Journal of Advanced Marine Engineering and Technology, Vol. 50, No. 1, pp. 80~88, 2026
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2026.50.1.80

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

Clustering-based data refinement considering normal data distribution in MIL-Based WSVAD

Jin-Se Lee[1] · Dong-Hoan Seo[†]

(Received December 16, 2025 ; Revised February 10, 2026 ; Accepted February 10, 2026)

Abstract: Multi-Instance Learning (MIL)–based Weakly Supervised Video Anomaly Detection (WSVAD) enables training with only video-level labels but often suffers from data imbalance and ambiguous normal–abnormal boundaries in real-world surveillance data. A key challenge arises from the heterogeneous structure of normal data, which consists of repetitive normal-majority patterns and rare, highly variable normal-minority patterns. In MIL settings, snippet representations constructed via mean pooling can dilute boundary-relevant characteristics of normal-minority patterns, leading to boundary ambiguity. To address this issue, we propose a Feature Norm Clustering–based under-sampling framework that explicitly reflects intra-video distributional structure. By clustering clip-level features based on their norm statistics, dominant clusters are identified and selectively removed during training, while temporal order is preserved through feature sequence reconstruction. Experimental results under a fixed backbone and MIL classifier show that emphasizing minor clusters improves AUC and EER and reduces overlap between normal and abnormal score distributions. These findings demonstrate that data reconfiguration guided by the major/minor normal structure is an effective data-centric strategy for improving MIL-based WSVAD.

Keywords: WSVAD, MIL, Data mining, Boundary ambiguity, Feature distribution

1. Introduction

Video Anomaly Detection (VAD) is a technique for automatically detecting abnormal behaviors or events in surveillance videos. Recently, Weakly Supervised Video Anomaly Detection (WSVAD), which relies only on weak video-level labels without requiring frame-level ground truth, has emerged as the dominant research paradigm. In particular, MIL-based VAD methods have the advantage of significantly reducing labeling costs, as they can be trained using only video-level labels. However, due to the nature of VAD data collected at the video level, normal situations are recorded far more frequently than abnormal ones in real-world environments, which leads to severe data imbalance during the training of MIL-based models.

A more critical issue is that normal data do not form a single homogeneous distribution. Because normal behaviors in real-world environments vary depending on people, locations, and time, the normal class contains both repetitive and frequently observed normal-majority patterns and normal-minority patterns that are visually similar to abnormal behaviors or exhibit high variability. However, the MIL framework does not account for such intra-class distributions within normal data, and all normal samples are learned as the same negative class. In MIL-based WSVAD, this limitation becomes more pronounced because learning is performed at the snippet level, where a video is divided into fixed-length temporal segments and each segment is treated as an individual training instance through feature aggregation. As a result, near-abnormal normal-minority patterns are learned in the same manner as normal-majority patterns, which blurs the distinction between normal and abnormal samples and leads to boundary ambiguity, manifested as overlapping anomaly scores between the two classes.

These limitations of MIL-based VAD are not only attributable to model architecture but are also closely related to the inherent data distribution characteristics of VAD datasets. In real-world surveillance environments, normal behaviors are recorded far more frequently than abnormal ones, which naturally leads to the formation

† Corresponding Author (ORCID: https://orcid.org/0000-0003-3610-0356): Professor, Division of Electronics & Electrical Information Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: dhseo@kmou.ac.kr, Tel: 051-410-4412

1 Ph. D. Candidate., Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: ljs171181@g.kmou.ac.kr, Tel: 051-410-4822

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright ⓒ The Korean Society of Marine Engineering

of normal-majority and normal-minority patterns within the normal class. However, MIL does not consider such intra-class distributions and treats all normal snippets as the same negative class. Since snippets are constructed by aggregating multiple clip-level features through mean pooling, subtle variations or abnormal-like characteristics contained in normal-minority patterns can be easily diluted by dominant normal-majority patterns. These structural properties of the data make it difficult to separate normal-minority patterns from abnormal patterns in the feature space, ultimately contributing to boundary ambiguity, where the decision boundary between normal and abnormal samples becomes unclear. In other words, the performance characteristics of MIL-based VAD depend not only on model design but also on how the heterogeneous distribution within normal data is reflected during the learning process.

