# Delving into CLIP latent space for Video Anomaly Detection and Recognition

[1]Luca Zanella*, [1]Benedetta Liberatori*, [1]Willi Menapace, [2]Fabio Poiesi, [2]Yiming Wang, [1,2]Elisa Ricci
[1] University of Trento, Trento, Italy [2] Fondazione Bruno Kessler, Trento, Italy

## Abstract

*We tackle detecting and recognising anomalies that deviate from normal patterns in videos, one of the classic Out-Of-Distribution (OOD) problem in computer vision. Our novel method AnomalyCLIP exploits Large Language and Vision (LLV) models, such as CLIP, with video-level supervision. The proposal specifically involves manipulating the latent CLIP feature space to identify the normal event subspace, which in turn allows us to effectively learn text-driven directions for abnormal events. When anomalous frames are projected onto these directions, they exhibit a large feature magnitude if they belong to a particular class. We also introduce a computationally efficient Transformer architecture to model short- and long-term temporal dependencies between frames, ultimately producing the final anomaly score and class prediction probabilities. We compare AnomalyCLIP against state-of-the-art methods considering two major anomaly detection benchmarks, i.e. ShanghaiTech and UCF-Crime, and empirically show that it outperforms baselines in recognising video anomalies. Code is available at* https://github.com/luca-zanella-dvl/AnomalyCLIP.

## 1. Introduction

Video anomaly detection (VAD) [21] is a widely studied task which aims to identify activities that deviate from normal patterns in videos [14, 33, 15, 1, 23, 5, 28]. VAD can be formulated as an Out-Of-Distribution (OOD) detection problem, and often being addressed by either learning one-class, i.e. normal, [16, 13, 11, 33], or with weak video-level supervision [22, 8, 9, 30, 24], as the data is typically highly imbalanced, *i.e.* normal events are many, whilst abnormal events are less represented due to their rare and sporadic occurrences. For the very same reason, Video Anomaly Recognition (VAR), i.e. detecting and recognising anomaly types (*e.g.* shooting vs. explosion), despite being desirable, has not been fully tackled by existing methods [22].

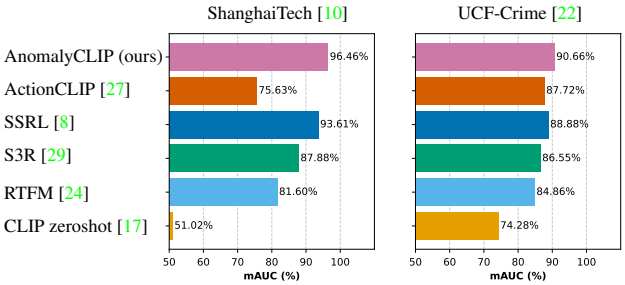



Figure 1: Comparison of various anomaly recognition methods on the ShanghaiTech and UCF-Crime datasets, measured by mAUC, the mean of the binary area under the curve (AUC) of the receiver operating characteristic (ROC) of all anomalous classes. Notably, our method, AnomalyCLIP, achieves the highest performance on both datasets.

The emergence of Large Language and Vision (LLV) models or foundation models [19, 17] that are trained on web-scale datasets [18], have shown strong generalisation capabilities in several downstream tasks and have become a key ingredient of modern computer vision systems. These pre-trained models are publicly available and can be seamlessly integrated into any recognition system. LLV models have also been effectively applied to videos and to action recognition tasks with video-level supervision [27, 32]. Thus, we are motivated to exploit LLV models to address VAR. We argue that by leveraging representations derived from LLV models, we can obtain more discriminative features for detecting and classifying abnormal behaviours. However, as supported by our experiments (Fig. 1), a naive application of LLV models to VAR cannot address the data imbalance and the subtle differences between normal and abnormal frames of the same video.

