Regularity Learning via Explicit Distribution Modeling for Skeletal Video Anomaly Detection

Shoubin Yu, Zhongyin Zhao, Haoshu Fang, Andong Deng, Haisheng Su, Dongliang Wang, Weihao Gan, Cewu Lu, Senior Member, IEEE, and Wei Wu

Abstract—Anomaly detection in surveillance videos is challenging but important for ensuring public security. Different from pixel-based anomaly detection methods, pose-based methods utilize highly-structured skeleton data, which decreases the computational burden and also avoids the negative impact of background noise. However, pose-based methods lack an alternative dynamic representation akin to the explicit motion features, such as optical flow, employed by pixel-based methods. In this paper, a novel Motion Embedder (ME), a label-efficient scheme without extra annotation efforts, is proposed to provide a pose motion representation for the structured posed data from a probability perspective. Furthermore, a novel task-specific Spatial-Temporal Transformer (STT) is deployed for self-supervised pose sequence reconstruction. These two modules are then integrated into a unified framework for pose regularity learning, which is referred to as Motion Prior Regularity Learner (MoPRL). MoPRL achieves competitive results on multiple challenging datasets while minimizing computational costs. Extensive experiments validate the versatility of the proposed modules and provide insights for future research.

Index Terms—Video anomaly detection, label-efficient motion prior, regularity learning.

I. INTRODUCTION

VIDEO Anomaly Detection (VAD) [1] represents a critical and challenging task within the realm of computer vision, aimed at identifying frames containing abnormal events, such as criminal activities and traffic accidents. The reliance on context, large variety, and rarity of abnormal events, make it extremely difficult to collect anomalous events for conventional supervised model training [2]. Consequently, un/self-supervised methods are usually employed to learn normal patterns, allowing anomaly detection to be approached as the identification of out-of-distribution [3] that deviates from the training distribution.

Pixel-based methods [4], [5], [6], [7], [8], have been extensively studied for VAD. Recently, thanks to the success of pose estimation, pose, as a clean and well-structured data, attracted the attention of researchers [9], [10], [11], [12]. Since pose is immune from background noise and also preserve the privacy related to human as compared to using the raw data, the pose-based method is viewed as a promising approach and is expected to compete with the pixel-based counterparts.

Despite the advantages of pose-based methods, as demonstrated in prior research, their performance gains remain relatively limited. We assert that this is due to the inherent differences between image anomaly detection, which relies solely on static features (e.g., appearance), and VAD, which is more dependent on dynamic features. For instance, as shown in the top half of Figure 1, the man in the red box may be judged as normal if only given the right frame, while the confidence of this assertion could be strengthened if the left frame is visible since he is jumping. The incorporation of motion features, such as optical flow, in pixel-based methods enhances these models’ sensitivity to dynamics and motion anomalies. That is why we can easily tell different motions by observing lighter areas in the optical flow map in Figure 1 top right.

In contrast, pose-based methods lack such an intuitive motion representation. Previous pose-based methods [9], [10], [11], [12], [13], [14] utilize the entire pose trajectory, comprised of a sequence of static coordinates, with motion implicitly incorporated.

The development of motion features for pose-based methods is both logical and crucial for enhancing pose-based VAD performance. However, two primary challenges warrant further consideration when conceptualizing such a pose motion feature: i) obtaining a dense motion descriptor analogous to optical flow for pose keypoints is difficult due to their highly structured, discrete nature and the higher semantics compared to pixel values. ii) replicating the optical flow approach to train a “pose motion flow” detector would be resource-intensive, as optical flow algorithms in previous pixel-based methods predominantly rely on supervised learning and extensive ground truth annotation, while also being susceptible to scene changes. As a result, our objectives are to: i) Develop label-efficient strategies to represent motion features for discrete poses. ii) ensure that these features capitalize on well-structured data and can be easily adapted to various VAD scenarios without necessitating an additional, burdensome learning process.
In this paper, we propose a novel Motion Prior Regularity Learner (MoPRL) to achieve the aforementioned goals in pose-based methods. MoPRL is composed of two sub-modules: Motion Embedder (ME) and Spatial-Temporal Transformer (STT). Specifically, Motion Embedder is designed to extract the spatial-temporal representation of input poses from the perspective of probabilistic. Inspired by the commonly used frame gradients and optical flow in pixel-based methods, we model the pose motion based on the first-order difference, or so-called displacement, between the center point of the poses among adjacent frames. However, directly applying such differences as the motion representation is oversimplified. Given the common assumption that the anomaly rarely happens, we further transform such displacement into the probability domain. In detail, we obtain the motion prior, which represents an explicit distribution of displacement on the training data, by statistics. In this way, as shown in the bottom half of Figure 1, to represent corresponding motion, every pose displacement is mapped to a certain probability based on the motion prior. Notably, such intuitive motion representation is directly deduced from the training data itself in a statistical manner, which avoids introducing extra annotations, making it suitable for large-scale VAD datasets. Spatial-Temporal Transformer is then deployed as a task-specific model to learn the regular patterns with the input of poses and their motion features from Motion Embedder. Different from previous Recurrent Neural Network (RNN) or Convolution Neural Network (CNN)-based frameworks, the transformer is adopted for its self-supervised and sequential structure, which naturally fits our task. To sum up, the contributions of this paper are four-fold:

- Motion Embedder is proposed as a label-efficient scheme to represent pose motion in the probability domain for video pose regularity learning.
- Motion Prior Regularity Learner (MoPRL) is proposed, which consists of the Motion Embedder and the Spatial-Temporal Transformer, to model regularity in motion-embedded poses.
- MoPRL delivers competitive performance on the two most challenging datasets while maintaining minimal computational costs.
- Comprehensive experiments are conducted to verify the effectiveness of each module, and more insights for future work based on the failure cases are provided.

II. RELATED WORKS

A. Self-Supervised Video Anomaly Detection

Different from supervised/weakly supervised [15], [16], [17], [18] video anomaly detection where abnormal classes labels are accessible for model training, in self-supervised video anomaly detection, anomalies are recognized as outliers of the distribution of normality. The pixel-based framework handles the problem by reconstruction and prediction. In [19], Hasan et al. reconstructed the normal appearance [20] and motion [21] features to learn video regularity. In [22], the authors leveraged sparsing coding to enforce adjacent frames to be encoded with similar reconstruction coefficients. In [4], Liu et al. proposed a future frame prediction framework with the optical flow [23] as additional input. In [24], Nguyen et al. proposed a cross-channel translation framework to learn the coherence between motion and appearance. In a recent work [25], Liu et al. exploited the Conditional Variational Autoencoder to capture the correlation between the frame and the optical flow. Recently, pose-based methods have been popular because of their efficiency and immunity from background noise. In [10], the authors proposed a connected RNN for learning pose regularity with decomposed keypoint. In [9], Rodrigues et al. deployed a multi-timescale prediction framework to model trajectories. Moreover, Markovitz1 et al. [11] learned poses graph embeddings with autoencoders and generated soft assignments via clustering. Moreover, Zeng et al. [12] proposes a hierarchical graph neural network capture scene semantics. However, it is rather no attempt to obtain a sophisticated dynamic feature in pose-based methods. In this work, we propose a novel Motion Embedder to generate pose motion representation from the probability domain.

