Spatio-temporal predictive tasks for abnormal event detection in videos

Yassine Naji, Aleksandr Setkov, Angelique Loesch, Michèle Gouiffès, Romaric Audigier

Conference: 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2022

Spatio-temporal predictive tasks for abnormal event detection in videos

Yassine Naji[1,2], Aleksandr Setkov[1], Angélique Loesch[1], Michèle Gouiffès[2], Romaric Audigier[1]
[1]Université Paris-Saclay, CEA, List, 91120, Palaiseau, France
firstname.lastname@cea.fr
[2] Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
firstname.lastname@universite-paris-saclay.fr

Abstract

Abnormal event detection in videos is a challenging problem, partly due to the multiplicity of abnormal patterns and the lack of their corresponding annotations. In this paper, we propose new constrained pretext tasks to learn object level normality patterns. Our approach consists in learning a mapping between down-scaled visual queries and their corresponding normal appearance and motion characteristics at the original resolution. The proposed tasks are more challenging than reconstruction and future frame prediction tasks which are widely used in the literature, since our model learns to jointly predict spatial and temporal features rather than reconstructing them. We believe that more constrained pretext tasks induce a better learning of normality patterns. Experiments on several benchmark datasets demonstrate the effectiveness of our approach to localize and track anomalies as it outperforms or reaches the current state-of-the-art on spatio-temporal evaluation metrics.

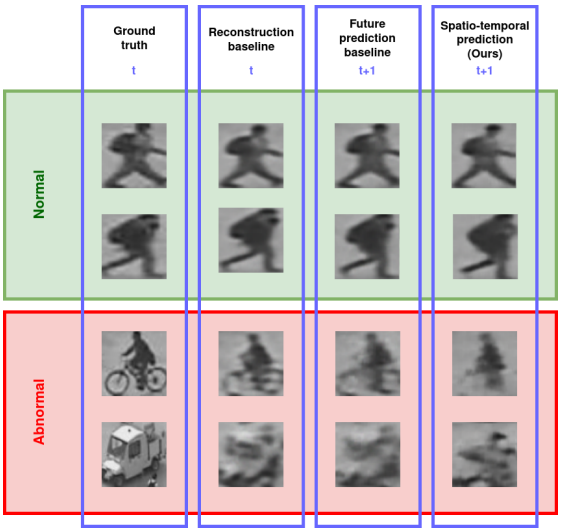

Figure 1. Normal and abnormal test samples with their corresponding predictions using 3 different anomaly detection pretext tasks: *spatio-temporal prediction* (ours), *future frame prediction* and spatial *reconstruction*. The use of spatio-temporal prediction tasks leads to a good recovery of normal objects and a worse recovery of abnormal samples compared to other pretext tasks.

1. Introduction

Video anomaly detection (VAD) is an open research problem which consists in detecting the occurrence of unexpected events. This problem has raised a lot of attention due to its critical applications such as video surveillance [30, 17, 19, 16, 23, 22] and autonomous driving [5]. For several reasons, this is a challenging problem. First, an event is considered as abnormal according to a set of normal events which define a context, therefore, the same event can be considered normal or abnormal in two different contexts. Second, the wide range of possible abnormal events and their rarity make it infeasible to collect enough annotations to train fully supervised models. Consequently, VAD is often regarded as a "one class" problem, when only normal data is used for training. Several approaches have been proposed in the literature to address this challenge. They can be grouped into three main paradigms: **Distance based methods** [11, 23, 24, 12, 26, 29] which are trained by learning a distance between samples. At inference time, the anomaly score is computed as the distance of a sample to normal data. **Probabilistic methods** [25] learn a distribution over normal data. Samples with low likelihood according to the estimated probability density are detected as anomalies. **Self-supervised methods** [8, 21, 9, 20, 33, 10, 2, 16, 18, 14, 7, 27, 31, 15, 3, 32] use some pretext tasks to learn normal appearance and motion features from training data. At inference time, anomalies are often detected by measuring the incapacity of the model to perform a pretext task on a test sample. The choice of

978-1-6654-6382-9/22/$31.00 ©2022 IEEE

the pretext task conditions the type of anomalies the model is able to detect. Therefore, these tasks must be chosen according to the anomalies of interest.

