VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Contrastive 3D Multimodal Feature Fusion for Abnormal Behavior Recognition in Low-Light Conditions

Hoai-Duy Nguyen, Van-Dong Huynh, Cuong D. H. Tran, Tien-Dung Cao*

School of Information Technology, Tan Tao University, Tay Ninh, Vietnam

Received 25th February 2026
Revised 14th May 2026; Accepted 29th June 2026

Abstract: Surveillance in low-light conditions faces challenges of poor visibility and limited computational resources for deployment. This paper presents a 3D multimodal feature fusion framework with contrastive learning for abnormal human behavior recognition. Our approach extracts spatiotemporal features from visible and thermal videos, using a weighted assembly strategy to fuse the most informative regions. Contrastive learning pre-trains the backbone model to enhance recognition performance. Experiments on the Thermal-LLAB dataset demonstrate that the backbone AMFCFB model achieves a recognition accuracy of 96.03% and a detection AUC of 92.60%. Contrastive learning pre-training yields a 5.3 percentage point improvement in detection accuracy over training from scratch. These results confirm that combining thermal and visible modalities under contrastive pre-training yields a practical framework for 24/7 surveillance in challenging lighting conditions. Additionally, we release Thermal-LLAB (Low-Light Anomalous Behavior Dataset) — a new collection of synchronized visible and thermal videos capturing abnormal behaviors in low-light indoor and outdoor environments — to support future research.

Keywords: Abnormal Behavior Detection, Abnormal Behavior Recognition, Contrastive Learning, Multimodal Feature Fusion, Thermal and Visible Images, Low-light Surveillance Video.

1. Introduction

The detection and recognition of abnormal human behaviors are crucial for threat prevention in public spaces. To achieve this, a lightweight, real-time monitoring system is essential. Edge computing platforms, which process data locally rather than in centralized data centers, enable real-time responses with reduced latency—critical for security applications. Additionally, local processing enhances data privacy and security while offering cost-effective deployment across diverse surveillance scenarios. Abnormal behaviors are deviations from typical human activities [1], such as fighting, running, sneaking, and leaving bags unattended. This process involves two phases: detection identifies abnormal events as the positive class against

*Corresponding author.
E-mail address: dung.cao@ttu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.7053

normal actions, while recognition categorizes these events into specific behavior classes.

In recent years, deep learning has emerged as a powerful technique for automatically extracting meaningful features from surveillance data [2, 3]. Deep learning models excel at detecting abnormal behaviors in video surveillance, but face significant challenges in low-light conditions. Poor visibility and noisy, grainy image quality make it difficult to distinguish human figures and their subtle details, while cluttered backgrounds created by shadows and occlusions can lead to misinterpretation. The reduced color information in low-light scenarios further complicates accurate identification and classification of human behaviors. Traditional methods, optimized for well-lit environments and visible cameras, often struggle to maintain reliable performance under these challenging conditions.

Thermal imaging offers a significant advantage in low-light surveillance by detecting human heat signatures without visible light. While combining thermal and visible imaging could enhance recognition accuracy, current fusion methods primarily focus on 2D inputs and lack temporal information [4–8]. Though optical flow can provide motion data [9], it struggles with noise and low-light conditions while being computationally intensive. The challenge is further complicated by the inherent class imbalance between normal and abnormal events in training datasets, making accurate behavior recognition highly context-dependent.

