

Narula Vivek

Delhi_Technological_University_M_Tech_Thesis_Vivek (2)_removed

Document Details

Submission ID

trn:oid:::27535:140546083

Submission Date

May 26, 2026, 3:30 PM GMT+0

Download Date

May 26, 2026, 3:34 PM GMT+0

File Name

Delhi_Technological_University_M_Tech_Thesis_Vivek (2)_removed.pdf

File Size

3.8 MB

40 Pages

12,968 Words

73,807 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Exclusions

- 3 Excluded Sources

Match Groups

-  **93 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
-  **9 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 6%  Publications
- 5%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 93 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
- 9 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5% Internet sources
- 6% Publications
- 5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	arxiv.org	<1%
2	Internet	www.mdpi.com	<1%
3	Student papers	Delhi Technological University on 2026-05-25	<1%
4	Internet	pure.tue.nl	<1%
5	Internet	irep.mbzuai.ac.ae	<1%
6	Student papers	Babes-Bolyai University on 2025-08-30	<1%
7	Publication	Guanghai Xu, Sifan Long, Hongwei Ge, Enxuan Gu, Zidi Li, Zhaoqin Wang. "Seman...	<1%
8	Student papers	University of Sydney on 2024-03-05	<1%
9	Student papers	Malaviya National Institute of Technology on 2025-08-29	<1%
10	Internet	pdffox.com	<1%

11	Student papers	Indian Institute of Technology, Bombay on 2026-02-18	<1%
12	Publication	Zhen Yang, Guodong Wang, Yuanfang Guo, Xiuguo Bao, Di Huang. "Anomaly-awa...	<1%
13	Internet	ebin.pub	<1%
14	Student papers	Imperial College of Science, Technology and Medicine on 2025-09-09	<1%
15	Publication	Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huan...	<1%
16	Publication	Shengyang Sun, Xiaojin Gong. "Long-Short Temporal Co-Teaching for Weakly Sup...	<1%
17	Student papers	UCL on 2025-08-28	<1%
18	Internet	ubir.buffalo.edu	<1%
19	Publication	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2022", ...	<1%
20	Student papers	Delhi Technological University on 2026-05-24	<1%
21	Publication	Helei Qiu, Biao Hou. "Temporal context and representative feature learning for w...	<1%
22	Publication	Yang Liu, Jing Liu, Wei Ni, Liang Song. "Abnormal Event Detection with Self-guidin...	<1%
23	Publication	Silas S. L. Pereira, José Everardo Bessa Maia. "MC-MIL: video surveillance anomaly...	<1%
24	Internet	etd.auburn.edu	<1%

25	Internet	vbn.aau.dk	<1%
26	Publication	Han, Mingfei. "Evolving Video Analysis: From Object Perception to Holistic Unders..."	<1%
27	Publication	Xiaojie Cai, Yucheng Qian, Chong Wang, Xiaohao Peng, Yuanbin Qian, Jiafei Wu. "...	<1%
28	Internet	doczz.net	<1%
29	Internet	gredos.usal.es	<1%
30	Internet	hal.science	<1%
31	Internet	ijeces.ferit.hr	<1%
32	Publication	Debi Prasad Senapati, Santosh Kumar Pani, Santos Kumar Baliarsingh, Prabhu Pr...	<1%
33	Publication	Kian Yu Gan, Yu Tong Cheng, Hung-Khoon Tan, Hui-Fuang Ng, Maylor Karhang Le...	<1%
34	Student papers	Leeds Beckett University on 2024-05-15	<1%
35	Student papers	Loughborough University on 2023-05-04	<1%
36	Publication	Peng Wu, Jing Liu, Mingming Li, Yujia Sun, Fang Shen. "Fast Sparse Coding Netwo..."	<1%
37	Internet	export.arxiv.org	<1%
38	Publication	"Communications and Networking", Springer Science and Business Media LLC, 20...	<1%

39	Publication	"Computer Vision – ECCV 2018", Springer Nature America, Inc, 2018	<1%
40	Student papers	University of Hong Kong on 2023-08-21	<1%
41	Student papers	University of Wollongong on 2017-08-25	<1%
42	Publication	Zhiwei Yang, Jing Liu, Guansong Pang, Peng Wu, Zhaoyang Wu. "Attention-driven ...	<1%
43	Internet	ualberta.scholaris.ca	<1%
44	Internet	zagan.unizar.es	<1%
45	Publication	"Computer Vision – ACCV 2024", Springer Science and Business Media LLC, 2025	<1%
46	Student papers	APJ Abdul Kalam Technological University, Thiruvananthapuram on 2024-11-22	<1%
47	Publication	Feiran Liu, Yan Sun, Liyan Ma, Xiangfeng Luo. "ProDe: Prototype-based pattern d...	<1%
48	Student papers	Flinders University on 2024-09-27	<1%
49	Publication	Hoangcong Le, Cheng-Kai Lu, Chen-Chien Hsu. "Video anomaly detection for edge...	<1%
50	Publication	Mouin Ben Ammar, Arturo Mendoza, Nacim Belkhir, Antoine Manzanera, Gianni F...	<1%
51	Publication	Qiyue Sun, Yang Yang, Yunxia Liu, Hongjun Wang. "RAG4VAD: A training-free retr...	<1%
52	Publication	Snehashis Majhi, Srijan Das, Francois Bremond. "DAM: Dissimilarity Attention Mo...	<1%

53	Student papers	University of Sydney on 2022-07-12	<1%
54	Student papers	University of Technology, Sydney on 2024-05-30	<1%
55	Publication	Yan Fu, Ting Hou, Ou Ye, Gaolin Ye. "A video anomaly detection and classification ...	<1%
56	Publication	Yufan Hu, Jie Fu, Mengyuan Chen, Junyu Gao, Jianfeng Dong, Bin Fan, Hongmin Li...	<1%
57	Internet	iris.imtlucca.it	<1%
58	Internet	theses.hal.science	<1%
59	Internet	webthesis.biblio.polito.it	<1%
60	Internet	www.ecva.net	<1%
61	Publication	"Computer Vision – ECCV 2020", Springer Science and Business Media LLC, 2020	<1%
62	Publication	Chen Yang, Xue-Song Tang, Kuangrong Hao, Yubing Li, Zhiqi Wang. "Explicit Kno...	<1%
63	Publication	Prabhu Prasad Dev, Pranesh Das, Raju Hazari. "Cross-modal uncertainty modelin...	<1%
64	Publication	Yaxin Yang, Chao Han, Jun Wan. "A short video anomaly detection benchmark wit...	<1%
65	Publication	Shalmiya Paulraj, Subramaniaswamy Vairavasundaram. " : Multiview multimod...	<1%
66	Student papers	University of Technology, Sydney on 2026-05-17	<1%

67

Publication

Ying Chen, Lianming Zhang, Jiusheng Li, Pingping Dong. "Hyperbolic graph repre..."

<1%

HIERARCHICAL PROGRESSIVE REFINEMENT FOR WEAKLY SUPERVISED VIDEO ANOMALY DETECTION

Narula Vivek

ABSTRACT

Weakly Supervised Video Anomaly Detection (WSVAD) is a technique that uses coarse video-level labels to find transient anomalous actions. Recent architectures are more likely to use heavy spatial-temporal graph networks or large Vision-Language Models (VLMs) to give semantic reasoning and address inherent pseudo-label noise. Nevertheless, constructing joint-embedding spaces directly between text and video modalities is computationally costly and often results in overfitting. More to the point, these methods are trying to learn the visual signature of anomalies, which is a false assumption to make since abnormal events are unlimited and heterogeneous in nature; normalcy, though, is circumscribed and extremely repetitive.

To address this, we present a hierarchical progressive refinement framework for weakly supervised video anomaly detection.

In the first stage, we present a lightweight Text-Conditioned Multiple Instance Learning (TC-MIL) model. Instead of enforcing a more complicated contrastive alignment, TC-MIL deploys semantic language as a literal gating mechanism[cite: 1]. We obtain motion-sensitive visual representations with a pre-trained I3D backbone and produce textual representations with a frozen CLIP encoder. The projection and multiplication of the text embedding of the temporal sequence results in the mathematical performance of the language as a semantic filter enhancing the use of visual channels of interest to the desired anomaly and actively suppressing non-harmful background noise.

In the second stage, we present CORE-MIL (Compact Orthogonal Residual Extraction), a simple, mathematically based framework that reverses this paradigm. In place of attempting to determine chaos, CORE-MIL explicitly measures the inability to rebuild order. We inculcate a small, learnable Normalcy Anchor Matrix to learn common background behaviors. This subspace receives the incoming video features and the reconstructed normal component is subtracted from the raw input. This gives an orthogonal (residual) which is flat to daily activities but steep in case of unforeseen events.

Extensive testing on the UCF-Crime and XD-Violence datasets demonstrates this framework delivers highly competitive, interpretable localization accuracy. It reduces computational costs by aggressively screening background noise before classification and avoiding cross-modal alignment, doing so while slashing computational overhead and training time by an order of magnitude.

LIST OF PUBLICATIONS

Conference Proceedings

- 9 1. **Narula Vivek**, and **Kavinder Singh**, 2026. TC-MIL: Text-Conditioned Multiple Instance Learning for Video Anomaly Detection. In: *Proceedings of the 2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*, Manila, Philippines, June 11-14, 2026. (Registered for Oral Presentation, Paper ID: 1571282709).
- 1 2. **Narula Vivek**, and **Kavinder Singh**, 2026. Subtracting the Norm: An Orthogonal Projection Framework for Weakly Supervised Video Anomaly Detection. In: *Proceedings of the 2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*, Manila, Philippines, June 11-14, 2026. (Registered for Oral Presentation, Paper ID: 1571283563).