A widely used self-supervised approach for anomaly detection consists in *reconstructing* normal samples from a low dimensional representations [8, 21, 9, 20, 33, 10, 2, 16, 18]. The underlying assumption is that a model, trained on normal data only, can not generalize well to abnormal samples. Other approaches [14, 7, 27, 31, 15, 3, 32] use *prediction* based pretext tasks to learn normal features from training data. Those approaches achieve good performances, however, we believe that it is important to further constrain those tasks to better learn normality patterns and thus better discriminate anomalies (figure 1). For instance, a common approach consists in training a neural network to predict the next frame from a series of past frames [14]. Given a video sequence of a person walking in a wrong direction, the model could infer the location of the person in the next frame by referring to the past. Even if this event is abnormal, it is potentially predictable if the model learns the concept of walking from normal data. As an alternative, we propose to discard past frames and to constrain the model to infer both the next and past frames *using only the current frame* by generalizing from training data and not by referring to past frames. The task is constrained further by *down-scaling* the input in order to recover as least as possible abnormal features. Thus, the network is forced to predict spatial features in addition to temporal ones. More precisely, we train our model to predict, in the original resolution, the appearance and optical flow of a given object while it is conditioned only on a down-scaled version of it. At test phase, abnormal samples are detected as those that produce prediction errors which are significantly higher than the normal training samples. We measure this discrepancy using a *Z-score* based normalization. In summary our contributions are:

• Introducing new constrained pretext tasks for anomaly detection based on spatio-temporal normality prediction. Specifically, by predicting jointly upscaled version of object appearance and motion.

• Showing empirically that our method outperforms or reaches the current state-of-the-art on spatio-temporal evaluation metrics.

Since our method is designed for object level anomalies, we tested it on the three common benchmark datasets: UCSDped2 [19], ShanghaiTech [17] and Avenue [16] for which the anomalies are related to object appearance and behaviour. The experiments conducted on those datasets show the relevance of the proposed approach.

2. Related works

Anomaly detection pretext tasks: Recently, deep learning approaches [8, 7, 11, 15, 14, 9, 21, 27, 23, 24, 3, 34, 32, 20, 12] have shown their effectiveness for detecting abnormal events in videos. Those approaches learn normality using some pretext tasks. Among those tasks, we can distinguish two major families: *reconstruction* and *prediction* tasks.

Many previous methods used *reconstruction* to learn normality [8, 9, 10, 21, 20, 33, 2, 16]. In a pioneering work [10], the authors proposed to train an auto-encoder to reconstruct normal handcrafted appearance and motion features. The reconstruction task is constrained further by adding a memory module which learns normality prototypes [9]. An extension of this approach was proposed by [21] which learns spatio-temporal patch prototypes.

Other methods such as [14, 3, 31, 27, 20, 31, 32, 35] trained GANs to perform *prediction* tasks. A common way consists in training a generator to predict future frame or equivalently the optical flow given few past frames [14, 3, 32]. Abnormal events are detected when the prediction error is high.

Similarly to adversarial approaches, we learn to predict normal appearance and motion patterns. Unlike those approaches, we do not use a discriminator during training since we observed that our method is already able to produce consistent predictions using our objective function.

Approaches [14, 3, 32] performed prediction at the temporal level. We propose to further constrain the normality learning task by imposing prediction at both the spatial and temporal levels. Our model learns up-scaling in conjunction with future and past prediction. Moreover, the normality is learned at the object level to ensure the invariance to scene changes, unlike previous works [14, 3, 32] which model normality at the frame level and therefore they are sensitive to background changes.

Object centric anomaly detection: Some recent works characterized abnormality at the object level [7, 8, 11, 15, 4, 34], by making use of a pretrained object detector as a preliminary step before anomaly scoring. Ionescu *et al.* [11] proposed to learn appearance and motion features using an auto-encoder and performed a clustering followed by a one versus rest classification. Recently, Georgescu *et al.* [8] introduced a new approach that uses out-of-domain observations to train a model to explicitly mis-reconstruct those pseudo-anomalies via adversarial training. This approach achieves very good performance on current benchmark datasets, nevertheless, it needs additional pseudo-abnormal data for training. We propose to train our model using normal data only without explicit assumptions on anomalies, which makes our method generalizable to multiple anomaly types.

Liu *et al.* [15] extended [14] by applying the future frame prediction task at the object level with a constraint on the reconstruction of optical flows via a multi-level memory mod-

ule. The addition of this module restricts the model ability to recover anomalies, however, it requires choosing the number of memory items which models the heterogeneity of normality. We propose not to discretize the normality space. Instead, we constrain anomaly prediction by discarding past frames and inputting only a down-scaling of the current frame. In addition, we perform optical flow prediction instead of reconstruction so that the model cannot reproduce abnormal motion features.

By observing that single objective optimization is generally suboptimal for anomaly detection, Georgescu *et al.* proposed in [7] a multitask architecture to learn normality through multiple pretext tasks such as middle frame prediction and arrow of time prediction. We propose to constrain further the pretext tasks to better learn normality patterns 1. Specifically, our model learns to restore normal appearance and motion from a down-scaled image of the object. To our knowledge, we are the first to propose *super-resolution* as a pretext task for video anomaly detection.

3. Method

3.1. Motivation

The main motivation behind our approach is the following: *self-supervised pretext tasks for which the target cannot be predicted directly from the input but only by generalizing from the training data manifold are more suited for anomaly detection*.

