Evidential Reasoning for Video Anomaly Detection

Che Sun
Beijing Institute of Technology, China
sunche@bit.edu.cn

Yunde Jia
Beijing Institute of Technology, China
Shenzhen MSU-BIT University, China
jiayunde@bit.edu.cn

Yuwei Wu*
Beijing Institute of Technology, China
Shenzhen MSU-BIT University, China
wuyuwei@bit.edu.cn

ABSTRACT

Video anomaly detection aims to discriminate events that deviate from normal patterns in a video. Modeling the decision boundaries of anomalies is challenging, due to the uncertainty in the probability of deviating from normal patterns. In this paper, we propose a deep evidential reasoning method that explicitly learns the uncertainty to model the boundaries. Our method encodes various visual cues as evidences representing potential deviations, assigns beliefs to the predicted probability of deviating from normal patterns based on the evidences, and estimates the uncertainty from the remained beliefs to model the boundaries. To do this, we build a deep evidential reasoning network to encode evidence vectors and estimate uncertainty by learning evidence distributions and deriving beliefs from the distributions. We introduce an unsupervised strategy to train our network by minimizing an energy function of the deep Gaussian mixed model (GMM). Experimental results show that our uncertainty score is beneficial for modeling the boundaries of video anomalies on three benchmark datasets.

CCS CONCEPTS

• **Computing methodologies** → **Scene anomaly detection**; *Activity recognition and understanding*.

KEYWORDS

Video Anomaly Detection, Evidential Reasoning, Uncertainty Estimation, Deep Gaussian Mixed Model

ACM Reference Format:
Che Sun, Yunde Jia, and Yuwei Wu. 2022. Evidential Reasoning for Video Anomaly Detection. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548091

1 INTRODUCTION

Video anomaly detection aims to discriminate events that deviate from the normal patterns, and it is increasingly being studied for various applications in ubiquitous surveillance videos [5, 6, 26, 45]. Due to lack of training anomaly data, it is difficult to use popular supervised deep learning to discriminate anomalies in videos. Hence, many researchers have investigated unsupervised deep learning for video anomaly detection.

Most existing unsupervised methods [12, 22, 44] model anomaly boundaries by calculating the probability of deviating from normal data distributions by using deep networks. However, these methods do not work well in discriminating anomalies that are context-specific and/or visually similar to normal patterns, due to the large uncertainty in the probability of deviating from the normal patterns. For example, anomaly scores representing the deviation probability as decision boundaries, which are estimated by using a reconstruction-based method [12] and a classification-based method [16], are not discriminative enough, as shown in Figures 1a and 1b. Figures 2a and 2b illustrate the average anomaly scores on the whole Avenue dataset, where the small score gaps also show less discrimination. The reason is that deep networks in these methods are encouraged to uniformly generate the low deviation probability for all normal data with high confidences, but they tend to wrongly extend their confidences to visually similar anomaly data. Fortunately, learning uncertainty is able to prevent deep networks from making overconfidence predictions [2, 3, 10]. In this paper, we propose to learn the uncertainty of deviation probabilities to model decision boundaries of anomalies in videos. As demonstrated in Figures 1c and 2c, using uncertainty to compute anomaly scores is beneficial for modeling the boundaries with good discrimination.

Using deep networks to learn the uncertainty of deviation probabilities is non-trivial. There are two challenges we address in

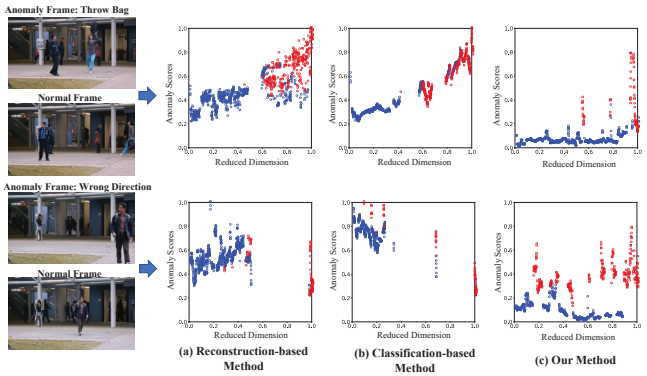

Figure 1: Frame-level anomaly scores of two example videos on the Avenue dataset [20]. Red and blue points represent anomaly and normal frames, respectively. The horizontal axis denotes the reduced 1-dimensional feature space by using t-SNE. The vertical axis denotes the computed anomaly scores. Our method produces more discriminated anomaly scores than the other two methods.

*Corresponding Author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548091

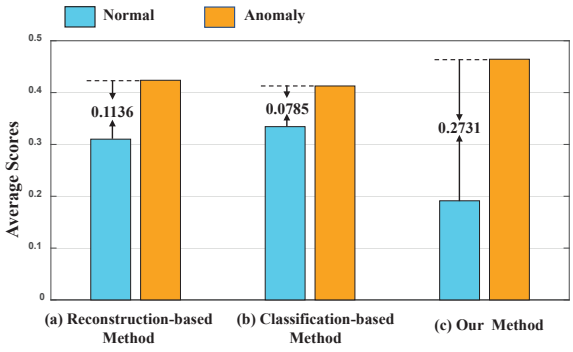

Figure 2: Average of anomaly scores on the whole Avenue dataset. Anomaly scores of our method between normal and anomaly frames indicate good discrimination.

particular. (1) The bias of visual cues is likely to mislead the uncertainty estimation, because the deviations of anomalies are often revealed in various visual cues in contexts (e.g., object appearance, motion and scene, etc.). Therefore, it is necessary to mine diverse visual cues in complex contexts for uncertainty estimation. (2) Lack of sufficient anomaly annotations will impede estimating the uncertainty in deep networks, since many deep uncertainty learning methods require supervised annotations to model uncertainty by penalizing wrong predictions with high uncertainty.

To address the two challenges, we propose a deep evidential reasoning method that learns the uncertainty to model the decision boundaries. Specifically, we mine visual cues in different types by multiple evidence learning processes, and estimate the uncertainty in an unsupervised manner by introducing a deep Gaussian mixed model (GMM) to penalize wrong cluster predictions. To this end, we build a deep evidential reasoning network consisting of an evidence encoder and an uncertainty learner. The evidence encoder extracts representative normal patterns in a memory auto-encoder, and selects relevant normal patterns to encode visual cues in different types, including object appearance, motion, visual relationship, and scene, into evidence vectors of deviations through a memory Transformer. The uncertainty learner is used to estimate the uncertainty by learning evidence distributions based on the evidence vectors and deriving both the beliefs and probabilities from the distributions. The uncertainty scores are used to calculate frame-level anomaly scores for video anomaly detection. We further introduce an unsupervised strategy to train our network by assigning the beliefs and probabilities to normal clusters of the deep GMM and penalizing wrong cluster predictions.