TABLE OF CONTENTS

Acknowledgements	ii
Candidate's Declaration	iii
Certificate by the Supervisor(s)	iv
Abstract	v
List of Publications	vi
List of Tables	ix
List of Figures	x
List of Symbols, Abbreviations and Nomenclature	xi
1 Introduction	1
1.1 Background and Motivation	1
1.2 The Multiple Instance Learning (MIL) Formulation and its Limitations	2
1.3 Problem Statement: The Modality Gap and Computational Bloat . . .	2
1.4 Proposed Hierarchical Framework	3
1.5 Objectives and Contributions of the Thesis	4
1.6 Organization of the Thesis	5
2 LITERATURE REVIEW	6
2.1 From Unsupervised Learning to Weakly Supervised Paradigms	6
2.2 The Multiple Instance Learning (MIL) Formulation	7
2.3 Label Noise and Feature Ambiguity	8
2.4 Memory-Augmented and Statistical Normality Modeling	9
2.5 Multi-Modal Fusion and Non-Euclidean Geometries	10
2.6 Vision-Language Models in Video Understanding	11
2.7 Overcoming Overhead via Hierarchical Refinement	11
2.8 Datasets and Evaluation Protocols in WS-VAD	12
2.9 Optimization Landscapes and MIL Loss Formulations	13
2.10 The Computational Bottleneck and Edge Deployment	14
3 METHODOLOGY	16
3.1 Proposed System Architecture	16

3.1.1	Problem Formulation	16
3.2	Multi-Scale Feature Aggregation	17
3.2.1	Visual Stream: Capturing Motion Dynamics	17
3.2.2	Semantic Stream: Semantic Context Generation	17
3.3	Scene Decomposition and Refinement	18
3.3.1	Stage 1: High-Level Semantic Gating Mechanism	18
3.3.2	Stage 2: Low-Level Orthogonal Normalcy Projection	18
3.4	Optimized Discriminative Learning	19
4	RESULTS AND DISCUSSION	21
4.1	Datasets and Evaluation Metrics	21
4.2	Implementation Details	22
4.3	Comparison with State-of-the-Art	22
4.4	Ablation Studies and Efficiency	23
4.4.1	Impact of Semantic Gating (Stage 1)	23
4.4.2	The Bottleneck Dimension (Stage 2)	23
4.4.3	Computational Efficiency	24
4.5	Qualitative Evaluation and Interpretability	24
4.5.1	Anomaly Scores and Residual Spikes Over Time	24
4.6	Hard-Negative Analysis (Failure Cases)	26
5	CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT	27
5.1	Conclusion	27
5.2	Future Scope	27
5.3	Social Impact	28
	References	32
	A Proof of Publication	33

List of Tables

3.1	Key Mathematical Notations used in the Hierarchical Framework . . .	18
4.1	Comparison with State-of-the-Art on UCF-Crime and XD-Violence. Parameter counts denoted with ($\sim$) are architectural estimates derived from respective literature.	22
4.2	Ablation Study using ROC-AUC demonstrating the impact of Semantic Gating	23
4.3	Ablation on Dictionary Size of Normalcy Matrix (k) on UCF-Crime and XD-Violence.	23

List of Figures

- 3.1 System architecture of the Stage 1 Semantic Gating Framework. Visual features from a frozen I3D backbone are modulated by a semantic gate derived from a projected CLIP text embedding. 17
- 3.2 The overall pipeline of the Stage 2 CORE-MIL framework. Features are projected onto a learned Normalcy Anchor Matrix to reconstruct the normal background component, isolating the Orthogonal Residual. 20
- 4.1 Extensive ablation study illustrating the impact of the bottleneck dimension (k) on both UCF-Crime (AUC) and XD-Violence (AP). A bottleneck of $k = 20$ provides the strict mathematical bound needed to prevent overfitting while capturing normal variance. 24
- 4.2 Qualitative comparison of anomaly scores over time for UCF-Crime (top) and XD-Violence (bottom). The baseline MIL struggles with false positives generated by harmless background movement. In contrast, our proposed framework effectively flattens the noise profile. . . 25
- 4.3 Interpretability on the highly diverse XD-Violence dataset: Magnitude of the Orthogonal Residual ($\|X_{res}\|$) and predicted Anomaly Score over time for the anomalous video *Black Hawk Down*. The residual successfully suppresses complex movie background dynamics, spiking dramatically precisely when the unforeseen anomalous action occurs. . . 25

LIST OF SYMBOLS

V	Untrimmed video (represented as a bag of instances)
v_t	Individual non-overlapping video snippet
T	Total number of temporal video snippets
Y	Video-level binary label ($Y \in \{0, 1\}$)
s_t	Predicted snippet-level anomaly score
X	Raw spatial-temporal visual feature matrix
d	Visual feature dimension
t	Dense semantic text embedding from CLIP
$\phi(\cdot)$	Linear projection layer for text embeddings
$\odot$	Hadamard product (element-wise multiplication)
$\tilde{X}$	Semantically gated visual feature matrix
M_n	Learnable Normalcy Anchor Matrix
k	Bottleneck dimension of the normal subspace
$\hat{X}$	Reconstructed normal component of the visual features
X_{res}	Orthogonal residual matrix (isolated anomalous features)
I	Identity matrix
$\mathcal{B}_a$	Batch of positive (anomalous) video bags
$\mathcal{B}_n$	Batch of negative (normal) video bags
$\mathcal{L}_{MIL}$	Multiple Instance Learning ranking loss
$\mathcal{L}_{ortho}$	Structural orthogonality penalty
λ	Hyper-parameter balancing the orthogonal projection constraint
$\mathcal{L}_{smooth}$	Temporal consistency constraint (smoothness prior)
$\ \cdot\ _F$	Frobenius norm
$\sigma(\cdot)$	Sigmoid activation function

24

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

With the rapid expansion of smart city infrastructure and continuous closed-circuit television (CCTV) surveillance, thousands of hours of video are generated daily. Monitoring this manually is practically impossible, driving the urgent need for automated intelligent systems. Video anomaly detection is a widely-known challenge to scale. Historically, due to the scarcity of training anomalies, the task was originally framed as an unsupervised outlier detection or one-class classification problem [1]. These approaches utilised auto-encoders or dynamic texture models to learn representations of normal behaviour, operating under the assumption that anomalies would naturally yield large reconstruction errors. However, lacking prior information regarding abnormal events, these methods consistently struggle to differentiate true anomalies from unseen, yet completely harmless, normal behaviours, heavily depending on manually curated normal training data.

It may sound easy to spot irregular activities, but it in reality takes huge, annotated datasets to establish a reliable system, which is impossible without such data, even though it is simple to spot anomalies in datasets of this size, as demonstrated by the neural network example [1]. The point is, real-life cameras capture a massive amount of uncontrolled background movement, and it is simply too expensive to pay human annotators to draw frame-level bounding boxes around rare events. To avoid this, the research community instinctively turned to **Weakly Supervised Video Anomaly Detection (WSVAD)**. With this configuration, models must determine the exact temporal location of an anomaly with just one, coarse label applied to the whole video [2]. The shift toward WS-VAD was brought about by the introduction of extensive, untrimmed video datasets that provide coarse video-level labels for training and fine-grained frame-level annotations for testing. Pioneering datasets such as UCF-Crime moved the field away from staged actor datasets toward complex, real-world surveillance footage, covering distinct anomaly categories alongside normal activities.

63

4

1.2 The Multiple Instance Learning (MIL) Formulation and its Limitations

In practice, WSVAD can be regarded as a Multiple Instance Learning (MIL) problem as stated in [2]. In this paradigm, the raw video becomes an “bag” with a number of smaller temporal fragments as instances. An ordinary bag of negative videos is a normal video, that is, all the snippets in it are guaranteed to be normal. On the other hand, an anomalous bag is a positive video that includes a mix of the normal background videos and at least one anomalous snippet somewhere inside the video. The main goal of a WS-VAD model is to learn a mapping function that can output a frame-level anomaly score based on a ranking loss metric. At the macro level, this binary classification puts the responsibility of localization directly on the learning algorithm, instead of on human annotators.

This structure can save vast amounts of annotation time, but results in a serious pseudo-label noise problem [3, 4, 5]. The fact that a surveillance video is inherently out of balance in favor of the abnormal makes it very probable that a given video is actually a very large proportion of normal frames. What you’re saying here is basically “Find the needle in the haystack” where the hay is a stack of regular frames that are tagged with no more than a video-level tag (which often falsely pushes regular background motion into the abnormal feature space). For example, a video that is supposed to be classified as “anomalous” may have just five seconds of a violent occurrence and the rest of the footage is completely normal pedestrian activity. If there is no temporal boundary, models can easily fall into the trap of concept drift, where they start to memorise normal (noisy) backgrounds and incorrectly identify them as anomalies during inference.

1.3 Problem Statement: The Modality Gap and Computational Bloat

To pierce this noise, new architectures have added more and more computational power to the issue. We observe the use of large spatial-temporal networks of graphs [6, 7], dual-memory attention systems to separate prototypes, and complex statistical checks relying on batch normalization divergence to weed out noisy data. While effective in research environments, dense joint-embedding spaces or attention estimation on thousands of temporal snippets require huge amounts of resources and often overfit with weak supervision.

As the field expanded, the realization that visual features alone are sometimes insufficient prompted the development of multi-modal architectures. The existence of real-world datasets such as XD-Violence has led to the development of strong audio-visual alignment systems. However, researchers immediately encountered the “modality gap”. The two modalities, i.e., audio and video, are fundamentally different: audio is quick and more susceptible to background noise, while video is consistent and detailed but can be easily occluded by objects in the scene. Furthermore, to handle the

complex boundaries between rare anomalous events and ubiquitous normal data, recent models have even mapped features into non-Euclidean hyperbolic spaces. In such representations, normal data is ubiquitous and immensely varied (and forms the base of the data), whereas abnormal data is uncommon and hang out on the branches.

51 Recently, the field has seen a massive shift toward using Large Vision-Language Models (VLMs) such as CLIP to semantically ground the visual data using natural language [8, 9]. The core concept is to provide the visual data with real semantic background using descriptive text prompts. However, there is one big snag. Normal VLMs are intrinsically designed to handle static images; they process frames individually, which effectively transforms a flowing video stream into a sequence of immobile images. Once you tear out the continuum of time, the model becomes totally blind to the movement which is in fact occurring in a violent event. On the other end, the traditional networks such as I3D [10] and C3D are designed to specifically follow that temporal flow, but they work in an absolute semantic vacuum. They are able to follow moving pixels flawlessly, but since they do not have language to contextualize which ones they are tracking, they do not actually know what they are looking at.

49 More to the point, these heavy architectures are based on a fundamentally flawed assumption: they are trying to learn the visual or semantic signature of an anomaly. Abnormal events are intrinsically unlimited. An arson, a street fight [11], a robbery, and a traffic accident have virtually no visual or time intersection. Direct mapping of this high-entropy space through cross-modal joint-embedding is computationally expensive beyond belief and often results in overfitting under weak supervision. Furthermore, this brute-force approach of scaling up model parameters creates a massive barrier to real-world deployment. Modern surveillance systems require rapid, real-time processing on resource-constrained edge devices, not just in massive, power-hungry server farms. However, normalcy is constrained, low-entropy and very repetitive. Therefore, the core problem is not just how to find the anomaly, but how to do so efficiently without attempting to learn an infinite dictionary of abnormal events.

1.4 Proposed Hierarchical Framework

To bring these modalities into coincidence without the heavy contrastive goal, we present a hierarchical progressive refinement framework. Rather than enforcing a network to explicitly point to high-entropy chaos, we mathematically remove the background in sequential stages. This means we flip the traditional paradigm on its head: if we can confidently strip away the mundane, routine activities that make up the vast majority of surveillance footage, whatever remains is naturally exposed as an anomaly.