Our proposal, AnomalyCLIP, is a novel solution for VAR based on the CLIP model [17], achieving state-of-the-art on two benchmark datasets, ShanghaiTech [10] and UCF-Crime [22], as shown in Fig. 1. AnomalyCLIP produces video representations that can be mapped to the textual description of the anomalous event. Rather than directly operating on the CLIP feature space, we re-centre it around a normality prototype, as shown in Fig. 2 (a). In this way, the space assumes important semantics: the magnitude of

This work was sponsored by EU ISFP PRECRISIS (ISFP-2022-TFI-AG-PROTECT-02-101100539) and the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. *Equal contribution.

the features indicates the degree of anomaly, while the direction from the origin indicates the anomaly type. To learn the directions that represent the desired anomaly classes, we propose a *Selector* model that employs prompt learning and a projection operator tailored to our new space to identify the parts in a video that better match the textual description of the anomaly. This ability is instrumental to address the data imbalance problem. We use the predictions of the Selector model to implement a semantically-guided Multiple Instance Learning (MIL) strategy that aims to widen the gap between the most anomalous segments of anomalous videos and normal ones. Differently from the features typically employed in VAD that are extracted using temporal-aware backbones [2, 12], CLIP visual features do not bear any temporal semantics as it operates at the image level. We thus propose a *Temporal* model, implemented as an Axial Transformer [6], which models both short-term relationships between successive frames and long-term dependencies between parts of the video.

## 2. Proposed method

AnomalyCLIP addresses VAR in a weakly-supervised setting. We consider a dataset of tuples in the form of $(\mathbf{V}, c)$, where $\mathbf{V}$ is a video and $c$ indicates the type of anomaly in the video ($c = \emptyset$ means no anomaly is present, thus being *Normal*). In the following, we omit the subscripts for the purpose of readability. We follow the Multiple Instance Learning (MIL) framework [22] to handle the video-level supervision and the imbalance between the normal and abnormal. MIL models each video as a bag of segments $\mathbf{V} = [\mathbf{S}_1, ..., \mathbf{S}_S] \in \mathbb{R}^{S \times F \times D}$, where $S$ is the number of segments, $F$ is the number of frames in each segment, and $D$ is the number of features associated to each frame. Each segment can be seen as $\mathbf{S} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_F] \in \mathbb{R}^{F \times D}$ where $\boldsymbol{x} \in \mathbb{R}^D$ is the feature corresponding to each frame. MIL computes a likelihood of each frame being anomalous, selects the most normal/anomalous ones based on it, and maximises the difference in the predicted likelihood between the most normal and abnormal.

**Selector model.** It is crucial for VAD and VAR to distinguish anomalous and normal frames in anomalous videos. Our Selector model $\mathcal{S}$ operates in the CLIP [17] feature space and exploits CoOp [38] to learn a set of directions in this space to represent the types of anomalies. The set of CLIP features for each frame in the dataset forms a space that is clustered around a central point, which we call the normality prototype. Therefore, the magnitude of the difference between a feature and the normality prototype reflects the likelihood of being abnormal, while its direction indicates the anomaly type. Following this intuition, we define the normal prototype $\boldsymbol{m}$ as the average feature extracted by the CLIP image encoder $\mathcal{E}_I$ on all $N$ frames $\mathbf{I}$ contained in videos labelled as normal in the dataset, i.e. $\boldsymbol{m} = \frac{1}{N} \sum_{j=1}^{N} \mathcal{E}_I(\mathbf{I}_j)$. For each frame $\mathbf{I}$ in the dataset, we produce frame features $\boldsymbol{x}$ by subtracting the normality prototype from the CLIP encoded feature, i.e., $\boldsymbol{x} = \mathcal{E}_I(\mathbf{I}) - \boldsymbol{m}$.