B. Vision Transformer

Transformer [26] has gradually become a mainstream framework for the computer vision community [27], [28], [29], [30], [31] for its tremendous potential in sequence modeling. It has achieved competitive even superior performance compared with CNN-based methods in image classification [32], [33], [34], object detection [35], semantic segmentation [36], etc. Dosovitskiy et al. [32] viewed an image as a patch sequence and constructed ViT to achieve effective recognition. Carion et al. [35] regarded the object detection problem as a set direction prediction task and solve it with a transformer-based encoder-decoder called DETR. Similarly, SETR [36] is proposed for image context modeling in semantic segmentation. Transformer is also deployed to estimate 3D human pose as in [37]. The authors build a divided temporal-spatial transformer to model the pose sequence. HOT-Net [38] fully exploits the correlation between joints and object corners to obtain a more accurate estimation. In video anomaly detection, Feng et al. [7] first apply transformer structure on the temporal dimension with pixel input. Different from related transformer works, MoPRL is proposed with a spatial-temporal transformer and motion-embedded pose input.
III. METHODS

In this section, we introduce our proposed pose-based video anomaly detection method called Motion Prior Regularity Learner (MoPRL), which introduces a new motion representation abstracted from the statistical distribution for label-efficient training purposes. As shown in the left of Figure 2, MoPRL consists of two sub-modules: Motion Embedder (ME) and Spatial-Temporal Transformer (STT). We first utilize a pose detector to obtain the pose trajectories. Unlike pixel-based methods, which adopt the widely-used optical flow as the motion representation, MoPRL models the posed-based motion representation as a probability distribution according to the statistical velocity and fuses spatial and temporal representation via the ME. Then, STT is applied to learn the spatial-temporal regularity with a self-supervised reconstruction task. In this way, the model learns the distribution of the normal samples; thus, the anomalies could be detected according to the Frame Anomaly Score.

A. Task Definition

Given the training set \( D_{train} = \{F_1, \ldots, F_m\} \) and the test set \( D_{test} = \{(F_1, L_1), \ldots, (F_n, L_n)\} \), where \( F_i \) represents the frames and \( L_i \in \{0, 1\} \) indicates the label of normality or anomaly. There are only normal samples in the training set and \( L_i \in \{0, 1\} \) indicates the label of normality or anomaly. There are only normal samples in the training set and checking set, respectively. According to the statistical velocity, \( L \) set reconstruction task. In this way, the model learns the distribution of the normal samples; thus, the anomalies could be detected according to the Frame Anomaly Score.

B. Pose Pre-Processing

Following the prepossessing operation proposed in [10], we decompose the original pose into a locally normalized pose and a global center point. We first calculate the center point \((\bar{x}_i, \bar{y}_i)\) and the size \((w_i, h_i)\) of the human box according to the maximum and minimum coordinates of the keypoints, and then normalize the pose as \(\tilde{P}_i = [\tilde{J}_{i,1}, \ldots, \tilde{J}_{i,k}]\) based on the human box size, where \(\tilde{J}_{i,j} = (\bar{x}_{i,j}, \bar{y}_{i,j})\) is the normalized coordinates. The normalized pose \(\tilde{P}_i\) unifies the scale in different distances, so even tiny changes in far can be captured.

C. Motion Embedder

Since dense motion features, such as optical flow, cannot be obtained, pose-based methods lack effective motion representations. In this work, we propose a multi-step approach to get the intuitive pose motion and embed it with the pose via the novel Motion Embedder (ME). We first calculate normalized displacement between adjacent poses in sequence and then obtain an explicit discrete distribution describing the training dataset displacement statistic. After this, we choose a predefined distribution (e.g., Rayleigh or Gaussian) to fit the discretized distribution and obtain its continuous version, which we refer to as motion prior. In the end, we leverage both the normalized pose and its motion probability, which represent spatial and temporal information, respectively, to obtain the motion-embedded pose. We will introduce ME in detail in the following subsections.

1) Displacement Calculation: The displacement is the first-order difference between each pose, and can be regarded as an average velocity during a short period. Thus, we consider utilizing displacement to construct the foundation of motion representation. Empirically, such displacement is:

\[
\begin{align*}
  v_i &= \sqrt{(\hat{x}_{i+1} - \hat{x}_i)^2 + (\hat{y}_{i+1} - \hat{y}_i)^2} \\
  \tau_i &= \frac{v_i}{(w_i + h_i)}
\end{align*}
\]

where \(v_i\) represents the average velocity from pose \(P_i\) to pose \(P_{i+1}\). Similar to what we have done to pose, we also normalize the velocity to obtain a normalized version \(\tau_i\) which eliminates the influence of perspectives. Nonetheless, directly leveraging \(\tau_i\) as the motion representation is oversimplified that the essence of a normal motion (normality is common) would be overlooked, leading to the spatial-temporal feature cannot be effectively represented. We resolve this issue from the perspective of probability.

2) Probabilities as Scaling Factors: We first obtain the statistic of the velocity, \(\Omega\), by counting the modes of their normalized version in the training set. Then, in order to obtain a continuous distribution, we fit this discretized data with \(f\). Intuitively, \(f\) can be regarded as parameter estimation for a certain distribution. In this paper, we have tried two different ways to fit the distribution. The one is non-linear least squares with the predefined scheme. The other is the neural networks. We name such fitted continuous distribution as Motion Prior \(\rho\). Based on the fact that different prior models its low-frequency part in different manners, we should carefully select an appropriate prior to ensure that it fits more with the real-world distribution, which is believed to be beneficial for the quality of the representation derived from Motion Embedder. As shown in the right bottom of Figure 2, we can tell that the real distribution of the displacement corresponds more to the Rayleigh distribution. And the further experimental results also demonstrate that prior fitted by Rayleigh distribution achieves the best performance. Next, in order to obtain a versatile representation that contains both temporal and spatial information, we expect to combine the normalized pose, which stands for spatial information, and the motion prior, which is actually the temporal representation, we consider employing the probability in motion prior as a scaling factor, and the scaling mechanism is as follows:

\[
\begin{align*}
  \rho &= f(\Omega) \\
  \hat{P}_i &= \frac{\tilde{P}_i}{\rho(\tau_i)}
\end{align*}
\]

where \(\hat{P}_i = [\hat{J}_{i,1}, \ldots, \hat{J}_{i,k}]\) is the pose feature after the scaling operation and represents motion embedded pose that fuses the spatial and temporal information for the \(i\)-th pose. This is exactly the reason that we call this module Motion.
Embedder. It is worth noting that, to avoid numerical error, we additionally deploy an affine transformation to the scaling factor, for it may be used as a denominator. Consequently, as shown in the right top of Figure 2, we can obtain a pose with a larger size if the emergence frequency is lower. \( \hat{P}_t \) will then be used as the input of the following module.