52 In the first stage, we constructed a Text-Conditioned Multiple Instance Learning (TC-MIL) framework to fill the semantic gap. The idea here is straightforward: a 3D CNN with the temporal accuracy of 3D CNN will use the semantic brains of a VLM, without the need to perform cross-modal contrastive training overhead. To achieve this, the raw spatio-temporal features are pulled using a pre-trained I3D backbone, and a frozen CLIP encoder generates a semantic embedding of a simple anomaly caption. As we extract both feature sets offline, the actual training footprint remains extremely small. Instead of using the text as an independent target that the video must be an exact

match of, we rely on the language embedding as a direct conditioning gate. We simply project the text embedding, broadcast it throughout the timeline, and do a simple element-wise multiplication with the visual features [8]. It essentially reduces the language to a semantic filter, proactively enhancing the visual channels associated with the target anomaly—such as tracking fire in the event of an explosion—and actively suppressing the channels that pick up random clutter of the background.

In the second stage, we present Compact Orthogonal Residual Extraction (CORE-MIL) to take advantage of the fact that normality is highly repetitive. We incorporate a learnable, very small Normalcy Anchor Matrix, within the network, to serve as a memory bank and reference point of normal motion. When the semantically gated spatial-temporal features obtained using conventional 3D CNNs [10, 12] are fed through the network, they are projected onto this normal subspace. And then we compute the orthogonal residual by removing this reconstructed normal component out of the original feature. When a snippet has daily background motion, the Normalcy Matrix is able to reconstruct the scene flawlessly, and the residual is brought down to nearly zero. Should an abrupt anomaly be present, the Normalcy Matrix is unable to re-create the high-entropy action, and the residual spontaneously spikes.

After filtering the features both semantically and mathematically, we just feed this mathematically filtered residual, which is completely isolated from the background noise, to a standard MIL lightweight classifier. Because the heavy lifting of cleaning up the inputs has already been done, we do not need a complex custom loss function. By optimizing a standard MIL ranking loss, the network essentially learns to balance these competing objectives: it trains itself to push the anomaly scores of abnormal frames toward 1, while keeping the scores for normal frames pinned down near 0.

1.5 Objectives and Contributions of the Thesis

We have the following major contributions:

- **Hierarchical Progressive Refinement Framework:** We assemble a light architecture that directly injects the open-world reasoning of Vision-Language Models into the temporal stream of 3D CNNs. This refocuses the WSVAD paradigm not on trying to classify unbounded anomalies, but on directly quantifying the inability to recreate bounded normalcy.
- **Semantic Gating Mechanism:** We do not use cross-modal standard alignment, but directly utilize text embeddings as a direct conditioning gate. The anomalous cues are intuitively salient when we multiply the features of the language along the visual timeline, flattening the background noises that occur in everyday life before the classification stage even initiates.
- **Orthogonal Normalcy Projection:** We propose a learnable projection that literally removes the remaining background motion with an efficient linear algebra operation. This physically executes feature decorrelation without involving computationally expensive cross-modal fusion, intricate attention maps, or non-Euclidean hyperbolic geometries.

- **Interpretability and Computational Efficiency:** We have cut training time and computational cost by taking the heavy feature extraction out of the primary training loop. This framework enables the use of a highly lightweight classifier to obtain competitive localization performance with respect to much heavier baselines, and ensures that the mathematical decision-making process of the network is fully transparent.

1.6 Organization of the Thesis

The rest of this thesis has been divided in a logical order to make it easy for the reader to follow the development and application of the proposed framework:

- Chapter 2 (Literature Review) reviews the current state-of-the-art in Weakly Supervised Video Anomaly Detection. It covers the history of the field from the traditional unsupervised learning and basic MIL approaches to recent breakthroughs in Vision-Language integration, statistical memory models and fusion of multiple modalities.
- The mathematical and structural formulations of the proposed hierarchical progressive refinement framework are explained in detail in Chapter 3 (Methodology). It explains the offline feature extraction process, the mathematics of the Stage 1 Semantic Gating Mechanism, and the linear algebra of the Stage 2 Orthogonal Normalcy Projection layer.
- In Chapter 4 (Results and Discussion), the experimental setup is described and a thorough analysis of the framework is conducted on the challenging UCF-Crime and XD-Violence benchmark datasets. This comprises a detailed ablation analysis of the bottleneck dimension, efficiency analysis of the computation, and qualitative visualisations to confirm the interpretability of the extracted anomalous residuals.
- The main conclusions of the thesis, the wider social impacts of democratising lightweight intelligent surveillance and suggestions for future research are summarised in Chapter 5 (Conclusion, Future Scope and Social Impact). Future research directions proposed include real-time edge streaming and prompt generation for explainable AI.

CHAPTER 2

LITERATURE REVIEW

2.1 From Unsupervised Learning to Weakly Supervised Paradigms

Traditionally, the concept of video anomaly detection was presented as an unsupervised outlier detection task or one-class classification task [1, 13, 14, 15, 16]. By default, video anomaly detection was an unsupervised learning problem where only normal patterns would be trained on and models would be expected to indicate spatial or temporal abnormalities [1]. The core mathematical premise here was that a model trained exclusively on standard patterns would inherently yield a high reconstruction error when fed an abnormal frame. For instance, early Sparse Coding approaches attempted to build a dictionary of normal events, reconstructing test frames as a linear combination of these dictionary atoms.

Historically, due to the scarcity of training anomalies, these approaches utilized auto-encoders or dynamic texture models to learn representations of normal behavior, assuming anomalies would naturally yield large reconstruction errors. However, lacking prior information regarding abnormal events, these methods struggle to differentiate true anomalies from unseen normal behaviors and heavily depend on manually curated normal training data. When a network is only trained to recreate a baseline, it lacks the discriminative boundaries required to understand what makes a deviation truly malicious versus just naturally uncommon.

The point is that real-life cameras capture much uncontrolled background movement, which is totally harmless. Since the models had no examples of real threats, they were always left confused as to whether the peculiarities were harmless or real. It became apparent that establishing a perfect, universally applicable threshold for reconstruction error is fundamentally impossible in open-world scenarios, as lighting changes, weather conditions, or simple camera jitter would routinely trigger false alarms.

To overcome the limitations of unsupervised learning while avoiding the prohibitive costs of frame-level annotations, the field shifted toward Weakly Supervised Video Anomaly Detection (WS-VAD). The need to handle big, un-cut surveillance data at a cost that was prohibitive by frame-level labels introduced the popularity of Weakly Supervised Video Anomaly Detection (WSVAD) [2, 17]. The progress in Weakly Su-

pervised Video Anomaly Detection (WS-VAD) is directly linked to the development of large-scale, real-world datasets. Prior to these, benchmarks like the UCSD Pedestrian or UMN datasets were incredibly constrained, featuring static cameras and predictable, homogeneous backgrounds.

The shift toward WS-VAD was brought about by the introduction of extensive, untrimmed video datasets that provide coarse video-level labels for training and fine-grained frame-level annotations for testing. Pioneering datasets such as UCF-Crime [2] moved the field away from staged actor datasets toward complex, real-world surveillance footage, comprising 1,900 long, untrimmed videos covering 13 distinct anomaly categories. This dataset fundamentally shifted the paradigm by introducing complex camera movements, varying illumination, and a diverse range of anomaly classes ranging from subtle shoplifting to explosive arson. Similarly, the XD-Violence dataset [11] introduced 4,754 untrimmed videos to address the lack of audio-visual benchmarks, pushing the community to finally reckon with the multi-sensory nature of true anomalous events.

2.2 The Multiple Instance Learning (MIL) Formulation

This problem is traditionally stated as a Multiple Instance Learning (MIL) problem [2, 18, 19]. The Sultani et al. paper on the subject simply established the norm in this area by presenting the entire process as a Multiple Instance Learning (MIL) problem. You simply cut the video into smaller, non-overlapping segments, and the network uses a ranking loss to rank them with just that single video-level label.

In the standard WS-VAD paradigm, anomaly detection is formalized as an MIL problem where only video-level annotations are available during training. A video sequence V is chopped up into T uniform, non-overlapping snippets, denoted as $\mathcal{V} = \{s_1, s_2, \dots, s_T\}$. We then run each snippet s_t through a pre-trained feature extractor like I3D [10] or C3D [12] which transforms it into a feature vector $x_t \in R^d$. This effectively turns the entire video into what we call a "bag of instances," represented as $X = \{x_1, x_2, \dots, x_T\}$. In this framework, each bag is assigned a coarse video-level label $Y \in \{0, 1\}$. An ordinary bag of negative videos is a normal video, and an anomalous bag with a combination of normal background videos and at least one anomalous snippet is an abnormal video. This mathematical abstraction is brilliant in its simplicity, but it inherently treats the video as a permutation-invariant set rather than a flowing temporal sequence, which forces the architecture to figure out the timeline entirely from scratch.

In order to process the visual data, pre-trained 3D convolutional networks which are normally I3D [10] and sometimes C3D [12] are usually used to do the heavy lifting of the extraction of spatio-temporal features. These networks, pre-trained on massive action recognition datasets like Kinetics-400, provide a robust initial mapping of pixel motion to high-dimensional vectors. The primary objective of a WS-VAD model is to learn a mapping function capable of generating a frame-level anomaly score by optimizing a ranking loss. Formally, the MIL ranking loss operates on the assumption

that the maximum anomaly score within a positive bag must be strictly greater than the maximum anomaly score in a negative bag. Essentially, it pushes the model to ensure that the highest anomaly score in a positive bag is noticeably higher than in a negative bag. However, the standard max-pooling approach used in early MIL models is notoriously brittle. Since only the snippet with the highest score contributes to the gradient during backpropagation, the model essentially ignores the rest of the video, leading to severe under-exploration of the feature space.

2.3 Label Noise and Feature Ambiguity

Although this MIL framework spares an enormous amount of manual annotation, it results in a serious pseudo-label noise issue. The positive bag automatically has an overwhelming majority of normal background frames, which puts the classifier to grapple with feature ambiguity [5, 20]. This phenomenon, often referred to as concept drift or label contamination, is the Achilles' heel of standard MIL. The fact of the matter is that even an abnormal video consists of 99% of normal background video, which is utterly confusing to the classifier during training.

Since the percentage of frames in a positive bag that is not harmless is negligible, the classifier is continually required to locate a needle in a hay stack, which frequently incorrectly places normal background motion into the abnormal feature space. When a network is constantly penalized for misclassifying a positive bag, but the gradient updates are driven by normal frames that happen to have transiently high scores, the decision boundary becomes highly distorted. The network inadvertently learns to memorize noisy backgrounds—like the swaying of trees or the flashing of a neon sign—and falsely reports them as anomalies during inference.

In an attempt to overcome the noise caused by weak labels, much attention has been paid by researchers to feature separation and normalcy suppression. To discard that noise and draw a hard line between what is normal and what is not, more recent work has relied on tricks such as temporal feature magnitude learning (RTFM) [3], self-training pipelines (MIST) [4] or employing clustering to suppress the everyday frames (CLAWS) [5]. RTFM, for example, hypothesizes that the L_2 norm (magnitude) of feature vectors can serve as a proxy for anomaly separability. Architectures such as RTFM [3] and SAS [21] rely on the magnitude of temporal features and anomalous snippet attention to impose boundaries between classes, and CLAWS [5] and GCL [22] repress everyday frames using clustering and cooperative learning. CLAWS approaches the issue from an unsupervised clustering perspective, grouping snippets into latent clusters and assigning pseudo-labels based on cluster purity, effectively creating a self-correcting feedback loop that progressively cleans the noisy video-level labels.