We believe that if the model do not access the necessary information to predict the target from the input, it is constrained to generalize from training data to perform the task. Since we assume that the model is trained on normal data only, it will induce a larger prediction error for abnormal samples. Motivated by this intuition, we propose to further constrain pretext tasks by training the model to perform future/past frame prediction using only a down-scaling of the current frame, which is more challenging than future frame prediction proposed in [14, 15, 3, 32] where past frames are given as input to perform the task. In addition to the common temporal prediction [14, 15, 3, 32], our model performs spatial prediction by up-scaling the input. Specifically, we train our method via four pretext tasks which can be grouped as follows:

Appearance tasks: consist in predicting the original resolution of an object image in both past and next frames.

Motion tasks: consist in predicting the forward optical flow magnitudes at both the past and the current frames.

3.2. Approach

3.2.1 Inputs and pre-processing

In order to localize anomalies at the object level, we first perform object detection similarly to other object-centric approaches [7, 8, 11]. We crop the detected objects using bounding boxes at frames F_{t-1}, F_t and F_{t+1}. We do not perform tracking since we want to capture the absolute object displacement. We suppose that the position change between times $t-1, t+1$ is small enough so that the majority of the object is still contained in the bounding box. In order to constrain further the prediction tasks, we propose to down-scale objects at frame F_t, which are fed as input to our architecture.

The forward optical flow is computed between the frames F_t and F_{t+1} and the magnitude map is extracted. We do not keep the orientation map since it is not precise enough for small magnitude displacements. The magnitude map is cropped using the bounding boxes and resized to the same objects resolution.

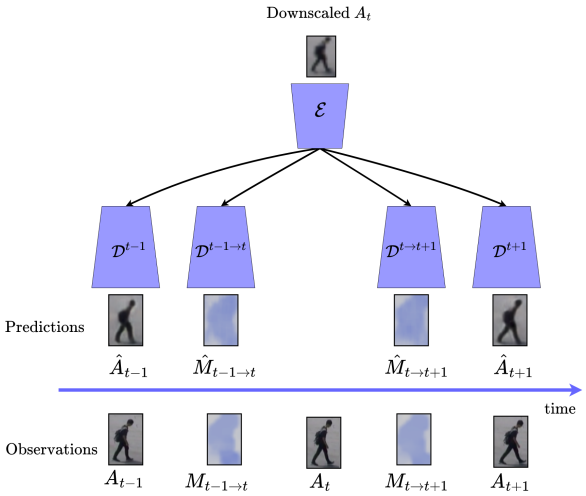

Figure 2. Overview of our architecture: Considering a frame t, we perform object detection and forward optical flow extraction: $t-1 \to t$ and $t \to t+1$. The object appearances at frames $t-1$, t and $t+1$: A_{t-1}, A_t, A_{t+1} with its corresponding optical flow magnitudes: $M_{t-1\to t}, M_{t\to t+1}$ are cropped. The object appearance at time step t: A_t is down-scaled and is then passed as an input to the encoder $\mathcal{E}$ that produces a latent representation which is given to the four decoding branches $\mathcal{D}$. Appearances: $\hat{A}_{t-1}, \hat{A}_{t+1}$ and optical flow magnitudes $\hat{M}_{t-1\to t}, \hat{M}_{t\to t+1}$ are predicted by those branches at the original resolution

3.2.2 Architecture

Since our model performs multiple pretext tasks, we choose an architecture with multiple predictive branches. More precisely, it is composed of a common encoder $\mathcal{E}$ which learns high level appearance and motion features and four decoding branches. $\mathcal{D}^{t-1}$, $\mathcal{D}^{t-1\to t}$ are dedicated to predict past context and the two other branches $\mathcal{D}^{t\to t+1}, \mathcal{D}^{t+1}$ are dedicated to future context. In particular, the model learns to recover from low-resolution input, normal appear-

ance and motion jointly via predicting the original resolution of the object in both previous and next frames using the branches $\mathcal{D}^{t-1}$ and $\mathcal{D}^{t+1}$. In order to emphasize learning motion patterns, we train the model to predict the magnitude of optical flow via the branches $\mathcal{D}^{t-1\to t}$ and $\mathcal{D}^{t\to t+1}$. The architecture takes as input a down-scaled object image and outputs past and future object appearances: $\hat{A}_{t-1}$ and $\hat{A}_{t+1}$ with the corresponding optical flow magnitudes $\hat{M}_{t-1\to t}$ and $\hat{M}_{t\to t+1}$. The figure 3.1 illustrates our network.