To alleviate these limitations and more effectively reflect the distributional characteristics within normal data, this study applies K-means clustering to the norms of clip-level features, demonstrating that features within a normal video tend to be concentrated in a small number of dominant clusters, while the remaining clusters can be categorized as minor clusters that occur less frequently and contain more diverse features. Based on this analysis, easy-case clusters are selectively removed during training, encouraging the MIL classifier to focus more on hard-case normal samples and abnormal segments. This data reconfiguration process mitigates repetitive normal patterns and contributes to forming a clearer boundary between normal and abnormal behaviors.

Experimental results reveal two key findings.

(1) Hard-case normal samples and anomaly clips predominantly reside in lower-ranked clusters and play an important role in forming the boundary with abnormal segments.

(2) Removing easy-case clusters reduces the dominant influence of majority normal patterns during MIL training, increases the relative importance of anomaly-relevant features, and consequently induces different performance trends in terms of AUC and EER depending on the data configuration.

In summary, this study empirically demonstrates that intra-class imbalance within normal data, characterized by the major/minor normal structure, is a critical factor influencing anomaly boundary formation and performance characteristics in MIL-based VAD. Furthermore, we propose a data reconfiguration–based learning approach that explicitly incorporates these structural properties into the training process.

2. Related Works

2.1 VAD

Defining anomalies in video data at the frame level or pixel level is inherently subjective and requires substantial human and temporal costs when applied to large-scale datasets. To alleviate these labeling challenges, Multi-Instance Learning (MIL) was proposed as an alternative learning paradigm. MIL assigns labels only at the level of a set of instances, referred to as a bag, rather than to individual instances, enabling effective learning even when instance-level ground truth is incomplete or unavailable.

Dietterich et al. first introduced MIL in the context of drug activity prediction, where multiple 3D molecular conformations associated with a single sample were represented as a bag, and learning was performed under the assumption that a positive bag contains at least one positive instance. This framework demonstrated that positive instances can be effectively identified even in partially labeled settings. Subsequently, MIL was extended to the field of computer vision by representing an image as a bag of patches and treating each patch as an instance. Chen *et al.* applied this formulation to natural scene classification and showed that image classification can be achieved without patch-level annotations.

With the development of strongly supervised object detection models such as R-CNN and YOLO, interest in MIL temporarily declined. However, as research attention shifted toward video data, MIL regained prominence. Videos contain a large number of frames and complex temporal structures, making precise instance-level annotation particularly challenging. To address this issue, Kuehne *et al.* proposed a method that estimates temporal action segments using weak action annotations based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs).

In the field of Video Anomaly Detection (VAD), Sultani et al. were the first to adopt MIL and proposed a learning framework that relies solely on video-level labels. This approach significantly reduced labeling costs while achieving superior performance compared to previous unsupervised methods, and it has since served as the foundation for numerous MIL-based VAD studies. Building on this line of work, Tian et al. proposed a model that combines self-attention mechanisms with Top-k instance selection to incorporate temporal information and partially alleviate learning imbalance. Additionally, Lv *et al.* introduced a strategy that selectively learns ambiguous frames based on changes in anomaly scores, while Yixuan *et al.* proposed

emphasizing anomaly-prone clips by measuring divergence from the feature mean (DFM).

Nevertheless, existing MIL-based VAD approaches primarily focus on snippet or instance selection strategies and do not explicitly consider the fine-grained distributional structure within normal videos. In real-world surveillance environments, normal behaviors are recorded far more frequently than abnormal ones, resulting in the coexistence of dominant repetitive patterns and relatively rare, highly variable patterns within normal data. If such intra-class structural differences are not considered during MIL training, repetitive normal patterns tend to dominate the learning of anomaly-relevant features, making it difficult to form a clear boundary between normal and abnormal samples. These limitations cannot be sufficiently mitigated by simple instance selection or attention weight adjustment alone.