Other approaches experiment with feature reconstruction. For example, the Fast Sparse Coding Network (FSCN) extracts Spatial-Temporal Fusion Features (STFF) and trains the network to minimize errors when recreating these features, effectively building a highly reliable dictionary of what "normal" actually looks like. Another angle is Domain Adaptation (DA-VAD), which bridges the gap between broad, video-level features and specific, clip-level details by treating the mismatch as a domain shift to weed out noisy data. By viewing the video-level label as a source domain and the

snippet-level predictions as a target domain, DA-VAD employs adversarial training to align the feature distributions, ensuring that the local representations do not drift too far from the global semantic constraints.

2.4 Memory-Augmented and Statistical Normality Modeling

The opposite approach is to get regularity by augmenting memory and sparse representations [23, 20, 24]. Previous MIL models were focused on modeling anomalies, with a lack of precision in defining what is "normal" behavior, that resulted in high false alarm rates when faced with unseen complex normal behaviors. The reasoning is that if the model has a clear and precise sense of what it considers normal, it does not have to imagine what an anomaly is, but just reject any deviation from the reference.

One example is the UR-DMU [25] system that is installed with two memory units and uncertainty control to prevent mistaken identification of observed normal behaviors in the model. Instead of selecting conventional network architecture, UR-DMU is equipped with a Global and Local Multi-Head Self-Attention module, and two memory banks, one for normal events and one for anomalies, with reference prototypes in each memory bank, so that the two types of events are clearly distinguished. Most of such memory banks are implemented using a read-write addressable cosine similarity memory addressing scheme. Most of the memory banks derive their addressing scheme from cosine similarity and are read-write. In the forward pass, the input feature visits the memory bank, and finds the most similar prototypes to reconstruct the signal. To reduce the reconstruction error a very robust dictionary is developed using encoding-decoding neural networks by FSCN [24]. What makes explicit memory modules hard to use is that they have fixed-sized memory (thus requiring storage space for a large number of memory modules) and that they need to be continuously refreshed during the training process of thousands of memory elements using moving averages or gradient descent for each one (which introduces a lot of computational overhead).

Likewise, BN-WVAD [26] detects outliers by exploiting the inherent statistical difference between Batch Normalization layers. This obviates the necessity of explicit memory banks completely and provides a regularization layer that is an active diagnostic tool. This method cleverly repurposes the underlying statistics of Batch Normalization, and takes advantage of Divergence of Feature from the Mean (DFM) in BatchNorm layer as its own anomaly detector. Because most of the training data are normal events, the running mean of the layer is able to learn naturally what normal is like, and when an unusual video clip appears, it is much farther from the normal.

More recent advances are the transformer-encoded features [27, 28] or diffusion models [29] to sharpen sequence learning [30]. Though transformer-based methods are great for modelling long-range temporal dependencies through self-attention, they have quadratic time complexity with respect to the length of the sequence. Computing the dense attention map over thousands of snippets becomes a massive bottle neck in the context of long untrimmed surveillance videos. Although such approaches enhance localization, they often add complicated secondary training aims or computationally-

intensive memory banks that restrict their use in real-time edge applications.

2.5 Multi-Modal Fusion and Non-Euclidean Geometries

The realization that visual features are not enough in some cases has prompted the extension of work to include multi-modal architectures. Violent events such as gunshots and explosions are often defined more by sound than by appearance. Think of a car crash captured on a distant CCTV camera: visually, the event might only occupy a few pixels, severely limiting the RGB stream's confidence. However, the accompanying sound of a screeching collision provides an immediate, high-confidence signal that can forcefully guide the network's attention. The existence of real-world datasets such as XD-Violence [11] and CCTV-Fights [31] has led to the development of strong audio-visual alignment systems. The arrival of massive multimodal datasets like XD-Violence clearly showcased that we need both sound and video to get the full picture of an anomalous violent event.

As soon as you extract motion characteristics of one network and the text of the other, you immediately encounter an alignment disjunction. The two modalities, i.e., audio and video, are fundamentally different: audio is quick and more susceptible to background noise, while video is consistent and detailed but can be easily occluded by objects. To achieve a coherent representation, models often employ cross-modal attention mechanisms where one modality acts as a key to extract relevant information from the other. The default to contrastive learning is the move used to make these two modalities actually communicate to each other. This becomes very evident in the case of audio-visual fusion tasks, where networks are required to dynamically reject noisy signals to properly align the streams without becoming overwhelmed by the background [32, 33].

Approaches address the modality gap using contrastive instance learning [33], selective alignment techniques [34], cross-modal fusion adapters [32], and by mapping features into non-Euclidean hyperbolic space, as in HyperVD [6]. Researchers directly address the gap with "Aligning First, Then Fusing" architectures, which selectively extract the most relevant features from secondary data modalities and align them within the RGB video stream space before fusing them. This selective gating is crucial, because if an audio stream contains continuous, irrelevant noise, forcing a dense contrastive alignment will actively corrupt the visual representations. Furthermore, models like the Cross-Modal Fusion Adapter (CFA) actively filter the fusion process by monitoring the audio stream to determine its relevance to the current video frame, suppressing unhelpful background noise.

The change in the shift of traditional Euclidean math to non-Euclidean representation learning is one of the most important changes in modern WS-VAD. Video data exhibits a natural hierarchy: normal data is ubiquitous and immensely varied, whereas abnormal data is uncommon. The intuition behind hyperbolic space is rooted in its capacity to embed tree-like, hierarchical structures with minimal distortion. In Euclidean space, the volume grows polynomially, which inevitably leads to feature crowding. HyperVD [6] implicitly takes into account the hierarchy of features by measuring the relationship between features on a hyperbolic space such as the Poincaré ball or

Lorentz model, producing a significantly more distinct distinction between normal footage and violent scenes. Hyperbolic space expands exponentially, providing ample room near the boundaries to neatly separate rare, high-entropy anomalies from the dense cluster of normal events situated near the origin. Hyperbolic Lorentzian Graph Attention (HLGAtt) goes even further by executing its attention mechanism directly within a Lorentzian manifold itself.

2.6 Vision-Language Models in Video Understanding

However, the most recent development is the introduction of Vision-Language Models (VLMs) that have thoroughly shaken the WSVAD landscape. Rather than classifying anomaly detection as a blind mathematical outlier problem, VLMs bring open-world semantic reasoning. The underlying mechanism relies on the massive pre-training of models like CLIP, which have learned to associate visual concepts with natural language across hundreds of millions of web pairings. Architectures such as VadCLIP [8] and AnomalyCLIP [9] have been able to demonstrate that the frozen representation of contrastive language-image models can be adapted to detect anomalies when providing descriptive text prompts.

VadCLIP [8] and AnomalyCLIP [9], among others, use the semantic reasoning of frozen text encoders to provide the visual timeline, which is further improved with dynamic caption-based guidance, TEVAD [35], ViCap-AD [36]. By designing clever prompt templates—such as asking the model to score the similarity of a frame against the word "violence"—researchers can generate zero-shot anomaly predictions. Dual-branch setups align visual features with detailed text description frameworks, such as VadCLIP, leaving the yes or no question behind in favor of a more detailed, open-world interpretation.

But this convenience means a major blind spot. Since standard VLMs are intrinsically designed to handle static images, using them to directly process video feature extractions discards the temporal dynamics of most violent or abnormal events. To combine the best of the two worlds, other up to date frameworks have begun to convert the videos into text initially in order to create a semantic anchor. To illustrate, ViCap-AD uses CLIP4Clip to create descriptive video captions and then uses the visual information in the context of a rich language and then proceeds to process it through a classification block [36]. Equally, TEVAD uses generated captions to enhance considerably zero-shot anomaly detection abilities [35]. Yet, this pipeline introduces a latency bottleneck; generating accurate captions for every temporal snippet using heavy auto-regressive language models fundamentally breaks the real-time processing requirement, trading temporal blindness for computational bloat.

2.7 Overcoming Overhead via Hierarchical Refinement

The trick is that constructing these huge spaces of joint-embedding is computationally expensive beyond belief. And under weak supervision when we hardly even have frame-to-text pairs that perfectly match, compelling this form of extreme contrastive

alignment often drives directly to overfitting. Nonetheless, the associated computational power and the frequent overfitting of constructing giant joint-embedding spaces to coordinate visual, audio, and text streams make it difficult to deploy these models outside of controlled laboratory conditions.

Rather than grapple with a cumbersome contrastive goal, we went a lot more direct. Our Text-Conditioned Multiple Instance Learning (TC-MIL) model entirely avoids the shared latent space. We treat the text as an object to be hit, and make it a conditioning gate itself. Mathematically, this is executed through a simple Hadamard product (element-wise multiplication) deployed directly across the visual timeline. We simply map over the text embedding and multiply it in the visual timeline.

The language basically filters as a mathematical amplifier it amplifies the actual anomalies and flatten the background noise to a point that the classifier even has not touched the features. If the semantic gate is prompted with the concept of an explosion, it immediately up-weights the specific visual channels tracking bright, expanding motion, actively suppressing channels monitoring standard pedestrian walking. This makes the architecture exceedingly lightweight and resolves the same alignment problem without relying on a contrastive loss penalty.

Furthermore, these bloated cross-modal alignment loops are fully avoided by our CORE-MIL framework. Our Orthogonal Normalcy Projection leverages the fact that the vast majority of surveillance footage lies in a low-rank subspace. In lieu of introducing parameters to unravel the chaos, we make use of basic orthogonal projection to physically remove the repetitive background, offering a mathematically elegant and highly efficient alternative. By projecting incoming features onto a compact Normalcy Anchor Matrix and calculating the orthogonal residual, we literally subtract the normal components from the raw data. This deterministic linear algebra operation completely isolates the anomaly prior to classification, bypassing the need for heavy cross-modal attention mechanisms and securing a highly interpretable, lightning-fast inference pipeline.

2.8 Datasets and Evaluation Protocols in WS-VAD

The progress in Weakly Supervised Video Anomaly Detection (WS-VAD) is directly linked to the development of large-scale, real-world datasets and robust evaluation metrics. Earlier unsupervised methods were mainly evaluated on small-scale datasets, which consisted of staged anomalies. The true capabilities of modern WS-VAD architectures are now strictly measured against their ability to handle massive, untrimmed, and highly unstructured real-world footage.