3.2.3 Loss functions

In order to learn normal appearance and motion patterns, we train our framework to predict the normal temporal context using a combination of two prediction losses:

$$\mathcal{L}_{context} = \mathcal{L}_{past} + \mathcal{L}_{future} \tag{1}$$

with:

$$\mathcal{L}_{past} = \mathcal{L}(\hat{A}_{t-1}, A_{t-1}) + \mathcal{L}(\hat{M}_{t-1\to t}, M_{t-1\to t}) \tag{2}$$

$$\mathcal{L}_{future} = \mathcal{L}(\hat{A}_{t+1}, A_{t+1}) + \mathcal{L}(\hat{M}_{t\to t+1}, M_{t\to t+1}) \tag{3}$$

where $\mathcal{L}$ is the logistic loss computed over all pixel locations i, j:

$$\mathcal{L}(\hat{X}, X) = \sum_{i,j} -X_{i,j}\log(\hat{X}_{i,j}) - (1-X_{i,j})\log(1-\hat{X}_{i,j}) \tag{4}$$

The past context loss $\mathcal{L}_{past}$ is composed of an appearance term which is a logistic loss between the prediction of the branch $\mathcal{D}^{t-1}$ prediction $\hat{A}_{t-1}$ and A_{t-1}. We use the same loss between the predicted optical flow magnitude $\hat{M}_{t-1\to t}$ at branch $\mathcal{D}^{t-1\to t}$ and the observation $M_{t-1\to t}$.

The future context loss $\mathcal{L}_{future}$ is defined in a similar way, between times t and $t+1$.

3.2.4 Inference: anomaly scoring

At inference time, in order to compute the object level anomaly scores given a test video sequence, we perform object detection with the same parameters as for training. Similarly to training data, we also extract the frame level optical flow magnitudes and we crop them at the object level using the predicted boxes. For each test sample O_t, we compute the ℓ_1 prediction errors for appearance and motion:

$$L(O_t) = \begin{pmatrix} L^{t-1}(O_t) \\ L^{t-1\to t}(O_t) \\ L^{t\to t+1}(O_t) \\ L^{t+1}(O_t) \end{pmatrix} = \begin{pmatrix} \|\hat{A}_{t-1} - A_{t-1}\|_1 \\ \|\hat{M}_{t-1\to t} - M_{t-1\to t}\|_1 \\ \|\hat{M}_{t\to t+1} - M_{t\to t+1}\|_1 \\ \|\hat{A}_{t+1} - A_{t+1}\|_1 \end{pmatrix} \tag{5}$$

In order to compare this vector with respect to the distribution of normal samples, we propose to compute the Z-scores across all branches and average them. More precisely, we estimate the mean μ and the covariance matrix Σ of L vectors of normal training objects.

If we denote $n = 4$, the dimension of the vector L, the aggregated score is given by:

$$S(O_t) = \frac{1}{n}\text{diag}(\Sigma)^{-\frac{1}{2}}(L(O_t) - \mu) \tag{6}$$

Let Δ_t be the set of N_t detected objects at frame F_t:

$$\Delta_t = \{O_t^{(i)}, \quad 1 \le i \le N_t\}$$

Since we consider a frame F_t abnormal if it contains at least one abnormal object, the frame level anomaly score is given by:

$$S(F_t) = \max_{O\in\Delta_t} S(O) \tag{7}$$

4. Experiments and results

4.1. Datasets

Several benchmarks exist for evaluating anomaly detection performances [19, 16, 17, 23, 30]. In order to compare our method with previous works, we performed experiments on the most common datasets.

UCSDped2 [19] contains 16 training and 12 testing videos of resolution 240x360. Anomalous events include riding a bike and driving a vehicle on a sidewalk. The original dataset is annotated at both the frame and the pixel level. Ramachandra *et al.* [23] provided region-level and track-level annotations.

ShanghaiTech [17] includes 330 training and 107 testing videos with the resolution of 480x856 with 13 different scenes. Each scene has a different background. Abnormal events include jumping, running, or stalking on a sidewalk. As UCSDped2, the original dataset is annotated at frame and pixel levels. The region-level and track-level annotations are provided by Georgescu et al [8].

CUHK Avenue [16] consists of 16 training and 21 test videos with the resolution of 360x640 with abnormal events such as running or walking towards the camera. We evaluated our method on all metrics using the ground truth region-level and track-level annotations provided by Ramachandra *et al.* [23] for this dataset. Unlike other datasets, some annotated anomalies are related to the position with respect to the camera. Therefore, we take into account object sizes for measuring abnormality as explained in section 4.4.