We evaluate our method on ShanghaiTech [21], Avenue [20], and UCSD Ped2 [23] datasets. Experimental results show that our method outperforms state-of-the-art methods. For a fair comparison, we use the same backbone (i.e., memory auto-encoder and memory Transformer) as our network to calculate anomaly scores in Figure 1 and Figure 2, and the inputs are cues of the scene type.

2 RELATED WORK

We review related work on the unsupervised deep anomaly detection that we are concerned about in this paper. Most existing methods roughly fall into two categories: reconstruction-based and classification-based.

Reconstruction-based methods assume that normal data can be better reconstructed from the feature space than anomalies, and use reconstruction errors to model the anomaly deviations from the patterns of normal data. Among these methods, auto-encoder networks are the commonly-used techniques [36, 38, 46]. For example, Hasan *et al.* [14] used one fully connected auto-encoder and another end-to-end convolutional auto-encoder to learn spatial normal patterns for modeling the deviations. Chong and Tay [8] introduced a spatio-temporal auto-encoder to learn spatio-temporal normal patterns in videos. Gong *et al.* [12] proposed a Memory-augmented Deep Auto-encoder (MemAE) to record prototypes of normal patterns for reducing noise in the learned patterns. These methods aim to represent common patterns of normal data and reconstruct them with low reconstruction errors uniformly, which may not work well when anomalies share somewhat similar patterns with the normal data, because the reconstruction task does not need to learn the discriminative information from normal data. In contrast, our method learns the uncertainty of deviation probabilities based on a deep GMM, and extracts discriminative information for assigning both beliefs and probabilities to different normal clusters of the deep GMM, thereby performing well in discriminating anomalies with a similar pattern of normal data.

Classification-based methods assume that normal data come from one or more abstract classes and anomalies do not conform to them. They usually use classification errors to model the deviations for anomaly discrimination [32, 35, 43, 44]. For example, Ionescu *et al.* [39] extracted deep features and adopted a one-class SVM to acquire classification errors as anomaly deviations. Xu *et al.* [42] extended one one-class SVM to three one-class SVMs based on fused deep appearance and motion features. Ionescu *et al.* [16] introduced a one-versus-rest SVM to classify samples into multiple normal classes for discriminating anomalies outside all classes. These methods classify normal data into known classes with high confidences, but are likely to wrongly classify unknown anomalies that fall near the classification boundaries into normal classes. In contrast, our method estimates the uncertainty by evidential reasoning to prevent making overconfidence predictions, obtaining the capacity to classify unknown anomalies near the boundaries successfully.

3 METHOD

Figure 3 illustrates the framework of our method. We parse input videos into different types of visual cues, and then build a deep evidential reasoning network to obtain the evidence vectors and uncertainty from visual cues in each type. The frame-level anomaly scores are computed according to the vectors and uncertainty. The deep evidential reasoning network consists of an evidence encoder and an uncertainty learner. We use the evidence encoder to encode evidence vectors of deviations, and use the uncertainty learner to estimate the uncertainty by learning evidence distributions and deriving both beliefs and probabilities from the distributions.

3.1 Video Parsing

Video anomalies are heterogeneous, and different anomalies may exhibit completely different expressions [11, 27]. Their deviations

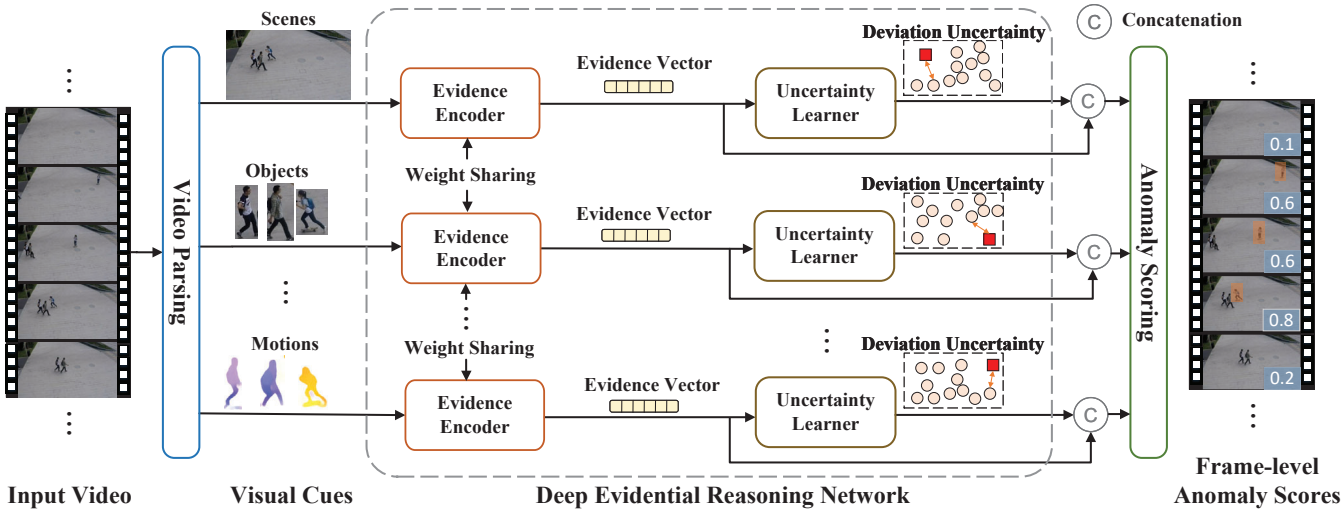

Figure 3: The framework of our method.

from the normal patterns will also involve visual cues in different types. Therefore, we parse input videos to acquire visual cues in different types for estimating the uncertainty of the deviation probability.

Given a video V with T frames $[I_1, I_2, \cdots, I_T]$, we use the region proposal network (RPN) [31] to generate object bounding boxes for each frame, where the top-K bounding boxes are selected in each frame. The pre-trained FlowNet-v2 [15] is used to generate optical-flow images between the current frame and the previous frame. We parse the video to obtain four typical types of visual cues as follows. (1) *Scene:* the whole image frames from the input video. (2) *Appearance:* the sub-images cropped by bounding boxes. (3) *Relationship:* the sub-images cropped by union bounding boxes over pairwise objects. (4) *Motion:* the optical-flow sub-images cropped by bounding boxes.

We use the notation $\mathbf{X} \in \mathbb{R}^{W*H*3}$ to represent a parsed sample of an image or a sub-image. W and H denote its width size and height size, and both of them are set to 64.

