

VLAVAD: Vision-Language Models Assisted Unsupervised Video Anomaly Detection

Changkang Li
lichangkang@buaa.edu.cn

The School of Electrical and
Information Engineering,
Beihang University,
Beijing 100191, China
Institute of Unmanned System,.
Beihang University,
Beijing 100191, China

Yalong Jiang
jiangyalong@buaa.edu.cn

The following is the supplementary appendix of this paper.

- Section **A** displays typical normal and abnormal samples from the UCSD Ped2, CUHK Avenue, and ShanghaiTech datasets used in the experiments of this study.
- Section **B** shows the algorithm's performance on the ShanghaiTech dataset is demonstrated. The graph illustrates abnormal events such as fighting and falling in the middle time frames of the video."
- Section **C** offers a thorough investigation of the algorithmic intricacy that is inherent in our approach. This comprises an in-depth analysis of the parameter and computational complexities that are linked to mainstream Vision Language Models and their quantized variants.
- Section **D** undertakes an extensive exploration of the constraints of our proposed methodology, drawing attention to areas in need of improvement and providing insights into promising avenues for future research initiatives.

A Datasets Overview

A.1 UCSD Ped2

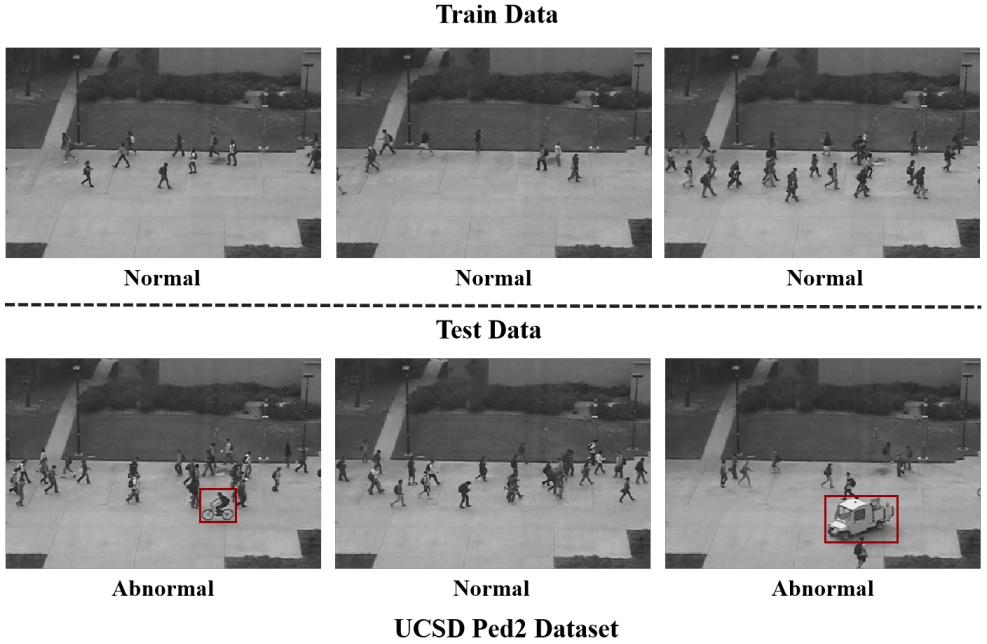


Figure 1: This presents the UCSD Ped2 dataset, with one scenario dedicated to both the training and testing sets. The training set exclusively captures normal pedestrian walking behavior, while the testing set encompasses abnormal activities like cycling and driving.

A.2 CUHK Avenue

Train Data



Normal



Normal



Normal

Test Data



Abnormal



Abnormal



Normal

Avenue Dataset

Figure 2: This is a presentation of the Avenue dataset, where both the training and testing sets consist of scenes captured exclusively at a subway entrance. The training set encompasses normal behaviors such as pedestrian flow, standing, and sitting at the subway station, while the testing set includes abnormal activities like dancing, cycling, and throwing objects.

A.3 ShanghaiTech



Figure 3: Samples from the ShanghaiTech dataset, which featuring videos from 13 distinct scenes. The training set captures typical behaviors like crowd movement, standing, and sitting from diverse perspectives. In contrast, the testing set includes abnormal activities such as car movements, baby stroller usage, skateboarding, cycling, and altercations.