2.2 Data Mining

In the field of data mining, representative approaches to address imbalanced learning are broadly categorized into over-sampling and under-sampling strategies. Over-sampling aims to increase the number of minority-class samples, and representative methods generate synthetic data through interpolation based on k-nearest neighbors (k-NN). In contrast, under-sampling adjusts class distributions by selectively removing samples from the majority class.

Video data inherently exhibit temporal structures, and artificial data augmentation processes may distort temporal continuity or dynamic patterns. Consequently, in video anomaly detection research, under-sampling strategies that remove existing samples are generally preferred over over-sampling methods that generate new data. This preference is particularly evident in Multi-Instance Learning (MIL) settings, where instance-level annotations are unavailable and only weak labels at the video level are provided. In this context, a snippet is defined as a short temporal segment extracted from a video and treated as an instance within a bag. Due to the absence of precise instance-level labels, simple random under-sampling is insufficient, and feature-based selective under-sampling strategies are required.

Motivated by this limitation, cluster-based under-sampling methods have been introduced. These approaches partition the data into multiple clusters using clustering algorithms such as k-means and compute the ratios between majority and minority classes within each cluster. Based on these ratios, majority-class samples are selectively removed to achieve a more balanced distribution. Gupta and Jivani further extended this idea by proposing a framework that quantitatively analyzes cluster-wise distributions and automatically adjusts the removal rate of majority-class samples to minimize information loss.

However, existing data-mining-based under-sampling techniques typically assume that explicit class labels are available for individual samples. This assumption does not hold in MIL-based video anomaly detection environments, where instance-level labels are absent and only weak video-level annotations are provided. Moreover, video data are temporally ordered time-series signals, making it difficult to directly apply conventional data mining techniques that rely on the independence assumption between samples.

Due to these constraints, MIL-based video anomaly detection requires alternative under-sampling strategies that do not directly depend on class ratios but instead analyze the underlying feature distributions. In particular, data reconstruction schemes that distinguish repetitive patterns from relatively rare patterns within normal data and regulate their influence during training can play a crucial role in improving the performance characteristics of MIL-based video anomaly detection models, especially in mitigating boundary ambiguity between normal and abnormal instances.

3. Proposed Method

In this study, we argue that the difficulty of MIL-based video anomaly detection (VAD) models in learning a clear normal–abnormal decision boundary is largely attributed to their inability to capture fine-grained distributional differences within normal data. In particular, such distributional variations can be observed in terms of feature norm statistics in the learned feature space, yet these characteristics are not effectively exploited in conventional MIL-based VAD frameworks.

To address this limitation, we propose a Feature Norm Clustering–based under-sampling framework designed to mitigate structural bias within normal data. By clustering normal samples according to their feature norm distributions and selectively regulating the influence of majority normal samples, the proposed method aims to alleviate learning bias near the normal–abnormal boundary.

3.1 Motivation

Conventional MIL-based video anomaly detection (VAD) methods utilize only video-level labels (normal = 0, abnormal = 1) and learn by selecting the most anomalous candidates at the snippet level. However, a snippet—defined as a short temporal

segment treated as an instance in MIL—is constructed via mean pooling over multiple clip-level features. As a result, informative but sparse patterns, such as a small number of normal or abnormal clips within a snippet, are easily diluted by the majority of normal clips.

This dilution effect leads to overlapping regions in the normal–abnormal feature space, making it difficult for MIL frameworks to learn a clear anomaly decision boundary. Consequently, it becomes necessary to structurally decompose the feature distribution of each video prior to snippet generation, in order to distinguish between majority normal and minority normal patterns. Such a process provides a prerequisite for effective under-sampling and boundary-aware learning in MIL-based VAD.

3.2 Overview

The overall procedure of the proposed framework is as follows. First, clip-level features are extracted from each video using a pre-trained feature extractor, I3D. For each video, K-means clustering is then performed based on the L2 norm of the clip-level features, decomposing the internal feature distribution of the video into KKK clusters.

Based on the clustering results, clusters are divided into two groups according to their sizes: clusters containing a relatively large number of features and those with lower frequencies. The former are referred to as majority normal clusters, while the latter are defined as minority normal clusters. This distinction is introduced to structurally separate repetitive patterns from relatively diverse patterns within normal data from the perspective of feature distributions.