4.2. Evaluation

In order to evaluate the frame level performance of our method, we adopt the area under the receiver characteristic

	Method	UCSDped2				ShanghaiTech				Avenue			
		AUC		RBDC	TBDC	AUC		RBDC	TBDC	AUC		RBDC	TBDC
		Micro	Macro			Micro	Macro			Micro	Macro		
Frame / Patch level	Liu *et al.* [14] (pred.)	95.4	-	-	-	72.8	-	-	-	85.1	-	-	-
	Nguyen *et al.* [20] (pred. & rec.)	96.2	-	-	-	-	-	-	-	86.9	-	-	-
	Gong *et al.* [9] (rec.)	94.1	-	-	-	71.2	-	-	-	83.3	-	-	-
	Park *et al.* [21] (pred. & rec.)	-	97.0	-	-	-	72.0	-	-	-	88.5	-	-
	Ramachandra *et al.* [23] (dist.)	88.3		62.5	80.5	-	-	-	-	72.0		35.8	**80.9**
	Ramachandra *et al.* [24] (dist.)	94.0		74.0	80.3	-	-	-	-	87.2		41.2	78.6
Object level	Ionescu *et al.* [11] (rec.)	94.3	97.8	52.76	72.88	78.7	84.9	20.7	44.5	87.4	90.4	15.8	27.0
	Yu *et al.* [34] (pred.)	97.3	-	-	-	74.8	-	-	-	89.6	-	-	-
	Liu *et al.* [15](pred. & rec.)	99.3		-	-	76.2		-	-	91.1		-	-
	Georgescu *et al.* [8] (rec.)	98.7	99.7	69.2	93.2	**82.7**	**89.3**	41.3	78.8	92.3	90.4	65.1	66.9
	Georgescu *et al.* [7] (multi.)	97.5	99.8	72.8	91.2	82.4	**89.3**	42.8	83.9	91.5	**91.9**	57.0	58.3
	Ours (pred.)	**98.9**	**99.9**	**77.2**	**98.5**	77.1	86.2	**51.6**	**84.6**	**92.6**	90.1	**75.3**	73.4

Table 1. Comparison of our approach with the state-of-the-art methods (%) on the metrics: Micro AUC, Macro AUC, RBDC and TBDC. The AUC scores for the methods which do not explicitly mention the type (Micro or Macro) are put in the middle of the column. Best results are in bold, second best are underlined. Methods are grouped according to whether they detect anomalies at the frame/patch or object level. We also classify approaches according to the used paradigm (pred.: prediction based, rec.: reconstruction based, dist.: distance based, and multi.: multiple pretext tasks)

curve (AUC ROC) which is a popular metric in the VAD literature. As pointed out in [8], many previous works do not specify if the AUC is computed at the video level and averaged across all the dataset (Macro AUC) or at the dataset level by concatenating all video frames (Micro AUC). We report both measures for clarity.

The AUC quantifies the anomaly detection performance at the frame level only and not the ability of the model to localize and track anomalies. In fact, if the model predicts a false positive in an abnormal frame, it will be counted as a true positive. Previous works such as [27] reported the pixel level AUC as an attempt to measure the anomaly localization performances. This metric considers a detection as a true positive if at least 40% of the ground truth pixels are detected. As pointed out in [23], this procedure is problematic since it takes into account only false positive detections in normal frames and not in abnormal ones. As an alternative to pixel level AUC, we adopt the region-based detection criterion (RBDC) introduced by Ramachandra *et al.* [23] since it takes into account false positives detections in all frames. We report also the track-based detection criterion (TBDC) [23] which quantifies the models ability to identify anomalies spatio-temporally. Since, the previous metrics evaluate the ability to localize and track anomalies, we consider them as the main evaluation criteria for abnormal event detection.

4.3. Implementation details

For the pre-processing step, we choose YOLOv3 [28] object detector as in [7, 8, 11] for a fair comparison with them. We use a MMDetection [1] implementation of YOLOv3 pretrained on MS COCO dataset. We set the confidence threshold to 0.7 for Avenue and ShanghaiTech. Since objects have a low resolution in UCSDped2 we decreased the confidence threshold to 0.5 as in [8]. For this dataset, we notice that the implementation of YOLOv3 produced very small false positive boxes, we filtered them out using an area criterion (300 pixels). Optical flow between consecutive frames is extracted using an OpenCV implementation of Gunnar Farneback's algorithm [6], which offers a good trade-off between accuracy, speed and generalisation across multiple contexts. In order to normalize the optical flow magnitudes between 0 and 1, we divide them by 50 which is an upper-bound over the maximum displacement in the training videos. For all datasets, we choose a time step $\delta t = 1$ for past and future prediction pretext tasks. As in [7, 8, 11], we resize objects to 64x64.

Regarding our architecture the encoder $\mathcal{E}$ consists of five 3x3 convolution layers followed by 2x2 max pooling and ReLU activation. We double the number of convolution filters after each layer so that we ensure that the architecture is not constrained to perform dimensionality reduction and therefore our approach is different from the reconstruction-based methods. The four decoding branches are identical except for the last layer that matches the output number of channels. They are composed of four transposed convolutions with a kernel size of 4x4 and a stride of 2 followed by ReLU activations.

For all datasets, we trained the network for 200 epochs using Adam optimizer [13] with a learning rate of 10^{-3}, a batch size of 640. The frame level scores are smoothed using a Gaussian filter as in [7, 8]. RBDC and TBDC are computed using the code provided by Georgescu *et al.* [8]. It is worth mentioning that our anomaly detection approach

achieves state-of-the-art performances with few hyperparameters.

