Context-aware Video Anomaly Detection in Long-Term Datasets

Zhengye Yang
Rensselaer Polytechnic Institute
Department of ECSE, Troy, NY, USA
yangz15@rpi.edu

Richard J. Radke
Rensselaer Polytechnic Institute
Department of ECSE, Troy, NY, USA
rjradke@ecse.rpi.edu

Abstract

Video anomaly detection research is generally evaluated on short, isolated benchmark videos only a few minutes long. However, in real-world environments, security cameras observe the same scene for months or years at a time, and the notion of anomalous behavior critically depends on context, such as the time of day, day of week, or schedule of events. Here, we propose a context-aware video anomaly detection algorithm, Trinity, specifically targeted to these scenarios. Trinity is especially well-suited to crowded scenes in which individuals cannot be easily tracked, and anomalies are due to speed, direction, or absence of group motion. Trinity is a contrastive learning framework that aims to learn alignments between context, appearance, and motion, and uses alignment quality to classify videos as normal or anomalous. We evaluate our algorithm on both conventional benchmarks and a public webcam-based dataset we collected that spans more than three months of activity.

1. Introduction

The goal of video anomaly detection (VAD) is to understand what is "normal" and then to identify anomalous events that deviate from the expected distribution. However, the definition of "normal" critically depends on *context* information such the time of day, day of week, or schedule of events. For example, a camera monitoring a campus would expect to see groups of people entering campus on a typical weekday morning, but the same scenario during a weekday evening or weekend would be considered anomalous. It would be impossible to accurately detect truly anomalous behaviour without referring to the date and time. While cameras operating over extended periods have the potential to capture and model such time-varying patterns [1], current VAD algorithms are generally focused on short, isolated clips and lack any sort of contextual awareness. This critical gap prevents current algorithms from being immediately applicable to real-world camera networks.

A naïve solution to this challenge is to develop separate VAD models for different contexts [26]. However, given that contextual elements like time are continuous, hard classification of contexts can be difficult and inevitably introduce bias. For example, a baseball stadium, as depicted in Figure 1, may have a varied game schedule, with differing game timings affecting the day's dynamics. Reducing these complex context-dependent dynamics to time-of-day variations by training multiple models is overly simplistic. Creating a context-sensitive VAD framework can aid in detecting anomalies unidentifiable by current techniques, a critical consideration for real-world camera systems, especially those monitoring soft targets such as malls, stadiums, and schools.

A few recent VAD datasets recognize the significance of context in anomalies. Leroux et al. [12] synthetically modified videos (e.g., adding a raindrop overlay) to simulate different conditions to improve the robustness of a VAD algorithm. The Street Scene dataset [25] highlights spatial anomalies like jaywalking and wrong-way incidents, while the NOLA dataset [4] collected videos from specific public areas over a week, accompanied by metadata. The latter emphasizes the continual learning aspect of VAD without explicitly correlating video and context. Both temporal and spatial contexts are crucial for comprehensive contextual awareness. To address the long-term contextual awareness requirement for a real-world VAD model, we collected a novel dataset comprising regularly captured videos for a period of three months, augmented with contextual data.

Most VAD benchmarks only contain object-level anomalies (e.g., a cyclist on a pedestrian sidewalk), and do not require holistic scene understanding. While object-centric methods perform highly on standard benchmarks [7, 10, 17], they primarily detect easily-segmented foreground objects. When applied to long-term real-world videos, issues such as complex lighting, diverse weather conditions, and occlusion in crowded settings (as illustrated in Figure 1) make detecting/tracking individuals difficult or impossible, which limits the application of this approach. The computational cost of processing a single clip also scales with the

11:30 AM Non-game day   **11:30 AM Game day**   **11:30 AM Unusual presence**

Figure 1. Snapshots at the same time of day at a baseball stadium. **Left:** A typical scenario during a non-game day; **Middle**: A typical scenario during a game day; **Right**: Unexpected group presence.

number of objects in a scene. Furthermore, neither object-centric nor frame-based methods are designed to detect absence anomalies (e.g., no one shows up on a school day).

In this work, our primary goal is to construct a VAD algorithm based on video-context correspondence. We hypothesize that context awareness can be built from the co-occurrence between visual features and their contexts. Inspired by CLIP [24], we project visual representations (appearance and motion) and a context representation into a joint embedding space. Appearance and motion can be considered as multiple perspectives of the same video clip, while the context information is an additional modality we want to align with the video clip.

To learn a robust representation in crowded environments and enable contrastive learning of video without context, we consider appearance and motion representations from video clips as multiple views for local alignment. Through contrastive learning, the context information will guide the model to learn a global representation of the current scenario and detect context anomalies by evaluating the alignment performance. Further, the strong correlation between appearance and motion can enable learning appearance-motion alignment at the local level. During inference, misalignment between appearance and motion can also be used to detect anomalies in traditional VAD benchmark datasets.

The contributions of this paper are as follows:

- We propose a real-world VAD task using an exemplar dataset addressing long-term context awareness.
- We propose a new contrastive-learning-based framework called Trinity to learn a joint embedding between modalities globally and locally, which uses alignment error to detect anomalies.
- We present experiments for detecting context-dependent and independent anomalies on both self-collected and benchmark datasets to illustrate the effectiveness of the Trinity framework.