This paper proposes a 3D multimodal feature fusion framework with contrastive learning for abnormal human behavior recognition in low-light environments. To address the challenge of acquiring multimodal data, we collected Thermal-LLAB (Low-Light Anomalous Behavior Dataset) containing synchronized thermal and visible videos. Our 3D fusion module extracts spatiotemporal features from both modalities and employs a weighted assembly strategy to identify the most informative regions for generating fused features. We leverage resource-efficient 3D CNNs [10] as feature extractors, combined with contrastive learning [11] to address data imbalance challenges. The main contributions of this paper are:

- A 3D multimodal feature fusion approach with contrastive learning that effectively combines thermal and visible features through efficient spatiotemporal extraction. This two-stage learning process first develops a pre-trained model using contrastive loss, then leverages this knowledge to fine-tune the recognition model, improving accuracy while reducing training time.
- Comprehensive experimental evaluation demonstrating effectiveness across various scenarios[1]: comparison with and without transfer learning, evaluation with and without multimodal fusion, and assessment using different loss functions.
- Our dataset[2] Thermal-LLAB, comprising synchronized thermal and visible videos of abnormal behaviors in low-light indoor and outdoor environments to support future research.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 introduces the proposed framework. Section 4 describes the data collection and implementation. Section 5 presents the experimental results. Section 6 reports ablation studies. Section 7 discusses the findings. Section 8 identifies the

[1]The source code is available at: https://github.com/DarkMagician000/Abnormal_CV

[2]The dataset is publicly available at: https://drive.google.com/drive/folders/1XXyXcESundsKvnDWyJTMBK2ZRv97U_iv?usp=drive_link

limitations of the current work. Section 9 outlines future research directions. Finally, Section 10 concludes the paper.

2. Related Works

2.1. Human Abnormal Behavior Detection

Human abnormal behavior detection has been significantly advanced by deep learning. Early unsupervised methods [12, 13] learned only normal patterns, flagging anomalies via high reconstruction error using Autoencoders (AEs) and Generative Adversarial Networks (GANs). Recent approaches [14, 15] integrate object detection for targeted analysis. For instance, STATE [14] combines object-level reconstruction with optical flow, while DSM-Net [15] uses a dual-stream memory architecture. A key limitation of these reconstruction-based methods, however, is their performance degradation in cluttered, low-light environments with high noise, where accurately modeling normal behavior becomes challenging

In parallel, research on supervised models has leveraged 3D CNNs to capture essential spatiotemporal features. Chen et al. [16] introduced a framework using 3D dense connections and a multi-instance learning (MIL) strategy to address weakly-labeled data, treating a video as a ”bag” of segments. Their model employs a ranking loss with sparsity constraints for better temporal localization and an adaptive soft-threshold mechanism for feature denoising, achieving high accuracy on benchmarks like UCF-Crime. However, such dense architectures remain computationally intensive for real-time edge deployment. Furthermore, their reliance on visible-spectrum data makes them vulnerable to performance degradation in low-light conditions

2.2. Human Abnormal Behavior Recognition

For this recognition task, the goal of behavior recognition is to classify human activities into specific predefined categories. Early approaches often relied on single data modalities. For instance, Ahn et al. [17] proposed the STAR-transformer to integrate spatio-temporal video and skeleton features via cross-modal attention. More recently, Munsif et al. [18] introduced AIR-Net, which leverages contextual features and attention mechanisms for action recognition in infrared videos.

A significant advancement in this domain is the use of 3D Convolutional Neural Networks (CNNs), such as C3D [19] and [20], which excel at learning spatiotemporal features directly from video sequences. While these models offer superior performance by capturing motion and appearance jointly, their high computational and memory requirements often render them impractical for real-time surveillance on resource-constrained devices.

To address the specific challenges of infrared imagery, such as blurry textures and low resolution, recent work has focused on sophisticated fusion architectures. Han et al. [21] proposed a Multilevel Features Cascade Fusion (MFCF) network that processes raw infrared video in an end-to-end manner. Their method utilizes an R(2+1)D network as its backbone, effectively decomposing 3D spatiotemporal feature learning into separate spatial and temporal modeling steps for more efficient processing. This foundation is enhanced by a Non-local attention module to capture global context and a deepwise features fusion module to cascade and refine features from multiple network levels, achieving state-of-the-art accuracy on infrared benchmarks.