4.4. Results

Table 1 reports quantitative evaluations of our method on UCSDped2 [19], ShanghaiTech [17] and Avenue [16] benchmarks, in comparison with state-of-the-art methods [7, 8, 15, 34, 11, 24, 23, 9, 20, 14]. Methods are grouped according to the used paradigm and whether they perform anomaly detection at the object or frame/patch level. It is important to mention that no method consistently outperforms the others in all metrics and in all datasets.

Qualitative results are provided in figure 3. We can observe that appearances and behaviours are well predicted for normal samples while it is not the case for anomalies, which allows to distinguish them.

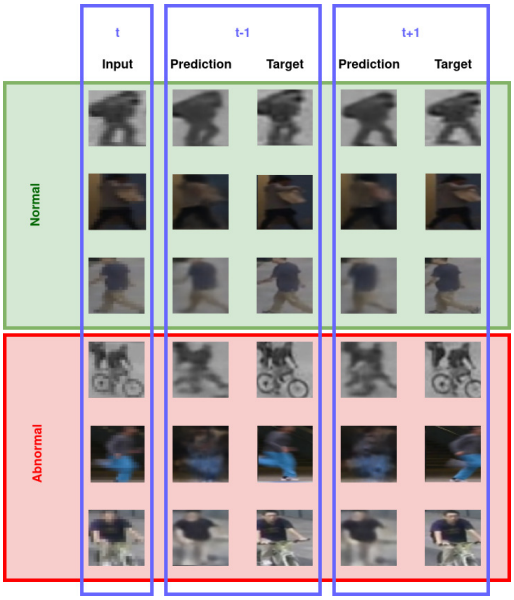

Figure 3. Normal and abnormal test samples with their corresponding predictions. Testing examples are taken from the datasets UCSDped2 [19], Avenue [16] and ShanghaiTech [17]. The first column indicates the down-scaled inputs at time t. The second column shows the past frame predictions and the corresponding ground truth at time $t-1$. The last column shows the future frame predictions with the ground truth $t+1$. Appearances and behaviours are well predicted for normal samples while it is not the case for anomalies.

Results on UCSDped2 Recent works reported AUC performances which are higher than 98% on this dataset, while, the metrics RBDC and TBDC show a better discrimination between approaches in terms of anomaly localization and tracking than the AUC. As shown in table 1, our approach is able to significantly outperform previous approaches on anomaly tracking by 3 p.p on the RBDC metric and 5 p.p on TBDC. Those improvements can be explained by the choice of pretext tasks which constrain further the anomaly prediction, at the same time, our pretext tasks leverage both motion and appearance which are relevant for anomalies present in this dataset. We observed that the optical flow prediction error is particularly relevant for this dataset since it achieves a Macro AUC of 99.8% as shown in the ablative study in section 5. Adding the appearance information allows the model to achieve state-of-the-art performances on all metrics. The figure 4 illustrates our model performances.

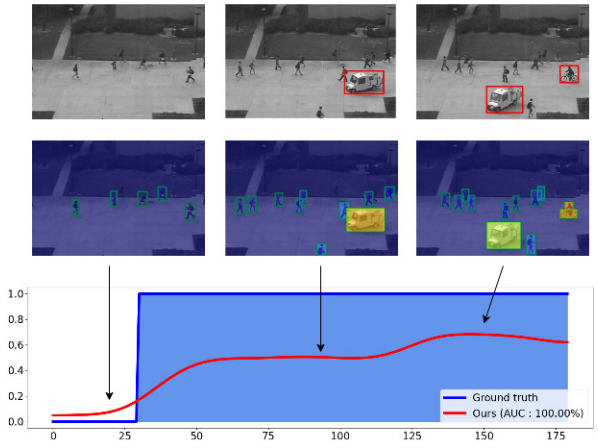

Figure 4. Qualitative results from the video Test004 taken from UCSDped2. Anomalies consist of a car and a cyclist introduced in a pedestrian area. Red boxes indicate the ground truth and green boxes indicate object detections

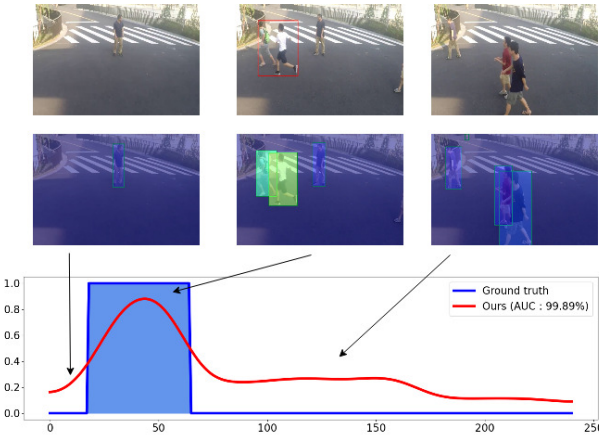

Figure 5. Qualitative results from the video 03_0061 taken from ShanghaiTech. The anomaly consists in one person running after another. The red box indicates the ground truth and green boxes indicate object detections