We then employ the prompt learning CoOp method [38] to learn the textual prompt embeddings whose directions can be used to indicate the anomalous classes. Given a class $c$ and the textual description of the label $\boldsymbol{t}^c$ expressed as token embeddings, we consider learnable context vectors $\boldsymbol{t}^{\text{ctx}}$ and derive the corresponding direction for the class $\boldsymbol{d}_c \in \mathbb{R}^D$ as $\boldsymbol{d}_c = \mathcal{E}_T([\boldsymbol{t}^{\text{ctx}}, \boldsymbol{t}^c]) - \boldsymbol{m}$, where $\mathcal{E}_T$ indicates the CLIP text encoder. Each class has a learnt direction.

The learnt directions serve as the base for our Selector model $\mathcal{S}$. As shown in Fig. 2(b), the magnitude of the projection of frame feature $\boldsymbol{x}$ on direction $\boldsymbol{d}_c$ indicates the likelihood of the anomalous class $c$:

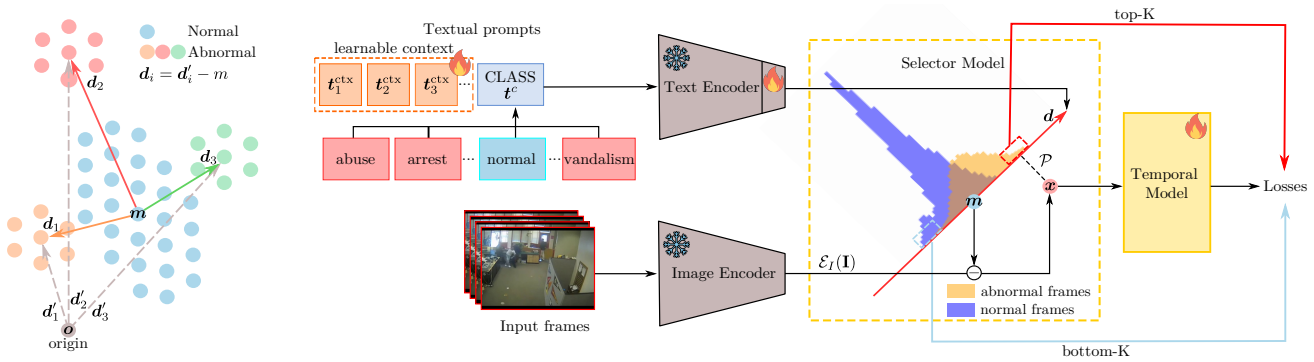
$$\mathcal{S}(\boldsymbol{x}) = [\mathcal{P}(\boldsymbol{x}, \boldsymbol{d}_1), ..., \mathcal{P}(\boldsymbol{x}, \boldsymbol{d}_C)] \in \mathbb{R}^C, \tag{1}$$

where $\mathcal{P}(\cdot)$ indicates our projection operation and $C$ is the number of classes. We further perform a batch normalisation [7] after the projection to normalise the magnitude scale, i.e. $\mathcal{P}(\boldsymbol{x}, \boldsymbol{d}_i) = \text{BN}\left(\frac{\boldsymbol{x} \cdot \boldsymbol{d}_i}{||\boldsymbol{d}_i||}\right)$, producing a distribution of projected features with zero mean and unitary variance.

The definition of likelihood can be extended to segments by summing the likelihoods of each frame, i.e. $\mathcal{S}(\mathbf{S}) = \sum_{i=1}^{F} \mathcal{S}(\boldsymbol{x}_i) \in \mathbb{R}^C$.

**Temporal Model.** As CLIP operates at the image level, the Selector model only learns a *time-independent* separation between anomalous and normal frames. To enrich the visual features with temporal information, we further propose the Temporal model $\mathcal{T}$ to model the relationships among frames in both short-term and long-term. We use a Transformer architecture to capture the short-term temporal dependencies between frames in a segment and the long-term temporal dependencies between all segments in a video, motivated by their success in relevant sequence modelling tasks [25]. As all the segments of $\mathbf{V}$ are received as the input, the large number of segments $S$ and frames $F$ increases the computational requirements for self attention. To reduce this cost, we implement $\mathcal{T}$ as an Axial Transformer [6] that computes attention separately for the two axes corresponding to the segments and the features in each segment. Finally, a sigmoid activation is applied to output the likelihood.

