Object-Centric And Memory-Guided Normality Reconstruction For Video Anomaly Detection

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

OBJECT-CENTRIC AND MEMORY-GUIDED NORMALITY RECONSTRUCTION FOR VIDEO ANOMALY DETECTION

Khalil Bergaoui[1] *, *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
[2]Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France

ABSTRACT

This paper addresses the anomaly detection problem for videosurveillance. Due to the inherent rarity and heterogeneity of abnormal events, this problem is tackled from a normality modeling perspective, where our model learns object-centric normal patterns without seeing anomalous samples during training. Our main contributions consist in coupling object-level action features with a cosine distance-based anomaly estimation function. We therefore extend previous methods by introducing explicit geometric constraints to the mainstream reconstruction-based strategy. Our framework leverages both appearance and motion information to learn object-level behavior and captures prototypical patterns within a memory module. Experiments on several well-known datasets demonstrate the effectiveness of our method as it outperforms current state-of-the-art on most relevant spatio-temporal evaluation metrics.

Index Terms— deep learning, abnormal event detection, video anomaly detection, object-centric normality modeling

1. INTRODUCTION

Video Anomaly Detection (VAD) is an open research problem which consists in detecting rare occurrences of abnormal events. This is a challenging problem due to two main reasons. Although anomalous events are generally defined as rare occurrences that deviate from normal patterns observed in familiar events [1], this definition does not differentiate anomalous events from rare normal ones. Secondly, abnormal events are inherently more difficult to collect and to learn, due to their few occurrences and the multiplicity of their nature. For these reasons, the VAD problem is often viewed within the *one-class* paradigm [2].

In a pioneering work [3], a model is trained to predict future "normal" frame, and anomalies are viewed as inaccurate predictions. The recent method [4] combines multiple proxy tasks (e.g. arrow of time prediction) to characterize anomalous events. Other approaches quantify the deviation from learned normal patterns including distance-based [5, 6, 7, 8, 9] and reconstruction-based methods [10, 11, 12, 4, 13]. Although they have been empirically shown to attain impressive performance levels on current standard benchmark datasets, the used strategy is not always in line with the nature of anomalous event detection. In fact, as pointed out in [4] , a car stopped in a pedestrian area should be labeled as an anomaly, yet the car is trivial to reconstruct (at a pixel-wise level) in a future frame, since it is still standing. This example shows that pixel-wise reconstruction error is suboptimal for anomaly detection. The recent work [13] addressed this challenge by learning object-level patterns of normal appearance and motion by training a discriminator network to classify (normal *vs.* abnormal samples) given pairs of reconstruction error maps. Despite being the current state of the art, this method includes out-of-domain observations during training, introducing a bias to the normality modeling. Instead, we propose to tackle these challenges by extending the mainstream reconstruction assumption on which most state-of-the-art methods [3, 11, 12, 13] are implicitly based: *Given a normality model, normal observations are easier to reconstruct from a low-dimensional representation than abnormal observations*. We propose to add geometric constraints in the reconstruction space in order to further narrow down this assumption to be more in line with the anomaly detection task. Different from prior works, we combine a cosine distance-based anomaly estimation function with pretrained object-level features. Additionally, we propose to constrain our model to reconstruct independent motion and appearance features from a single embedding space. This way, our network has fewer degrees of freedom to perform the training task, which is in line with the aforementioned reconstruction assumption. Following [5, 9, 14, 4, 13] we apply an object detector allowing to localize anomalies at the object level, which is semantically more relevant than at the pixel-level. Similarly to [11, 12], we incorporate a memory block in our framework in order to model diverse normality patterns. In summary, our contributions are:

- Imposing geometric constraints in the reconstruction space using cosine distance.
- Introducing object-level action prototypical features.
- State-of-the-art results on the most relevant metrics.