During the training stage, clip-level features belonging to the majority normal clusters are selectively removed for each video. The remaining features are then rearranged while preserving their original temporal order. From this reconstructed clip feature sequence, snippets of the same length as those used in conventional approaches are generated. Finally, a MIL-based classifier is trained using the reconstructed snippets as input.

The proposed data reconstruction process does not alter the original video-level labels (normal or abnormal), and the overall model architecture remains identical to that of existing MIL frameworks. By adjusting the internal feature distribution of videos at the data level without modifying the model design, the proposed approach provides a data-centric strategy for influencing the formation of the normal–abnormal decision boundary.

3.3 Top Under-sampling

In this section, we define a Feature-level Top Under-sampling strategy that selectively removes clip-level features with low contribution to learning, based on the analysis of intra-video feature distributions. The proposed under-sampling approach does not modify the snippet partitioning scheme or the MIL training architecture; instead, it aims to reconstruct the input distribution for learning solely by adjusting the composition of the clip-level feature set.

For each video v, let the extracted clip-level feature set be $X^{(v)} = \{x_1, x_2, \dots, x_N\}(x_i \in R^d)$ To decompose the internal feature distribution of a video, K-means clustering is applied based on the L2 norm of clip-level features by iteratively updating cluster assignments and centers to minimize the following objective:

$$\min_{C_k^{(v)}{}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{x_i \in C_k^{(v)}} \left\| \|x_i\|_2 - \mu_k^{(v)} \right\|^2 \quad (1)$$

where $C_k^{(v)}$ denotes the set of clip-level features assigned to the k-th cluster of video v, and $\mu_k^{(v)}$ represents the mean of the feature norms within that cluster. The size of each cluster is defined as the number of features contained in the cluster.

Empirically, a skewed distribution is consistently observed across most videos, where a small number of clusters account for a substantial portion of the overall features. Based on this observation, clusters ranked at the top according to cluster size are designated as majority normal clusters, while the remaining clusters are defined as minority normal clusters. This distinction is not based on predefined action categories or semantic labels, but is determined solely by the statistical characteristics of the feature distribution.

In the Top Under-sampling stage, clip-level features belonging to the majority normal clusters are removed from the training data, while features associated with minority normal clusters are retained. This procedure mitigates the influence of dominant, repeatedly occurring patterns within normal data and allows less frequent and more dispersed features to have a greater impact during subsequent training stages. The process does not alter the original video-level labels (normal or abnormal) and is performed as a feature-level data refinement strategy in the preprocessing stage.

3.4 Snippet Reformation with Temporal Order Preservation

This section describes the snippet reformation process that enables the clip-level features remaining after Top Under-sampling to be used as input for MIL-based training. The proposed method

preserves the snippet partitioning length employed in conventional MIL-based video anomaly detection (VAD) frameworks, while differing in that the input feature set has been modified through under-sampling. This design allows the effect of input distribution reconstruction to be evaluated independently of changes to the MIL architecture or snippet segmentation protocol.

After Top Under-sampling, the clip-level feature set is filtered while preserving temporal order. Since certain clip features are removed during the under-sampling process, each resulting snippet maintains the original temporal ordering but may consist of a non-contiguous subsequence in which intermediate clips are missing. Thus, a snippet is no longer defined as a set of temporally adjacent frames, but rather as an ordered subset of clip-level features with preserved temporal sequence. This representation remains compatible with MIL-based learning, as the snippet is treated as a set-level instance rather than a strictly continuous temporal segment.

The filtered clip feature sequence is then divided into snippets of fixed length following the same partitioning scheme as in existing approaches. For each snippet S_j , mean pooling is applied to the included clip-level features to obtain the snippet representation:

$$x'_j = \frac{1}{|S_j|}\sum_{x_i \in S_j} x_i \tag{2}$$

where x'_jdenotes the representation of the reconstructed j-th snippet. Mean pooling is intentionally retained to isolate the effect of input feature distribution reconstruction, ensuring that any observed performance differences arise from the proposed data-level modification rather than changes in the pooling operation or MIL learning structure. Because the feature set S_j is constructed after removing features associated with over-represented clusters, the dilution of anomaly-adjacent features by repeatedly occurring normal patterns during averaging is mitigated, thereby facilitating clearer boundary formation between normal and abnormal representations.