3.2 Deep Evidential Reasoning Network

Evidential reasoning is to represent and combine evidences to allocate belief masses to subsets of a frame-of-discernment [1] according to Dempster-Shafer theory [9, 34], which contributes to detecting out-of-distribution samples by modeling the uncertainty. We design a deep evidential reasoning network consisting of an evidence encoder and an uncertainty learner, to learn the uncertainty of deviation probabilities. As shown in Figure 4, the evidence encoder extracts normal patterns from visual cues in a memory auto-encoder, and jointly represents and combines evidences by encoding the evidence vector in a memory Transformer. The uncertainty is estimated by deriving both belief masses and probabilities in the uncertainty learner. The frame-of-discernment in our evidential reasoning is defined as a finite set $\Theta = \{\theta_1, \theta_2, \cdots, \theta_K\}$ of potential normal clusters of a deep Gaussian mixed model (GMM). Allocating belief masses to the clusters is realized by learning an evidence distribution in the uncertainty learner.

3.2.1 Evidence Encoder. As illustrated in Figure 4, the evidence encoder is composed of a memory auto-encoder and a memory Transformer. We use the memory auto-encoder for explicitly learning representative normal patterns, providing supports of potential deviation references. We introduce a self-attention Transformer network for mining the relationships between input samples and the learned normal patterns, generating evidence vectors representing deviations.

Memory Auto-encoder. The memory auto-encoder is used to obtain the encoding feature $\mathbf{f}$ and memory items $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_N]^\top$, where $\mathbf{M}$ records the patterns of normal samples. We encode a parsed sample $\mathbf{X}$ into an encoding feature $\mathbf{f}$ in a convolutional network $\mathbf{f} = \text{Conv}(\mathbf{X})$, and randomly initialize all memory items $\mathbf{m}_i$ by using a multidimensional standard Gaussian distribution. The encoding feature $\mathbf{f}$ and memory items $\mathbf{m}_i$ are used to calculate attention values for memory addressing:

$$d(\mathbf{f}, \mathbf{m}_i) = \frac{\mathbf{f}\mathbf{m}_i^\top}{\|\mathbf{f}\|\|\mathbf{m}_i\|}, \tag{1}$$

$$\omega_i = \frac{\exp(d(\mathbf{f}, \mathbf{m}_i)}{\sum_{j=1}^{N} \exp(d(\mathbf{f}, \mathbf{m}_j))}, \tag{2}$$

where $d(\cdot, \cdot)$ denotes the cosine similarity. The memory addressing is performed to construct the decoding feature $\hat{\mathbf{f}}$ by soft weighting $\hat{\mathbf{f}} = \sum_{i=1}^{N} \omega_i \mathbf{m}_i$. We use the decoding feature $\hat{\mathbf{f}}$ for reconstructing $\mathbf{X}$ in a de-convolutional network, and obtain the reconstruction output $\hat{\mathbf{X}} = \text{DeConv}(\hat{\mathbf{f}})$. The reconstruction error $\|\mathbf{X} - \hat{\mathbf{X}}\|^2$ is treated as the loss function for optimizing the network parameters. All memory items $\mathbf{M}$ are also optimized together with the network parameters by using the gradient descent algorithm. The network parameters are shared across different visual cues in each type,

[1] In Dempster-Shafer theory, the frame-of-discernment denotes the set of exclusive assumed states, e.g., assumed classes for a sample.

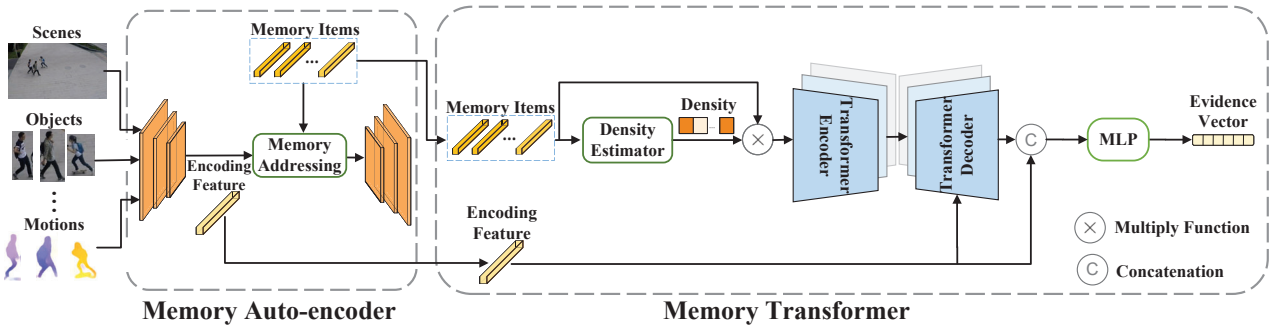

Figure 4: Illustration of the evidence encoder. The evidence encoder consists of a memory auto-encoder and a memory Transformer.

while the memory items are not shared. Once the network parameters and memory items are optimized, the encoding feature $\mathbf{f}$ and memory items $\mathbf{M}$ can be obtained by feeding a parsed sample $\mathbf{X}$ to the auto-encoder.

Memory Transformer. The evidence vector $\mathbf{e}$ is acquired in a memory Transformer, and its inputs include the encoding feature $\mathbf{f}$ and memory items $\mathbf{M}$. Considering the redundancy of $\mathbf{M}$ in recording normal patterns, we perform memory weighting to select more important normal patterns in $\mathbf{M}$, and then transform both the encoding feature $\mathbf{f}$ and the weighted memory items into the evidence vector $\mathbf{e}$ in support of estimating the uncertainty.

The memory items $\mathbf{M}$ record representative normal patterns. We argue that different patterns in memory items contribute differently to estimating the uncertainty of deviation probabilities, because anomalies are discriminated by modeling deviations from common patterns instead of uncommon patterns. Therefore, we use a density estimator to weight memory items based on their local density. A neural density estimator of autoregressive models [28] is used to generate densities as soft weights $\boldsymbol{\beta} = [\beta(\mathbf{m}_1), (\mathbf{m}_2), \cdots, \beta(\mathbf{m}_N)]$. The final weighted memory items are $\hat{\mathbf{M}} = [\hat{\mathbf{m}}_1, \hat{\mathbf{m}}_2, \cdots, \hat{\mathbf{m}}_N]^\top$, where $\hat{\mathbf{m}}_i = \beta(\mathbf{m}_i)\mathbf{m}_i$.