2. Related Work

Video anomaly detection. Frame-based approaches [3, 8, 9, 15, 16, 20, 21, 23, 39] and object-centric approaches [2, 7, 10, 17, 31] are the two main branches of detecting video anomalies. In frame-based methods, detecting anomalies is typically based on reconstruction error. For example, Hasan *et al.* [9] proposed to use a convolutional autoencoder (AE) in which reconstruction failures indicate anomalies. Gong *et al.* [8] proposed MemAE, using an external memory array that is jointly optimized with an AE to further enlarge the reconstruction error between normal and abnormal objects. Another line of work also takes an AE as its main component to predict the next frame and uses the prediction discrepancy as the anomaly indicator [15, 20, 23, 27]. MNAD [23] concatenated encoded features with the closest prototypical pattern to pass onto the decoder. Recently, DMAD [16] added a deformable model to capture nuances between normal variations and anomalous events to improve performance. The recent progress of unsupervised VAD has slowly shifted to object-centric methods [7, 10, 17, 31], which use pre-trained object detection or tracking algorithms to segment objects used as the input instead of the whole frame. HF^{2}VAD [17] proposed object detection and a multi-level memory module with shortcuts to memorize prototypical flow patterns and used its reconstruction to further condition on the next frame prediction. These newly proposed methods use the strong prior that only moving objects are considered as anomalous region candidates, which cannot detect absence/presence type anomalies.

Context modeling. Prior work on building scene awareness into VAD [2, 31] can be considered as capturing a special case of context. For example, Sun *et al.* [30] proposed a spatial-temporal context (STC) graphical model to capture context information about the current video tube to improve the performance of VAD. It treated bounding boxes from object detection as nodes for building the context graphic model. Sun and Gong [31] used scene segmentation, tracklets, and a pretrained vision transformer [5] to parse the
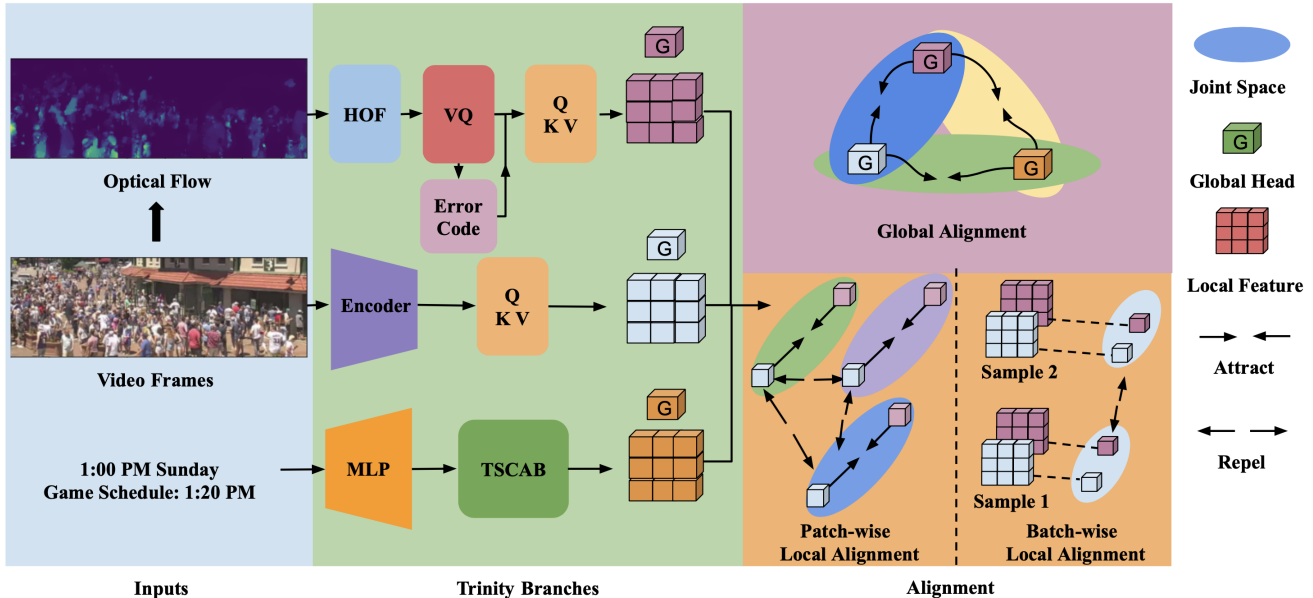
Figure 2. Pipeline for the proposed Trinity algorithm, which takes video frames, optical flow, and contextual information as inputs. The Trinity branches extract global and local representation from each input stream. Global and local alignment are used to learn joint embeddings at different scales and later used to determine anomalies by evaluating misalignments between branches.

video. They used memory modules to memorize the prototypical patterns of scene-appearance and scene-motion features with contrastive learning to enlarge the difference between different scene representations.

Context-aware anomaly detection in more general modalities has been studied [13, 14, 29]. For example, Song *et al*. [29] proposed a method called Context Anomaly Detection (CAD) that uses a Gaussian mixture model to model context feature and behavior feature distributions separately with regression mapping as a linkage to detect context anomalies. As mentioned in Tian *et al*. [32], a predictive objective is less effective than using a contrastive objective for learning a better representation under the multi-view setting. In this paper, we use contrastive learning to align the context information and visual representations to build contextual awareness without using auxiliary pre-trained models for context modeling.

Long term datasets. The LOST dataset [1] provided videos from 19 cameras from the same half-hour each day over the course of a year, along with metadata. Through trajectory clustering, the aggregated statistics show strong correlations with external metadata, but most of the following work focused on trajectory prediction or trajectory abnormality detection [6, 36] without investigating the context anomalies. The NOLA dataset [4] consistently collected videos in specific public areas over a week, accompanied by metadata. Although it contains context-dependent anomalies like loitering, the random camera position change makes it unsuitable for frame-based methods.