3.5 MIL Training

The MIL training procedure using reconstructed snippets follows the same architecture as conventional MIL-based video anomaly detection (VAD) methods. Each video is annotated only with a video-level label $y \in \{0,1\}$, and no ground-truth labels are provided at the snippet level. In the MIL setting, the snippet with the highest anomaly score within a video is regarded as the representative instance for that video during training.

Specifically, the anomaly score of a reconstructed snippet x'_j is computed by a classifier $f(\cdot)$, and the prediction for a video v is defined as

$$\hat{y} = \max_j f\left(x'_j\right) \tag{3}$$

For MIL training, we employ a magnification-based MIL loss formulation commonly used in prior work. This loss consists of a video-level classification term and an additional feature magnification term designed to enhance the separability between normal and abnormal features. The overall loss function is defined as follows.

$$\mathcal{L} = \mathcal{L}_{MIL} + \lambda \mathcal{L}_{MAG} \tag{4}$$

where λ is a weighting factor that balances the two loss term

The MIL loss $\mathcal{L}_{MIL}$is defined as a binary classification loss between the video-level prediction $\hat{y}$and the ground − truth label y:

$$\mathcal{L}_{MIL} = -[\mathrm{y}\log\hat{y} + (1-\mathrm{y})\log(1-\hat{y})] \tag{5}$$

The magnification loss$\mathcal{L}_{MAG}$ is designed to encourage anomaly-relevant features to be sufficiently emphasized compared to normal features in abnormal videos. Let z_a denote the feature representation of the snippet with the highest anomaly score, and z_n denote the feature representation of a normal snippet. The magnification loss is defined as

$$\mathcal{L}_{MAG} = \max(0,\ \mathrm{m} - ||z_a||_2 + ||z_n||_2) \tag{6}$$

where m is a margin hyperparameter. This term enforces the norm of abnormal snippet features to be sufficiently larger than that of normal snippet features, thereby strengthening normal–abnormal separation in the feature space.

The key effect of the proposed Feature-level Top Under-sampling lies in adjusting the input feature distribution so that this max-based MIL loss operates more stably. When majority clusters are not removed, repeatedly occurring normal patterns may dominate the anomaly score distribution, preventing abnormal snippets from being selected as the maximum-scoring instances. In such cases, the gradient of $\mathcal{L}_{MIL}$ is insufficiently propagated to abnormal snippets, and the effect of $\mathcal{L}_{MAG}$ is weakened due to the excessive influence of normal features.

In contrast, after the proposed reconstruction process, snippets

are primarily composed of features from minority clusters. As a result, normal videos tend to converge to low anomaly scores, while anomaly-relevant features are relatively emphasized in abnormal videos. Consequently, the probability that the snippet selected by the max-based aggregation contains genuine abnormal patterns increases, allowing the magnification-based loss to more effectively promote feature separation between normal and abnormal instances.

In summary, the proposed Top Under-sampling strategy adjusts the input feature distribution without modifying the MIL architecture or loss formulation. By doing so, it enables the core assumption commonly adopted in max-based MIL that an abnormal video contains at least one abnormal instance to be more faithfully reflected during training.

4. Experiment

This chapter evaluates the effectiveness of the proposed Feature Norm Clustering–based Under-sampling framework from three perspectives. First, a cluster distribution analysis is conducted to examine the structural characteristics of intra-video feature distributions. Second, the impact of different data refinement strategies on MIL-based learning performance is compared for both normal and abnormal videos. Third, the distributions of anomaly scores obtained from normal and abnormal videos are analyzed to visually investigate how the proposed method influences anomaly boundary formation.

All experiments are conducted using the same backbone (I3D) and the same MIL classifier. The model architecture and training conditions are kept fixed, while only the composition of the training data is altered. This experimental design aims to ensure that the observed performance differences arise from data distribution adjustment rather than changes in model design.