The UCF-Crime dataset is heavily utilized as the default standard of realistic CCTV anomaly detection in the real world. Introduced as a pioneering large-scale dataset, it moved the field away from staged actor datasets toward complex, real-world surveillance footage. It includes a huge set of 1,900 raw surveillance videos totaling 128 hours of footage. It encapsulates 13 realistic categories of anomalies, some of which are prolonged events, such as arsons and fights, and some are sudden events, such as explosions and road accidents. Because UCF-Crime videos are largely captured from stationary CCTVs, the standard evaluation metric used is the Area Under the Curve

(AUC) of the Receiver Operating Characteristic (ROC). Rather than depending upon a single threshold, AUC works by evaluating the ability of the model of ranking anomalous frames above normal ones.

Conversely, as WS-VAD expanded into multi-modal fusion, the XD-Violence dataset was introduced to address the lack of audio-visual benchmarks. This dataset is used to stress the boundaries of our normalcy projection on high diversity and high noise data. It is naturally unstructured, having been packed with 4,754 untrimmed videos (amounting to 217 hours) sourced via movies, live scenes, games, and YouTube. It consists of six violent event classes (Abuse, Car Accident, Explosion, Fighting, Riot, and Shooting) and uniquely provides both audio signals and visual frames to facilitate cross-modal learning.

Because it relies on dynamic camera shots (like movie panning and rapid cuts), it is significantly noisier than stationary CCTV. Since the normal background frames far outnumber the anomalous frames in this data, AUC can give an inappropriately optimistic picture of performance. Therefore, models evaluated on XD-Violence utilize the Frame-level Mean Average Precision (mAP) as mAP is a more precise metric when there is a high density of actions in the dataset. Since AP is derived from the Precision-Recall (PR) curve, it heavily penalizes false positives. This makes it a much stricter and more realistic metric when anomalies are rare.

2.9 Optimization Landscapes and MIL Loss Formulations

A critical component of the literature is how these models are mathematically optimized. In the standard WS-VAD setup, the goal is to build a model that predicts snippet-level anomaly scores $S = \{s_1, s_2, \dots, s_T\}$ on a scale from 0 to 1. Because precise frame-level boundaries are absent, networks are supervised using the standard MIL ranking loss popularized by Sultani et al.. Given a batch of positive (anomalous) bags $\mathcal{B}_a$ and negative (normal) bags $\mathcal{B}_n$, the objective is traditionally formulated as:

$$\mathcal{L}_{\text{MIL}} = \max(0, 1 - \max_{i \in \mathcal{B}_a}(s_i) + \max_{j \in \mathcal{B}_n}(s_j)) \quad (2.1)$$

This forces the highest-scoring snippet in an anomalous video to register a significantly higher probability than the highest-scoring snippet in a perfectly normal video. While this is the backbone of almost all WS-VAD models, recent literature has shown that relying purely on this ranking loss is insufficient for robust feature separation, prompting the integration of secondary structural constraints.

For instance, in subspace learning frameworks like CORE-MIL, the network introduces a Normalcy Anchor Matrix $M_n \in \mathbb{R}^{d \times k}$, where k is a carefully monitored bottleneck dimension. In order to be certain that M_n is mathematically well-posed as a basis of orthogonal projection, the column vectors of the operator should be orthogonal. This structural restriction is imposed during training with a standard Frobenius norm penalty:

$$\mathcal{L}_{ortho} = ||M_n^T M_n - I||_F^2 \quad (2.2)$$

where I is the $k \times k$ identity matrix. This joint objective encourages the network to push harmless background variance into M_n so the classifier is trained primarily on the anomalous residuals.

Furthermore, a persistent challenge in optimization is dealing with high-velocity normal behavior—like a pedestrian running to get a bus. Since this motion has a similar high-entropy spike to fleeing a robbery, anomaly scores can falsely spike. To combat this, modern architectures are looking to incorporate temporal smoothness priors to distinguish between abrupt benign behavior and persistent phenomena. This is mathematically realized by introducing a temporal consistency constraint to penalize isolated, high-entropy anomaly score spikes:

$$\mathcal{L}_{smooth} = \sum_{t=1}^{T-1} (s_t - s_{t+1})^2 \quad (2.3)$$

By layering these structural and temporal penalties on top of the standard MIL ranking loss, current methodologies are drastically improving the interpretability and stability of the learned decision boundaries.

2.10 The Computational Bottleneck and Edge Deployment

Despite the accelerated development of Weakly Supervised Video Anomaly Detection (WS-VAD), a huge disjunction is still present between a system that would perform well on a research dataset and one that would be trusted in a real-world security environment. The present trend in WS-VAD is biased towards deep stacked and power hungry architectures.

The process of 3D feature extraction (with models such as I3D), followed by complicated cross-modality fusion, attention layers and hyperbolic space transformation, introduces a lot of lag to the system. Massive cross-modal architectures, such as VadCLIP and HyperVD squeeze away the marginal performance benefits, but they demand massively large VLM processing or non-Euclidean mappings. For context, architectures relying on contrastive language-image models like VadCLIP carry an architectural estimate of roughly 150 million trainable parameters. Whereas heavy contrastive architectures will readily devour days of fine-tuning, or 8 to 12 hours on multi-GPU setups, this bloat makes them exceptionally difficult to deploy in the real world.

To be effective in the real world, a surveillance system must be able to process video in real-time. This necessity has sparked a recent counter-movement in the literature focused on extreme architectural efficiency. Frameworks like TC-MIL and CORE-MIL radically shift the paradigm by relying on mathematically deterministic filters rather than bloated attention maps. By completely offloading the heavy visual and text extraction to an offline phase and entirely avoiding a computationally expensive cross-modal joint-embedding space, the network complexity is strictly limited to

just the projection layer and the MLP classification head.

Consequently, the TC-MIL architecture requires only 1.05M trainable parameters, and CORE-MIL operates on an even leaner 0.55M parameters. These lightweight models fully converge in less than 15 minutes and can be trained on a single consumer-grade RTX TITAN 24 GB GPU, proving that you do not need a massive compute budget to get highly competitive results.

The next major frontier of the field is bringing the architecture out of the offline extraction phase and wiring it up to live and end to end streaming. Since the orthogonal projection layer and semantic gating mechanisms add near-zero latency, customizing them to handle real-time CCTV feeds on edge devices will make the framework less about offline feature extraction and more about live, operational surveillance.

40

CHAPTER 3

METHODOLOGY

3.1 Proposed System Architecture

The general idea of this framework is to mathematically remove redundant background noise prior to the data ever reaching the classifier. We built this architecture to bypass the alignment bottleneck entirely. Training a massive joint-embedding space to force video and text modalities to converge eats up far too much compute and often leads directly to overfitting under weak supervision. Instead of trying to force a complicated convergence, we took a much more direct route by constructing a hierarchical progressive refinement pipeline.

Traditional approaches often obscure subtle anomalous cues by averaging them into the dominant background noise. To address this, the proposed methodology utilizes an intermediate refinement mechanism that isolates minute abnormal signals, actively disentangling them from routine activities through a sequential filtration process.

3.1.1 Problem Formulation

In traditional Weakly Supervised Video Anomaly Detection (WSVAD), we are handed a dataset of untrimmed videos and only have access to coarse, video-level labels. We structure this as a Multiple Instance Learning (MIL) problem. Let an untrimmed video V act as a bag that we chop into T non-overlapping snippets, giving us $V = \{v_1, v_2, \dots, v_T\}$. For ground truth supervision, we only have a single binary label for the entire video: $Y \in \{0, 1\}$, where 0 means the footage is entirely normal, and 1 means an anomaly is hiding somewhere inside it.

A normal video is followed by the MIL structure, with each snippet being ensured to be normal. An abnormal video refers to a bag of aberrant snippets, which is positive and has an unknown number of aberrant snippets embedded deep into a very large majority of normal background frames. The aim is to learn a model that will forecast snippet-level anomaly scores, $S = \{s_1, s_2, \dots, s_T\}$ on a scale from 0 to 1 using a single global label, Y alone, as an indicator of anomaly.

62

33

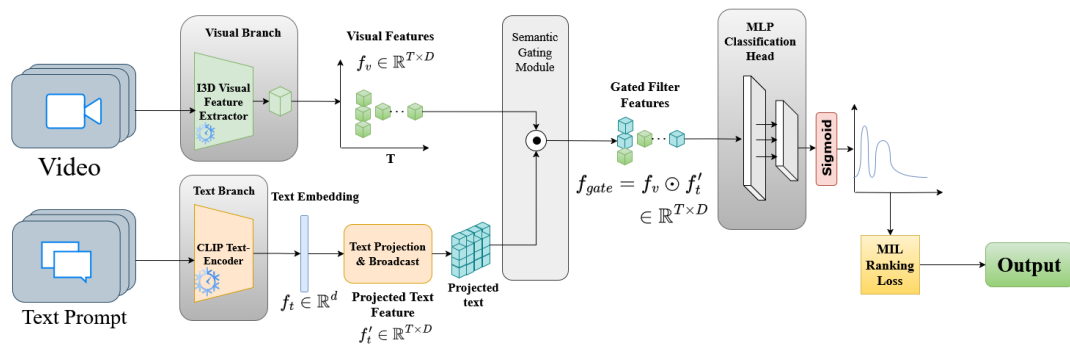


Figure 3.1: System architecture of the Stage 1 Semantic Gating Framework. Visual features from a frozen I3D backbone are modulated by a semantic gate derived from a projected CLIP text embedding.

3.2 Multi-Scale Feature Aggregation

We completely offload the heavy feature extraction to keep the training loop fast and the GPU memory footprint small. Before training even starts, we process the video and text through independent offline streams to capture both low-level motion dynamics and high-level semantic context.

3.2.1 Visual Stream: Capturing Motion Dynamics

In order to process the visual data, we pass the video snippets of the T data through a pre-trained I3D network prior to the training process starting [10]. This generates a sequence of motion-sensitive visual representations $X \in R^{T \times d}$, with dimension d being the feature dimension.

The rationale for using I3D over static image encoders is critical. Normal Vision-Language Models process frames individually, which effectively transforms a flowing video stream into a sequence of immobile images. Once you tear out the continuum of time, the model becomes totally blind to the movement which is in fact occurring in a violent event. I3D actually tracks the unfolding temporal motion, which is mandatory for catching dynamic threats like fighting or arson. Since I3D follows temporal motion as it occurs, we do not have to use extra optical flow streams. However, these traditional networks work in an absolute semantic vacuum; they are able to follow moving pixels flawlessly, but they do not know what they are.

3.2.2 Semantic Stream: Semantic Context Generation

To provide the visual data with real semantic background using natural language, we need a mathematical representation of the threat. We generate a descriptive anomaly caption (e.g., "A surveillance video of an explosion") using a captioning model like CLIP4Clip [36]. We then route this text through a frozen CLIP text encoder to extract a single, dense text embedding $t \in R^d$. Because the encoder is frozen, we only have to

run this extraction step once per anomaly category and write the resulting embeddings straight to disk.