*Equal contributions

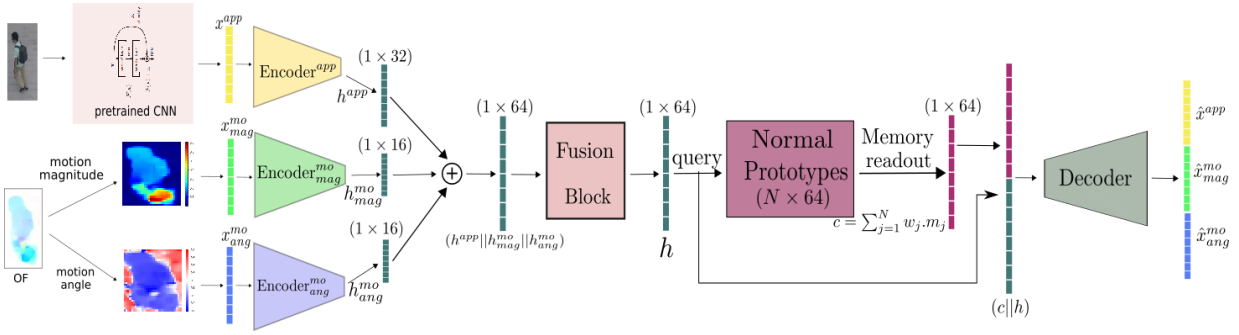

Fig. 1. Overview of OMAE. At the preprocessing step, object bounding boxes and Optical Flow (OF) are computed. Object appearance features, extracted using a pretrained CNN, motion magnitude and angle maps are fed to the corresponding autoencoders. The encoded representations are fused (h) and sent as a query to the memory module which fetches similar memory items. Their linear combination concatenated with h is sent to the decoder to obtain object appearance and motion reconstructions. The anomaly score is defined based on the dissimilarity between the input features and their reconstructions as well the dissimilarity between the query h and its neighbors.

2. METHOD

2.1. Overview

The architecture of the proposed method: OMAE which stands for object centric memory-guided auto-encoder is displayed in Fig. 1.First, we detect objects and compute optical flow for each frame. Next, appearance x^{app} and motion x^{mo} features are extracted for each object. The former are obtained using a pretrained CNN, whereas the latter consist of motion magnitude and angle maps. We denote the input features $\mathcal{X} = \{(x^{app}_{,} x^{mo})_j \quad ; 1 \leq j \leq O\}$, where O is the total number of objects in the training set. Then, the object representations are encoded and fused into a single embedding h. The three encoders have the same structure: two successive blocks, each block is a shallow fully connected network of two layers. The fusion block consists of a single shallow fully connected network of two layers that learns new embeddings from the concatenation of the three encoders' bottlenecks. Thus, we obtain a single hidden representation combining motion and appearance, which can be interpreted as an object-level action feature vector. We will use $\mathcal{H}$ to denote the set of these hidden features corresponding to input vectors in $\mathcal{X}$. This action feature is then used as a query to the *memory of normal patterns* to extract similar existing prototypes in the memory module $\mathcal{M} = \{m_i \ , 1 \leq i \leq N\}$, where N is the total number of memory items. Similar to [12] the most similar memory item m_k to the query h is defined as the soft nearest neighbor:

$$k = \underset{1 \leq i \leq N}{\operatorname{argmax}} (w_i) \, ; w_i = \frac{exp(h^T m_i)}{\sum_{j=1}^{N} exp(h^T m_j)} ; 1 \leq i \leq N.$$

After the memory readout step, and following an attention-mechanism strategy, a linear combination of the memory items is computed as $c = \sum_{j=1}^{N} w_j.m_j$ and then concatenated to the query h to obtain an augmented hidden representation $z = (c||h)$ that will be used as input to the decoder network. Finally, a single fully connected decoder network learns to reconstruct x given z. This way, our auto-encoder model is trained to reconstruct object-centric features $\hat{x}$ under two major constraints:

1. The auto-encoder learns normality patterns (memory items) that allow the reconstruction of both appearance and motion features from a single embedding space.

2. The decoder's reconstructive capacity is limited by the set of memory items thus its generalization ability is reduced, which is useful for detecting anomalies (via poor outlier reconstructions).

During inference, an *anomaly score* is attributed to each object (cf. Section 2.3) based on the dissimilarity between x and $\hat{x}$ as well as the dissimilarity between h and its neighbors in the space of prototypical patterns ($(m_i)_{1 \leq i \leq N}$).

2.2. Loss functions

To take into account the above-mentioned learning constraints, we combine different loss terms in our objective function. We incorporate a reconstruction term $\mathcal{L}_{rec}$ to minimize the discrepancy between the input x and its reconstruction and a memory term $\mathcal{L}_{mem}$ to capture normal prototypical patterns observed in the training set. Hence, the total loss is given by: $\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{mem}$.