**D. Spatial-Temporal Transformer**

To learn the regularity of human pose trajectories, we proposed to utilize a transformer to process the motion embedding aforementioned, because of its acknowledged advantage of modeling sequential data. However, the orthodox transformer model results in a computational complexity of \( O((N \times T)^2) \) (where \( N \) is the number of joints in a single pose, \( T \) is the pose number in a single trajectory) and grows exponentially with the increasing \( N \) and \( T \). Thus, following [39], we divided the attention mechanism into spatial and temporal parts to decrease the computational complexity to \( O(N^2 + T^2) \). We call this variant of the transformer a Spatial-Temporal Transformer (STT). Specifically, STT contains an \( L_s \)-layer spatial transformer and an \( L_t \)-layer temporal transformer. Aiming to fully exploit the potential of STT, we view \( L_s \) and \( L_t \) as hyperparameters and experimentally identify their value, which will be shown in the experiment section. As shown in Figure 3, we also illustrate details of the proposed Spatial-Temporal Transformer. In the following subsections, we introduce the details of STT.

**1) Masked Pose Embedding:** Before the transformer blocks, we first obtain the embedding of joints. Specifically, for joint \( \hat{J}_{i,j} \), we adopt the mask operation [40] on pose sequences as data augmentation. With masking, the model can reconstruct masked joints with nearby joints information only. This masking is designed to help the model learn contextual information and avoid overfitting. We randomly select joints in the whole input trajectory sequence with a certain probability and set them to zero only during training (as shown in Figure 3, the gray patches in the input represent masked pose embedding.). And then we map it into the embedding space to obtain the joint vector \( z_{i,j} \in \mathbb{R}^C \), where \( C \) is the embedding dimension, as the following equation:

\[
z_{i,j} = E \cdot \text{mask}(\hat{J}_{i,j}) + E_{s}^{j}
\]

where mask(\cdot) is the mask function that operated on \( \hat{J}_{i,j} \) with a certain probability, \( E \in \mathbb{R}^{C \times 2} \) the learnable embedding matrix. Moreover, \( E_{s}^{j} \in \mathbb{R}^{C} \) represents a learnable spatial position embedding (SPE) deployed to encode the spatial position of the \( j \)-th joint in a single pose. Notably, the mask operation only works during the training. We then obtain the embedding of the \( i \)-th pose as \( Z_i = [z_{i,0}, \ldots, z_{i,N}] \). As a result, the embedding matrix for the trajectory is \( Z = [Z_1, \ldots, Z_T] \).
2) **Spatial Transformer:** We manage to model the trajectory \( Z \in \mathbb{R}^{T \times N \times C} \) on a spatial domain with a \( L_s \)-layer Spatial Transformer. To be noticed, we conduct self-attention on the dimension of joint number, i.e., \( N \). Without loss of generality, we denote the input trajectory of the \( l \)-th layer as \( Z^l \), where \( l \in [1, L_s] \). The multi-layer attention operation is given by:

\[
Q = Z_{in}^l \cdot W_Q, K = Z_{in}^l \cdot W_K, V = Z_{in}^l \cdot W_V,
\]

\[
Z^{l+1} = \text{softmax}(QK^T / \sqrt{C})V + Z^{l},
\]

\[
Z^{l+1} = \text{LayerNorm}(fc(Z^{l+1}) + Z^{l+1}),
\]

where \( Q, K, V \) is the query, key and value matrix, \( W_Q, W_K, W_V \in \mathbb{R}^{C \times C} \), the corresponding project heads. The subscript \( ln \) indicates a tensor after layer normalization. \( \text{softmax} \) and \( fc \) represent softmax operation and fully-connected layer, respectively. Actually, we leverage multi-head self-attention as our attention operation for stronger representation. As it has been a common structure, we ignore its formulation here for simplicity. Please refer to [26] for more details.

3) **Temporal Transformer:** We then model pose trajectories on the temporal domain with a \( L_t \)-layer Temporal Transformer. Taking the output \( Z^{t-1} \) of Spatial Transformer as input, we first incorporate temporal position information for each joint embedding as follows:

\[
z = z_{iL_t}^j + E_{ipe}^l,
\]

where \( z_{iL_t}^j \) represents the \( j \)-th joint embedding in the \( i \)-th frame of \( Z^{t-1} \), \( E_{ipe} \) a learnable temporal position embedding (TPE). Thus, the trajectory embedding matrix can be obtained accordingly. Following the same steps in Equation 6, 7 and 8 in the aforementioned spatial transformer, we finally obtain the spatial-temporal output \( Z^o \).

**E. Self-Supervised Training**

In this subsection, we introduce our self-supervised training to learn the regularity in human pose trajectories. Specifically, we achieve this via the commonly used reconstruction method. With the help of a reconstruction head, we are able to learn the distribution of the normal samples.

1) **Reconstruction Head:** As shown in the left side of Figure 2, taking motion-embedded trajectory \([\tilde{P}_1, \ldots, \tilde{P}_t] \) as input, the Reconstruction Head, a single linear layer with layer normalization following [37], recovers the normalized trajectory \( \tilde{S} = [\tilde{P}_1, \ldots, \tilde{P}_t] \) from the output \( Z^o \).

\[
\tilde{S} = \text{Head}_{rec}(Z^o)
\]

2) **Objective Functions:** The final objective function can be described as follows, where \( \omega^{i,j} \) is the confidence score of each pose joint coming from the pose detector, and \( \tilde{J}_{i,j} \) is the reconstructed joint in \( \tilde{S} \). Similar to the operation in [9], we also normalize the raw confidence.

\[
\text{Loss} = \sum_{i=1}^{T} \sum_{j=1}^{N} \omega^{i,j} ||\tilde{J}_{i,j} - J_{i,j}||_2
\]

**F. Inference**

We introduce the mechanism, Frame Anomaly Score, that the proposed method detects frame-level anomaly with human pose trajectories. Firstly, an anomaly score \( A_{m,n} \), where \( n \) and \( m \) represent the \( n \)-th trajectory in the \( m \)-th frame, will be obtained from each pose trajectory via MoPRL. The anomaly score \( A_{m,n} \) is the \( L_1 \) norm of the difference between \( \tilde{S} \) and \( S \). Then, since each frame may contain multiple trajectories, we select the highest \( A_{m,n} \) as the frame-level anomaly score \( A_m \). The higher frame anomaly score \( A_m \) suggests the higher possibility for the current frame to be abnormal.

\[
A_{m,n} = ||\tilde{S} - S||_1.
\]

\[
A_m = \text{Max}(A_{m,n})
\]

**IV. Experiments**

In this section, we first introduce the experimental details of the proposed MoPRL. Extensive experiments are then conducted on two challenging datasets to evaluate the effectiveness and superiority of MoPRL with qualitative examples. More experiments and reports, like hyperparameters setting, pose estimator analysis, score strategy analysis, qualitative examples, and limitations are also discussed.