It emphasizes the continual learning aspect of VAD without explicitly correlating video and context. The LTD dataset [22] collected 8 months of videos from a thermal camera with context information like weather and time for investigating concept drifts like seasonal changes. In this paper, to better investigate the video-context correspondence, we collect a new dataset containing videos with context information targeting a baseball stadium that has a complex game schedule and potentially crowded scenarios.

3. Method

The overview of our proposed method, which we call Trinity, is shown in Fig. 2. Trinity contains three branches: a context branch, an appearance branch, and a motion branch. In this section, we first introduce the design of each branch and describe the training scheme of the model. We then discuss how to use global and local alignment to detect context anomalies.

3.1. Branches of the Trinity Model

Consider a batch of video clips $\{\mathbf{x}_i, \ldots, \mathbf{x}_B\} \in \mathbb{R}^{C \times T \times H \times W}$, where B, C, T, H, W denote batch size, channel size, number of frames, height, and width respectively. For later processing, each frame is split into a $k \times k$ grid of nonoverlapping patches. The corresponding context information is $\{\mathbf{c}_1, \ldots, \mathbf{c}_B\} \in \mathbb{Z}^Y$, where Y represents the dimension of concatenated one-hot encodings of context information, discussed further below. Each branch takes different source information as input but generates the same-

sized representation as output, containing the concatenation of a global head token and a local token resulting in $\mathbf{h} \in \mathbb{R}^{B \times N \times D}$, where $N = \frac{H \times W}{k^2} + 1$ and D represents the dimension of the joint embedding space. Detailed structure information is provided in the Supplementary Material Sec. 7.

3.1.1 Context Branch

The context information is treated as a concatenation of one-hot encodings of each type of context. For example, in our stadium dataset, different types of context include the hour of day, day of week, a flag for game/non-game day and the corresponding game schedule. We use a simple two-layer MLP to project the context information into a context embedding: $\{\mathbf{P}_1^{\mathbf{t}}, \ldots, \mathbf{P}_B^{\mathbf{t}}\} \in \mathbb{R}^{d_t}$. To enable the local alignment between context and other modalities, we use a learnable positional encoding: $\mathbf{P}_i^{\mathbf{s}} \in \mathbb{R}^{d_p}, i \in \{1, \ldots, N\}$ embedding as the context prototypical representation.

Temporal/Spatial Context Attention Block. To learn the relationship between temporal and spatial context in an unsupervised fashion, we propose a Temporal/Spatial Context Attention Block (TSCAB) that is able to adjust the spatial embedding $\mathbf{P}^{\mathbf{s}}$ based on contextual information $\mathbf{P}_j^{\mathbf{t}}$ at batch index j. The key ingredient here is to intertwine information between the positional embedding and the temporal context embedding through Multi-head Cross Attention (MCA) [34] as below:

$$\hat{\mathbf{P}}_i^{\mathbf{s}} = W_{proj} \cdot \text{MultiHead}$$

$$\text{MultiHead} = Softmax\left(\frac{Q \cdot K^T}{\sqrt{d_p}}\right) \cdot V \quad (1)$$

$$Q = W_Q \cdot \mathbf{P}_j^{\mathbf{t}}, \quad K = W_K \cdot \mathbf{P}_i^{\mathbf{s}}, \quad V = W_V \cdot \mathbf{P}_i^{\mathbf{s}}$$

Through a series of TSCAB operations, we obtain the augmented positional embedding $\hat{\mathbf{P}}_i^{\mathbf{s}}$ that serves as the basis to align features from other branches via a projection $W_{proj}^{cxt} \in \mathbb{R}^{d_p \times D}$ shown in (2), where $[\cdot]$ is the concatenation operation. The context representation is $\mathbf{h}_{cxt} \in \mathbb{R}^{B \times N \times D}$.

$$\mathbf{h}_{cxt} = [\hat{\mathbf{P}}_1^{\mathbf{s}}, \ldots, \hat{\mathbf{P}}_N^{\mathbf{s}}] \cdot W_{proj}^{cxt} \quad (2)$$

3.1.2 Motion Branch

Feature preparation. We use the Histogram of Optical Flow (HOF) as the representation of motion for its simplicity. We use the TV-L1 method [38] to extract a pixel-wise flow field and divide the input into non-overlapping patches for calculating the histogram of flow. Following [28], each patch contains 12 orientation bins, a background ratio representing the proportion of pixels with magnitude below a threshold, and the average magnitude for each orientation bin, resulting in a total of 25 channels for each local patch.

Vector Quantization. In our initial attempts, we found that using a convolution block to extract flow information tends to downplay the importance of flow magnitude. To better preserve the magnitude information, we use a vector quantization layer [33] to formulate a vocabulary $\mathbf{z}_n \in \mathbb{R}^{25}, n \in \{1, \ldots, M\}$ with vocabulary size M after extracting the HOF features $\mathbf{z}_{mot}^e$ shown in (3). The index of the quantization vector is used as the input token to the following transformer [34] blocks.

$$\mathbf{z}_{mot}^q = \arg\min_{\mathbf{z}_n} ||\mathbf{z}_{mot}^e - \mathbf{z}_n||_2 \quad (3)$$

Error Code Usage. Using the vector quantization layer for tokenizing motion has one problem: it is unable to pass unseen motions into the motion branch. To identify unseen motions, we propose a simple method called Error Code Usage. For the given HOF feature $\mathbf{z}_{mot}^e$, we obtain the quantized feature $\mathbf{z}_{mot}^q$, and further encode the residual between $\mathbf{z}_{mot}^e$ and $\mathbf{z}_{mot}^q$ using the same vocabulary to get the error word $\mathbf{z}_{mot}^{err}$. The intuition behind the error code is normally that the VQ layer can find a $\mathbf{z}_{mot}^q$ that is close to $\mathbf{z}_{mot}^e$, leaving a residual that is nearly the zero vector. The following model will learn this as a normal word pair. If an unseen motion happens, the residual word contains a large motion and the resulting word pair is unseen, causing misalignment between the branches.