**Predictions Aggregation.** We combine the predictions from $\mathcal{S}$ and $\mathcal{T}$ to obtain the final output: the probabilities indicating whether a frame is normal or anomalous ($p_N(\boldsymbol{x})$ and $p_A(\boldsymbol{x})$) and the probability that a frame presents an anomaly of a certain class ($p_{A,c}(\boldsymbol{x})$). Given an input frame feature $\boldsymbol{x}$, we define its probability of being anomalous $p_A(\boldsymbol{x})$ as its corresponding output from the Temporal model $\mathcal{T}$. The probability of the frame being normal is $p_N(\boldsymbol{x}) = 1 - p_A(\boldsymbol{x})$. To obtain the probability distribution of the frame to present an anomaly of a specific




(a) (b)

Figure 2: (a) Illustration of the CLIP space and the effects of the re-centring transformation with the normal prototype $\boldsymbol{m}$. (b) Our proposed framework. The Selector model estimates the likelihood of each feature $\boldsymbol{x}$ being an anomalous class, which is exploited by the MIL selection of the top-$K$ and bottom-$K$ abnormal segments. A Temporal model aggregates the visual features over time to produce the final prediction.

class $p_{A,c}(\boldsymbol{x})$, we employ the predictions of the Selector model that can be seen as the conditional distribution over the anomalous classes $p_{c|A}(\boldsymbol{x}) = \texttt{softmax}(\mathcal{S}(\boldsymbol{x}))$. From the definition of conditional probability it follows that $p_{A,c}(\boldsymbol{x}) = p_A(\boldsymbol{x}) * p_{c|A}(\boldsymbol{x})$.

**Training.** We train the model following the MIL framework. MIL considers a batch with an equal number of normal and anomalous videos, uses the predicted likelihoods to identify the top-$K$ most abnormal segments in anomalous videos, and imposes separation from the normal ones [22]. Due to the higher capacity of $\mathcal{T}$ with respect to $\mathcal{S}$ and its initial random initialisation, $\mathcal{T}$ can not directly perform this selection since the predicted likelihoods would be excessively noisy. Instead, we use the likelihood predictions from $\mathcal{S}$ to perform MIL segment selection.

Table 1: Comparison on VAD and VAR with ShanghaiTech.

| Supervision | Method | Features | VAD | VAR | AUC(%) | mAUC(%) |
|---|---|---|---|---|---|---|
| One-class | MNAD (2020) [16] | | ✓ | | 70.50 | |
| | MPN (2021) [13] | | ✓ | | 73.80 | |
| | HF$^2VAD$ (2021) [11] | | ✓ | | 76.20 | |
| | Zaheer *et al.* (2022) [35] | ResNext | ✓ | | 79.62 | |
| Unsupervised | Zaheer *et al.* (2022) [35] | ResNext | ✓ | | 78.93 | |
| Zero-shot | CLIP [17] | ViT-B/16 | | ✓ | 49.17 | 51.02 |
| Weakly-supervised | Sultani *et al.* (2018) [22] | C3D-RGB | ✓ | | 86.30 | |
| | IBL *et al.* (2019) [36] | C3D-RGB | ✓ | | 82.50 | |
| | Zaheer *et al.* (2022) [35] | ResNext | ✓ | | 86.21 | |
| | GCN (2019) [37] | TSN-RGB | ✓ | | 84.44 | |
| | MIST (2021) [4] | I3D-RGB | ✓ | | 94.83 | |
| | CLAWS (2021) [34] | C3D-RGB | ✓ | | 89.67 | |
| | RTFM (2021) [24] | I3D-RGB | ✓ | | 97.21 | 81.60 |
| | Wu and Liu (2021) [30] | I3D-RGB | ✓ | | 97.48 | |
| | MSL (2022) [9] | I3D-RGB | ✓ | | 96.08 | |
| | MSL (2022) [9] | VideoSwin-RGB | ✓ | | 97.32 | |
| | S3R (2022) [29] | I3D-RGB | ✓ | | 97.48 | 87.88 |
| | SSRL (2022) [8] | I3D-RGB | ✓ | | 97.98 | 93.61 |
| | ActionCLIP (2021) [27] | ViT-B/16 | | ✓ | 96.36 | 75.63 |
| | AnomalyCLIP (ours) | ViT-B/16 | ✓ | ✓ | **98.07** | **96.46** |