**A. Datasets and Setup**

**Datasets.** We evaluate our method on the three most challenging datasets: ShanghaiTech [22] contains 330 training videos and 107 testing ones. It consists of 13 training scenes and 12 testing scenes. And HR-ShanghaiTech [10] is a subset of ShanghaiTech containing only Human-Related anomalies with 101 testing videos. Corridor [9], a recent large-size dataset for video anomaly detection, contains 10 abnormal classes in a single scene. UCF-Crime [41] contains 1900 untrimmed videos with a total duration of 128 hours from real-world surveillance scenes (e.g., street, indoor). This dataset encompasses 13 anomaly categories, representing a diverse range of backgrounds. We further curated a subset, dubbed Human-Related-UCF-Crime (HR-UCF-Crime), specifically for pose-based methods by excluding the ‘Explosion’ and ‘Accident’ classes, as they aren’t human-related.

1) **Pose Estimator:** We adopt the same pose estimator with other compared methods to avoid the variance caused by pose quality. Specifically, we use the tools [42], [43] to obtain the trajectories as in [11] on ShanghaiTech. While for the Corridor dataset, we extract trajectories with tools [44], [45] as in [9]. For UCF-Crime [41], we adopt tools [42], [43] to extract trajectories at 5 fps. Each pose joint is provided with a confidence score.

2) **Implementation Details:** We apply AdamW [46] optimizer with an initial learning rate of \( 5 \times 10^{-5} \) and adopt a warm-up schedule with 1000 steps. Empirically, the layers number of Spatial Transformer and Temporal Transformers are both set to 2. The batch size is 256 and the dimension of vector embedding is 128. Each trajectory contains 8 poses (\( T = 8 \)) and each pose contains 17 joints for AlphaPose [42] results and 25 joints for OpenPose [44] results (e.g., \( N = 17 \) or 25). To obtain the pose sequences, we sample the pose trajectory using sliding windows with a window size of 16 and stride of
2. Following BERT [40], the mask ratio of poses is set to 0.15. We normalize the frame-level anomaly scores in each scene for final evaluation as in [10] and [24]. And all experiments are conducted on the entire dataset without division by scenarios.

3) Evaluation Metrics: Following the conventions, the frame-level Area Under Curve (AUC) is calculated as the evaluation metric. Since there is no tracklet-level label in both datasets, all related methods report frame-level only. All anomaly scores are concatenated into one list and only one ROC curve is on the entire dataset. If there is scene normalization, scores will be concatenated after normalization.

B. Comparison With the Posed-Based State-of-the-Arts

Table I illustrates the comparison results of our MoPRL with other state-of-the-art on two popular datasets respectively. We can observe that our method can obtain competitive performance improvement compared with other pure pose-based methods on all conducted datasets. Furthermore, we conduct a comparison of the running costs among pose-based methods, as shown in Table II. It becomes evident that our proposed MoPRL significantly reduces resource consumption, utilizing approximately 1/95 of the model parameters and 1/6 of the running time compared to the previous state-of-the-art. This finding indicates that our label-efficient motion representation strategy enables the tiny yet efficient model to perform well on these tasks, even rivaling the capabilities of larger models.

Besides, we find that our method can not only detect obvious anomalous events (e.g., “Biking”), but also is capable of observing some subtle activities, such as “Robbing” happening at a far distance away from the camera. Besides, MoPRL is sensitive to the anomalies with extreme movements (e.g., “running” and “pushing”) rather than the appearance-based anomaly events (e.g., “holding the suspicious object”), which is reasonable owing to the fact that original RGB information (e.g., clothing and belongings) is absent. However, more object-related anomaly classes (e.g., “wearing a mask”) are included in the Corridor dataset, which accounts for the reason why performance achieved in the ShanghaiTech dataset is higher than the Corridor dataset by a large margin (+11.27%). In short, MoPRL achieves non-trivial performance boosts from both efficiency and effectiveness.

C. Comparison With Pixel-Based Methods

In this section, we list recent representative weakly/self-supervised pixel-based methods as well as self-supervised pose-based methods in Table III. For a long time, self-supervised video anomaly detection (only normal data available during the training) depends on detector-free algorithms. Researchers aim to model better normal data distribution under finer constraint designs. HF2-VAD [25] achieves 76.20% on ShanghaiTech without any detector. Even so, compared with early detector-based methods like OCAE [5], it still lags a large gap (8.70%). Recently, HSC [64] leverage both pose and pixel information into use, it increases the performance but also highly increases the model complexity. Those researches demonstrates detectors benefit video anomaly detection by addressing the location problem. Compared with object detectors [59], [66], [67], pose detectors [42], [44], [65] address human location as well, and extract lighter and well-structured information, pose, for next stage. However, such poses discard rich RGB information like scene and light inevitably as well as optical flow [23] and deep RGB features [32], [60] extraction. Therefore, pose-based methods should outperform their pixel-based and detector-free rivals easily, while competing with pixel-based and detector-based ones. MoPRL achieves comparable performance on two mainstream datasets among pose-based methods and is comparable with OCAE [5] on ShanghaiTech. Furthermore, we observe that Weakly-supervised methods outperform self-supervised methods significantly due to the availability of coarse-grained labels. Nonetheless, self-supervised methods are still crucial for studying more efficient video anomaly detection as they eliminate the need for annotations during training and can handle various anomalies through out-of-distribution detection.

D. Ablation Studies

In this section, we control all hyper-parameters and conduct module-level ablation. As shown in Table IV, our baseline is a linear auto-encoder that includes only an encoder and a decoder. The proposed Motion Embedder is applied to other pose-based methods [10] to show the generalization ability and necessity of motion representation.

