang, J., Tu, E., Chen, L. and He, X., "Anomaly3D: Video anomaly detection based on 3D-normality clusters," *J. Vis. Commun. Image Represent.,* vol. 75, no. 103047, 2021.

[25] Chen, C., Xie, Y., Lin, S., Yao, A., Jiang, G., Zhang, W. and Ma, L., "Comprehensive Regularization in a Bi-directional Predictive Network for Video Anomaly Detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2022.

[26] Wu, P. and Liu, J., "Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection," *IEEE Trans. Image Process.,* vol. 30, pp. 3513-3527, 2021.

[27] X. Li, M. Chen and Q. Wang, "Quantifying and Detecting Collective Motion in Crowd Scenes," *IEEE Trans. Image Process.,* vol. 29, pp. 5571-5583, 2020.

[28] Liu, T., Lam, K. M., Zhao, R. and Qiu, G., "Deep Cross-Modal Representation Learning and Distillation for Illumination-Invariant Pedestrian Detection," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 32, no. 1, pp. 315-329, 2022.

[29] Guo, C., Wang, H., Xia, Y. and Feng, G., "Learning Appearance-Motion Synergy via Memory-guided Event Prediction for Video Anomaly Detection," *IEEE Trans. Circuits Syst. Video Technol.,* 2023.

[30] Zaheer, M. Z., Mahmood, A., Astrid, M. and Lee, S. , "CLAWS: Clustering Assisted Weakly Supervised Learning with Normalcy Suppression for Anomalous Event Detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 358-376.

[31] Zhou, J. T., Zhang, L., Fang, Z., Du, J., Peng, X. and Xiao, Y., "Attention-driven loss for anomaly detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 30, no. 12, pp. 4639-4647, 2019.

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

[33] Lu, C., Shi, J. and Jia, J., "Abnormal event detection at 150 fps in matlab," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2720--2727.

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

[35] Adam, A., Rivlin, E., Shimshoni, I. and Reinitz, D., "Robust real-time unusual event detection using multiple fixed location monitors," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 30, no. 3, pp. 555-560, 2008.

[36] Sultani, W. , Chen, C. and Shah, M., "Real-world Anomaly Detection in Surveillance Videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6479-6488.

[37] Kwolek, B. and Kepski, M., "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Comput. Methods Programs Biomed.,* vol. 117, no. 3, pp. 489-501, 2014.

[38] R. Mehran, A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 935-942.

[39] Rodrigues, R., Bhargava, N., Velmurugan, R. and Chaudhuri, S., "Multitimescale trajectory prediction for abnormal human activity detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2020, pp. 2626-2634.

[40] Doshi, K. and Yilmaz, Y., "Continual learning for anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 254-255.

[41] Georgescu, M. I., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M. and Shah, M., "Anomaly Detection in Video via Self-Supervised and Multi-Task Learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12742-12752.

[42] Liu, W., Luo, W., Li, Z., Zhao, P. and Gao, S. , "Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies," in *Proc. Int. Jt. Conf. Artif. Intell. (IJCAI)*, Aug. 2019, pp. 3023-3030.

[43] Ionescu, R. T., Khan, F. S., Georgescu, M. I. and Shao, L., "Object-centric Auto-encoders and Dummy Anomalies for Abnormal Event Detection in Video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7842-7851.

[44] Sun, C., Jia, Y., Hu, Y. and Wu, Y., "Scene-Aware Context Reasoning for Unsupervised Abnormal Event Detection in Videos," in *Proceedings of the 28th ACM International Conference on Multimedia*, Oct. 2020, pp. 184-192.

[45] M. Z. Zaheer, J. H. Lee, M. Astrid and S. I. Lee, "Old is Gold: Redefining the Adversarially Learned One-Class Classifier Training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14183-14193.

[46] Stanković, M., Mirza, M. M. and Karabiyik, U., "UAV Forensics: DJI Mini 2 Case Study," *Drones,* vol. 5, no. 2, p. 49, 2021.

[47] Guo, X., Gao, L., Liu, X. and Yin, J., "Improved deep embedded clustering with local structure preservation," in *Proc. Int. Jt. Conf. Artif. Intell. (IJCAI)*, Aug. 2017, pp. 1753-1759.

[48] Sabokrou, M., Khalooei, M., Fathy, M. and Adeli, E., "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3379-3388.

[49] Sun, Q., Liu, H. and Harada, T. , "Online growing neural gas for anomaly detection in changing surveillance scenes," *Pattern Recognition,* vol. 64, pp. 187-201, 2017.