We use the same embedding to project both $\mathbf{z}_{mot}^q$ and $\mathbf{z}_{mot}^{err}$ and concatenate the representation, feeding word pairs of the whole clip into k_{mot} transformer blocks. After the transformer blocks, we use a simple projection to generate the motion representation $\mathbf{h}_{mot} \in \mathbb{R}^{B \times N \times D}$ for joint embedding alignment.

3.1.3 Appearance Branch

Since Trinity is agnostic to the appearance feature extraction model, we simply adopt the basic U-net structure from MNAD [23], which is widely used in the VAD community. The U-net framework contains an encoder $f(\mathbf{x})$ and a decoder $g(\cdot)$ with shortcuts in between to perform the reconstruction task shown in (4).

$$L_{recon} = ||\mathbf{x} - g(f(\mathbf{x}))||_2 \quad (4)$$

The latent representation $\mathbf{z}$ is extracted through the encoder $f(\mathbf{x}), \mathbf{z} \in \mathbb{R}^{d' \times h' \times w'}$. To avoid adding disturbance to the reconstruction, we apply the stop gradient to the extracted features before feeding into the transformer blocks. To tokenize the extracted feature map, we use a 2D convolution layer with the same kernel size and stride $(r_h, r_w) = \left(\frac{h'}{h}, \frac{w'}{w}\right)$ to embed features from the deepest layer of the encoder. Similar to the motion branch, after k_{app} transformer blocks and a final projection layer, we obtain the appearance representation $\mathbf{h}_{app} \in \mathbb{R}^{B \times N \times D}$.

The three Trinity branches result in $\mathbf{h}_{app}$, $\mathbf{h}_{mot}$, and $\mathbf{h}_{cxt}$, the output representations for the alignment training described in the next section.

3.2. Aligning Feature Representations

The following alignment process contains both global and local alignment. The global representation is denoted as $\mathbf{h}^g \in \mathbb{R}^{B \times 1 \times D}$, and the local representation is denoted as $\mathbf{h}^l \in \mathbb{R}^{B \times N-1 \times D}$, where $\mathbf{h} = [\mathbf{h}^g, \mathbf{h}^l]$. Our focus is on context-dependent anomalies; in Section 4.5 we describe how the framework can be modified to also detect context-free anomalies (i.e., the traditional VAD problem).

3.2.1 Global Alignment

The global representation gathers the information from local patches through transformer blocks to provide a holistic representation of the current video clip. This is vital for formulating context awareness. Global alignment treats all samples within the batch except the target sample as negative samples. We use the extracted global tokens from selected branch pairs $\mathbf{h}_1^g, \mathbf{h}_2^g$ as inputs. The global alignment loss for a pair $L_{1,2}^{global}$ is calculated through the symmetric binary cross entropy loss as employed in CLIP [24] shown in (5), where $+$ is the indicator of paired sample and τ is a learnable temperature parameter.

$$L_{1,2}^{global} = -\sum_{i=1}^{B} \left(\log \text{NCE}(\mathbf{h}_1^g, \mathbf{h}_2^g) + \log \text{NCE}(\mathbf{h}_2^g, \mathbf{h}_1^g) \right)$$

$$\text{NCE}(\mathbf{h}_1^g, \mathbf{h}_2^g) = \frac{\exp(\mathbf{h}_1^g \cdot (\mathbf{h}_2^g)^+ / \tau)}{\sum_{i=1}^{B} \exp(\mathbf{h}_1^g \cdot \mathbf{h}_2^g / \tau)} \quad (5)$$

For context video anomaly detection, the motion-context and appearance-context pairs are selected to build the global alignment resulting in $L_{global} = L_{app,cxt}^{global} + L_{mot,cxt}^{global}$.

3.2.2 Local Alignment

At the local level, the reference negative samples can be separated into two categories: batch-wise samples and patch-wise samples. These two categories address different aspects of information alignment by applying different negative sampling strategies.

Batch-wise local alignment treats all the local patches at the same location across a batch as negative samples. It tries to align the information from two selected branches $\mathbf{h}_1^l, \mathbf{h}_2^l$ at the local level, which is similar to the global alignment. After getting similarity logits with shape ($N \times B \times B$), the batch-wise local alignment loss $L_{1,2}^{batch}$ is calculated similarly as $L_{1,2}^{global}$ with average N batch alignment loss performed locally. This type of alignment is not suitable for cross-modality alignment (e.g., context-motion alignment) but can be applied between multiple views (e.g.,

motion-appearance alignment) since cross-modality information might not have a consistent correlation at the local level. In the later ablation study, we further validate that cross-modality local alignment causes a performance drop.