1) Motion Representation: As shown in Table IV, with the help of ME, significant AUC improvement can be observed even without sequence modeling (+11.45%), which demonstrates that motion representation is essential for pose-based methods. Without loss of generality, we continue to evaluate the necessity of motion representation in a classical multi-task pose-based method named MPED-RCNN [10], which adopts RNN for sequence modeling of input pose embedding. To be clarified, we reproduce the MPED-RCNN method to fit the proposed motion prior of ME, and only the reconstruction task...
TABLE III
COMPARISON INCLUDING WEAKLY-SUPERVISED PIXEL-BASED METHODS, SELF-SUPERVISED PIXEL-BASED METHODS, AND SELF-SUPERVISED POSED-BASED METHODS OF THEIR SETTING AND FRAME-LEVEL AUC ON TWO POPULAR DATASETS. HERE, WE OMIT TRACKING ALGORITHMS IN THE POSE-DETECTOR FOR SIMPLIFICATION. MIX: WE USE BOTH ALPHAPOSE AND OPENPOSE TO ALIGN WITH OUR MAIN COUNTERPARTS. *: THE RESULTS REPRODUCED ON THE NEW DATASET BY US

| Methods       | Publication | Detector | Feature Extraction | ShanghaiTech | Corridor | UCF-Crime |
|---------------|-------------|----------|--------------------|--------------|----------|-----------|
|               |             |          |                    |              |          |           |
| Weakly-Supervised Pixel-based |             |          |                    |              |          |           |
| CLAWS [17]    | ECCV20      | -        | C3D [50]           | 89.67        | -        | 83.03     |
| RTFM [15]     | ICCV21      | -        | I3D [51]           | 97.21        | -        | 84.30     |
| GCLw2 [16]    | CVPR22      | -        | ResNext3d [52]     | 86.21        | -        | 79.84     |
| STMSL [18]    | CVPR22      | -        | VideoSwin [53]     | 97.32        | -        | 85.62     |
| MGFN [16]     | AAAI23      | -        | I3D [51]           | -            | -        | 86.98     |
| Self-Supervised Pixel-based |             |          |                    |              |          |           |
| ConvAE [19]   | CVPR16      | -        | -                  | 70.40        | -        | 50.60     |
| FramePred [4] | CVPR18      | -        | FlowNet2.0 [23]    | 72.80        | 64.65    | -         |
| Mem-AB [54]   | ICCV19      | -        | -                  | 71.20        | -        | -         |
| OCAE [5]      | CVPR19      | S3D [55] | -                  | 84.90        | -        | -         |
| SACR [56]     | MM20        | RPN [29] | -                  | 74.70        | -        | 72.70     |
| CDAE [57]     | ECCV20      | -        | -                  | 73.30        | -        | -         |
| HF2-VAAD [25] | ICCV21      | -        | FlowNet2.0 [23]    | 76.20        | -        | -         |
| AMCM [58]     | AAAI21      | -        | FlowNet2.0 [23]    | 73.70        | -        | -         |
| SSMT [25]     | CVPR21      | YOLOv3 [59] | ResNetS0 [60]      | 90.20        | 76.97*   | -         |
| BPDPN [61]    | AAAI22      | R-CNN [62] | FlowNet2.0 [23]    | 78.10        | 71.34*   | -         |
| S3R [63]      | ECCV22      | -        | -                  | 80.47        | -        | 79.58     |
| GCLv2 [16]    | CVPR22      | -        | ResNext3d [52]     | 71.04        | -        | 78.93     |
| HSC [64]      | CVPR23      | HRNet [65], YOLOv3 [59] | ViT [32] | 83.40 | - |
| Self-Supervised Pose-based |             |          |                    |              |          |           |
| MPRD-RNN [10] | CVPR19      | AlphaPose [42] | -                  | 73.40        | 64.27    | 55.65*    |
| GEPC [11]     | CVPR20      | AlphaPose [42] | -                  | 75.50        | -        | 58.10*    |
| MTIP [9]      | WACV20      | OpenPose [44] | -                  | 76.03        | 67.12    | -         |
| PoseCVAE [47] | ICP21       | AlphaPose [42] | -                  | 74.90        | 67.34    | -         |
| HSTGCNN [12]  | TCSVT1      | HRNet [65] | -                  | 81.80        | 70.46    | -         |
| STGformer [48] | MM22       | HRNet [65] | -                  | 82.90        | 76.60    | -         |
| MSTA-FCN [49] | JVCI23      | AlphaPose [42] | -                  | 75.90        | -        | -         |
| MoPRl (Ours)  | -           | Mix [42], [44] | -                  | 83.35        | 71.63    | 64.13     |

TABLE IV
COMPARISON WITH [10] AND STUDY OF EACH PROPOSED MODULE ON SHANGHAI TECH. ME REPRESENTS THE MOTION EMBEDDER, AND STT IS THE SPATIAL-TEMPORAL TRANSFORMER. *: THE RESULTS ARE REPRODUCED BY US. †: ONLY RECONSTRUCTION TASK WITH THE NORMALIZEDPOSE

| Methods       | Appearance Feature | Motion Feature | Sequence Modeling | AUC   |
|---------------|--------------------|----------------|-------------------|-------|
| MPRD-RCNN† [10] | -                  | ME | RNN               | 71.54* |
| Ours          | Pose Embedding      | -  | -                 | 72.94*
|               |                     | ME | STT               | 66.01*
|               |                     | ME | -                 | 67.29*
|               |                     | ME | STT               | 77.46*
|               |                     | ME |                 | 83.35* |

with a locally normalized pose is applied to align the setting of our MoPRl for a fair comparison. As expected, the proposed method can bring consistent performance improvement as listed in Table IV. It confirms that ME can truly benefit other pose-based methods, and the motion representation should be accounted as an essential factor in developing pose-based methods. However, the huge contrast of performance gain in different frameworks (11.45% vs. 1.40%) also raises a concern that such a distribution-based hand-crafted feature extractor is still not an optimal way for all pose-based methods.

2) Sequence Modeling: We explore the impact of sequence modeling which is considered as the core of spatial-temporal regularity learning. STT can only bring slight performance improvement (+1.28%) without motion representation. When there is no ME, the input and output of STT are both normalized poses. We observe the training curve down extremely fast, and the evaluation result goes down. It means the model is overfitting even with a 15% pose mask. Under such a setting, we argue that STT actually learns an identical mapping which is a shortcut, rather than the normal data distribution. Combining with ME, STT can boost the overall performance by a great margin (+5.89%). It reveals that STT module can further model the temporal inter-dependencies with discriminative motion clues provided by ME. We also observe that RNN [68] adopted in [10] actually leads to the performance declination (−4.52%) when ME is applied. It demonstrates that the step-by-step RNN model actually limits the performance gain from such motion features. We hypothesize that, unlike transformer-based STT benefiting from big data, the cascaded and history-dependent RNN underfits such a large data size with limited model capacity. Thus, we argue that the normalized poses are inferior to describing the motion
dynamics of pose trajectories, which hinders the potential of such sequence modeling. This is also observed in Natural Language Processing domain, RNN is weak to capture long-term changes which leads to the creation of LSTM [69] and transformer. The visualization of reconstructed poses from STT and RNN in the following parts further confirms this assumption. Compared with STT, the RNN model reconstructs poses with huge deviations even for normal samples. And such deviation actually shadows the distinction between normality and anomaly brought by the motion prior. Conclusively, both motion representation and sequence modeling are important and indispensable.