We introduce a self-attention Transformer network [40] to generate the evidence vector $\mathbf{e}$ for mining the relationships between the encoding feature $\mathbf{f}$ and its normal references (i.e., weighted memory items $\hat{\mathbf{M}}$). The weighted memory items $\hat{\mathbf{M}}$ are fed into a Transformer encoder to capture their internal relationships. A Transformer decoder captures co-contextual representations between the input memory items and the encoding feature, generating the transformed feature as

$$\mathbf{z} = \text{Transformer\_1}(\hat{\mathbf{M}}, \mathbf{f}). \tag{3}$$

The evidence vector $\mathbf{e} = [e_1, e_2, \cdots, e_K]$ is generated via a multilayer perceptron (MLP) with the input of the concatenation vector $[\mathbf{z}, \mathbf{f}]$:

$$\mathbf{e} = \text{ReLU}(\text{MLP\_1}([\mathbf{z}, \mathbf{f}]), \tag{4}$$

where $\text{ReLU}(\cdot)$ represents the ReLU activation function ensuring non-negative outputs. The network architectures of the Transformer and MLP_1 as well as the optimization of their learnable parameters are detailed in the experiment section.

3.2.2 Uncertainty Learner. We estimate the uncertainty in the uncertainty learner by learning an evidence distribution and deriving belief masses and probabilities assigned to a frame-of-discernment based on the distribution.

Belief Mass and Probability. We define the frame-of-discernment as K mutually exclusive singletons, i.e., $\Theta = \{\theta_1, \theta_1, \cdots, \theta_K\}$, which is similar to the assumed classes of a classification task. Since none of supervised information is available during training, we introduce the deep GMM to generate Gaussian clusters as potential classes of normal samples, and treat each Gaussian cluster as the singleton. The belief mass of the k-th singleton is denoted as b_k and the overall mass is u. We use the notation p_k to represent the probability with which instance is predicted to be the k-th Gaussian cluster. Both the probabilities of all membership predictions $\mathbf{p} = [p_1, p_2, \cdots, p_K]^\top$ and the overall mass u represent the uncertainty.

We calculate the belief masses b_k, overall mass u, and probabilities $\mathbf{p}$ by learning an evidence distribution instead of using conventional Dempster's rule [9, 34], thereby ensuring differentiable calculations in deep evidential reasoning. Learning an evidence distribution is an equivalent form of Dempster's rule given some constraint conditions [33], which is summarized in Theorem 1. The proof can be found in *Supplementary materials.*

Theorem 1. *When the frame-of-discernment of evidential reasoning is in the form of singletons, subjective logic formalizes evidential reasoning's notion of belief assignments as a Dirichlet distribution.*

Implementation. We establish a Dirichlet distribution as the evidence distribution based on the evidence vector $\mathbf{e}$ in Eq. (4). According to the work of Sensoy *et al.* [33], the belief assignment corresponds to a Dirichlet distribution with parameters $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_K]$:

$$\boldsymbol{\alpha} = \mathbf{e} + 1. \tag{5}$$

The belief mass b_k and overall mass u can be derived from the parameters $\boldsymbol{\alpha}$ by

$$\begin{aligned} b_k &= \frac{e_k}{\sum_{k=1}^{K} \alpha_k}, \\ u &= 1 - \sum_{k=1}^{K} b_k = \frac{K}{\sum_{k=1}^{K} \alpha_k}. \end{aligned} \tag{6}$$

The expected probability p_k with which an instance is predicted to be the k-th normal Gaussian cluster is computed by

$$p_k = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}. \tag{7}$$

Both p_k and u are calculated in a differentiable form with the input of the evidence vector $\mathbf{e}$, so the network parameters of both the memory Transformer and memory auto-encoder can be optimized jointly. To this end, we use the notation $\mathbf{P} \in \mathbb{R}^{M*K}$ to denote the probabilities with which M input instances are predicted to be K Gaussian clusters. We use the notation $\mathbf{z}_i$ to represent the transformed feature of the i-th instance in Eq. (3). An MLP is used to perform dimension reduction as $\mathbf{g}_i \in \mathbb{R}^D = \text{MLP\_2}(\mathbf{z}_i)$. $\mathbf{P}$ and $\mathbf{g}_i$ are computed in a differentiable feedforward process, and the network parameters are optimized by using the loss function

$$\begin{aligned}\mathcal{L} =& \frac{1}{M}\sum_{i=1}^{M} -\ln\Big(\sum_{k=1}^{K}\phi_k \frac{e^{-\frac{1}{2}(\mathbf{g}_i-\boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\mathbf{g}_i-\boldsymbol{\mu}_k)}}{\sqrt{|2\pi\Sigma_k|}}\Big) \\ &+ \lambda \sum_{k=1}^{K}\sum_{j=1}^{D}\frac{1}{\Sigma_k(j,j)},\end{aligned} \tag{8}$$

where

$$\phi_k = \sum_{i=1}^{M}\frac{\mathbf{P}(i,k)}{M}, \quad \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{M}\mathbf{g}_i\mathbf{P}(i,k)}{\sum_{i=1}^{M}\mathbf{P}(i,k)}, \tag{9}$$

$$\Sigma_k = \frac{\sum_{i=1}^{M}\mathbf{P}(i,k)(\mathbf{g}_i-\boldsymbol{\mu}_k)^\top(\mathbf{g}_i-\boldsymbol{\mu}_k)}{\sum_{i=1}^{M}\mathbf{P}(i,k)}. \tag{10}$$

The number of Gaussian clusters K is 10 and the trade-off parameter λ is 0.005.

3.3 Anomaly Scoring

We calculate frame-level anomaly scores based on the overall mass and probabilities from visual cues in different types for anomaly detection. As shown in Figure 5, anomaly scores are estimated in an anomaly scoring network. In each type, the feature vector $\mathbf{z}$ is concatenated with the probabilities $\mathbf{p} = [p_1, p_2, \cdots, p_K]^\top$ and overall mass u. The concatenated features are fed into a Transformer encoder and an MLP, modeling a new Dirichlet distribution with parameters $\tilde{\boldsymbol{\alpha}} = [\tilde{\alpha}_1, \tilde{\alpha}_2, \cdots, \tilde{\alpha}_K]$:

$$\tilde{\boldsymbol{\alpha}} = \text{ReLU}(\text{MLP\_3}(\text{Transformer\_2}([\mathbf{z}, \mathbf{p}, u]))) + 1. \tag{11}$$

We introduce the self-attention Transformer encoder to capture the internal relationships among all visual cues of features, probabilities, and overall mass for aggregation. Different from the memory Transformer in Eq. (3), we do not use the Transformer decoder because capturing co-contextual relationships is unnecessary for frame-level anomaly scoring. The loss function of the Transformer_2 and MLP_3 networks is similar to $\mathcal{L}$ in Eq. (8) for unsupervised clustering and network training. Once the networks are optimized, the frame-level anomaly score is derived from the Dirichlet distribution:

$$s = \frac{K}{\sum_{k=1}^{K}\tilde{\alpha}_k} - \max_j\left(\frac{\tilde{\alpha}_j}{\sum_{k=1}^{K}\tilde{\alpha}_k}\right). \tag{12}$$

Since anomalies significantly deviate from the normal patterns, they should not belong to any normal clusters, indicating that anomaly scores should have a lower assignment probability, i.e., $-\max_j \frac{\tilde{\alpha}_j}{\sum_{k=1}^{K}\tilde{\alpha}_k}$. Anomalies are unseen during training and their scores should have a larger overall mass $\frac{K}{\sum_{k=1}^{K}\tilde{\alpha}_k}$. The anomaly score should vary smoothly among frames, so we use a Gaussian filter to enforce temporal smoothness of the final frame-level anomaly scores.