Additionally, research has expanded into multi-modal frameworks to address complex, real-world scenarios. Tsiktsiris et al. [22] extended the SlowFast architecture for abnormal event detection in public transportation by incorporating RGB, depth, and audio modalities. Their approach employs dedicated fast pathways for depth and audio to capture rapid spatial variations and auditory cues, demonstrating that

multi-modal fusion can significantly enhance detection accuracy in dynamic environments.

2.3. Multimodal Fusion

Multimodal image fusion has proven to be effective for video anomaly detection and recognition [23]. This technique combines data from various sources or modalities, such as RGB, infrared images, optical flow, and audio. Deep learning-based approaches, including CNNs and RNNs, are commonly employed for multimodal image fusion [24, 25]. For instance, SeAFusion [4] leverages semantic information in the gradient residual dense block to learn fine-grained details from infrared and visible images. Ma et al. [5] have proposed long-range learning techniques across domains for the Swin Transformer, allowing the incorporation of information from distant regions of input images, which is crucial to capture relevant features in the fused image. Tang et al. [6] introduced the Y-shape Dynamic Transformer (YDTR) to extract both local and global features from infrared and visible images, enhancing the production of high-quality fusion. Additionally, Wei et al. [26] proposed the Multimodal Supervise-Attention Enhanced Fusion (MSAF) framework, a versatile tool for video anomaly detection capable of harnessing various modalities, including RGB, optical flow, and audio. Recently, Liu et al. [27] proposed HATF, a cross-attention Transformer framework for infrared and visible image fusion. The framework employs residual U-Nets with mixed receptive fields to extract multi-scale features, followed by a cross-attention Transformer and hybrid attention fusion to integrate complementary information from different modalities. Adaptive loss functions are then used to optimize the fusion of these multimodal features.

In comparison, Nguyen and Kong [9] proposed multimodal feature fusion (MFF) extracted from the C3D model along with optical flow information for both thermal and visible images to recognize abnormal human behaviors. This approach is ill-suited for real-time surveillance: computing dense optical flow alongside C3D features introduces substantial latency that exceeds practical deployment budgets on edge hardware. Instead, our framework uses resource-efficient 3D networks as feature extractors, keeping inference lightweight while retaining full spatiotemporal representation. Critically, the fused features feed directly into a contrastive learning objective that exploits the class structure of the training data — separating normal from abnormal embeddings — which is particularly effective under the severe class imbalance typical of surveillance datasets, where normal segments far outnumber abnormal ones.

Although Jiang et al. [28] proposed a multimodal fusion network focused on object detection from visible and thermal infrared images rather than abnormal behavior detection, their approach can serve as the foundation for human detection in our proposed system.

2.4. Contrastive Learning

Contrastive learning was first introduced in [29] with the principle of using positive pairs in contrast with negative pairs. The objective is to maximize the similarity among positive pairs while minimizing the similarity between positive and negative pairs. The loss can be extended to a fully supervised setting [11], enabling the model to effectively utilize label information during training. Contrastive learning has recently been applied in video anomaly detection and recognition. Zheng et al. [30] proposed contrastive learning on a graph neural network to construct different contextual subgraphs for video anomaly detection. Kopuklu et al. [31] applied supervised contrastive learning to detect anomalous driving behaviors. Sun et al. [32] introduced a novel hierarchical semantic contrast strategy that leverages both scene-level and object-level contrastive learning to enhance video

anomaly detection. The method enforces encoded latent features to maintain compactness within semantic classes while ensuring separability between different classes, effectively improving the discrimination of normal patterns. Hoang et al. [33] used both knowledge distillation and contrastive learning to recognize abnormal human behaviors in surveillance videos.