4.1 Cluster Distribution Analysis

For each video, Feature Norm Clustering is performed using the K-means algorithm with K = 10, and the distribution of clip-level features across clusters is analyzed. The value of K was chosen to provide a sufficient level of granularity to distinguish dense feature groups from more sparsely distributed ones, while maintaining a manageable clustering structure for subsequent analysis.

Across the dataset, it is observed that, in many videos, the top three clusters account for approximately 50% or more of the total features. The detailed distribution statistics are summarized in **Table 1**. Based on this empirical observation, the top clusters are defined as majority normal clusters in this study.

Table 1: Comparison of the Number of Clips in Top3 and Bot7 for Feature Norm Clustering–Based Data Refinement

Item	Normal Video (clips)	Abnormal Video(clips)
Top3	291,699	104,915
Bot7	298,901	92,601

Table 2: Performance Comparison of MIL-Based WSVAD under Different Top3/Bot7 Data Configurations (AUC/EER)

Item	AUC_O	EER_O	AUC_A	EER_A
Bot7-Top3	0.780	0.277	0.593	0.433
Top3-Top3	0.789	0.280	0.616	0.428
Base line	0.794	0.264	0.650	0.390
Top3-Bot7	0.811	0.243	0.663	0.378
Bot7-Bot7	0.812	0.245	0.663	0.374

Table 1 quantitatively summarizes this distribution by reporting the number of clip-level features belonging to the top three clusters (Top3) and the remaining seven clusters (Bot7) for both normal and abnormal videos. As shown in the table, a small number of clusters dominate the overall feature distribution, while the remaining clusters contain a comparable but more fragmented portion of the data.

In contrast, the remaining seven clusters contain relatively fewer features and tend to exhibit more diverse visual and motion characteristics. Accordingly, in normal videos, these clusters are referred to as minority normal clusters, while in abnormal videos, they are regarded as abnormal-relevant clusters with a higher likelihood of containing anomaly-related features.

These observations demonstrate the existence of intrinsic distributional heterogeneity (i.e., skewed distributions) within normal data. Moreover, they suggest that treating all normal snippets equally during MIL training may allow such structural bias to adversely affect the learning of the normal–abnormal decision boundary.

4.2 Effect of Data Refinement on MIL Performance

To analyze the impact of the proposed data refinement strategy, four training data configurations are constructed by selecting either the top three clusters (Top3) or the bottom seven clusters (Bot7) for normal and abnormal videos, respectively. Specifically, four training data combinations are evaluated. The Top3–Top3 combination uses majority clusters for both normal and abnormal videos. The Bot7–Bot7 combination uses minority clusters for both normal and abnormal videos. The Top3–Bot7

combination uses majority clusters for normal videos and abnormal-relevant clusters for abnormal videos. The Bot7–Top3 combination uses minority clusters for normal videos and majority clusters for abnormal videos.

Table 2 reports the quantitative performance of each configuration in terms of overall test AUC (AUC_O), abnormal-video (AUC_A), and the corresponding EER values. For each configuration, the performance is compared against the baseline model trained on the original dataset. The results indicate that different data compositions lead to distinct performance characteristics. The Top3–Top3 configuration yields the lowest overall performance, as it relies primarily on repetitive and less informative majority patterns in both normal and abnormal videos. This result suggests that majority normal patterns provide limited discriminative information for anomaly detection.

In contrast, the Bot7–Bot7 configuration improves both AUC_O and AUC_A relative to the baseline by leveraging clusters with higher diversity in both normal and abnormal videos. This finding implies that minority normal clusters and abnormal-relevant clusters play a more critical role in forming the normal–abnormal decision boundary.

The Top3–Bot7 configuration demonstrates an improvement in overall AUC by maintaining stability in normal data while

emphasizing abnormal-relevant clusters. Conversely, the Bot7–Top3 configuration exhibits a trade-off: while sensitivity to abnormal videos increases, false positives also increase due to training that focuses on hard normal samples.

Figure 1 further illustrates these trends by presenting the ROC curves corresponding to different data refinement configurations. In particular, configurations that emphasize minority normal clusters and abnormal-relevant clusters exhibit consistently higher true positive rates in low-to-mid false positive rate regions, highlighting the effectiveness of the proposed data refinement strategy in modulating anomaly sensitivity and shaping the normal–abnormal decision boundary under a fixed MIL setting.