E. Analysis on Motion Embedder

1) Motion Prior Type: ME is designed to represent intuitive pose motion by probability prior. Such prior directly controls the probability of pose motion, so the selection of prior matters to the final performance. In this section, we compare different priors. As shown in Figure 4 a, blue charts represent prior fitted by non-linear least squares with different schemes. The result shows the larger difference between the selected prior and the statistical distribution on the training dataset is, the more the performance of MoPRL declines (16.34% with a uniform prior and 12.78% with a Gaussian prior). It demonstrates that the model can indeed benefit from appropriate prior. And the gain will increase if the discrepancy between the prior and the real-world distribution decreases. We also use neural networks to fit the prior. We build a network with three linear layers, and an activation function, and train it with statistical data. The network is frozen, and only output probability for scaling the input. It improves only 2.52% with the heaviest computation. Moreover, observing the declination compared with spatial

2) Spatial-Temporal Fusion Type: In order to obtain the spatial-temporal input for STT, ME fuses the motion prior into poses. Hence, the fusion operation mentioned in Section III-C should be thoroughly explored to verify its effectiveness. In this section, besides the division, we also deploy other common operations to conduct such fusion. As shown in Figure 4 b, the results show that MoPRL is less sensitive to different scale-related operations (multiply with 81.54% and division with 83.35%). While the common fusion strategy in pixel-based methods, e.g., addition, does not work in pose-based methods, impairing the model performance compared with baseline (61.36% versus 67.29%). In this case, poses

may just be moved from their original spatial location and perturbed by the fusion. It demonstrates the effectiveness of the proposed fusion operation and also shows that appropriate fusion is essential for spatial-temporal regularity learning.

F. Analysis on Spatial-Temporal Transformer

1) Pose Masking and Position Embedding: We verify the effectiveness of pose masking and position embedding. We leverage pose masking as an operation of data augmentation. As shown in Figure 4 c, model performance first improves as the pose mask ratio increases and then drops to about 79.01%, which demonstrates that the model benefits from appropriate masking augmentation. However, the model would fail to handle regularity learning under a too-large ratio. Furthermore, we also evaluate the performance change brought by different position embedding strategies. As shown in Figure 4 d, different position embedding utilized on MoPRL bring consistent performance increment.

2) Attention Mechanism: A major concern of transformer-based models is their exponentially increasing complexity with input size. As shown in Table V, we list comparative results among different attention mechanisms for quantitative evaluation. We establish the baseline as the model without any attention but with motion prior. The joint attention representation of such sequence modeling. This is also observed in Natural Language Processing domain, RNN is weak to capture long-term changes which leads to the creation of LSTM [69] and transformer. The visualization of reconstructed poses from STT and RNN in the following parts further confirms this assumption. Compared with STT, the RNN model reconstructs poses with huge deviations even for normal samples. And such deviation actually shadows the distinction between normality and anomaly brought by the motion prior. Conclusively, both motion representation and sequence modeling are important and indispensable.

E. Analysis on Motion Embedder

1) Motion Prior Type: ME is designed to represent intuitive pose motion by probability prior. Such prior directly controls the probability of pose motion, so the selection of prior matters to the final performance. In this section, we compare different priors. As shown in Figure 4 a, blue charts represent prior fitted by non-linear least squares with different schemes. The result shows the larger difference between the selected prior and the statistical distribution on the training dataset is, the more the performance of MoPRL declines (16.34% with a uniform prior and 12.78% with a Gaussian prior). It demonstrates that the model can indeed benefit from appropriate prior. And the gain will increase if the discrepancy between the prior and the real-world distribution decreases. We also use neural networks to fit the prior. We build a network with three linear layers, and an activation function, and train it with statistical data. The network is frozen, and only output probability for scaling the input. It improves only 2.52% with the heaviest computation. Moreover, observing the declination compared with spatial

may just be moved from their original spatial location and perturbed by the fusion. It demonstrates the effectiveness of the proposed fusion operation and also shows that appropriate fusion is essential for spatial-temporal regularity learning.

F. Analysis on Spatial-Temporal Transformer

1) Pose Masking and Position Embedding: We verify the effectiveness of pose masking and position embedding. We leverage pose masking as an operation of data augmentation. As shown in Figure 4 c, model performance first improves as the pose mask ratio increases and then drops to about 79.01%, which demonstrates that the model benefits from appropriate masking augmentation. However, the model would fail to handle regularity learning under a too-large ratio. Furthermore, we also evaluate the performance change brought by different position embedding strategies. As shown in Figure 4 d, different position embedding utilized on MoPRL bring consistent performance increment.

2) Attention Mechanism: A major concern of transformer-based models is their exponentially increasing complexity with input size. As shown in Table V, we list comparative results among different attention mechanisms for quantitative evaluation. We establish the baseline as the model without any attention but with motion prior. The joint attention represents the vanilla transformer taking the entire sequence as input. It improves only 2.52% with the heaviest computation. Moreover, observing the declination compared with spatial

may just be moved from their original spatial location and perturbed by the fusion. It demonstrates the effectiveness of the proposed fusion operation and also shows that appropriate fusion is essential for spatial-temporal regularity learning.

F. Analysis on Spatial-Temporal Transformer

1) Pose Masking and Position Embedding: We verify the effectiveness of pose masking and position embedding. We leverage pose masking as an operation of data augmentation. As shown in Figure 4 c, model performance first improves as the pose mask ratio increases and then drops to about 79.01%, which demonstrates that the model benefits from appropriate masking augmentation. However, the model would fail to handle regularity learning under a too-large ratio. Furthermore, we also evaluate the performance change brought by different position embedding strategies. As shown in Figure 4 d, different position embedding utilized on MoPRL bring consistent performance increment.

2) Attention Mechanism: A major concern of transformer-based models is their exponentially increasing complexity with input size. As shown in Table V, we list comparative results among different attention mechanisms for quantitative evaluation. We establish the baseline as the model without any attention but with motion prior. The joint attention represents the vanilla transformer taking the entire sequence as input. It improves only 2.52% with the heaviest computation. Moreover, observing the declination compared with spatial
only (2.15%) or temporal only (2.28%) cases, we claim that the indifferent attention among the entire sequences would actually harm the performance. Furthermore, the temporal modeling matters more to the performance with the most significant boosting (3.89%), which demonstrates that motion information is essential to video anomaly detection. The highest performance with Spatial-Temporal Attention shows the effectiveness of our design.