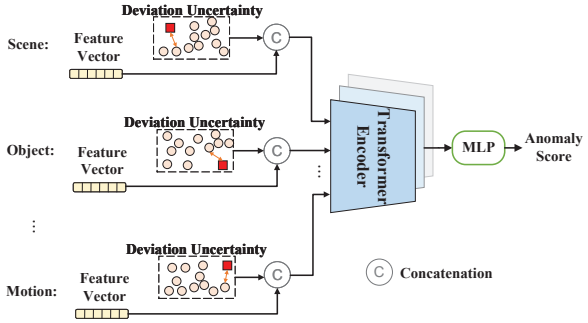

Figure 5: Illustration of frame-level anomaly scoring.

4 EXPERIMENTS

4.1 Datasets

We evaluate our method on the ShanghaiTech [21], Avenue [20], and UCSD Ped2 [23] datasets. The **ShanghaiTech** dataset contains 13 campus scenes with various viewpoints. It has over 310k video frames and 130 anomalies such as jumping, biking, and fighting. The **Avenue** dataset has 16 training and 21 testing videos with about 35k frames. Each video lasts about 2 minutes long. Anomalies include running, walking in opposite direction, throwing objects, and loitering. The **UCSD Ped2** dataset has 16 training and 12 testing videos with about 5k frames of pedestrian walkways. Anomalies include cars, person skating, bicycling, and so on.

4.2 Implementation Details

Architecture. We construct the memory auto-encoder proposed by the work of Gong *et al.* [12], generating the memory items $\mathbf{M} \in \mathbb{R}^{2000*128}$ and feature $\mathbf{f} \in \mathbb{R}^{128}$. The encoder and decoder in Transformer_1 both have 2 layers and 4 attention heads, and the embedding dimension is 32. The MLP_1 has the architecture of (256, 64, ReLU)-(64, 32, ReLU)-(32, 10, None), where (a, b, f) means a fully-connected layer with a trainable weight matrix $\mathbf{W} \in \mathbb{R}^{a*b}$ and an activation function f. The architecture of MLP_2 is (128, 64, ReLU)-(64, 32, None). Transformer_2 and MLP_3 have similar architectures as those in the memory Transformer.

Hyper-parameter. We set the memory size N to 2000 among all datasets, guaranteeing that most of the representative normal patterns are recorded. The cluster number K is set to 10 to cover common behavior classifications in the scenes of the three datasets (i.e., avenues, subways, and campuses).

Training. The training process is summarized in Algorithm 1. The training batch is set to 32, and we use the RMSprop optimizer with a 0.001 learning rate.

Algorithm 1 The training process.

Input: Training videos V.

Output: Optimized parameters $\{\Phi_M, \Phi_T, \Phi_A\}$ of the memory auto-encoder, memory Transformer and anomaly scoring network.

Pre-process V into the images or cropped sub-images $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ and $\mathbf{X}_4$ as visual cues in four types;
Randomly initialize the parameters $\{\Phi_M, \Phi_T, \Phi_A\}$;
for each type i **do**
    Reconstruct $\mathbf{X}_i$ as $\hat{\mathbf{X}}_i$ through the memory auto-encoder;
    Use the stochastic gradient descent (SGD) algorithm with the loss function of $\|\mathbf{X}_i - \hat{\mathbf{X}}_i\|^2$ to optimize parameters Φ_M;
    Extract memory items $\mathbf{M}$ and feature vectors $\mathbf{f}$ from the memory auto-encoder. Calculate the evidence vectors $\mathbf{e}$, probabilities $\mathbf{p}$, and overall mass u from the memory Transformers through Eqs. (3), (4) and (6);
    Use the SGD algorithm with the loss function in Eq. (8) to optimize parameters Φ_T;
end for
Concatenate $\mathbf{z}$, $\mathbf{p}$ and u into $[\mathbf{z}, \mathbf{p}, u]$ in all types, and calculate the frame-level scores in Eqs. (11) and (12) through the anomaly scoring network;
Use the SGD algorithm with a similar loss function in Eq. (8) to optimize parameters Φ_A;
return Network parameters $\{\Phi_M, \Phi_T, \Phi_A\}$.

Table 1: Anomaly detection results in terms of the frame-level AUROC on the ShanghaiTech dataset.

Method	AUROC↑
Hasan *et al.* ([14])	60.85%
Morais *et al.* ([25])	73.4%
Park *et al.* ([29])	72.8%
Markovitz *et al.* ([24])	76.1%
Cai *et al.* ([4])	73.7%
Szymanowicz *et al.* ([37])	70.4%
Astrid *et al.* ([1])	73.7%
Hao *et al.* ([13])	73.8%
Liu *et al.* ([19])	76.2%
Chen *et al.* ([7])	78.1%
Ours	**79.3**%

4.3 Evaluation Metric

We generate frame-level anomaly scores, and compute the Area Under the Receiver Operation Characteristic (AUROC ↑) by gradually changing the threshold of anomaly scores for evaluation. We also compute the Equal Error Rate (EER ↓) for evaluation. A higher AUROC value and a lower EER value indicate a better performance.

4.4 Comparisons

Results on the ShanghaiTech Dataset. We report the AUROC performance of our method and state-of-the-art methods in Table 1. The performances of all compared methods are taken from their original paper. Our method outperforms the state-of-the-art method of Chen *et al.* [7] on the ShanghaiTech dataset, gaining an improvement of 1.2% on the AUROC evaluation. Since the ShanghaiTech dataset is recognized as a challenging benchmark of video anomaly detection, the performance demonstrates the superiority of our method.

Table 2: Anomaly detection results in terms of the frame-level AUROC and EER on the Avenue dataset.