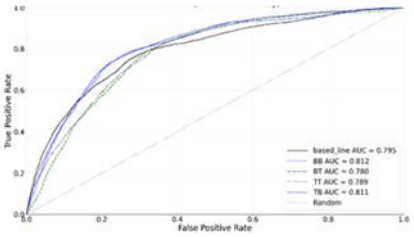

Figure 1: ROC Curve Comparison of MIL-Based WSVAD under Baseline and Top3/Bot7 Data Refinement Strategies

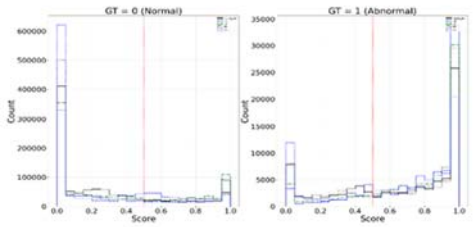

Figure 2: Anomaly Score Distributions for Normal (GT = 0) and Abnormal (GT = 1) Samples under Different Data Refinement Strategies

4.3 Analysis of Normal and Abnormal Score Distributions

Finally, anomaly scores obtained during testing are analyzed in histogram form by separating normal and abnormal videos for each experimental configuration. **Figure 2** visualizes the score distributions for normal (GT = 0) and abnormal (GT = 1) samples under different data refinement strategies.

In the baseline setting, the score distributions of normal and abnormal videos exhibit substantial overlap, reflecting ambiguity in the anomaly boundary. In particular, a large portion of normal samples receive intermediate anomaly scores rather than being confidently assigned low values, while some abnormal samples are not sufficiently separated toward high-score regions.

In contrast, the Bot7–Bot7 and Top3–Bot7 configurations show a clear shift in score distributions. Normal scores are concentrated toward lower ranges, while abnormal scores are more strongly pushed toward higher ranges, thereby reducing overlap between the two distributions. Notably, the Bot7–Bot7 configuration exhibits reduced variance and an increased median in abnormal scores, indicating improved stability and consistency in anomaly detection.

These distributional analyses, as illustrated in **Figure 2**, provide visual evidence that complements quantitative performance metrics such as AUC and EER. They demonstrate that the proposed data refinement strategy effectively mitigates boundary ambiguity by reshaping the score distributions of normal and abnormal samples from a feature distribution perspective.

5. Conclusion

This study investigated why MIL-based WSVAD often fails to form a clear normal–abnormal decision boundary in real-world surveillance settings. Beyond model architecture, we examined how intra-class heterogeneity within normal data, characterized by a major/minor normal structure, influences anomaly boundary formation and performance behavior. In particular, snippet

construction via mean pooling may dilute boundary-relevant patterns in normal-minority samples, potentially contributing to boundary ambiguity.

To reflect these distributional characteristics in a data-centric manner, we proposed a Feature Norm Clustering–based under-sampling framework. By clustering clip-level features based on their norm statistics, we observed that a small number of dominant clusters concentrate a substantial portion of features, while minor clusters tend to contain relatively rarer and more diverse patterns. Based on this observation, selectively reducing dominant clusters and reconstructing clip sequences prior to snippet formation can encourage the MIL classifier to place greater emphasis on hard-case normal samples and anomaly-relevant segments.

Experimental results under a fixed backbone and MIL classifier indicate that data composition significantly influences MIL learning dynamics. Configurations that reduce the influence of dominant clusters showed improved AUC/EER performance and reduced overlap between normal and abnormal score distributions compared to majority-focused training. These findings suggest that explicitly managing the major/minor normal structure through data refinement can contribute to mitigating boundary ambiguity in MIL-based WSVAD.

However, several limitations should be noted. First, the clustering strategy relies on norm-based feature statistics, which may not fully capture complex semantic relationships among snippets. Second, the effectiveness of the proposed under-sampling framework may depend on the specific feature extractor and dataset characteristics. Third, the current study focuses on fixed K clustering and static data refinement, without adaptive mechanisms that dynamically adjust to different video contexts.