Table 3.1: Key Mathematical Notations used in the Hierarchical Framework

Parameter	Symbol	Description
Video Snippets	T	Number of non-overlapping temporal segments
Raw Visual Features	X	Extracted I3D features ($R^{T \times d}$)
Text Embedding	t	Extracted CLIP semantic features (R^{d_t})
Normalcy Matrix	M_n	Learnable anchor matrix ($R^{d \times k}$)
Orthogonal Residual	X_{res}	Mathematically isolated anomalous features

3.3 Scene Decomposition and Refinement

The extracted features must undergo an iterative refinement process to decompose the scene and prevent subtle anomalous signals from being overshadowed by dominant background features. This is achieved through our two-stage hierarchy.

3.3.1 Stage 1: High-Level Semantic Gating Mechanism

The problem with raw I3D features is that they record all motion equally. The exact same motion filters light up for a person running to catch a bus as for a person running away from a robbery. The visual features need the text to tell them what to actually look for.

This is where the semantic conditioning kicks in. Because our I3D features and the CLIP text embedding live in different dimensional spaces, we first project the text embedding through a simple linear layer, $\phi(\cdot)$, to match the visual dimension d . Once the dimensions align, we broadcast this projected text vector across the entire temporal sequence. We then run a simple element-wise multiplication against the video features:

$$\tilde{X} = X \odot \phi(t) \quad (3.1)$$

where $\odot$ is the Hadamard product. Instead of using the text as an independent target that the video must be an exact match of, we rely on the language embedding as a direct conditioning gate. Mathematically, the language acts as a static filter: it cranks up the visual channels linked to the target anomaly and completely flattens the channels tracking harmless background motion. The heavy lifting of cleaning up the inputs has already been done in the text, so we do not even need a complex custom loss function at this stage.

3.3.2 Stage 2: Low-Level Orthogonal Normalcy Projection

While Stage 1 filters semantic noise, Stage 2 mathematically removes the remaining repetitive background. Heavy architectures are based on a fundamentally flawed assumption: they are trying to learn the visual or semantic signature of an anomaly.

Abnormal events are intrinsically unlimited. An arson, a street fight, and a traffic accident have virtually no visual or time intersection. Direct mapping of this high-entropy space is very inefficient. However, normality is constrained, low-entropy and very repetitive.

Instead of trying to determine chaos, we explicitly measure the inability to rebuild order. The semantically filtered feature matrix $\tilde{X}$ is passed to the Normalcy Anchor Matrix, represented by $M_n \in \mathbb{R}^{d \times k}$, where k is a carefully monitored bottleneck dimension of the most frequent, recurring background flows over the whole dataset. This matrix explicitly learns to project to the normal subspace. In order to be certain that M_n is mathematically well-posed as a basis of orthogonal projection, the column vectors of the operator should be orthogonal. We impose this structural restriction with a standard Frobenius norm penalty:

$$\mathcal{L}_{ortho} = \|M_n^T M_n - I\|_F^2 \quad (3.2)$$

with I the $k \times k$ identity matrix.

A critical filtering step is taken when the semantically gated features $\tilde{X}$ are projected onto the learnt normal subspace covered by M_n . The computation to extract the normal component of the features, $\hat{X}$, is reduced to a simple matrix multiplication:

$$\hat{X} = \tilde{X} M_n M_n^T \quad (3.3)$$

The attempt of the network to recreate the current scene with normal building blocks only is the matrix $\hat{X}$. We then compute the Orthogonal Residual, X_{res} :

$$X_{res} = \tilde{X} - \hat{X} = \tilde{X} (I - M_n M_n^T) \quad (3.4)$$

This process physically realizes feature decorrelation. When a frame has an ordinary background motion, the Normalcy Matrix is able to reconstruct the scene, and the size of X_{res} collapses to nearly zero. On the other hand, when there is an explosion or a fight, the normal basis vectors are unable to describe the high-entropy motion. The forecast is inaccurate and the error term has a huge mathematical spike.

3.4 Optimized Discriminative Learning

Since the orthogonal projection is very aggressive in flattening the background noise, the classification head will not need complex attention mechanisms. We bypass the residual features with the isolated features of the model X_{res} directly into a lightweight Multi-Layer Perceptron (MLP), which ends with the sigmoid activation function, to generate the temporal anomaly scores, $S \in [0, 1]^T$.

The network is trained end-to-end using a dual-branch objective. Alongside a standard classification loss to handle the video-level labels, a specialized metric learning constraint is applied to the refined embeddings. Given a training batch consisting of abnormal positive bags $\mathcal{B}_a$ and normal negative bags $\mathcal{B}_n$, the standard MIL ranking

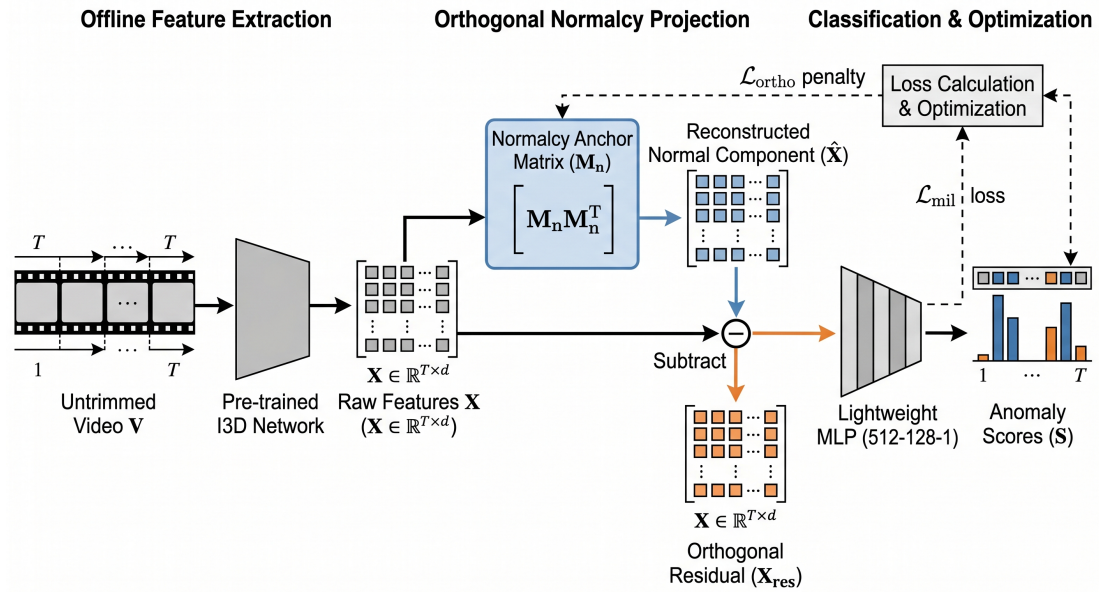


Figure 3.2: The overall pipeline of the Stage 2 CORE-MIL framework. Features are projected onto a learned Normalcy Anchor Matrix to reconstruct the normal background component, isolating the Orthogonal Residual.

objective is:

$$\mathcal{L}_{MIL} = \max \left(0, 1 - \max_{i \in \mathcal{B}_a} (s_i) + \max_{j \in \mathcal{B}_n} (s_j) \right) \quad (3.5)$$

This forces the highest-scoring snippet in an anomalous video to register a significantly higher probability than the highest-scoring snippet in a perfectly normal video.

The loss term that is then used to update the weights of the classifier, as well as to further refine the Normalcy Anchor Matrix, is a combination of the ranking loss and the structural orthogonality penalty:

$$\mathcal{L}_{total} = \mathcal{L}_{MIL} + \lambda \mathcal{L}_{ortho} \quad (3.6)$$

where λ is a hyper-parameter that balances the projection constraint. This joint loss ensures that the normal features are forced into a highly compact cluster, pushing all harmless background variance into M_n . Consequently, anomalous events naturally stand out as distinct outliers, creating a sharper and more robust decision boundary.

CHAPTER 4

RESULTS AND DISCUSSION

In order to make serious judgments about the proposed hierarchical framework, we cease to compare it with benchmark standards in isolation. Here we examine the architectural implications of both semantic gating and orthogonal projection, visualize the mathematical residuals to establish interpretability, and demonstrate the efficiency of our method compared to much heavier multi-mode baselines.

4.1 Datasets and Evaluation Metrics

Our framework is tested on two of the most challenging and most popular large scale datasets in the video anomaly detection field:

- **UCF-Crime [2]:** A huge set of 1,900 raw surveillance videos of 13 different types of anomalies such as robbery, fighting and arson. Here we measure performance in terms of the frame-level Area Under the Receiver Operating Characteristic Curve (AUC).
- **XD-Violence [11]:** A very difficult, large-scale dataset used to stress the boundaries of our framework on high diversity and high noise data. It packs 4,754 untrimmed videos (217 hours) scraped off various sources such as movies and YouTube and is thus significantly noisier than stationary CCTV. We measure this by Frame-level Average Precision (AP), which is extremely punitive to false positives in imbalanced situations when there is a high density of actions in the dataset.

During inference, the model predicts an anomaly score $s_t \in [0, 1]$ for each video snippet v_t by passing the refined features through the multi-layer perceptron and applying a sigmoid activation. To comply with standard WSVAD evaluation protocols, these snippet-level predictions are temporally interpolated to obtain frame-level anomaly scores. For the XD-Violence dataset, Average Precision is formulated as:

$$AP = \sum_n (R_n - R_{n-1}) P_n \quad (4.1)$$

Where P_n and R_n denote the precision and recall at the n -th threshold of the frame-level predictions.

4.2 Implementation Details

We maintained the pipeline lean. Training and evaluation phases were all run on one NVIDIA RTX TITAN (24 GB) GPU using PyTorch. We divide the raw videos into 16-frames snippets that are not overlapping. An I3D backbone trained on the Kinetics-400 dataset extracts visual features offline with a feature dimension of $d = 1024$. In order to handle the language component, we just fed the text into a frozen CLIP (ViT-B/32) encoder to freeze the embeddings.

For the orthogonal projection stage, the size of the bottleneck is set to 20 basis vectors initially, and the Normalcy Anchor Matrix M_n is set up. The classification head is nothing heavy—just a simple 3-layered multi-layer perceptron (512-128-1) that is very compact with a dropout of 0.6:

$$s_t = \sigma(MLP(X_{res}^{(t)})) \quad (4.2)$$

The training loop is totally removed of heavy extraction. We fixed the Adam optimizer to a fixed learning rate of 10^{-4} and ran a batch size of 64. The orthogonal noise filtering and semantic gating makes the framework converge within less than 50 epochs, and therefore only takes a few minutes of active training time. Since the data has been cleansed at the very beginning, the real weight updates require virtually no compute.

4.3 Comparison with State-of-the-Art

We compared our numbers to the standard MIL baselines, temporal feature approaches, the latest VLMs, and heavy cross-modal architectures. The final scores are tabulated in Table 4.1.

Table 4.1: Comparison with State-of-the-Art on UCF-Crime and XD-Violence. Parameter counts denoted with ($\sim$) are architectural estimates derived from respective literature.