3. Proposed Framework

This section outlines our framework for detecting and recognizing abnormal human behaviors using 3D multimodal feature fusion and contrastive learning. As shown in Figure 1, our framework consists of three phases: i) training a pre-train model for abnormal behavior detection using the contrastive learning loss; ii) fine-tuning the pre-train model for abnormal behavior recognition using the Softmax cross-entropy loss; and iii) inferring the abnormal behavior detection by defining the abnormal score. Both detection and recognition models share the same architecture of 3D feature extractors and 3D multimodal feature fusion, as well as taking human segments from thermal and visible sources as input data. During pre-training, the contrastive loss pulls normal segments closer together in the embedding space while pushing them away from abnormal ones — a formulation that suits surveillance data well, given that normal behavior is consistent and repeatable whereas abnormal events are varied and unpredictable.

The fusion process involves two main phases: weight map generation and fusion. In weight map generation, separate weight maps (M_v, M_t) are generated for each source, leveraging attentive information from each modality. These weight maps serve as filters, extracting essential abstract features from F_v and F_t while eliminating irrelevant ones. The fusion step combines the 3D feature maps and their respective weight maps (M_v, M_t) through weighted assembly, determining the regions where thermal or visible information is more informative. The resulting 3D fused feature map (F_{fused}) encapsulates the most prominent abstract features from each modality, serving as the input for training the detection model with contrastive loss. After pre-training the detection model, the training normal segments are fed into the pre-trained model to generate the normal template vector v_n, which is used to calculate similarity scores with the test video segments for abnormal event detection. Subsequently, pre-trained detection weights are further employed to fine-tune the recognition model for classification head C to categorize distinct abnormal behaviors. The information learned from contrastive learning aids downstream supervised learning for the recognition model [34], resulting in accelerated convergence and enhanced training. The subsequent sections provide comprehensive details on the system's core components: the 3D feature extraction architecture, multimodal fusion mechanisms for combining thermal and visible data, and the complete training and inference pipelines.

3.1. 3D Feature Extractors

3D CNNs offer advantages in capturing both spatial and temporal information simultaneously using 3D kernels [20]. This enables them to analyze complex motion patterns, temporal dynamics, and object interactions over time, making them well-suited for tasks like behavior recognition. By learning hierarchical spatiotemporal features, 3D CNNs can detect simple elements (e.g., edges and textures) in early layers and more complex patterns (e.g., object motion and interactions) in deeper layers, leading to improved performance on complex video recognition tasks.

Our proposed framework utilizes 3D CNNs as feature extractors for each modality, taking a 4-dimensional input size of $C \times L \times W \times H$, where C denotes the image channels, L is the sequence length, and W and H are the width and height of