Future research may explore more adaptive clustering criteria, alternative snippet aggregation strategies beyond mean pooling, and integration with end-to-end trainable data selection mechanisms. Additionally, validating the framework across diverse surveillance datasets and backbone architectures would further clarify its generalizability and practical applicability.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT). (RS-2024-00352187)

This work was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE). (RS-2024-00424595, Regional Residency Program for Cultivating Advanced Research Talent in Next-Generation Marine Mobility Industry Innovation).

Author Contributions

Conceptualization, D. H. Seo; Methodology, J. S. Lee; Software, J. S. Lee; Formal Analysis, J. S. Lee; Investigation, J. S. Lee; Resources, D. H. Seo; Data Curation J. S. Lee; Writing-Original Draft Preparation, J. S. Lee; Writing-Review & Editing, D. H. Seo; Visualization, J. S. Lee; Supervision, D. H. Seo; Project Administration, D. H. Seo; Funding Acquisition, D. H. Seo.

References

[1] 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 AAAI Conference on Artificial Intelligence, vol. 38, no. 6, pp. 6074-6082, 2024.

[2] M. Jiang, C. Hou, A. Zheng, X. Hu, S. Han, H. Huang, X. He, P. S. Yu, & Y. Zhao, “Weakly supervised anomaly detection: A survey,” arXiv preprint arXiv:2302.04549, 2023.

[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 Proceedings of the IEEE/CVF international conference on computer vision, pp. 4955-4966, 2021.

[4] M. Abdalla, S. Javed, M. Al Radi, A. Ulhaq, and N. Werghi, “Video anomaly detection in 10 years: A survey and outlook,” Neural Computing and Applications, vol. 37, pp. 26321-26364, 2025.

[5] S. Park, H. Kim, M. Kim, D. Kim, and K. Sohn, “Normality guided multiple instance learning for weakly supervised video anomaly detection,” In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2664-2673, 2023.

[6] H. Lv, Z. Yue, Q. Sun, B. Luo, Z. Cui, and H. Zhang, “Unbiased multiple instance learning for weakly supervised video anomaly detection,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8022-8031, 2023.

[7] C. Cao, X. Zhang, S. Zhang, P. Wang, and Y. Zhang, “Weakly supervised video anomaly detection based on cross-batch clustering guidance,” In 2023 IEEE

International Conference on Multimedia and Expo (ICME), pp. 2723-2728, 2023.

[8] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” Artificial intelligence, vol. 89, no. 1, pp. 31-71, 1997.

[9] O. Maron and A. L Ratan, “Multiple-instance learning for natural scene classification,” In Proceedings of the Fifteenth International Conference on Machine Learning, vol. 98, pp. 341-349, 1998.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” In Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, pp. 580-587, 2014.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.

[12] H. Kuehne, A. Richard, and J. Gall, “Weakly supervised learning of actions from transcripts,” Computer Vision and Image Understanding, vol. 163, pp. 78-89, 2017.

[13] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6479-6488, 2018.

[14] Y. Zhou, Y. Qu, X. Xu, F. Shen, J. Song, and H. T. Shen, “Batchnorm-based weakly supervised video anomaly detection,” IEEE Transactions on Circuits and Systems for Video Technology, 2024.

[15] X. Gao, D. Xie, Y. Zhang, Z. Wang, C. Chen, C. He, H. Yin & W. Zhang, “A comprehensive survey on imbalanced data learning,” arXiv preprint arXiv:2502.08960, 2025.

[16] M. Bekkar and T. A. Alitouche, “Imbalanced data learning approaches review,” International Journal of Data Mining and Knowledge Management Process, vol. 3, no. 4, pp. 15-33, 2013.

[17] S. Gupta and A. Jivani, “A cluster-based under-sampling solution for handling imbalanced data,” International Journal on Emerging Technologies, vol. 10, no. 4, pp. 160-170, 2019.

[18] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, “Clustering-based undersampling in class-imbalanced data,” Information Sciences, vol. 409, pp. 17-26, 2017.

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725-1732, 2014.

[20] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, “Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes,” Computer Vision and Image Understanding, vol. 172, pp. 88-97, 2018.