## 3. Experiments

We validate our method in terms of VAD and VAR performances against baselines taken from state-of-the-art VAD and action recognition methods which we adapt to the VAR task, using two benchmark VAD datasets, *i.e.*, ShanghaiTech [10] and UCF-Crime [22] following the protocol in [8]. We measure the performance regarding VAD using the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC) as it is agnostic to thresholding for the detection task. Regarding the VAR performance, we measure the AUC for each anomalous class, by considering the anomalous frames of the class as positive and all other frames as negatives. Successively, the mean AUC (mAUC) is computed over all the anomalous classes.

Please refer to Supp. Mat. for more qualitative results and ablations regarding our approach in terms of losses and feature representation.

**Comparisons.** Regarding VAD, we compare AnomalyCLIP against state-of-the-art methods with different supervision setups, including one-class [16, 13, 11], unsupervised [35] and weakly-supervised [24, 29, 8]. As no existing method specifically addresses VAR, we compare against i) the best-performing VAD methods including RTFM [24], S3R [29] and SSRL [8], that we re-purposed by appending a trainable classification head using a cross entropy loss on the top-$K$ most anomalous segments selected by their method, and ii) CLIP-based baselines including *Zero-shot CLIP* [17] and weakly-supervised ActionCLIP [27].

Tab. 1 presents the results on ShanghaiTech [10], a rather saturated dataset for VAD due to its simplicity in scenarios. AnomalyCLIP scores the state-of-the-art results on both VAD and VAR, with $+0.09\%$ and $+2.85\%$ in terms of AUC ROC and mAUC ROC, respectively. ActionCLIP [27] performs poorly in terms of mAUC, which we attribute to the low proportion of abnormal events in ShanghaiTech that makes the MIL selection important to avoid incorrect supervisory signals on normal frames of abnormal videos.

Table 2: Comparison on VAD and VAR with UCF-Crime.

| Supervision | Method | Features | VAD | VAR | AUC(%) | mAUC(%) |
|---|---|---|---|---|---|---|
| One-class | SVM Baseline [22] | | ✓ | | 50.00 | |
| | SSV (2018) [20] | | ✓ | | 58.50 | |
| | BODS (2019) [26] | I3D-RGB | ✓ | | 68.26 | |
| | GODS (2019) [26] | I3D-RGB | ✓ | | 70.46 | |
| | Zaheer *et al.* (2022) [35] | ResNext | ✓ | | 74.20 | |
| Un-supervised | Zaheer *et al.* (2022) [35] | ResNext | ✓ | | 71.04 | |
| Zero-shot | CLIP [17] | ViT-B/16 | | ✓ | 58.63 | 74.28 |
| Weakly-supervised | Sultani *et al.* (2018) [22] | C3D-RGB | ✓ | | 75.41 | |
| | Sultani *et al.* (2018) [22] | I3D-RGB | ✓ | | 77.92 | |
| | IBL (2019) [36] | C3D-RGB | ✓ | | 78.66 | |
| | Zaheer *et al.* (2022) [35] | ResNext | ✓ | | 79.84 | |
| | GCN (2019) [37] | TSN-RGB | ✓ | | 82.12 | |
| | MIST (2021) [4] | I3D-RGB | ✓ | | 82.30 | |
| | Wu *et al.* (2020) [31] | I3D-RGB | ✓ | | 82.44 | |
| | CLAWS (2021) [34] | C3D-RGB | ✓ | | 83.03 | |
| | RTFM (2021) [24] | VideoSwin-RGB | ✓ | | 83.31 | |
| | RTFM (2021) [24] | I3D-RGB | ✓ | | 84.03 | 84.86 |
| | Wu and Liu (2021) [30] | I3D-RGB | ✓ | | 84.89 | |
| | MSL (2022) [9] | I3D-RGB | ✓ | | 85.30 | |
| | MSL (2022) [9] | VideoSwin-RGB | ✓ | | 85.62 | |
| | S3R (2022) [29] | I3D-RGB | ✓ | | 85.99 | 86.55 |
| | MGFN (2023) [3] | VideoSwin-RGB | ✓ | | 86.67 | |
| | MGFN (2023) [3] | I3D-RGB | ✓ | | 86.98 | |
| | SSRL (2022) [8] | I3D-RGB | ✓ | | **87.43** | 88.88 |
| | ActionCLIP (2021) [27] | ViT-B/16 | | ✓ | 82.30 | 87.72 |
| | AnomalyCLIP (ours) | ViT-B/16 | ✓ | ✓ | 86.36 | **90.66** |