Reconstruction Loss: We constrain the reconstruction to not only be in the Euclidean neighborhood of the input but also to lie on the same spatial direction. The geometrical constraint is applied via a cosine distance loss and controlled via

the hyperparameter λ_{cos} such that:

$$\mathcal{L}_{rec} = (\|x^{app} - \hat{x}^{app}\|_2 + \|x^{mo} - \hat{x}^{mo}\|_2) + \lambda_{cos} \times \left(1 - \frac{< x^{app}, \hat{x}^{app} >}{\|x^{app}\|_2 \|\hat{x}^{app}\|_2}\right)$$

Memory Loss: This loss is obtained as a combination of three terms:

$$\mathcal{L}_{mem} = \lambda_{comp}.\mathcal{L}_{comp} + \lambda_{tr}.\mathcal{L}_{tr} + \lambda_{OLE}.\mathcal{L}_{OLE}$$

Similarly to [12], discriminative normality action features prototypes are learnt based on nearest neighbor distances within the memory space via a loss that favors compactness of data samples around prototypes:

$$\mathcal{L}_{comp} = \sum_{k=1}^{N} \sum_{j \in \mathcal{U}_k \ ; \ \mathcal{U}_k \neq \emptyset} \|h_j - m_k\|_2$$

where $\mathcal{U}_k \subset \{1, ..., O\}$ is the subset of training object indices which have the memory item m_k as their first nearest neighbor.

We also use the same triplet loss $\mathcal{L}_{tr}$ introduced in [12]. Contrarily to [12], we incorporate a third term $\mathcal{L}_{OLE}$ that adds orthogonality constraints within the memory space. This is achieved through the geometric loss formulation proposed in [15] for supervised classification. We adapt the OLE (Orthogonal Low-rank Embedding) loss to our setting by formulating the memory query step as a classification problem:

$$\mathcal{L}_{OLE} = \sum_{c=1}^{N} max(\Delta, \|H_c\|_*) - \|H\|_*$$

where Δ is a positive real number that we set to 1 in all our experiments, H_c is the sub-matrix of object hidden representations within the batch that are attributed to memory item m_c. $\|.\|_*$ denotes the nuclear norm i.e the sum of matrix singular values.

2.3. Inference: Abnormality score

At test time, given the t^{th} frame, a set of abnormality scores is denoted as $\mathcal{S}_t = \{s_t^j \ , 1 \leq j \leq O_t\}$, where O_t is the number of detected objects in frame t.

Each score s_t^j is computed as follows:

$$s_t^j = s = \frac{1}{3}\left(s_{L_2}^{rec} + s_{cos}^{rec} + s^{mem}\right)$$

where:

$s_{L_2}^{rec} = g\left(\|x^{app} - \hat{x}^{app}\|_2 + \|x^{mo} - \hat{x}^{mo}\|_2\right)$

$s_{cos}^{rec} = g\left((1 - \frac{<x^{app}, \hat{x}^{app}>}{\|x^{app}\|_2 \|\hat{x}^{app}\|_2}) + (1 - \frac{<x^{mo}, \hat{x}^{mo}>}{\|x^{mo}\|_2 \|\hat{x}^{mo}\|_2})\right)$

$s^{mem} = g\left(1 - \frac{<h, m_k>}{\|h\|_2 \|m_k\|_2}\right)$

with $g(.)$ a normalization function:

$g(\epsilon) = \frac{\epsilon - \epsilon_{min}}{\epsilon_{max} - \epsilon_{min}}$

where ϵ_{min} and ϵ_{max} are the lowest and the highest object level scores respectively across the entire video.

3. EXPERIMENTS AND RESULTS

Datasets. Several benchmarks had been proposed for evaluating anomaly detection methods [16, 17, 18, 19, 20, 7, 21], In order to compare our methods with existing approaches, we performed experiments on the most common datasets : **UCSD ped2** [16] includes 16 training and 12 testing videos of resolution 240x360. Anomalous events include riding a bike and driving a vehicle on a sidewalk. **CUHK Avenue** [17] consists of 16 training and 21 test videos of resolution 360x640 with abnormal events such as running, walking towards the camera, or throwing papers. We use the annotations provided by [7]. **ShanghaiTech** [18] contains 330 training and 107 testing videos of resolution 480x856 with 13 different scenes. Each scene has a different background or camera angle. Abnormal events include jumping, running, or stalking on a sidewalk.