B Algorithm Performance

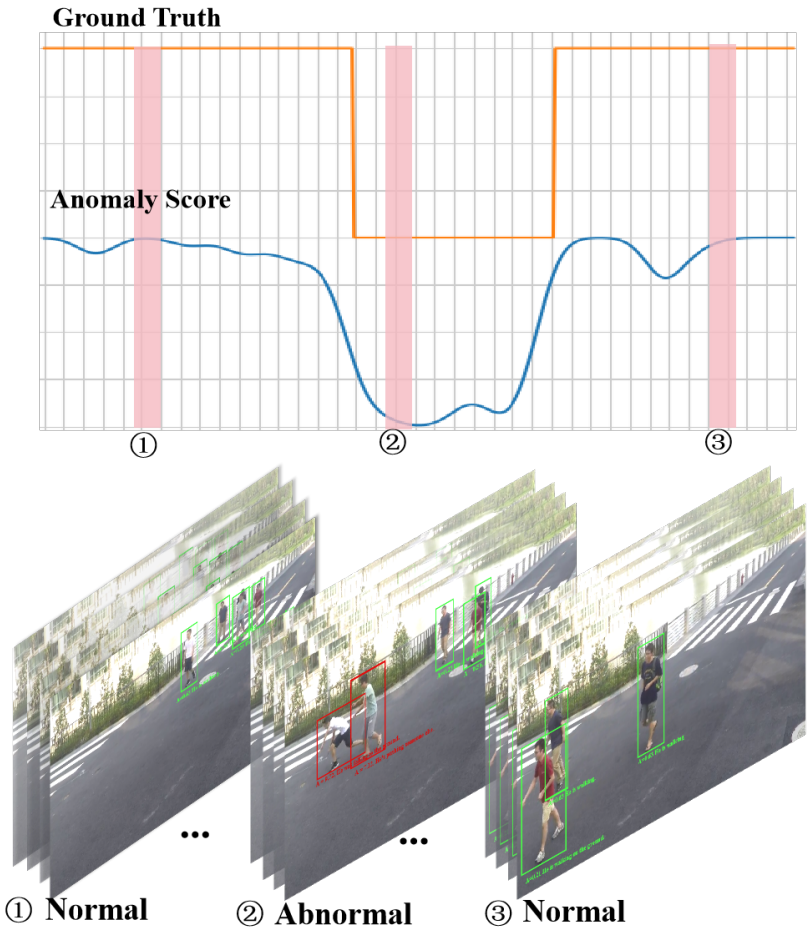


Figure 4: Figure 4 depicts the outcomes of anomaly detection for a video segment from Scene 07 of the ShanghaiTech test dataset. The specific period when the two individuals engage in combat is recognized as an anomalous occurrence. The two individuals involved in the altercation are indicated by red bounding boxes, while the others who are acting normally are denoted by green boxes. The anomaly score notably increases during the altercation, effectively identifying the anomaly.

Models	Vision Models	Language Models	Total Params
BLIP-2	ViT-L (304M)	OPT-2.7B	3.1B
	ViT-g (1011M)	OPT-2.7B	3.8B
	ViT-g (1011M)	OPT-6.7B	7.8B
	ViT-L (304M)	FlanT5 _{XL} - 3B	3.4B
	ViT-g (1011M)	FlanT5 _{XXL} - 11B	12.1B

Table 1: Comparison of Model Parameters.

C Model computational complexity analysis

The recent surge in the popularity of Large Language Models (LLMs) has led to a significant increase in both the parameter count and the computational requirements of deep learning models. While this growth has placed a significant burden on hardware computation and increased the demand for advanced hardware devices, it has also enhanced the representational capacity of these models, which is why LLMs exhibit exceptional generalization abilities. Although our utilized Blip-2 model is no longer at the cutting edge of vision-language understanding models, our method still achieved very good performance on three datasets. By leveraging models with larger parameter counts for text feature space mapping, our approach could further improve accuracy in cross-scene scenarios.

It’s worth noting that quantized versions of these models are available, which have smaller parameter counts. The [I](#) below illustrates the parameter counts for each module of Blip-2 as well as other Vision Language Models. This comparison suggests potential for further improvements. As the field of vision-language models continues to advance, we anticipate that integrating more sophisticated models could yield even more impressive results, particularly in challenging cross-domain tasks.

D Limitations analysis

Our current approach has certain limitations that we must acknowledge. Firstly, we rely on object detection for object-level anomaly detection and precise localization, which makes our method sensitive to the accuracy of the detection, potentially overlooking anomalies that are not object-level in nature, such as sudden crowd gatherings. Secondly, we base our anomaly detection on the frequency of text features, which is data-dependent and may be affected if the distribution of features in the training data is diverse.

In light of these limitations, the development of video understanding with large language models such as miniGPT4 or CogVLM may offer new approaches to anomaly detection. These models have the potential to handle multiple objects and temporal relationships within a single model, thereby enhancing our understanding of video content.