AnomalyCLIP achieves a large improvement of $+45.44\%$ in terms of mAUC against zeroshot CLIP, demonstrating the importance of our transformations of the CLIP space albeit our proposal shares the same CLIP backbone. Tab. 2 reports the results on UCF-Crime [22]. Our method exhibits the best discrimination of the anomalous classes, achieving the highest mAUC ROC among baselines. Similar to ShanghaiTech, it also achieves an improvement in terms of mAUC against zeroshot CLIP, verifying the importance of our proposed adaptation of the CLIP space. Compared to ActionCLIP [27], our AnomalyCLIP obtains $+2.94\%$ in terms of mAUC, highlighting the need for a MIL framework to mitigate mis-assignment of anomalous class labels to normal frames of anomalous videos. It is also worth noting that the higher mAUC obtained by ActionCLIP does not result in a competitive AUC ROC on VAD, which implicates a worse separation between normal and abnormal frames. When compared to the best performing method SSRL [8] on VAD, our method obtains an improvement of $+1.78\%$ in terms of mAUC on VAR.

**Ablation** of our main design choices with UCF-Crime:

*i) Representation and learning of the directions.* We evaluate the choice of prompt learning with CoOp. As shown in Tab. 3, when CoOp is removed, we learn the directions from randomly initialised points in the CLIP space (Row 1) or engineered prompts of the form *"a video from a CCTV camera of a* $\{class\}$*"* (Row 2). Both choices lead to worse results, indicating that text-guided initialisation and directions finetuning are necessary. Furthermore, we show that unfreezing the last projection of the text encoder (Row 4) enables a greater freedom in finetuning the discovered directions, yielding the best results.

*ii) CLIP latent space for likelihood estimation in Selector model.* How the extracted CLIP features are transformed and the likelihood estimation play a crucial role in the MIL segment selection. As shown in Tab. 4, directly using the CLIP space and cosine similarities with the learned directions (Row 1) produces the worst results, indicating that the use of the normality prototype $\boldsymbol{m}$ is important. Second, Row 2 shows that only using the feature magnitude without accounting for the direction for estimating the likelihood is not as effective, given that the large magnitude could be attributed to directions representing irrelevant factors.

*iii) Temporal model architecture.* We ablate different architectures of $\mathcal{T}$ *i.e.* a 3-layer MLP, a Transformer Encoder [25] and the employed Axial Transformer. As shown in Tab. 5, MLP (Row 1) leads to worst performance, indicating the necessity of considering temporal relationships. Compared to the Transformer, the Axial Transformer, despite having comparable capacity, produces better results. This is possibly due to the fact that it computes the attention separately for the two axes, thus being less prone to overfitting.