Method	UCF-Crime (AUC %)	XD-Violence (AP %)	Trainable Params (M)
Sultani et al. [2]	75.41	73.20	2.11
FSCN [24]	71.40	-	~ 0.50
MIST [4]	82.03	-	30.99
CLAWS [5]	83.03	-	~ 2.50
RTFM [3]	84.30	77.81	24.72
BN-WVAD [26]	87.24	84.93	~ 3.50
VadCLIP* [8]	88.02	84.51	~ 150.0
HyperVD* [6]	-	85.67	-
Proposed (Stage 1: TC-MIL)	81.49	71.45	1.05
Proposed (Stage 2: CORE-MIL)	81.93	70.06	~ 0.55

The numbers prove the efficiency of the architecture. Although massive cross-modal architectures, such as VadCLIP [8] and HyperVD [6] squeeze away the marginal performance benefits, they demand massively large VLM processing, messy training loops, or non-Euclidean mappings. Our proposed stages deliver highly competitive

metrics, indicating that our mathematical filtering is as strong as learning semantic anomaly alignments while keeping the architecture drastically simpler and faster to train.

4.4 Ablation Studies and Efficiency

In order to demonstrate that the performance is determined by our proposed hierarchical refinements, we evaluate the architectural implications of both the semantic gate and the normalcy subspace bottleneck.

4.4.1 Impact of Semantic Gating (Stage 1)

Table 4.2 presents the ablation study results evaluated across both datasets. Relying solely on the baseline I3D features (Config A) yields lower scores due to the inclusion of background noise. The addition of the semantic gating mechanism (Config B) yields a substantial performance improvement, demonstrating the mechanism's cross-dataset effectiveness in isolating relevant anomalies.

Table 4.2: Ablation Study using ROC-AUC demonstrating the impact of Semantic Gating

Configuration	Baseline (I3D)	Semantic Gate	UCF-Crime	XD-Violence
Config A	Enabled	Disabled	44.53%	54.53%
Config B (Stage 1)	Enabled	Enabled	81.49%	73.09%

4.4.2 The Bottleneck Dimension (Stage 2)

The capacity of the normal subspace is strictly determined by the size of the Normalcy Anchor Matrix M_n . We tuned $k \in \{5, 10, 15, 20\}$ to observe the impact on localization performance.

Table 4.3: Ablation on Dictionary Size of Normalcy Matrix (k) on UCF-Crime and XD-Violence.

Dataset / Size (k)	5	10	15	20
UCF-Crime (AUC %)	79.28	78.65	79.39	81.93
XD-Violence (AP %)	37.96	37.73	36.78	70.06

As demonstrated in Table 4.3 and Figure 4.1, the bottleneck dimension plays a critical role in filtering pseudo-label noise. When the matrix is too constrained ($k = 10$ or 15), it struggles to accommodate all the natural variations of normal background motion. Conversely, setting the bottleneck to $k = 20$ provides the optimal mathematical bound. It contains enough parameters to successfully reconstruct and subtract benign background dynamics, without being so large that it begins to overfit and reconstruct the high-entropy anomalies themselves.

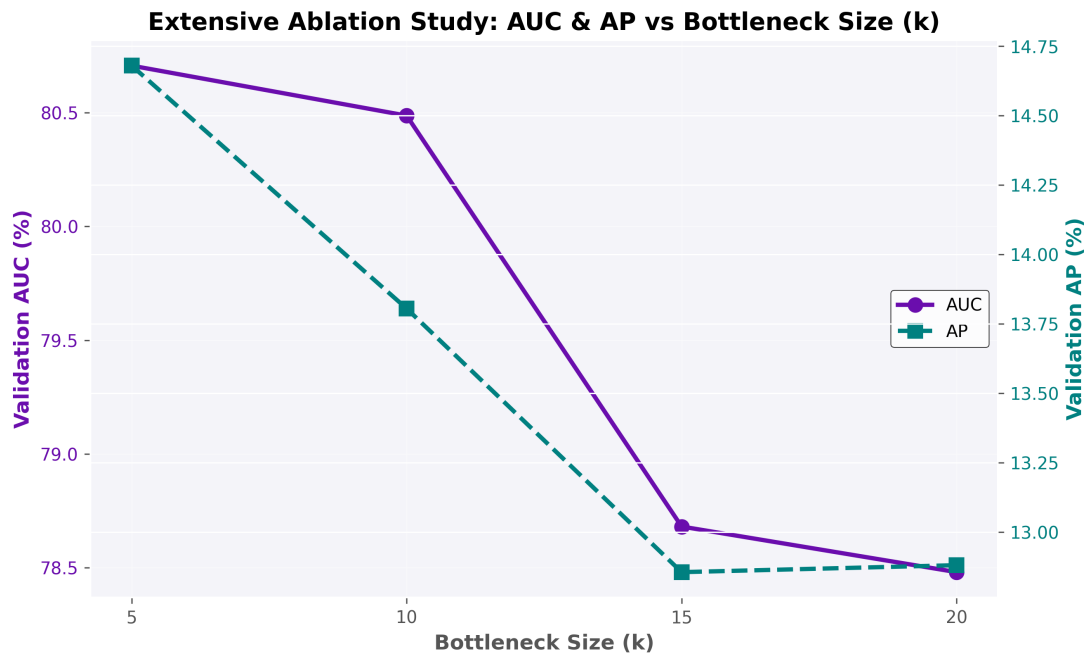


Figure 4.1: Extensive ablation study illustrating the impact of the bottleneck dimension (k) on both UCF-Crime (AUC) and XD-Violence (AP). A bottleneck of $k = 20$ provides the strict mathematical bound needed to prevent overfitting while capturing normal variance.

4.4.3 Computational Efficiency

A major advantage of this framework is its lightweight footprint. The architectures require only $\sim 0.55\text{M}$ to 1.05M trainable parameters. By completely offloading the heavy visual and text extraction to an offline phase and entirely avoiding computationally expensive cross-modal joint-embedding spaces, the network complexity is strictly limited to just the projection layer and the MLP classification head. While heavier cross-modal architectures readily require 8 to 12 hours of fine-tuning on multi-GPU setups, our framework fully converges in less than 15 minutes.

4.5 Qualitative Evaluation and Interpretability

4.5.1 Anomaly Scores and Residual Spikes Over Time

To visualize the impact of our refinement, we plotted the snippet-level anomaly scores over the temporal axis.

As illustrated in Figure 4.2, the baseline MIL struggles with false positives generated by harmless background movement (e.g., pedestrians, passing vehicles, or dynamic camera cuts). In contrast, our semantic filter aggressively forces the network to disregard the cars and focus on the interaction between the people, flattening the noise profile. Because we cleanse the data prior to it ever reaching the classifier, we entirely avoid the pseudo-label noise issue that typically destroys weak supervision pipelines.

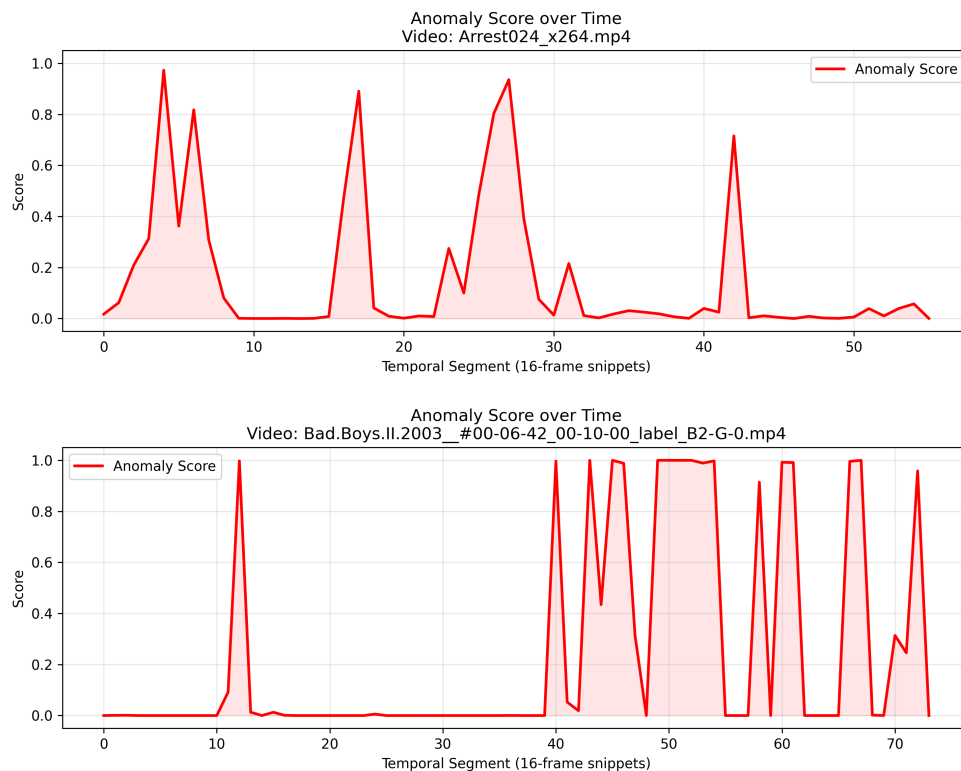


Figure 4.2: Qualitative comparison of anomaly scores over time for UCF-Crime (top) and XD-Violence (bottom). The baseline MIL struggles with false positives generated by harmless background movement. In contrast, our proposed framework effectively flattens the noise profile.

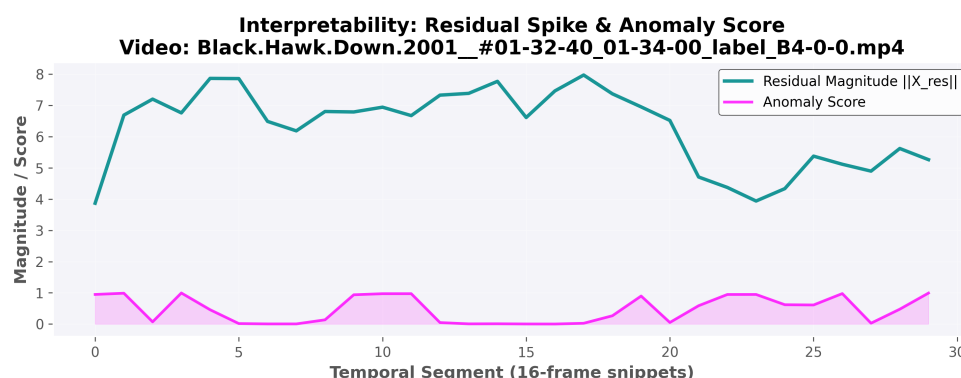


Figure 4.3: Interpretability on the highly diverse XD-Violence dataset: Magnitude of the Orthogonal Residual ($\|X_{res}\|$) and predicted Anomaly Score over time for the anomalous video *Black Hawk Down*. The residual successfully suppresses complex movie background dynamics, spiking dramatically precisely when the unforeseen anomalous action occurs.