Patch-wise local alignment treats the rest of the local patches within a video clip as negative samples. It aims to articulate the spatial context, pushing the feature representations from different positions away from each other. The patch-wise local alignment loss $L_{1,2}^{patch}$ is calculated by averaging the B alignment results using the same symmetric binary cross entropy loss of logits with shape ($B \times N \times N$). For context video anomaly detection, the local level alignment loss L_{local} is:

$$L_{local} = L_{app,mot}^{batch} + L_{cxt,mot}^{patch} + L_{cxt,app}^{patch} + L_{app,mot}^{patch} \quad (6)$$

3.3. Training and Inference

Training. We pretrain the vector quantization layer before training the full Trinity model using the simple MSE loss. Given a video with provided context, the training loss combines the reconstruction loss, the global alignment loss, and the local alignment loss with weighing parameters:

$$L = L_{recon} + \beta_1 \cdot L_{local} + \beta_2 \cdot L_{global} \quad (7)$$

Inference. During inference, context anomalies can be detected by evaluating the global context-motion similarity and context-appearance similarity. The global context similarity $\mathbf{S}_g$ is the equally weighted average of the alignment logits after applying the sigmoid function σ:

$$\mathbf{S}_g = \frac{1}{2} \left(\sigma \left(\frac{\mathbf{h}_{cxt} \cdot \mathbf{h}_{mot}^T}{\tau} \right) + \sigma \left(\frac{\mathbf{h}_{cxt} \cdot \mathbf{h}_{app}^T}{\tau} \right) \right) \quad (8)$$

We combine this with the widely adopted next frame prediction quality from the U-net using the min-max normalized PSNR $\mathbf{S}_r$ with a weighing parameter α for detecting non-context anomalies. We compute the final normalcy score as $\mathbf{S} = \alpha \cdot \mathbf{S}_r + (1 - \alpha) \cdot \mathbf{S}_g$ (and in practice graph the anomaly score $1 - \mathbf{S}$).

4. Experiments and Analysis

4.1. The WF Dataset

Benchmark VAD datasets mainly target anomalies that solely depend on individual objects and don't consider temporal context. We therefore collected a new dataset, denoted as the WF dataset, from a public webcam looking at a baseball stadium to address the video-context correspondence problem. We collected 2000 1920×1080 resolution videos of approximately 1 minute length at 15 frames per second across a period of three months. The collection process has a 30-minute gap between each run, resulting in a large set of videos of the same location under various lighting, weather,

and crowding conditions. We also collected corresponding metadata including the time of the day, day of week, and scheduled game time. To address the potentially populated region of interest, we cropped a 400×1200 region out of the original frames for both training and testing. More details can be found in the Supplementary Materials Sec. 8.

Anomalies in the wild. The anomalous behaviors in the WF dataset are a mixture of context-dependent anomalies (i.e., behaviors whose anomalousness depends on time of day, game schedule, etc.) and context-free anomalies (i.e., behaviors that would be anomalous regardless of when they occurred). Context-dependent anomalies we found included people running in the middle of the night, an ambulance driving against normal traffic flow, and unexpected group presence in front of the stadium. This resulted in 12 videos containing 64 anomalous events with bounding box annotations, which we call True Anomalies.

Pseudo-contextual anomalies. A well-known problem in real-world VAD is the lack of natural anomalous events in proportion to normal observations. To compensate, we introduced what we term pseudo-contextual anomalies, which are simply existing videos in the dataset given an incorrect context label. For example, a video recorded just before a game starts should contain frames of people entering quickly into the stadium. If we modify the context indicating the game has already started two hours ago, the model should flag the whole video as anomalous. Following this approach, we selected 20 normal videos covering various contexts. During testing, each video is evaluated twice: once with the original context and again with an incorrect context. The ideal result is that all the frames in the first case are normal and all the frames in the second case are anomalous. In contrast, as Fig. 4 illustrates, a context-unaware algorithm will have a 50.0 AUC score in this evaluation (since it has the same ratio of positive detections to false alarms for each double trial).
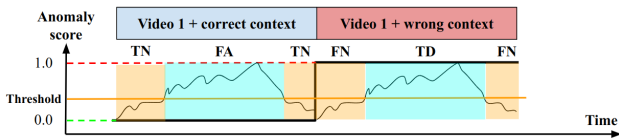

Figure 4. Example of pseudo context anomaly evaluation. The algorithm is expected to identify whether the given context matches the input video. TN, TD, FA, FN are true negative, true detection, false alarm and false negative respectively.

4.2. Benchmark Datasets and Evaluation Metrics

In addition to the new dataset we introduced above, we also evaluate the model performance in the widely used benchmark datasets UCSD Ped2 [35], Avenue [19] and ShanghaiTech [15]. All three benchmarks treat walking pedestrians as normal events, while other objects like bikers and abnormal motions like dropping bags or running

are considered as anomalies. ShanghaiTech contains 13 different scenes, while Ped2 and Avenue are single scene datasets.

We created a context-augmented dataset called the Biker Day dataset based on the Ped2 dataset to verify the effectiveness of our context-dependent approach. The dataset is created by adding a context called Biker Day that considers both pedestrians and bikers as normal objects. To create this dataset, we included test videos [Test001, Test002, Test003] that contain bikers from the Ped2 test set into the training data and removed these videos during evaluation. If we turn off the Biker Day context, the bikers will again become anomalous as in the original Ped2 dataset. Since all selected videos also contain frames in which bikers are not present, those frames should be considered normal on a Biker Day.

We use the widely adopted area under the curve (AUC) of the receiver operating characteristics (ROC) as the frame-level metric to report the model performance in the following experiments.