Evaluation metrics. Since we focus on spatio-temporal anomaly detection, we adopt the Region-Based Detection Criterion (RBDC) and the Track-Based Detection Criterion (TBDC) metrics introduced in [7] as an alternative to the flawed pixel-level AUC metric. We also report the Area Under of the ROC Curve (AUC) obtained with respect to the frame-level ground-truth annotations. Yet, it gives only a global frame score and, therefore, doesn't reflect the model capacity to localize anomalies. Emphasis is given to the parts of the ROC curve where false positive rate is too high for a practical use [22]. Hence, AUC is the least relevant metric in our study.

Parameters and implementation details. The first step of our framework is to perform object detection using Yolov3 [23] pretrained on COCO dataset as in [4, 13, 5]. We set the detection confidence to 0.7 on ShanghaiTech and Avenue for both the training and testing sets. Since the image resolution of UCSD ped2 is lower, we reduced the threshold to 0.5. We used ResNet101 to precompute appearance features of detected objects and Farneback's algorithm to compute optical flows [24]. Similarly to [4], we trained the network for 30 epochs on each dataset using Adam optimizer with a learning rate of 10^{-3}. We use a batch size of 256 object-level action features for the smallest dataset (UCSD ped2) and 512 for ShanghaiTech and Avenue. For all experiments, we set λ_{cos} = 0.1 so that the cosine term has a similar order of magnitude as the L_2 reconstruction term; as well as fixed empirical values for the number of memory items and the memory loss weights: $N = 40$, λ_{comp} = 1.6 , λ_{tr} = 0.2, λ_{OLE} = 0.3. The epoch achieving the lowest loss value in training is used for inference. As a post-processing step and similarly to [13], we apply a spatio-temporal mean filtering to smooth object level scores and a Gaussian filter at the frame level. In the case of Avenue dataset, scale change is taken into account by an anomaly score adjustment: the anomaly score of an object is multiplied by its bounding box width. This post-processing allows an increase of 26% in RBDC and 2% in TBDC without degrading AUC.

Approach	Method	UCSD Ped2			ShanghaiTech			Avenue		
		AUC	RBDC	TBDC	AUC	RBDC	TBDC	AUC	RBDC	TBDC
frame-level	MemAE [11]	94.1	-	-	71.2	-	-	83.3	-	-
	MNAD [12]	97.0	-	-	72.5	-	-	88.5	-	-
	SSMT [4]	92.4	-	-	83.5	-	-	86.9	-	-
video patch-level	StreetScene [7]	88.3	62.50	80.50	-	-	-	72.0	35.80	**80.90**
	Siamese [8]	94.0	<u>74.0</u>	89.3	-	-	-	87.2	41.20	<u>78.60</u>
frame & object level	SSMT [4]	**99.8**	-	-	**90.2**	-	-	<u>92.8</u>	-	-
object-level	dummyAE [5]	82.2	-	-	78.6	20.65	44.54	88.9	-	-
	SSMT [4]	**99.8**	72.8	91.2	<u>89.3</u>	-	-	91.9	-	-
	BAF [13]	<u>98.7</u>	69.23	<u>93.15</u>	82.7	<u>41.43</u>	<u>78.79</u>	92.3	<u>65.05</u>	66.85
	OMAE (ours)	96.46	**80.07**	**95.39**	79.18	**51.51**	**82.19**	**93.56**	**75.83**	70.02

Table 1. Comparison with the state-of-the-art methods (%). Best results in bold and second best results are underlined

Features	Reconstruction loss (AE)		Anomaly score		Evaluation metrics		
	MSE	COS	MAE	COS	AUC	RBDC	TBDC
appearance	✓	✓	✓	✓ (AE+Mem)	70.21	35.68	49.82
appearance + motion	✓	-	✓	-	70.48	35.15	69.16
	✓	-	-	✓ (Mem)	71.61	27.72	57.70
	✓	-	-	✓ (AE)	76.66	43.02	68.13
	✓	-	✓	✓ (AE+Mem)	75.98	39.53	67.49
	✓	✓	✓	-	70.71	35.96	71.27
	✓	✓	-	✓ (Mem)	69.50	31.18	74.47
	✓	✓	-	✓ (AE)	76.68	43.99	69.72
	✓	✓	✓	✓ (AE+Mem)	77.81	49.37	**83.61**
+ 3D smoothing	✓	✓	✓	✓ (AE+Mem)	**79.18**	**51.51**	82.19

Table 2. Frame-AUC, RBDC and TBDC scores (in %) obtained on ShanghaiTech by making gradual design changes to the baseline method, until the final framework. (*AE* stands for the auto-encoder component; *Mem* for the nearest-neighbor memory item.)