Table 3: Ablation on representation and learning of the directions of abnormality.

| **Text encoder** | **Directions** | **AUC** | **mAUC** |
|---|---|---|---|
| No | Direct Optimisation | 84.98 | 69.86 |
| Frozen | Engineered Prompts | 84.66 | 81.35 |
| Frozen | CoOp | 85.88 | 87.39 |
| Finetuning | CoOp | **86.36** | **90.66** |

Table 4: Ablation on CLIP latent space for likelihood estimation in the Selector model.

| **Likelihood** | **Features** | **MIL Selection** | **AUC** | **mAUC** |
|---|---|---|---|---|
| cosine sim. | CLIP | cosine sim. | 85.59 | 83.69 |
| $\mathcal{S}$ | CLIP - $\boldsymbol{m}$ | feature magnitude | 84.92 | 89.82 |
| $\mathcal{S}$ | CLIP - $\boldsymbol{m}$ | $\mathcal{S}$ | **86.36** | **90.66** |

Table 5: Ablation on Temporal model architectures.

| **Temporal Model** | **AUC** | **mAUC** |
|---|---|---|
| MLP | 74.86 | 84.46 |
| Transformer | 84.69 | 88.38 |
| Axial Transformer | **86.36** | **90.66** |

## 4. Conclusions

We proposed $AnomalyCLIP$, the first method that leverages LLV models to address the challenging task of VAR, which extends the scope of VAD by further requiring the classification of the anomalous activities. Our work shed light on the fact that a naive application of existing LLV models [17, 27] to VAR leads to unsatisfactory performance and we demonstrated that several technical design choices are required to build a multi-modal deep network for detecting and classifying abnormal behaviours. Our extensive evaluation showed that $AnomalyCLIP$ scores the new state-or-the-art VAR on ShanghaiTech [10] and UCF-Crime [22]. As future work, we plan to extend our method in open-set scenarios where anomalies are not pre-defined.

## References


[1] 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 *ACM Multimedia*, 2022. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[3] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. *arXiv*, 2022. 4

[4] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, 2021. 3, 4

[5] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. In *ACM Multimedia*, 2021. 1

[6] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv*, 2019. 2

[7] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 2

[8] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *ECCV*. Springer, 2022. 1, 3, 4

[9] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI*, 2022. 1, 3, 4

[10] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *CVPR*, 2018. 1, 3, 4

[11] 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*, 2021. 1, 3

[12] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 2

[13] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, 2021. 1, 3

[14] Tao Mei and Cha Zhang. Deep learning for intelligent video analysis. In *ACM Multimedia*, 2017. 1

[15] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106, 2021. 1

[16] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, 2020. 1, 3

[17] 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*, 2021. 1, 2, 3, 4

[18] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv*, 2021. 1

[19] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 1

[20] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *ICPR*, 2018. 4

[21] Jessie James P Suarez and Prospero C Naval Jr. A survey on deep learning techniques for video anomaly detection. *arXiv*, 2020. 1

[22] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 1, 2, 3, 4

[23] Che Sun, Yunde Jia, and Yuwei Wu. Evidential reasoning for video anomaly detection. In *ACM Multimedia*, 2022. 1

[24] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *ICCV*, 2021. 1, 3, 4

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2, 4

[26] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *ICCV*, 2019. 4

[27] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv*, 2021. 1, 3, 4

[28] Ziming Wang, Yuexian Zou, and Zeming Zhang. Cluster attention contrast for video anomaly detection. In *ACM Multimedia*, 2020. 1

[29] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *ECCV*, 2022. 1, 3, 4

[30] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 2021. 1, 3, 4

[31] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, 2020. 4

[32] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021. 1

[33] Ke Xu, Tanfeng Sun, and Xinghao Jiang. Video anomaly detection and localization based on an adaptive intra-frame classification network. *IEEE Transactions on Multimedia*, 2019. 1

[34] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *ECCV*, 2020. 3, 4

[35] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *CVPR*, 2022. 3, 4

[36] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *ICIP*, 2019. 3, 4

[37] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, 2019. 3, 4

[38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022. 2