Results on ShanghaiTech This is one of the most challenging datasets due to its diversity of anomalies and scene

Ablative	pretext tasks		Down-scaling ratios			Evaluation metrics		
Variant	Appearance	Flow Mag.	$\frac{1}{2}$	$\frac{1}{4}$	1	Micro AUC	RBDC	TBDC
Best model	✓	✓	✓			**98.9**	**77.2**	**98.5**
	✓	✓		✓		98.5	77.1	96.3
w/o down-scaling	✓	✓			✓	98.1	73.3	95.0
w/o Flow Mag.	✓	✗				95.0	64.4	77.2
w/o Appearance	✗	✓	✓			96.9	73.1	98.0

Table 2. Micro AUC, RBDC, TBDC scores in (%) obtained on UCSDped2 by removing various component of our method. Best performances are in bold.

changing. Our method achieves significant improvement of 8 p.p in terms of RBDC and slightly outperforms other methods on TBDC. We notice that our method surpasses approaches that use the future frame prediction pretext task such as [14, 15, 34] in terms of AUC which shows the relevance of further constraining those tasks by discarding past frames and performing down-scaling. In addition, we observe that approaches surpassing our method in terms of AUC [7, 8] introduce an additional supervision such as pseudo anomalies in the case of [8] or model distillation [7] where the anomaly detection model is trained on YOLO class prediction. Our approach requires less supervision since 1) object detection is performed only for localisation and not for classification, 2) no pseudo-abnormal data are used. Figure 5 shows an example of anomaly localization and tracking on this dataset. Our scores follow well the temporal and spatial ground truth annotations.

Results on Avenue Our method outperforms state-of-the-art on Micro AUC while being competitive on Macro AUC. In terms of localization, we achieve a significant improvement of 13 p.p on RBDC over the state-of-the-art, while being close in terms of TBDC. The difference in tracking performances between the current state-of-the-art [23, 24] and our method can be explained by the fact that the temporal context used in [23, 24] is larger than ours which leads to a better tracking performances. Nevertheless, compared to object centric approaches such as [8, 7], which uses an equivalent temporal context, we improve the TBDC by 6 p.p. We explain those improvements by two main reasons. First, we found that down-scaling the input considerably enhances RBDC by 10 p.p. Second, we noticed that adding the scale information to the anomaly scoring is beneficial for this dataset since anomalies in Avenue depend on their location with respect to the camera pose (scene layout) unlike in the UCSDped2 and Shanghaitech datasets for which the location information is less relevant to detect anomalies. Therefore, we take into account scale change by multiplying object level anomaly scores S by the bounding boxes width, which leads to a substantial improvement of 12 p.p on RBDC. The figure 6 shows an illustration of our model performances.

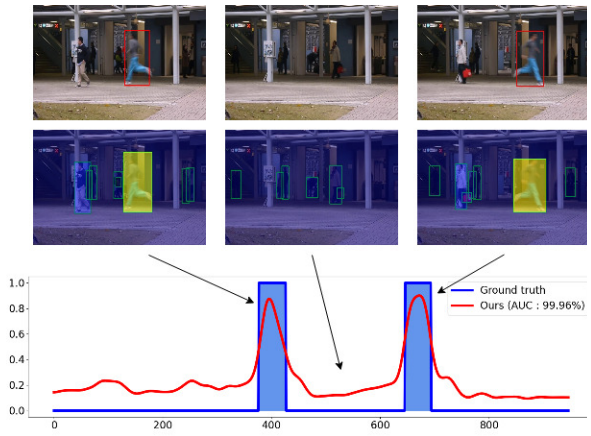

Figure 6. Qualitative results from the video 04 taken from Avenue, the anomaly is represented by a running person. Red boxes indicate the ground truth and green boxes indicate object detections

5. Ablation study

We performed an ablation study to evaluate the impact of each component of our method (Table 2). First, we tested the impact of removing the pretext tasks related to appearance and motion respectively. We notice that in both cases, the performances decrease, showing the complementary roles of appearance and motion in detecting abnormal events. In order to assess the impact of input down-scaling, we tried two down-scaling ratios : $\frac{1}{2}$ and $\frac{1}{4}$. We observe that the ratio $\frac{1}{2}$ gives slightly better performances. One possible explanation is that it offers a good compromise in terms of predictability. In fact, too much down-scaling makes the prediction hard even for normal samples, while without down-scaling it is easier for the model to recover anomalies. Indeed, we observe a global drop in performances which is particularly high for RBDC (4.9p.p) when no down-scaling is performed. This results show that constraining further the prediction task by down-scaling the input is useful.