Ablation study. We conduct an ablation study on ShanghaiTech dataset to assess the importance of each component in our framework. The corresponding results are shown in Table 2. We can see that both appearance and motion features are necessary to model usual actions to better detect anomalies. Indeed, the baseline model takes only appearance features as input and performs lower than the current state of the art on all metrics. Including the motion information through optical flow improves the frame-AUC by 8.97% and significantly increases RBDC and TBDC by 15.83% and 32.37% respectively allowing OMAE to outperform the current state of the art on these metrics. We also note that including the cosine similarity together with MSE or MAE in the reconstruction loss as well as in the abnormality scores improves all metrics significantly, showing the importance of the orthogonality constraints.

Comparison with state of the art. In Table 1, we present our results in comparison with the state-of-the-art methods on the 3 benchmark datasets. Our framework significantly outperforms current state-of-the-art methods on the most relevant evaluation metrics RBDC and TBDC which quantifies the model's ability to localize anomalies spatially and to track them temporally. It's important to highlight that, given a single dataset, there is not a single best method that outperforms the other approaches on the three metrics. On ShanghaiTech dataset, we outperform best previous work by a margin of 10.08% on RBDC and 3.4% on TBDC while remaining competitive with respect to other object-centric approaches in terms of frame-level AUC. Our model also reaches new state-of-the-art performances in terms of RBDC and TBDC on UCSD ped2 with significant margins of 6.07% and 2.24% respectively. On Avenue, our model improves the current state of the art by 10,78% on the RBDC metric and outperforms the recent object-centric approach [13] in terms of TBDC by a margin of 3.17% while remaining competitive on the AUC metric. Advantageously, inference time takes only 20ms for a batch of 256 precomputed object action features. The preprocessing time required for computing those features is the following: optical flow extraction (170ms), object detection (50ms) and feature extraction (40ms). In addition, training [3] is much faster than in [12]: only 70 min on ShanghaiTech, 9 min on Avenue and 4 min on UCSD ped2 with a single NVIDIA TITAN X (PASCAL) GPU.

4. CONCLUSIONS

In this work we introduced OMAE, an object-centric VAD framework that uses a memory module for object-level appearance and motion features with a new abnormality scoring strategy based on cosine distance. In our experiments, OMAE reaches superior results on localization and tracking metrics while remaining competitive on the frame-level AUC. This shows the effectiveness of our approach to localize anomalies better than the current state of the art.

5. REFERENCES

[1] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai, "A survey of single-scene video anomaly detection," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[3]This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council

[2] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, 2021.

[3] W. Liu, D. Lian W. Luo, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] 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*, 2021.

[5] 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 *Proceedings of CVPR*, 2019, pp. 7842–7851.

[6] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe, "Detecting abnormal events in video using narrowed normality clusters," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1951–1960.

[7] Bharathkumar Ramachandra and Michael Jones, "Street scene: A new dataset and evaluation protocol for video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[8] Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai, "Learning a distance function with a siamese network to localize anomalies in videos," in *WACV*, 2020.

[9] Keval Doshi and Yasin Yilmaz, "Any-shot sequential anomaly detection in surveillance videos," in *Proceedings of CVPRW*, 2020, pp. 934–935.

[10] Trong-Nguyen Nguyen and Jean Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[11] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[12] Hyunjong Park, Jongyoun Noh, and Bumsub Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14372–14381.

[13] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah, "A background-agnostic framework with adversarial training for abnormal event detection in video," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, pp. 1–1.

[14] Keval Doshi and Yasin Yilmaz, "Continual learning for anomaly detection in surveillance videos," in *Proceedings of CVPRW*, 2020, pp. 254–255.

[15] José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro, "Olé: Orthogonal low-rank embedding, a plug and play geometric loss for deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[16] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos, "Anomaly detection in crowded scenes," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.

[17] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of ICCV*, 2013, pp. 2720–2727.

[18] Weixin Luo, Wen Liu, and Shenghua Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 341–349.

[19] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[20] Kun Liu and Huadong Ma, "Exploring background-bias for anomaly detection in surveillance videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, 2019, MM '19, p. 1490–1499, Association for Computing Machinery.

[21] Mantini Pranav, Li Zhenggang, and Shah Shishir K, "A day on campus - an anomaly detection dataset for events in a single camera," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[22] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real, "Auc: a misleading measure of the performance of predictive distribution models," *Global ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.

[23] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," in *arXiv preprint arXiv:1804.02767*, 2018.

[24] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.