Method	AUROC↑	EER↓
Hasan *et al.* ([14])	70.2%	25.1%
Liu *et al.* ([18])	84.9%	-
Wang *et al.* ([41])	85.3%	23.9%
Morais *et al.* ([25])	86.3%	-
Park *et al.* ([29])	88.5%	-
Cai *et al.* ([4])	87.4%	-
Li *et al.* ([17])	83.5%	23.5%
Hao *et al.* ([13])	86.6%	-
Astrid *et al.* ([1])	87.1%	-
Chen *et al.* ([7])	90.3%	-
Liu *et al.* ([19])	91.1%	-
Ours	**92.7**%	**20.0**%

Table 3: Anomaly detection results in terms of the frame-level AUROC and EER on the UCSD Ped2 dataset.

Method	AUROC↑	EER↓
Hasan *et al.* ([14])	90.0%	21.7%
Chong and Tay ([8])	87.4%	12.0%
Ravanbakhsh *et al.* ([30])	95.5%	14.0%
Park *et al.* ([29])	97.0%	-
Cai *et al.* ([4])	96.6%	-
Hao *et al.* ([13])	96.9%	-
Chen *et al.* ([7])	98.3%	-
Astrid *et al.* ([1])	98.4%	-
Liu *et al.* ([19])	**99.3**%	-
Ours	97.1%	**10.9**%

Results on the Avenue Dataset. Experimental results in Table 2 show that our method outperforms all compared methods on both the AUROC and EER evaluations on the Avenue dataset. The state-of-the-art work of Park *et al.* [19] achieves the AUROC values of 91.1%, and our method gains an improvement of 1.6%, showing the superiority of our method.

Results on the UCSD Ped2 Dataset. We compare our method with existing methods on the UCSD Ped2 dataset in Table 3. Our method achieves slightly worse results on the UCSD Ped2 dataset compared with state-of-the-art methods of [1, 7, 19]. Nevertheless, our method outperforms them on the other two datasets, with improvements of 5.6%, 1.2%, and 3.1% on the ShanghaiTech dataset as well as improvements of 5.6%, 2.4%, and 1.6% on the Avenue dataset, respectively. The reason for the dropped performance is that the UCSD Ped2 dataset is relatively small so that our deep networks are prone to overfitting, and the overfitting problem may be solved by performing data augmentation in the future study.

Table 4: The AUROC and the EER of different components of our method on the Avenue dataset.

	Method	AUROC↑	EER↓
Visual Cues	w/ scene	61.8%	34.7%
	w/ appearance	87.6%	23.3%
	w/ relationship	78.0%	27.2%
	w/ motion	83.6%	26.5%
	w/o scene	91.2%	20.9%
	w/o appearance	78.8%	25.4%
	w/o relationship	88.2%	23.1%
	w/o motion	83.6%	24.4%
Memory	w/o memory	82.4%	25.5%
	w/o weights	85.2%	24.5%
Uncertainty	w/o uncertainty	89.7%	21.3%
Scoring	w/ max-pooling	88.1%	21.5%
	w/ average-pooling	84.4%	24.0%
	w/o GMM	88.5%	22.4%
	Ours	**92.7%**	**20.0%**

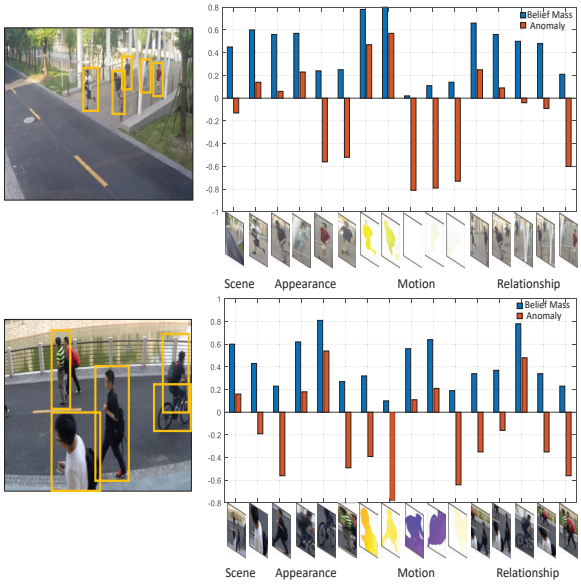

Figure 6: Qualitative results of overall mass scores and anomaly scores from visual cues in four types.

4.5 Ablation Study

We compare the contributions of different components of our method in Table 4, including the visual cues, the memory items, the uncertainty, and the scoring network.

Analysis of Visual Cues. Table 4 shows the quantitative results about four types of visual cues. "w/ scene", "w/ appearance", "w/ relationship" and "w/ motion" denote that the uncertainty is only estimated based on visual cues in one corresponding type. "w/o scene", "w/o appearance", "w/o relationship" and "w/o motion" means that removing the belief mass and probability from visual cues in the corresponding type. It can be seen that: (1) The visual cue of object appearance plays the most important role because when discarding it, the largest decrease of the AUROC is obtained. This represents that anomalies in videos are often object-centric. (2) When discarding any cues, the AUROC value is reduced by 1.5%-13.9%, which further verifies the importance of various visual cues.

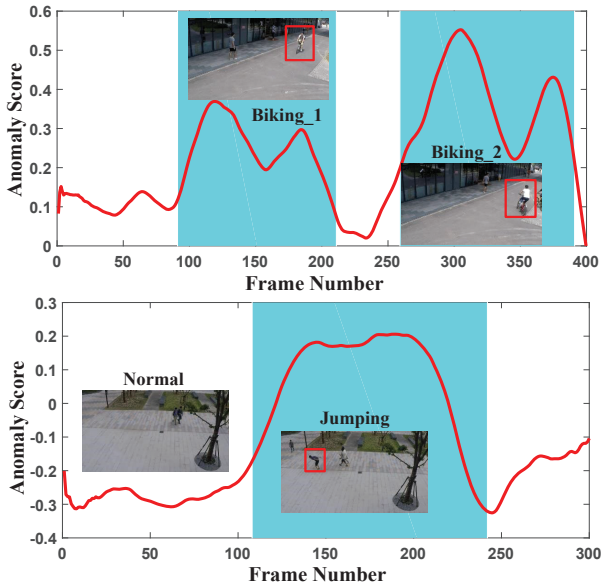

Figure 7: Qualitative results on the ShanghaiTech dataset. Frames in white and blue windows are the ground-truth normal and anomaly events, respectively. Anomaly scores in our method match well with the ground-truth annotations.

Figure 6 visualizes the overall mass score $\frac{K}{\sum_{k=1}^{K}\tilde{\alpha}_k}$ and anomaly scores s in Eq. (12) of two anomaly examples with visual cues in different types. It can be seen that: (1) The high scores of "running" are acquired from the optical-flow images because the behavior pattern tends to be discriminated based motions. (2) The "biking" can be detected from object appearance images with higher scores.