[50] Dhiman, C. and Vishwakarma, D. K., "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *IEEE Trans. Image Process.,* vol. 29, pp. 3835-3844, 2020.

[51] Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W. and Carneiro, G., "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4975-4986.

[52] Li, J., Xiong, C. and Hoi, S. C., "Learning from noisy data with robust representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9485-9494.

[53] Feng, C. and Hu, P., "Learning Invariant Rules from Data for Interpretable Anomaly Detection," in *arXiv preprint*, 2022, arXiv: 2211.13577.

[54] Gu, A., Dao, T., Ermon, S., Rudra, A. and Ré, C., "Hippo: Recurrent memory with optimal polynomial projections," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, 33, 1474-1487.

[55] Pavlova, U. and Rakitskiy, A., "Time Series Forecasting Method Based on Finite State Machine," in *IEEE 22nd International Conference of Young Professionals in Electron Devices and Materials (EDM)*, 2021, pp. 533-536.

[56] Rangapuram, S.S., Seeger, M.W., Gasthaus, J., Stella, L., Wang, Y. and Januschowski, T., "Deep state space models for time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2018, 31.

[57] Lin, Y., Koprinska, I. and Rana, M., "SSDNet: State space decomposition neural network for time series forecasting," in *IEEE International Conference on Data Mining (ICDM)*, 2021, pp. 370-378.

[58] Gu, A., Goel, K. and Ré, C., "Efficiently Modeling Long Sequences with Structured State Spaces," in *arXiv preprint*, 2021, arXiv:2111.00396.

[59] Sukhbaatar, S., Grave, E., Bojanowski, P. and Joulin, A., "Adaptive attention span in transfomers," in *arXiv preprint*, 2019, arXiv:1905.07799.

[60] Yu, F., Zhang, M., Dong, H., Hu, S., Dong, B. and Zhang, L., "DAST: Unsupervised Domain Adaptation in Semantic segmentation based on discriminator attention and self-training," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, Vol. 35, No. 12, p. 10.

[61] Lim, S.K., Loo, Y., Tran, N.T., Cheung, N.M., Roig, G. and Elovici, Y., "Doping: Generative data augmentation for unsupervised anomaly detection with gan," in *IEEE international conference on data mining (ICDM)*, 2018, pp. 1122-1127.

[62] Sun, S. and Gong, X., "Hierarchical Semantic Contrast for Scene-aware Video Anomaly Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22846-22856.

[63] Ding, M., Chen, Z., Du, T., Luo, P., Tenenbaum, J. and Gan, C., "Dynamic visual reasoning by learning differentiable physics models from video and language," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2021, 34.

[64] Yao, Y., Wang, X., Xu, M., Pu, Z., Atkins, E. and Crandall, D., "When, Where, and What? A New Dataset for anomaly detection in driving videos," in *arXiv preprint* , 2020, arXiv:2004.03044.

[65] Jin, P., Mou, L., Xia, G. S. and Zhu, X. X., "Anomaly Detection in Aerial Videos With Transformers," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-13, 2022.

[66] Ramachandra, B. and Jones, M., "Street Scene: A new dataset and evaluation protocol for video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 2569-2578.

[67] Arfken, George B., Hans J. Weber and Frank E. Harris., Mathematical methods for physicists: a comprehensive guide, Academic press, 2011.

[68] Wu, N., Green, B., Ben, X. and O'Banion, S., "Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case," in *arXiv preprint*, 2020, arXiv:2001.08317.

[69] Motta, D., Casaca, W. and Paiva, A., "Vessel optimal transport for automated alignment of retinal fundus images," *IEEE Trans. Image Process.,* vol. 28, no. 12, pp. 6154-6168, 2019.

[70] Yoon, S., Koo, G., Kim, D. and Yoo, C. D., "SCANet: Scene Complexity Aware Network for Weakly-Supervised Video Moment Retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023, pp. 13576-13586.

[71] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I., "Learning Transferable Visual Models From Natural Language Supervision," in *PMLR*, 2021, pp. 8748-8763.

[72] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. and Torralba, A., "Places: A 10 million Image Database for Scene Recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 40, no. 6, pp. 1452-1464, 2017.

[73] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A. and Brox, T., "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 2462-2470.

[74] Reiss, T. and Hoshen, Y., "Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection," in *arXiv preprint* , 2022, arXiv:2212.00789.

[75] Van der Maaten, L. and Hinton, G., "Visualizing data using t-sne," *Journal of machine learning research,* vol. 9, no. 11, 2008.