4.3. Implementation Details

For Ped2, Avenue, and ShanghaiTech, all frames are resized to 256×256 and the frame sampling gap is 1. For the WF dataset, all frames are resized to 128×384 and the sampling frame gap is 4. The patch size for all videos is set to 16. The RGB values of all frames are normalized to [-1,1]. The length of the input frame tube is 4. We use 1 and 2 TSCAB blocks in the Biker Day and WF datasets respectively. The size of the vector quantization codebook is 512. The initial temperature τ is set to 0.07. β_1 and β_2 are set to 1.0. The learning rate used for the pretraining procedure of the vector quantization layer is 1e-3. The model is optimized by Adam [11] and the learning rate is 2e-4 with the cosine annealing [18] strategy. The total training epochs for Ped2, Avenue, ShanghaiTech, and WF are 60, 60, 10, and 20 respectively. The weighting parameter α for WF dataset and benchmark datasets is 0.3 and 0.7 respectively. We adopt a simple median filter to smooth the normalcy result with a kernel size of 17 following [17]. We provide more details in Supplementary materials Sec. 8.

4.4. Detecting Context Anomalies

WF Dataset True Anomalies. We applied several widely-used VAD algorithms (previously evaluated on common benchmarks) to our WF dataset to compare to our proposed method. We use officially released codebases and follow all the hyperparameter settings used in the original implementations except the resolution. The evaluation performance is shown in Table 1. These referenced algorithms' overall performance on the real anomalies shows relatively similar ranking as in benchmark datasets. Since the WF real anomalies contain a significant portion of context-free

Method	Real Anomaly	Pseudo Anomaly
ConvAE [9]	69.4	50.0
MemAE [8]	71.3	50.0
MNAD [23]	72.3	50.0
Ours	**75.5**	**92.0**

Table 1. Contextual anomaly detection results on the WF dataset.

anomalies, the referenced algorithms can perform well on detecting them. However, they perform poorly at detecting out-of-place presence/absence/motion that only seem unusual in context. With the help of global context alignment, Trinity is able to achieve the best performance with a noticeable performance margin. We visualize two testing samples containing real context anomalies with the prediction score using Trinity shown in Fig. 4. The global alignment is quite robust in detecting contextual anomalies.

WF Dataset Pseudo Anomalies. Out-of-context detection is the main purpose of the pseudo anomaly test. However, previous VAD methods are not equipped with the ability to detect out-of-context anomalies, finding each double-length video to be completely normal. On the other hand, Trinity performs extremely well in detecting out-of-context anomalies such as presence, absence, and counter-flow direction, with an AUC of 92.0. We illustrate several pseudo anomaly detection results in Fig. 5. These types of anomalies are impossible to detect using traditional VAD methods. Here, the context-appearance and context-motion branches complement each other very well. If only use a single branch, the pseudo anomaly performance is 76.5 and 87.9 AUC for the context-appearance and context-motion branch respectively. This is because different context anomalies depend on different types of information. For example, the only cue for detecting an anomaly in game schedule (lower

left of Fig. 5) are some large sandwich boards set out on the pavement "too early", which cannot be detected by motion information. Similarly, for the anti-flow anomaly introduced at the lower right of Fig. 5, it is not easy to distinguish whether the crowd is entering or exiting the stadium from only the static image.
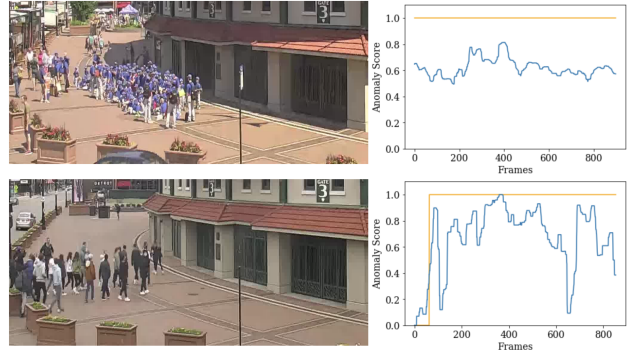

Figure 4. Real context anomalies in the WF dataset. These two sample videos contain unexpected crowds at the front of the stadium (presence anomalies). The blue line is the prediction result and the orange line is the ground truth.

Biker Day experiment. Without Trinity global alignment, solely using the U-net from Section 3.1.3 on the augmented dataset resulted in 92.4 AUC, compared to 94.3 in the original Ped2 dataset, which is a significant drop due to the mixture dataset containing bikers as normal objects. By learning the Biker Day context, the global alignment can differentiate the context and boost the performance from 97.9 in Ped2 to **98.5** AUC. To further validate that the Trinity model really learns the Biker Day context, we visualize the attention map for the global head of the context-

outflow night game ➡ game canceled (presence anomaly) **during game ➡ adjust 2 hours earlier (absence anomaly)**