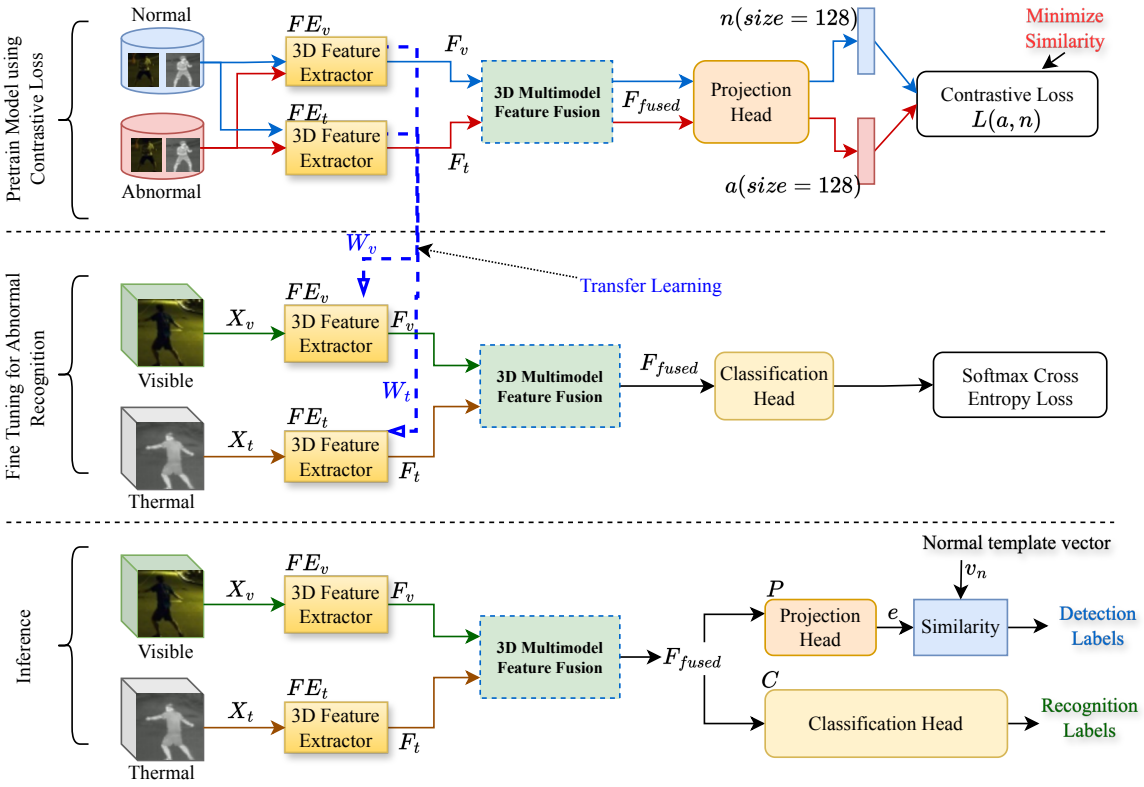

Figure 1. A schematic diagram of our proposed approach. The detection and recognition models share the same architecture. Pre-trained weights (W_v, W_t) for each feature extractor (FE_v, FE_t) are initially acquired through contrastive training for the detection model and are subsequently fine-tuned for the recognition model.

the video segment. Specifically, the architecture consists of two feature modules: FE_v for the visible input image and FE_t for the thermal input image. The extracted 3D feature maps F_v and F_t are expressed by (Eq. 1):

$$F_s = FE_s(X_s), s = [v, t] \quad (1)$$

where X_v and X_t denote the input sources for the feature extractors FE_v and FE_t, respectively.

3.2. 3D Multimodal Feature Fusion

The proposed 3D multimodal feature fusion module enhances human appearance and motion analysis by integrating thermal and visible inputs, focusing on both spatial and temporal information in videos. Unlike traditional 2D fusion methods [4–8], this module captures dynamic patterns and motion cues crucial for tasks like behavior recognition. Fusing the two modalities in a 3D space allows the module to jointly encode motion, appearance, and temporal relationships — information that 2D methods discard by processing each frame in isolation.

Instead of treating each 3D feature map equally, our proposed module uses a weighted assembling strategy that estimates a fusion weight for each pixel to determine which modality is more informative in a given region. Thermal and visible features may play different roles in different areas of the frames — for instance, thermal is more reliable in dark regions while visible provides finer texture detail in well-lit areas — and our goal is to retain the most discriminative signal from each. This module

therefore captures spatiotemporal information directly from the 3D input feature maps, including complex and abstract features, which reduces the impact of modality-specific noise on the final representation. Figure 2 illustrates our 3D multimodal feature fusion comprising two main phases: weight map generation and fusion.

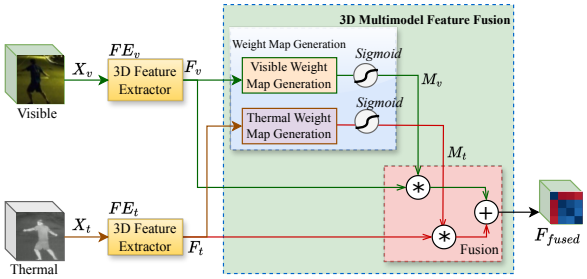

Figure 2. The proposed 3D multimodal feature fusion. The fusion module includes two phases: weight map generation and fusion.