3) Model Depth: In this section, we ablate several combinations of layer depth in both spatial and temporal dimensions. The results listed in Table VI demonstrate that MoPRL does not actually benefit from a deeper attention structure. The deepest model brings a 2.02% decrease compared with no STT. Compared with the deepest spatial attention which brings an average 1.26% decrease, the deepest temporal attention leads to a more significant average drop (-1.59%).

4) Embedding Vector Dimension: The pose embedding is an essential component of MoPRL, which encodes the core information of the input data and plays as a bedrock for the following spatial-temporal transformer. According to Table VII, we find that too small an embedding dimension impedes the performance, which indicates that a larger dimension is necessary to encode pose coordinates completely.

G. Diverse Evaluation

In this section, we conduct evaluations with different scoring strategies, metrics, and datasets to further explore and evaluate MoRPL under different conditions.

1) Scoring Strategies: It should be noticed that scene normalization is not applicable in real-world applications, and more discussions are needed. To address this concern, we re-evaluated MoPRL with different strategies, including $L_1$ with norm, $L_2$ with norm, $L_1$ without norm, and $L_2$ without norm (where “norm” represents scenario normalization). As shown in Table VIII, we conclude that scenario normalization benefits performance on the whole dataset, but MoPRL still shows its stable and comparable performance without such operation. It also demonstrates that our proposed method, MoPRL, can distinguish anomalies in normal data well.

2) Region-level Evaluation: Besides Frame-level evaluation, region-level metrics benefit anomaly locating. However, only ShanghaiTech provides pixel-level labels. Besides, only frame-level results are reported in all related pose-based methods. Herein, we re-evaluate MoPRL and baseline with the region-level metric in [70]. On ShanghaiTech, MoPRL achieves 78.26%, and MPED-RCNN (Rec) achieves 68.49%.

3) Cross-dataset Evaluation: We re-evaluate Corridor with the model trained on ShanghaiTech only. The results are 70.16% under the ShanghaiTech motion prior. Further, we evaluate the model trained on merged data (ShanghaiTech + Corridor) with a merged motion prior. MoPRL achieves 82.45% on SHT and 70.93% on Corridor. The cross-dataset results further indicate that MoPRL has scene adaptation which is regarded as a challenge in pixel-based methods as mentioned in [71]. And the performance decrease may come from pose changes in different camera angles in various scenes.

4) Robustness to Abnormal Noise During Training: In this section, we evaluate the robustness of the proposed MoPRL to abnormal samples using experiments on ShanghaiTech. We rank video samples by their IDs and create new training datasets by combining the original normal training data with the original abnormal testing data. The remaining test samples are used for evaluation. We vary the ratio of abnormal videos using multiple thresholds. Table IX shows that when introducing 20% noise in the training data, the performance of MoPRL only decreases by 1.41%. This demonstrates the robustness of MoPRL to abnormal noises during training, making it suitable for real-world scenarios where perfect filtering of abnormal data is difficult.

H. Visualization

This section provides straightforward evidence to demonstrate both the weakness and strength of MoPRL via several reconstructed poses visualization and anomaly score visualization. We also provide visualization of the Spatial-Temporal Transformer structure for better understanding. We further offer a comparison among different methods via visualization to explain the results in Table IV.

1) Reconstructed Pose: In this section, we compare different reconstruction results between MPED-RNN [57] and our MoPRL here to elucidate the difference in model capacities. To show the inner reconstructed errors intuitively, we use the mean error of two endpoints to represent an edge error. And a warmer color indicates a higher error. Our ground truth comes from pose estimator [42]. All poses are extracted from the ShanghaiTech dataset. We hypothesize that the model capacity of RNN actually limits performance gain from the motion prior. As shown in Figure 5, MoPRL can reconstruct normal poses with less reconstruction error, while MPED-RCNN offers a poorer ability in recovering normal poses.

We further assume that such deviation on normal samples shadows the distinction between normality and anomaly brought by the motion prior. We also provide abnormal cases for intuitive comparison. Compared with the reconstruction of
Fig. 5. Comparison among ground truth and reconstructed normal pose trajectories from different methods. We select both abnormal and normal samples for comparison. A warmer color represents a higher reconstruction error. Best viewed in color.

Fig. 6. Frame-level anomaly scores (blue lines) and the ground truth labels (red lines) in ShanghaiTech, and their corresponding AUC under the scene. We calculate AUC in each scene separately for this visualization. Best viewed in color.

Fig. 7. Frame-level anomaly scores (blue lines) and the ground truth labels (red lines) in Corridor, and their corresponding AUC under the scene. We calculate AUC in each scene separately for this visualization. Best viewed in color.

abnormal cases by MPED-RCNN [57], the outputs of MoPRL differ from the ground truth more in the pose scale rather than shape since we embed motion into the pose via scaling without changing the pose shape. Meanwhile, compared with RNN taking step-by-step inputs, MoPRL adopts the transformer, which is good at global information modeling and takes all...
inputs simultaneously. In this case, the abnormal poses are reconstructed with a larger size as well as error.

For quantitative results, we report the mean reconstruction error on abnormal and normal poses in different models. Since trajectory-level labels are not available, we use frame-level labels to label all poses in a frame. MoPRL gets 3.08 on anomaly and 2.20 on normality (40.0% relative gap), while MPED-RNN (Rec) gets 21.72 on anomaly and 16.13 on normality (34.7% relative gap). It reveals that MoPRL can distinguish anomalies better even with less reconstruction error.

2) Anomaly Score: As shown in Figure 6, in ShanghaiTech Scene 08_0157 and 04_0050, both anomalies (skateboarding and balance biking) are related to motion more and correspond to a relatively normal appearance. The performances in both scenes achieve over 98%. On the contrary, in ShanghaiTech Scene 08_0179, MoPRL fails to capture the skateboarder with a slow speed. And in ShanghaiTech Scene 01_0053, the slow vehicle cheats MoPRL successfully with both occluded pose and regular motion speed. The performances on those two scenes dropped to about 60%, which verifies our discussion in the paper. It demonstrates MoPRL benefits from the proposed Motion Embedder, which strengthens the motion features but still lacks the diversity of representation. As shown in Figure 7, besides the similar conclusion that MoPRL is sensitive to the motion speed anomaly (like chasing) but fails to capture the motion direction anomaly (like wandering), we observed an interesting difference between Corridor Scene 000276 and 000287. Both scenes contain the same type of abnormal events in that people carry a suspicious object (box) with normal motion. MoPRL achieves high performance when the object does not occlude the human. When the human is hidden behind the object, the performance decreases. It demonstrates MoPRL can capture the appearance-related anomaly, and the quality of poses matters to the final result.