Analysis of Memory. We analyze the importance of the memory items in Table 4. "w/o memory" represents that we only use the MLP_1 with inputs of the encoding feature to obtain the belief and probability, where the memory Transformer is removed. "w/o weight" denotes removing the memory weighting. It can be seen that: (1) When discarding the memory items, the performance of the AUROC drops sharply from 92.7% to 82.4%, which verifies the necessity of explicitly adopting the memory items to record the reference samples. (2) The performance is significantly improved by employing the weights, demonstrating the effectiveness of dynamically representing normal patterns.

Analysis of Uncertainty. "w/o uncertainty" in Table 4 means that we remove the evidential distributions and directly obtain the prediction probability by $\mathbf{p} = \text{softmax}(\mathbf{e})$, and the removal of uncertainty decreases the performance of anomaly detection. The possible reason is that unknown anomalies may be wrongly assigned to a certain class without considering the uncertainty. We show two qualitative results of our uncertain-aware anomaly scores in Figure 7. Blue windows show ground-truth labels of

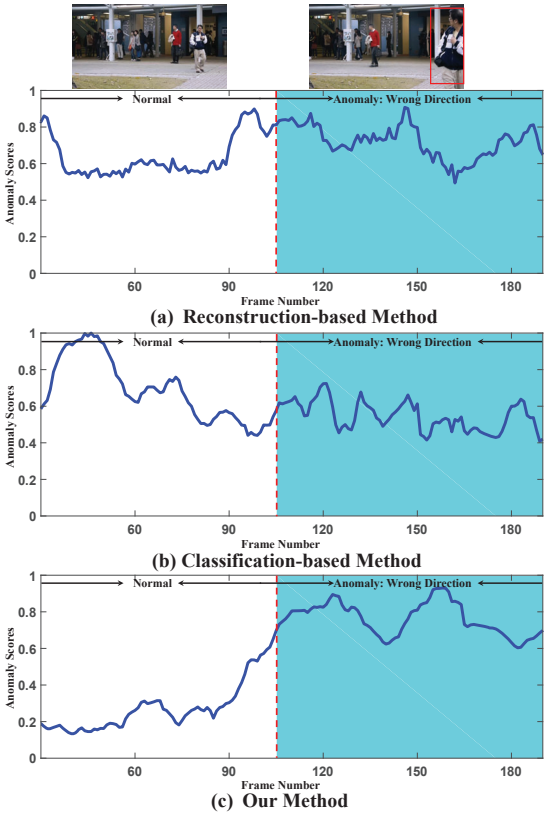

Figure 8: Frame-level anomaly scores of an example video. Frames in white and blue windows are the ground-truth normal and anomaly events, respectively. Red dotted lines represent their ground-truth boundary. It is easy to construct the correct decision boundary between normal and anomaly frames by using our method to compute anomaly scores.

anomalies, and red curves represent the smoothed anomaly scores of our method. The scores match well with the anomaly annotations, and the score gaps between normal and anomaly frames are large, indicating good discrimination. We also compare our method with a reconstruction-based method [12] and a classification-based method [16] in Figure 8. For a fair comparison, we use the same backbone (i.e., memory auto-encoder and memory Transformer) to calculate anomaly scores in [12, 16], and the inputs are visual cues of the scene type. The results in Figure 8 verify the benefits of learning the uncertainty for modeling the boundaries of video anomalies. Furthermore, we project scene features $\mathbf{z}$ in Eq. (3) on the Avenue dataset onto a 2D space by using t-SNE. The results of two example videos are shown in Figure 9, where red and blue points represent anomaly and normal frames, respectively. It can be seen that the uncertainty estimation will separate normal and anomaly frames clearly for better discrimination.

Analysis of Scoring. We use a Transformer encoder to aggregate visual cues for scoring. We compare the Transformer encoder with two aggregation methods of max-pooling and average-pooling. "w/ max-pooling" and "w/ average-pooling" denote replacing the

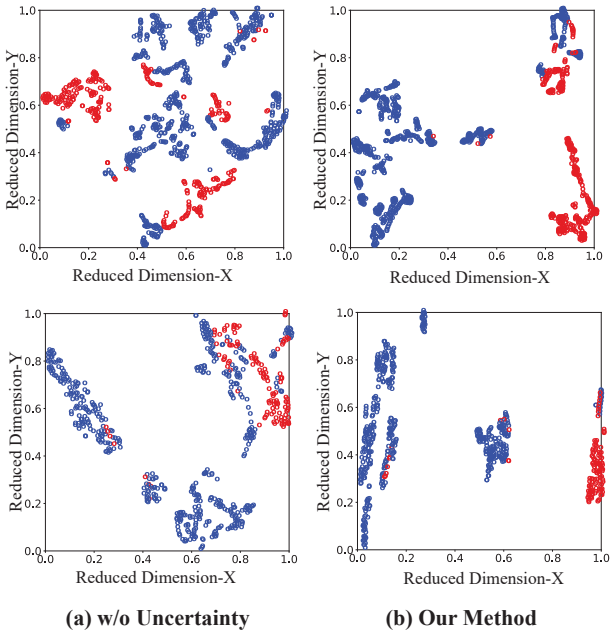

Figure 9: Visualizations of learned features of two example videos by using t-SNE on the Avenue dataset. Red and blue points represent anomaly and normal frames, respectively.

Transformer_2 in Eq. (11) with pooling operations. The results show that the self-attention aggregation in the Transformer encoder is more effective than the pooling operation. In addition, we use the deep GMM to train our scoring network, and perform an ablation study of "w/o GMM" to verify its effectiveness. "w/o GMM" means replacing the end-to-end training strategy of the deep GMM by using another learning strategy of pseudo-labels [16, 35], where we use the k-means algorithm to obtain pseudo-labels for training our scoring network by establishing a classifier. The AUCROC performance of using the deep GMM is 92.7%, and is higher than the 88.5% of using the pseudo-label strategy, which verifies the contribution of the GMM in our method.

5 CONCLUSION AND DISCUSSION

We have presented a deep evidential reasoning method that can learn the uncertainty in the probability of deviating from normal patterns to model the boundaries of video anomalies without anomaly annotations. We build a deep evidential reasoning network that can both encode evidences by mining various visual cues and estimate the uncertainty by deriving beliefs and probabilities from the evidence distribution. We assign the beliefs and probabilities to normal clusters of the deep GMM, which can train the network in an unsupervised manner. Experimental results demonstrate the benefits of learning the uncertainty for modeling the decision boundaries of video anomalies.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (NSFC) under Grants No. 62176021 and No. 62172041.