In the weight map generation phase, the extracted feature vectors F_v and F_t from the 3D feature extractors are fed into the fusion module to establish distinctive maps (A_v, A_t) for thermal and visible sources, respectively. Given two 3D feature maps $F_v \in \mathbb{R}^{(C\times L\times W\times H)}$ and $F_t \in \mathbb{R}^{(C\times L\times W\times H)}$, their corresponding focus maps are computed across the channel dimension by (Eq. 2)

$$A_s = \sum_{c=1}^{C} \alpha_c^s |F_s|, s = [v, t] \tag{2}$$

where (F_v, F_t) represents the input 3D feature maps from different sources, (A_v, A_t) are the attention distinctive maps, and (α_c^v, α_c^t) are the per-channel weights calculated by (Eq. 3).

$$\alpha_c^s = \frac{1}{L\times W\times H} \sum_{k=1}^{L}\sum_{i=1}^{W}\sum_{j=1}^{H} F_s(c,k,i,j), s = [v, t] \tag{3}$$

The distinctive attention maps (A_v, A_t) are calculated based on their discriminative representation, indicated by the (α_c^v, α_c^t) value applied to the feature maps (F_v, F_t). Global contextual information is gathered through global average pooling across spatial dimensions, aiding the network in obtaining discriminative features and producing improved fusion results. In the next step, the weight maps M for each source are calculated by (Eq. 4).

$$M_s = \frac{A_s}{\sum A_s}, s = [v, t] \tag{4}$$

The Sigmoid function is then applied to normalize the weight maps (M_v, M_t) to a consistent range, ensuring that neither modality dominates simply due to differences in feature magnitude rather than actual informativeness. The normalized maps act as soft masks: they scale each modality's contribution region-by-region before the final weighted sum is computed. In the fusion phase, the fused vector $F_{fused} \in \mathbb{R}^{(C\times L\times W\times H)}$ is computed as the weighted average of input feature maps (F_v, F_t) and their corresponding weight maps through Eq. 5:

$$F_{fused} = M_v * F_v + M_t * F_t \tag{5}$$

where $*$ denotes element-wise product of two matrices. The resulting F_{fused} serves as the input to the contrastive detection model. In low-light conditions, visible features degrade with ambient lighting while thermal features remain stable; by weighting each modality's contribution per region, the fusion explicitly exploits this complementarity rather than assuming both sources are equally reliable.

3.3. Training Models

3.3.1. Training a pre-train model

The objective of this pre-train model is to acquire the weights of the 3D feature extractors employed for extracting feature maps from both abnormal (X^a) and normal (X^n) human segments, as depicted in the contrastive training phase of Figure 1, that will be used later for the detection model. Given the abnormal and normal human segments from each modality,

the 3D feature extractors generate 3D feature maps (F_v, F_t) from both visible and thermal sources. These extracted 3D feature maps (F_v, F_t) for each type of human segment are then fed into the 3D multimodal feature fusion module to produce the 3D fused feature map (F_{fused}) combining essential information from each modality. Specifically, the weight maps (M_v, M_t) are computed inside the fusion module and multiplied with the extracted features (F_v, F_t) to form the fused vector (F_{fused}).

The projection head P maps the 3D fused feature map (F_{fused}) to another latent space vector e. The latent space is of lower dimensionality, typically set as a default 128-dimensional space, where the fused feature vectors can be compared with other fused feature vectors to learn useful representations. The projection head P is implemented as a Multilayer Perceptron (MLP) network with a ReLU activation function. The latent feature vector e is then normalized to a comparable range using l_2-regularization. The latent feature vector e is computed at projection head P by (Eq. 6).