6. Conclusions

In this work, we introduced a new way of approaching the VAD problem, by imposing constrained pretext tasks to learn appearance and motion normality patterns. The experiments conducted on benchmark datasets show the effectiveness of our methodology. In future work, we will explore more pretext tasks in order to further improve VAD performances and address new types of anomalies.

References

[1] K. Chen, J. Wang, and *et al. MMDetection: Open mmlab detection toolbox and benchmark.* arXiv preprint arXiv:1906.07155, *2019.* 6

[2] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. CVPR, *2011.* 2, 3

[3] F. Dong, Y. Zhang, and X. Nie. Dual discriminator generative adversarial network for video anomaly detection. IEEE Access, *8, 2020.* 2, 3, 4

[4] K. Doshi and Y. Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In CVPR Workshop, *2020.* 3

[5] J. Fang, D. Yan, J. Qiao, and J. Xue. Dada: A large-scale benchmark and model for driver attention prediction in accidental scenarios. ArXiv, *abs/1912.12148, 2019.* 2

[6] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In Image Analysis, *2003.* 6

[7] M.-I. Georgescu, A. Bărbălău, R. T. Ionescu, F. S. Khan, M. C. Popescu, and M. Shah. Anomaly detection in video via self-supervised and multi-task learning. CVPR, *2021.* 2, 3, 4, 6, 7, 8

[8] M. I. Georgescu, R. Ionescu, F. S. Khan, M. Popescu, and M. Shah. A background-agnostic framework with adversarial training for abnormal event detection in video. TPAMI, *2021.* 2, 3, 4, 5, 6, 7, 8

[9] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In ICCV, *2019.* 2, 3, 6, 7

[10] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, *2016.* 2, 3

[11] R. T. Ionescu, F. S. Khan, M.-I. Georgescu, and L. Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In CVPR, *2019.* 2, 3, 4, 6, 7

[12] R. T. Ionescu, S. Smeureanu, M. Popescu, and B. Alexe. Detecting abnormal events in video using narrowed normality clusters. In WACV, *2019.* 2, 3

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, *abs/1412.6980, 2015.* 6

[14] W. Liu, D. L. W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In CVPR, *2018.* 2, 3, 4, 6, 7, 8

[15] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In ICCV, *2021.* 2, 3, 4, 6, 7, 8

[16] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. ICCV, *2013.* 2, 3, 5, 7

[17] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In ICCV, *2017.* 2, 3, 5, 7

[18] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. ICCV, *2017.* 2, 3

[19] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In CVPR, *2010.* 2, 3, 5, 7

[20] T.-N. Nguyen and J. Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In ICCV, *2019.* 2, 3, 6, 7

[21] H. Park, J. Noh, and B. Ham. Learning memory-guided normality for anomaly detection. In CVPR, *2020.* 2, 3, 6

[22] M. Pranav, L. Zhenggang, and S. S. K. A day on campus - an anomaly detection dataset for events in a single camera. In ACCV, *2020.* 2

[23] B. Ramachandra and M. Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In WACV, *2020.* 2, 3, 5, 6, 7, 8

[24] B. Ramachandra, M. J. Jones, and R. R. Vatsavai. Learning a distance function with a siamese network to localize anomalies in videos. WACV, *2020.* 2, 3, 6, 7, 8

[25] B. Ramachandra, M. J. Jones, and R. R. Vatsavai. A survey of single-scene video anomaly detection. TPAMI, *2022.* 2

[26] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In WACV, *2018.* 2

[27] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe. Abnormal event detection in videos using generative adversarial nets. In ICIP, *2017.* 2, 3, 6

[28] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, *2018.* 6

[29] V. Saligrama and Z. Chen. Video anomaly detection based on local statistical aggregates. In CVPR, *2012.* 2

[30] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In CVPR, *2018.* 2, 5

[31] S. Szymanowicz, J. Charles, and R. Cipolla. Discrete neural representations for explainable anomaly detection. In WACV, *2022.* 2, 3

[32] Y. Tang, L. Zhao, S. Zhang, C. Gong, G. Li, and J. Yang. Integrating prediction and reconstruction for anomaly detection. Pattern Recognit. Lett., *2020.* 2, 3, 4

[33] H. T. Vu, T. D. Nguyen, T. Le, W. Luo, and D. Q. Phung. Robust anomaly detection in videos using multilevel representations. In AAAI, *2019.* 2, 3

[34] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, and M. Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In ACM Multimedia, *2020.* 3, 6, 7, 8

[35] Z. Zhang, S. hua Zhong, A. Fares, and Y. Liu. Detecting abnormality with separated foreground and background: Mutual generative adversarial networks for video abnormal event detection. CVIU, *2022.* 3