day game morning ➡ change schedule (background anomaly) **inflow day game ➡ adjust 3 hours later (anti flow anomaly)**
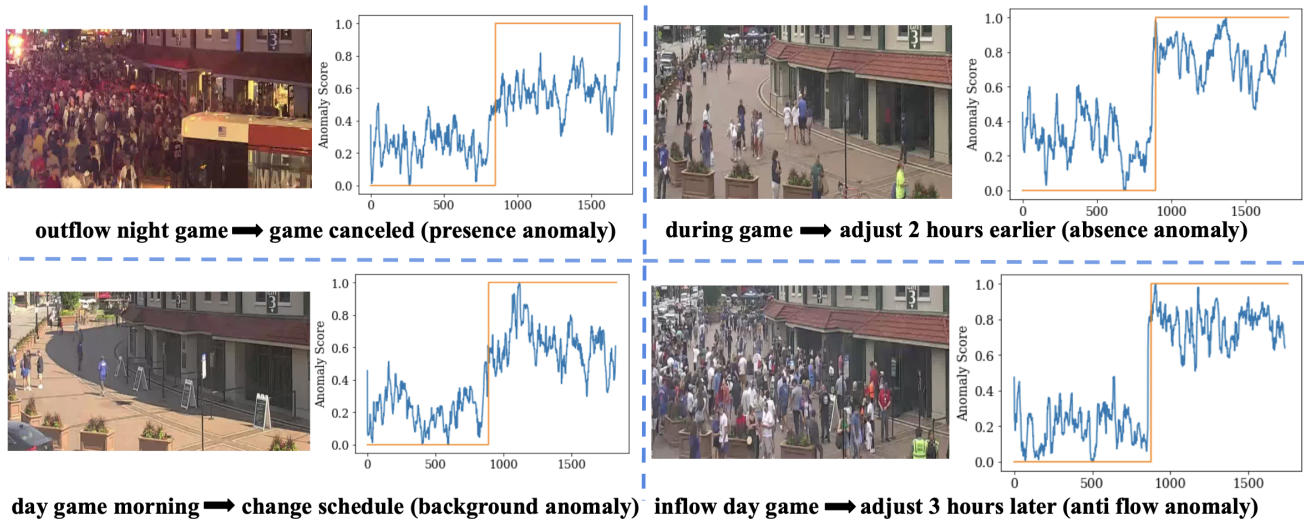
Figure 5. Selected results of pseudo anomalies in the WF dataset. **Top:** Selected pseudo anomaly detection results with exemplar frames. The orange line indicates ground truth and the blue line is corresponding prediction. The caption indicates the change from the true context (left of the arrow) to the pseudo context (right of the arrow). The parentheses indicate the type of context anomaly.

appearance pair and the context-controlled anomaly detection result of 'test video010' from the Ped2 dataset in Fig. 6. We see that Trinity highlights the biker region with high attention to differentiate context. The 'test video010' only contains a biker as an anomaly across the entire video. By switching the input context, Trinity assigns a high anomaly score under the Non-Biker Day context and a low anomaly score under the Biker Day scenario.
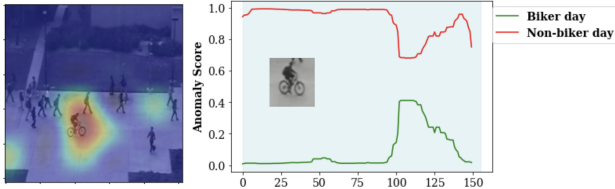

Figure 6. **Left:** Visualization of global context-appearance head attention on the Biker Day dataset. Biker regions have been assigned highest attentions across all local patches. The attention map has been enlarged for image overlay. **Right:** The anomaly score under Biker Day and Non-Biker Day context on 'Test Video010' from Ped2. The biker region has been highlighted using light blue.

Ablation studies. We perform ablation studies on the impact of different combinations of alignment schemes and the importance of vector quantization, summarized in Table 2. The vector quantization and patch-wise local alignment improve the context detection performance significantly while adding the batch-wise local alignment causes the performance drop. This indicates that the context information of the video does not have strong mutual information with motion at the local level, which suggests that the cross-modality alignment should not focus on local precise alignment but on finding a good global representation.

GA	VQ	LB	LP	AUC_{mot}	AUC
✓				83.8	89.3
✓	✓			86.4	91.2
✓	✓	✓		82.9	85.8
✓	✓	✓	✓	82.7	86.3
✓	✓		✓	**87.9**	**92.0**

Table 2. Ablation studies in the context-motion branch on the WF Pseudo Anomaly dataset. **GA** is global alignment; **VQ** is vector quantization; **LB** is local batch-wise alignment; **LP** is local patch-wise alignment; AUC_{mot} is reported only using a single branch; **AUC** is the final pseudo anomaly score.

4.5. Detecting context-free anomalies

To apply Trinity to VAD problems in which there is no context information/dependence, we modify the loss functions in Section 3.2 by removing any term involving context (cxt). Since there is no context information, the global term in Section 3.2.1 is replaced by $L_{global} = L_{app,mot}^{global}$ and computed similarly.

For inference, we use patch-wise local alignment between motion and appearance as the indicator for determining anomalies. After computing the patch-wise local alignment logits $\mathbf{A}^{patch}$ with the softmax function applied, a normal frame is expected to be similar to the identity matrix $\mathbf{I}$. Hence we use the Frobenius norm of the residual $\mathbf{S}_l = \|softmax(\mathbf{A}^{patch}) - \mathbf{I}\|_F$ as the anomaly indicator. We combine the next frame prediction quality from the U-net using PSNR and patch-wise local alignment as the final indicator of normalcy $\mathbf{S} = \alpha \cdot \mathbf{S}_r + (1 - \alpha) \cdot \mathbf{S}_l$.

Following the strategy described above, we evaluate Trinity without the context part on three traditional VAD benchmarks, and show that we can achieve comparable performance to state-of-the-art methods that are tailored to these benchmarks, as shown in Table 3. Again, we emphasize that Trinity was not designed for short, context-free datasets in the first place. We provide these results to indicate that Trinity can also be applied successfully to VAD in isolated video clips.

Methods	Ped2 [35]	Avenue [19]	SHT [15]
MemAE [8]	94.1	83.3	71.2
MNAD [23]	97.0	88.5	72.8
USTN-DSC [37]	98.1	89.9	73.8
HF^2VAD [20]	99.3	91.1	76.2
DMAD [16]	**99.7**	92.8	78.8
HSC [31]	98.1	**93.7**	**83.4**
Ours	97.9	88.5	74.1

Table 3. Video anomaly detection results on benchmark datasets Ped2 [35], Avenue [19], and ShanghaiTech (SHT) [15].