REFERENCES

[1] Marcella Astrid, Muhammad Zaigham Zaheer, and Seung-Ik Lee. 2021. Synthetic Temporal Anomaly Guided End-to-End Video Anomaly Detection. In *Proc. IEEE Int. Conf. Comput. Vis.* 207–214.

[2] Wentao Bao, Qi Yu, and Yu Kong. 2021. Evidential Deep Learning for Open Set Action Recognition. In *Proc. IEEE Int. Conf. Comput. Vis.* 13349–13358.

[3] Abhijit Bendale and Terrance E Boult. 2016. Towards open set deep networks. In *Proc. Int. Conf. Pattern Recognit.* 1563–1572.

[4] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. 2021. Appearance-Motion Memory Consistency Network for Video Anomaly Detection. In *Proc. Conf. Artif. Intell.*, Vol. 35. 938–946.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41, 3 (2009), 15:1–15:58.

[6] Yunpeng Chang, Zhigang Tu, Wei Xie, Bin Luo, Shifu Zhang, Haigang Sui, and Junsong Yuan. 2022. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognit.* 122 (2022), 108213.

[7] Chengwei Chen, Yuan Xie, Shaohui Lin, Angela Yao, Guannan Jiang, Wei Zhang, Yanyun Qu, Ruizhi Qiao, Bo Ren, and Lizhuang Ma. 2022. Comprehensive Regularization in a Bi-directional Predictive Network for Video Anomaly Detection. In *Proc. Conf. Artif. Intell.* 1–9.

[8] Yong Shean Chong and Yong Haur Tay. 2017. Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder. In *Proc. Adv. Neural Net.* 189–196.

[9] Arthur P. Dempster. 2008. Upper and Lower Probabilities Induced by a Multivalued Mapping. In *Classic Works of the Dempster-Shafer Theory of Belief Functions.* Vol. 219. 57–72.

[10] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. 2022. Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 13678–13688.

[11] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. 2021. Convolutional Transformer based Dual Discriminator Generative Adversarial Networks for Video Anomaly Detection. In *Proc. ACM Conf. Multimedia.* 5546–5554.

[12] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *Proc. IEEE Int. Conf. Comput. Vis.* 1705–1714.

[13] Yi Hao, Jie Li, Nannan Wang, Xiaoyu Wang, and Xinbo Gao. 2022. Spatiotemporal consistency-enhanced network for video anomaly detection. *Pattern Recognit.* 121 (2022), 108232.

[14] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. 2016. Learning Temporal Regularity in Video Sequences. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 733–742.

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 1647–1655.

[16] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. 2019. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 7842–7851.

[17] Nanjun Li, Faliang Chang, and Chunsheng Liu. 2021. Spatial-Temporal Cascade Autoencoder for Video Anomaly Detection in Crowded Scenes. *IEEE Trans. Multim.* 23 (2021), 203–215.

[18] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future Frame Prediction for Anomaly Detection - A New Baseline. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 6536–6545.

[19] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. In *Proc. IEEE Int. Conf. Comput. Vis.* 13588–13597.

[20] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal Event Detection at 150 FPS in MATLAB. In *Proc. IEEE Int. Conf. Comput. Vis.* 2720–2727.

[21] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In *Proc. IEEE Int. Conf. Comput. Vis.* 341–349.

[22] Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. 2021. Future Frame Prediction Network for Video Anomaly Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).

[23] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2009. Anomaly detection in crowded scenes. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 1975–1981.

[24] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. 2020. Graph Embedded Pose Clustering for Anomaly Detection. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 10536–10544.

[25] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Reda Mansour, and Svetha Venkatesh. 2019. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 11996–12004.

[26] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. 2021. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.* 54, 2 (2021), 38:1–38:38.

[27] Guansong Pang, Chunhua Shen, and Anton van den Hengel. 2019. Deep anomaly detection with deviation networks. In *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 353–362.

[28] George Papamakarios, Theo Pavlakou, and Iain Murray. 2017. Masked autoregressive flow for density estimation. In *Proc. Adv. Neural Inf. Process. Syst.*, Vol. 30.

[29] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. 2020. Learning Memory-Guided Normality for Anomaly Detection. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 14360–14369.

[30] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo S. Regazzoni, and Nicu Sebe. 2017. Abnormal event detection in videos using generative adversarial nets. In *Proc. Int. Conf. Image Process.* 1577–1581.

[31] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proc. Adv. Neural Inf. Process. Syst.* 91–99.

[32] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. 2018. Adversarially Learned One-Class Classifier for Novelty Detection. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 3379–3388.

[33] Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *Proc. Adv. Neural Inf. Process. Syst.* 3183–3193.

[34] Glenn Shafer. 1976. *A mathematical theory of evidence.* Princeton university press.

[35] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. 2020. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proc. ACM Conf. Multimedia.* 184–192.

[36] Che Sun, Yunde Jia, Hao Song, and Yuwei Wu. 2021. Adversarial 3D Convolutional Auto-Encoder for Abnormal Event Detection in Videos. *IEEE Trans. Multimedia* 23 (2021), 3292–3305.

[37] Stanislaw Szymanowicz, James Charles, and Roberto Cipolla. 2021. X-MAN: Explaining multiple sources of anomalies in video. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 3224–3232.

[38] Hanh Tran and David C. Hogg. 2017. Anomaly Detection using a Convolutional Winner-Take-All Autoencoder. In *Proc. Brit. Mach. Vis. Conf.*

[39] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. 2017. Unmasking the Abnormal Events in Video. In *Proc. IEEE Int. Conf. Vis. Pattern Recognit.* 2895–2903.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proc. Adv. Neural Inf. Process. Syst.* 5998–6008.

[41] Siqi Wang, Yijie Zeng, Qiang Liu, Chengzhang Zhu, En Zhu, and Jianping Yin. 2018. Detecting Abnormality without Knowing Normality: A Two-stage Approach for Unsupervised Video Abnormal Event Detection. In *Proc. ACM Conf. Multimedia.* 636–644.

[42] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.* 156 (2017), 117–127.

[43] Ke Xu, Xinghao Jiang, and Tanfeng Sun. 2018. Anomaly Detection Based on Stacked Sparse Coding With Intraframe Classification Strategy. *IEEE Trans. Multimedia* 20, 5 (2018), 1062–1074.

[44] Ke Xu, Tanfeng Sun, and Xinghao Jiang. 2020. Video Anomaly Detection and Localization Based on an Adaptive Intra-Frame Classification Network. *IEEE Trans. Multimedia* 22, 2 (2020), 394–406.

[45] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David Crandall. 2022. DoTA: unsupervised detection of traffic anomaly in driving videos. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022).

[46] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. 2017. Spatio-temporal autoencoder for video anomaly detection. In *Proc. ACM Conf. Multimedia.* 1933–1941.