[76] Zhong, J., Li, N., Kong, W., Liu, S., Li, T. H. and Li, G., "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1237–1246.

[77] Redmon, J. and Farhadi, A., "Yolov3: An incremental improvement," in *arXiv preprint* , 2018, arXiv:1804.02767 .

[78] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan and C. L. Zitnick, "Microsoft coco: Common objects in context.," in *Proc. ECCV, Aug.*, 2014.

[79] Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X. and Huang, D., "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles," in *arXiv preprint*, 2022, arXiv:2207.10172.

[80] Yang, Z., Liu, J., Wu, Z., Wu, P. and Liu, X., "Video event restoration based on keyframes for video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14592–14601.

[81] Fang, H. S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y. L. and Lu, C., "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 6, pp. 7157-7173, 2023.

[82] Markovitz, A., Gilad, S., Itamar, F., Lihi, Z. M. and Shai, A., "Graph embedded pose clustering for anomaly detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10539--10547.

[83] Wu, C., Shao, S., Tunc, C., Satam, P. and Hariri, S., "An explainable and efficient deep learning framework for video anomaly detection," in *Cluster Computing*, 2021, pp. 1-23.

[84] Huang, C., Liu, Y., Zhang, Z., Liu, C., Wen, J., Xu, Y. and Wang, Y., "Hierarchical graph embedded pose regularity learning via spatio-temporal transformer for abnormal behavior detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 307-315.

[85] Yang, Z., Wu, P., Liu, J. and Liu, X. , "Dynamic local aggregation network with adaptive clusterer for anomaly detection.," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 404–421.

[86] Ristea, N. C., Madan, N., Ionescu, R. T., Nasrollahi, K., Khan, F. S., Moeslund, T. B. and Shah, M., "Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection," in *arXiv preprint*, 2021, arXiv:2111.09099.

[87] Acsintoae, A., Florescu, A., Georgescu, M. I., Mare, T., Sumedrea, P., Ionescu, R. T. and Shah, M., "Ubnormal: New benchmark for supervised open-set video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20143-20153.

[88] Georgescu, M. I., Ionescu, R. T., Khan, F. S., Popescu, M. and Shah, M., "A background-agnostic framework with adversarial training for abnormal event detection in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2021.

[89] Jain, Y., Sharma, A. K., Velmurugan, R. and Banerjee, B., "Posecvae: Anomalous human activity detection," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Jun. 2021, pp. 2927-2934.

[90] Zhang, Y., Wang, C., Wang, X., Zeng, W. and Liu, W., "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision,* vol. 129, pp. 3069-3087, 2021.

[91] Cao, C., Lu, Y., Wang, P. and Zhang, Y., "A New Comprehensive Benchmark for Semi-supervised Video Anomaly Detection and Anticipation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 20392-20401.

[92] Cho, M., Kim, T., Kim, W. J., Cho, S. and Lee, S., "Unsupervised video anomaly detection via normalizing flows with implicit latent features," *Pattern Recognition,* vol. 129, p. 108703, 2022.

Yalong Jiang received the Ph.D. degree in the Department of Electronic and Information Engineering, Hong Kong Polytechnic University. Since April 2020, he has been with Beihang University, where he is now an assistant professor. His research interests include pattern recognition, computer vision, and machine learning.

Changkang Li received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2022. He is currently pursuing the master degree with Beihang University. His current research interests include video anomaly detection, temporal action detection, weakly supervised learning, and deep learning.

Wenrui Ding received the doctorate degree in electrical and information engineering from Beihang University. She is currently in charge of the Unmanned System Research Institute in Beihang University. Her research interests include the command and control of aerial vehicles and image processing.

Jinzhi Xiang received the B.S. and Ph.D. degrees in electronic engineering from the Beijing University of Posts and Telecommunications and the Beijing Institute of Technology. He is currently with Beijing Institute of Technology. His research interests include moving target detection and signal processing.

Zheru Chi received the BEng and MEng degrees from Zhejiang University, in 1982 and 1985, respectively, and the PhD degree from the University of Sydney, in March 1994, all in electrical engineering. Between 1985 and 1989, he was on the faculty of the Department of Scientific Instruments, Zhejiang University. He worked as a senior research assistant/research fellow in the Laboratory for Imaging Science and Engineering, University of Sydney, from April 1993 to January 1995. Between February 1995 and April 2021, he was with Hong Kong Polytechnic University, where he was an associate professor in the Department of Electronic and Information Engineering. He was an associate editor of the IEEE Transactions on Fuzzy Systems, and is currently an editor of the International Journal of Information Acquisition. His research interests include image processing and computational intelligence.