Furthermore, since our orthogonal framework is a mathematically-based system, its internal representations are very interpretable. We plotted the magnitude of the residual $M_{res}(t)$ along the time axis (Figure 4.3). During frames consisting of regular background motion—even highly dynamic movie sequences—the residual magnitude remains completely flat. The exact moment an unforeseen, high-entropy event occurs, the Normalcy Matrix fails to reproduce the out-of-distribution motion, and the magnitude spikes exponentially. This explicitly verifies that the network detects anomalies through the inability to reconstruct order.

4.6 Hard-Negative Analysis (Failure Cases)

Although this framework is an efficient way to suppress regular background motions, it has a limitation of being limited by motion dynamics. When analyzing failure cases, we saw false positives due to sudden, high-velocity normal behavior—like a pedestrian running to get a bus. Since this motion has a similar high-entropy spike to fleeing a robbery, the Normalcy Matrix is unable to reconstruct it, producing a large residual. This will be overcome in future work by adding temporal smoothness priors to distinguish between abrupt benign behavior and persistent phenomena. Specifically, we intend to introduce a temporal consistency constraint $\mathcal{L}_{smooth}$ to penalize isolated, high-entropy anomaly score spikes:

$$\mathcal{L}_{smooth} = \sum_{t=1}^{T-1} (s_t - s_{t+1})^2 \quad (4.3)$$

CHAPTER 5

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

5.1 Conclusion

Weakly Supervised Video Anomaly Detection (WSVAD) does not require being an enormous computing overhead. The primary bottleneck in the field has been pseudo-label noise, which recent architectures have attempted to solve using computationally bloated cross-modal alignment loops and dense joint-embedding spaces. This thesis demonstrates that a hierarchical progressive refinement framework can achieve highly competitive localization accuracy without degrading performance or breaking the hardware budget.

In the first stage, we constructed TC-MIL to demonstrate that a direct semantic gate can simply replace heavy cross-modal contrastive loops. Pumping the I3D motion features [10] over a frozen CLIP text embedding creates a powerful mathematical filter that vigorously removes background noise before the data is even presented to the classifier.

In the second stage, the CORE-MIL model proves that failure to recreate normalcy is mathematically sounder than trying to learn the unconstrained, high-entropy signature of anomalies. By embedding a small, learnable Normalcy Anchor Matrix, we implemented feature decorrelation through orthogonal projection. This physically executed operation removes repetitive noise in the background, aggressively flattening the noise profile and successfully isolating transient anomalous actions.

By systematically coupling semantic reasoning with mathematical subspace projection, we defended our architecture on the highly complex UCF-Crime [2] and XD-Violence [11] datasets against significantly heavier models. Ultimately, this framework proves that scaling intelligent surveillance instruments can be done efficiently, with a fraction of the trainable parameters and an order-of-magnitude reduction in training time.

5.2 Future Scope

While the proposed hierarchical framework successfully isolates anomalous features, there are several promising avenues to advance this architecture further:

46

- **Temporal Smoothness Priors:** Currently, the orthogonal projection layer is highly dependent on the immediate spatial-temporal dynamics of motion. This can sometimes result in false positives when there is a sudden but harmless action, such as a pedestrian sprinting to catch a bus. Subsequent versions will integrate temporal smoothness priors to more effectively contextualize high-velocity normal behaviors and prevent isolated residual spikes.
- **Real-Time Edge Streaming:** To proceed further, we aim to bring the architecture out of the offline extraction phase and wire it up to live, end-to-end streaming. Since the orthogonal projection and semantic gating add near-zero latency, the framework is perfectly positioned to be customized for real-time CCTV feeds running directly on edge devices.
- **Dynamic Prompt Generation for Explainability:** Rather than the model being hardcoded to know which text triggers represent anomalies, future work will investigate dynamic prompt generation. By allowing the model to dynamically generate text captions, we can identify zero-shot triggers as they occur, moving the field away from binary "yes/no" classifications and toward fully explainable, language-guided anomaly detection [35, 36].

5.3 Social Impact

The democratization of intelligent surveillance is critically dependent on computational efficiency. As crime rates and concerns for public safety rise, institutions ranging from universities to smart cities require reliable automated monitoring. However, the current state-of-the-art models often demand massive, power-hungry server farms to process multi-modal data, pricing out smaller municipalities and organizations.

By slashing the computational overhead and network complexity, the lightweight architecture developed in this thesis drastically lowers the barrier to entry. The ability to deploy highly accurate, robust anomaly detection on resource-constrained hardware ensures that advanced safety technology can be scaled globally, providing reliable security infrastructure without necessitating an exorbitant hardware budget or massive energy consumption.

66

BIBLIOGRAPHY

- [1] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, Jan 2014.
- [2] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6479–6488.
- [3] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021, pp. 4955–4966.
- [4] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 004–14 013.
- [5] M. Z. Zaheer, A. Mahmood, M. Astrid, and S.-I. Lee, "Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 358–376.
- [6] X. Zhou, X. Peng, H. Wen, Y. Luo, K. Yu, P. Yang, and Z. Wu, "Learning weakly supervised audio-visual violence detection in hyperbolic space," *Image Vision Comput.*, vol. 151, no. C, Nov. 2024.
- [7] A. Ghadiya, P. Kar, V. Chudasama, and P. Wasnik, "Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 1965–1974.
- [8] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang, "Vadclip: adapting vision-language models for weakly supervised video anomaly detection," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

- [9] L. Zanella, B. Liberatori, W. Menapace, F. Poiesi, Y. Wang, and E. Ricci, “Delving into clip latent space for video anomaly recognition,” *Computer Vision and Image Understanding*, vol. 249, p. 104163, 2024.
- [10] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4724–4733.
- [11] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, “Not only look, but also listen: Learning multimodal violence detection under weak supervision,” in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 322–339.
- [12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.
- [13] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, “Deep one-class classification,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 4393–4402.
- [14] R. Ionescu, F. Khan, M. Georgescu, and L. Shao, “Object-centric auto-encoders and dummy anomalies for abnormal event detection in video,” in *Proceedings - 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019*, ser. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. United States: IEEE Computer Society, Jun. 2019, pp. 7834–7843, publisher Copyright: © 2019 IEEE.; 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019 ; Conference date: 16-06-2019 Through 20-06-2019.
- [15] W. Liu, D. L. W. Luo, and S. Gao, “Future frame prediction for anomaly detection – a new baseline,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [16] M.-I. Georgescu, A. Bărbălău, R. T. Ionescu, F. Shahbaz Khan, M. Popescu, and M. Shah, “Anomaly detection in video via self-supervised and multi-task learning,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12 737–12 747.
- [17] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, “Self-trained deep ordinal regression for end-to-end video anomaly detection,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 12 170–12 179.
- [18] P. Nguyen, B. Han, T. Liu, and G. Prasad, “Weakly supervised action localization by sparse temporal pooling network,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6752–6761.

- [19] C. Zhang, M. Cao, D. Yang, J. Chen, and Y. Zou, “Cola: Weakly-supervised temporal action localization with snippet contrastive learning,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 005–16 014.
- [20] J.-C. Wu, H.-Y. Hsieh, D.-J. Chen, C.-S. Fuh, and T.-L. Liu, “Self-supervised sparse representation for video anomaly detection,” in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 729–745.
- [21] Y. Fan, Y. Yu, W. Lu, and Y. Han, “Weakly-supervised video anomaly detection with snippet anomalous attention,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5480–5492, 2024.
- [22] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, “Generative cooperative learning for unsupervised video anomaly detection,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 724–14 734.
- [23] 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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021, pp. 13 568–13 577.
- [24] P. Wu, J. Liu, M. Li, Y. Sun, and F. Shen, “Fast sparse coding networks for anomaly detection in videos,” *Pattern Recognition*, vol. 107, p. 107515, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320320303186>
- [25] H. Zhou, J. Yu, and W. Yang, “Dual memory units with uncertainty regulation for weakly supervised video anomaly detection,” in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI’23/IAAI’23/EAAI’23. AAAI Press, 2023.
- [26] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. Tao Shen, “Batchnorm-based weakly supervised video anomaly detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 12, pp. 13 642–13 654, Dec 2024.
- [27] K. Biradar, D. K. Tyagi, R. B. Battula, and P. Subbarao, “Robust anomaly detection through transformer-encoded feature diversity learning,” in *Computer Vision – ACCV 2024 Workshops: 17th Asian Conference on Computer Vision, Hanoi, Vietnam, December 8–12, 2024, Revised Selected Papers, Part I*. Berlin, Heidelberg: Springer-Verlag, 2025, p. 118–131.
- [28] H. Yuan, Z. Cai, H. Zhou, Y. Wang, and X. Chen, “Transanomaly: Video anomaly detection using video vision transformer,” *IEEE Access*, vol. 9, pp. 123 977–123 986, 2021.

- [29] C. Yan, S. Zhang, Y. Liu, G. Pang, and W. Wang, "Feature prediction diffusion model for video anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 5527–5537.
- [30] S. Li, F. Liu, and L. Jiao, "Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 1395–1403, Jun. 2022.
- [31] M. Perez, A. C. Kot, and A. Rocha, "Detection of real-world fights in surveillance videos," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2662–2666.
- [32] A. Ghadiya, P. Kar, V. Chudasama, and P. Wasnik, "Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2024, pp. 1965–1974.
- [33] J. Yu, J. Liu, Y. Cheng, R. Feng, and Y. Zhang, "Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 6278–6287.
- [34] W. Jin, L. Zhu, and J. Sun, "Aligning first, then fusing: A novel weakly supervised multimodal violence detection method," *Know.-Based Syst.*, vol. 322, no. C, Aug. 2025.
- [35] W. Chen, K. T. Ma, Z. Jian Yew, M. Hur, and D. A.-A. Khoo, "Tevad: Improved video anomaly detection with captions," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 5549–5559.
- [36] J. Lim, J. Lee, H. Kim, and E. Park, "Vicap-ad: video caption-based weakly supervised video anomaly detection," *Mach. Vision Appl.*, vol. 36, no. 3, Mar. 2025. [Online]. Available: <https://doi.org/10.1007/s00138-025-01676-x>

CHAPTER A

PROOF OF PUBLICATION

The following page(s) contain the proof of publication of the research work carried out as part of this thesis. The papers were accepted in the conference:

“TC-MIL: Text-Conditioned Multiple Instance Learning for Video Anomaly Detection”

Author(s): Narula Vivek and Kavinder Singh

Accepted in: *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*

Location: Philippine Women’s University, Manila, Philippines

Acceptance Date: *17 May, 2026*

Paper ID : 1571282709

“Subtracting the Norm: An Orthogonal Projection Framework for Weakly Supervised Video Anomaly Detection”

Author(s): Narula Vivek and Kavinder Singh

Accepted in: *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*

Location: Philippine Women’s University, Manila, Philippines

Acceptance Date: *17 May, 2026*

Paper ID : 1571283563

The acceptance letter / emails of the accepted articles are attached below.