V. DISCUSSION

A. Failure Cases

Pose-based methods depend heavily on the quality of the estimated poses. Therefore, when the pose estimator fails to extract the pose structure, MoPRL would have poor performance (e.g., when the objects are occluded or fast-moving). Besides, the tracking algorithm often captures inaccurate trajectories in a crowd scene, which directly leads to a serious impediment for MoPRL to detect. Moreover, failure cases are also observed in object-relative anomalies (e.g., human with a trailer or a mask, etc.) and motion direction anomalies (e.g., sudden turning around). The absence of RGB information probably causes this, and Motion Embedder alone cannot capture directional features.

B. Limitations

Although this work adopts the high-level pose features extracted from the video, we still find it too challenging to only take the displacement for pose motion modeling. Other essential characteristics, like motion direction, are ignored. Thus, in the future, we suggest fully considering all possible motion features to construct a completed pose motion representation for related tasks. The motion prior is influenced by varying camera angles and surveillance scenes, necessitating the generation of individual motion priors for each scene to enhance anomaly detection effectiveness. This constraint hinders the generalization of the proposed method. Moreover, the speed bottleneck of pose-based pipelines lies in the feature extraction, i.e., pose estimation and tracking process, which usually runs at 15 fps. Thus, the end-to-end speed could be ~14 fps.

VI. CONCLUSION

In this paper, we propose a novel Motion Prior Regularity Learner (MoPRL) for pose-based video anomaly detection. Intuitively, MoPRL takes pose motion probability from the prior statistics on the training dataset as motion representation through a label-efficient module, Motion Embedder (ME). Then, MoPRL models pose trajectories regularity with a Spatial-Temporal Transformer (STT) equipped with divided attention. It should be noted that via introducing the novel motion prior as an additional representation, we do not cost extra annotation efforts, which makes our model competitive in large-scale VAD datasets. MoPRL achieves competitive results on two challenging mainstream datasets with minimal computational cost. Ablation studies and failure case analyses provide more insights for future works.

REFERENCES

[1] R. Chalapathy and S. Chawla, “Deep learning for anomaly detection: A survey,” 2019, arXiv:1901.03407.
[2] M.-I. Georgescu, A. Barbulescu, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, “Anomaly detection in video via self-supervised and multi-task learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 12737–12747.
[3] J. Yang, K. Zhou, Y. Li, and Z. Liu, “Generalized out-of-distribution detection: A survey,” 2021, arXiv:2110.11334.
[4] W. Liu, W. Luo, D. Lian, and S. Gao, “Future frame prediction for anomaly detection—A new baseline,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6536–6545.
[5] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao, “Object-centric auto-encoders and dummy anomalies for abnormal event detection in video,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7834–7843.
[6] M. Zaigham Zaheer, J.-H. Lee, M. Astrid, and S.-I. Lee, “Old is gold: Redefining the adversarially learned one-class classifier training paradigm,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 14171–14181.
[7] X. Feng, D. Song, Y. Chen, Z. Chen, J. Ni, and H. Chen, “Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection,” in Proc. 29th ACM Int. Conf. Multimedia, Oct. 2021, pp. 5546–5554.
[8] S. Leroux, B. Li, and P. Simoens, “Multi-branch neural networks for video anomaly detection in adverse lighting and weather conditions,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2022, pp. 3027–3035.
[9] R. Rodrigues, N. Bhargava, R. Velumurugan, and S. Chaudhuri, “Multi-timescale trajectory prediction for abnormal human activity detection,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2020, pp. 2615–2623.
[10] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, “Learning regularity in skeleton trajectories for anomaly detection in videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 11988–11996.
[11] A. Markovitz, G. Shariat, I. Friedman, L. Zelnik-Manor, and S. Avidan, “Graph embedded pose clustering for anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10536–10544.
[60] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[61] C. Chen et al., “Comprehensive regularization in a bi-directional predictive network for video anomaly detection,” Proc. AAAI Conf. Artif. Intell., vol. 36, no. 1, pp. 230–238, Jun. 2022.

[62] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[63] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, “Self-supervised sparse representation for video anomaly detection,” in Computer Vision–ECCV, Tel Aviv, Israel: Springer, Oct. 2022, pp. 729–745.

[64] S. Sun and X. Gong, “Hierarchical semantic contrast for scene-aware video anomaly detection,” 2023, arXiv:2303.15031.

[65] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 5693–5703.

[66] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 936–944.

[67] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6154–6162.

[68] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[69] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[70] N. Li, F. Chang, and C. Liu, “Spatial–temporal cascade autoencoder for video anomaly detection in crowded scenes,” IEEE Trans. Multimedia, vol. 23, pp. 203–215, 2021.

[71] Y. Lu, F. Yu, M. K. K. Reddy, and Y. Wang, “Few-shot scene-adaptive anomaly detection,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 125–141.

Andong Deng received the B.S. degree from Sichuan University, China, and the M.S. degree from Shanghai Jiao Tong University, China. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Central Florida. His research interests include computer vision, video understanding, and multimodal machine learning.

Haisheng Su received the M.S. degree from the Department of Automation, Shanghai Jiao Tong University, in 2020, supervised by Prof. X. Zhao. He is currently a Researcher with SenseTime Group Ltd. His research interests mainly include deep learning, video understanding, and 3D perception.

Dongliang Wang received the degree (Hons.) from Xi’an Jiaotong University (XJTU) and the joint Ph.D. degree from MSR, Asia, and XJTU. He is currently a Researcher with SenseTime, specializing in computer vision with a particular focus on video understanding, 3D content generation, and autonomous driving. He has contributed to the field through the publication of several academic papers on these topics.

Weihao Gan received the B.E. degree in automation from the Huazhong University of Science and Technology, Wuhan, China, in 2012, and the M.S. and Ph.D. degrees in electrical engineering from the Media Communication Laboratory, University of Southern California, Los Angeles, USA, in 2014 and 2017, respectively. His research interests include computer vision and deep learning, including object detection/recognition, target tracking, video analysis, and interaction. He has published more than 20 journals and conference papers in these areas and holds several patents.

Cewu Lu (Senior Member, IEEE) received the Ph.D. degree from The Chinese University of Hong Kong, supervised by Prof. Jiaya Jia. He is currently a Professor with Shanghai Jiao Tong University (SJTU). Before he joined SJTU, he was a Research Fellow with Stanford University working under Prof. Fei-Fei Li and Prof. Leonidas J. Guibas. He was a Research Assistant Professor with The Hong Kong University of Science and Technology with Prof. Chi Keung Tang. He was selected as MIT TR35 in 2019 and served as an Area Chair for CVPR’20 and ICCV’21. His research interests include computer vision, deep learning, and robotics.

Wei Wu received the B.S. and M.S. degrees from the Department of Mathematics, Nanjing University, China, in 2011 and 2016, respectively. He is currently a Researcher with SenseTime Group Ltd. His current research interests include computer vision, robotics, and connected and automated vehicle driving.