5. Conclusions and Future Work

We proposed a context-aware video anomaly detection algorithm, introducing a new consideration for real-world VAD. We carefully defined the task of detecting context-related anomalies, collected a novel dataset to enable our investigation, and proposed a contrastive learning framework to successfully detect context anomalies that cannot be found by traditional VAD methods. In future work, we plan to investigate a better negative sampling strategy for global alignment to avoid using negative samples that may have similar contexts.

6. Acknowledgment

References

[1] Austin Abrams, Jim Tucek, Joshua Little, Nathan Jacobs, and Robert Pless. LOST: Longterm Observation of Scenes (with Tracks). In *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, pages 297–304, 2012. 1, 3

[2] Qianyue Bao, Fang Liu, Yang Liu, Licheng Jiao, Xu Liu, and Lingling Li. Hierarchical Scene Normality-Binding Modeling for Anomaly Detection in Surveillance Videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6103–6112. ACM, 2022. 2

[3] MyeongAh Cho, Taeoh Kim, Woo Jin Kim, Suhwan Cho, and Sangyoun Lee. Unsupervised video anomaly detection via normalizing flows with implicit latent features. *Pattern Recognition*, 129:108703, 2022. 2

[4] Keval Doshi and Yasin Yilmaz. Rethinking Video Anomaly Detection - A Continual Learning Approach. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3036–3045, Waikoloa, HI, USA, 2022. 1, 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[6] Dario Dotti, Mirela Popa, and Stylianos Asteriadis. Unsupervised Discovery of Normal and Abnormal Activity Patterns in Indoor and Outdoor Environments:. In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 210–217, 2017. 3

[7] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly Detection in Video via Self-Supervised and Multi-Task Learning. In *CVPR*, pages 12737–12747, 2021. 1, 2

[8] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton Van Den Hengel. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *ICCV*, pages 1705–1714, 2019. 2, 7, 8

[9] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning Temporal Regularity in Video Sequences. In *CVPR*, pages 733–742, 2016. 2, 7

[10] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *CVPR*, pages 7834–7843, 2019. 1, 2

[11] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 6

[12] Sam Leroux, Bo Li, and Pieter Simoens. Multi-branch Neural Networks for Video Anomaly Detection in Adverse Lighting and Weather Conditions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3027–3035. IEEE, 2022. 1

[13] Zhong Li and Matthijs Van Leeuwen. Explainable contextual anomaly detection using quantile regression forests. *Data Mining and Knowledge Discovery*, 37(6):2517–2563, 2023. 3

[14] Jiongqian Liang and Srinivasan Parthasarathy. Robust contextual outlier detection: Where context meets sparsity. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 2167–2172, 2016. 3

[15] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, pages 6536–6545, 2018. 2, 6, 8

[16] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *CVPR*, pages 12147–12156, 2023. 2, 8

[17] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *ICCV*, pages 13588–13597, 2021. 1, 2, 6

[18] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[19] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal Event Detection at 150 FPS in MATLAB. In *ICCV*, pages 2720–2727, 2013. 6, 8

[20] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, pages 15425–15434, 2021. 2, 8

[21] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *CVPR*, pages 1273–1283, 2019. 2

[22] Ivan Nikolov, Mark Philip Philipsen, Jinsong Liu, Jacob Velling Dueholm, Anders Skaarup Johansen, Kamal Nasrollahi, and Thomas B. Moeslund. Seasons in drift: A long term thermal imaging dataset for studying concept drift. In *NeurIPS*, 2021. 3

[23] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning Memory-Guided Normality for Anomaly Detection. In *CVPR*, pages 14360–14369, 2020. 2, 4, 7, 8

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 5

[25] Bharathkumar Ramachandra and Michael J. Jones. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2558–2567, 2020. 1

[26] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai. A Survey of Single-Scene Video Anomaly Detection. *IEEE TPAMI*, 44(5):2293–2312, 2020. 1

[27] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive con-

volutional attentive block for anomaly detection. In *CVPR*, pages 13576–13586, 2022. 2

[28] Ashish Singh, Michael J Jones, and Erik G Learned-Miller. Eval: Explainable video anomaly localization. In *CVPR*, pages 18717–18726, 2023. 4

[29] Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. Conditional Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):631–645, 2007. 3

[30] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-Aware Context Reasoning for Unsupervised Abnormal Event Detection in Videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 184–192, 2020. 2

[31] Shengyang Sun and Xiaojin Gong. Hierarchical Semantic Contrast for Scene-aware Video Anomaly Detection. In *CVPR*, pages 22846–22856, 2023. 2, 8

[32] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. In *ECCV*, pages 776–794, 2020. 3

[33] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural Discrete Representation Learning. In *NeurIPS*, 2017. 4

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 4

[35] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly Detection and Localization in Crowded Scenes. *IEEE TPAMI*, 36(1):18–32, 2014. 6, 8

[36] Dan Xie, Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Learning and inferring "dark matter" and predicting human intents and trajectories in videos. *IEEE TPAMI*, 40 (7):1639–1652, 2017. 3

[37] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, pages 14592–14601, 2023. 8

[38] C. Zach, T. Pock, and H. Bischof. A Duality Based Approach for Realtime TV-L 1 Optical Flow. In *Pattern Recognition*, pages 214–223. Berlin, Heidelberg, 2007. 4

[39] Bin Zhao, Li Fei-Fei, and Eric P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, pages 3313–3320, 2011. 2