$$e_c = P(F^c_{fused}), c = [n, a] \tag{6}$$

where n, a represents normal and abnormal human segments, respectively; P is the projection head and F_{fused} is the fused feature map of the input. The supervised contrastive loss is used to ensure that normalized embeddings obtained from normal behaviors are more similar to each other compared to embeddings obtained from abnormal behaviors. The training dataset contains both normal and abnormal video segments. For each mini-batch, we selected K normal video segments and M abnormal video segments with the segment index $i = 1, 2, ..., K + M$. The corresponding embedding latent vector at index i is denoted as n_i for the normal segment and a_i for the abnormal segment. During training, there are $K(K-1)$ positive pairs and KM negative pairs fed into the detection model. We construct positive pairs using only normal samples, which reflects the nature of anomaly detection: since abnormal behaviors are unpredictable and can take countless forms, defining a fixed set of anomalous positive pairs would be both impractical and incomplete. The model uses the back-propagation algorithm to minimize the contrastive loss L defined in (Eq. 7) and (Eq. 8).

$$L_{ij} = -log \frac{exp(n_i^T n_j/\tau)}{exp(n_i^T n_j/\tau) + \sum_{m=1}^{M} exp(n_i^T a_m/\tau)} \tag{7}$$

$$L = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{K} L_{ij} | i \neq j \tag{8}$$

where $\tau \in (0, 1)$ is a scalar temperature that controls the similarity between samples. The dot product $(n_i^T n_j)$ in the numerator of (Eq. 7) measures similarity between normal pairs, while the denominator accumulates dissimilarity terms $(n_i^T a_m)$ between each normal segment and all abnormal segments in the mini-batch. Minimizing (Eq. 8) jointly with the 3D feature extractors and fusion module tightens the cluster of normal embeddings while increasing their separation from abnormal ones across each training batch.

3.3.2. Fine-tuning the pre-train model for recognition

The recognition model and the detection model share the same architecture, including the 3D feature extractors and the 3D multimodal feature fusion module. Thus, the trained weights from the detection model are used as pre-training weights for the recognition model. The projection head P is now replaced by a classification head C that applies a softmax function to produce the probability distribution over all behavior classes, and this classification head is trained using the softmax cross-entropy loss function, which is minimized through back-propagation to adjust the weights and enable the model to learn discriminative features for accurate behavior classification.

3.4. Computing Abnormal Score

To compute the abnormal score, we used the pre-trained detection model to encode a set of 128-dimensional latent vectors e_i , where $i = 1, ..., N$, from N normal video segments of the training set. Then, the average of these l_2 normalized e_i is used to generate the normal template vector v_n via (Eq. 9).

$$v_n = \frac{1}{N} \sum_{i=1}^{N} \frac{e_i}{\|e_i\|_2} \tag{9}$$

where e_i is the i^{th} normal latent vector generated from N normal training video segments. The abnormal score of test video segments is computed by (Eq. 10), measuring Cosine similarity between its latent vector h_j and the normal template vector v_n. If the similarity score o_j is lower than a predefined threshold γ, then the test video segment is identified as abnormal. A higher abnormal score indicates that the testing segment deviates significantly from the typical behavior captured in the normal video segments, indicating the presence of abnormal behavior.

$$o_j = v_n^T \frac{h_j}{\|h_j\|_2} \tag{10}$$

where h_j is the 128-dimensional latent vector of each test video segment encoded by the trained detection model.

4. Data Collection and Framework Implementation

4.1. Data Collection and Processing

We constructed a multimodal image dataset, named Thermal-LLAB, featuring both thermal and visible imaging to recognize abnormal human behaviors. This dataset was independently collected by our team. The thermal videos have resolutions of 640×480 and 1280×720 at 25 FPS, captured using the device HIKMICRO Pocket2. Visible videos were recorded using two devices: a secondary smartphone Redmi 12C (providing 1920×1080 at 30 FPS) and iPhone 11 Pro Max (1920×1080 at 26 FPS). Scenarios for each anomaly action were thoughtfully defined in areas with limited and no lighting conditions. We categorized our scenarios into indoor and outdoor activities, where human subjects performed abnormal behaviors in five categories, as described in Table 1.

We analyzed the histogram of each visible video and determined if the visible video has a low-light level based on a predefined threshold (defaulting to 100) to the accumulated histogram. If more than 40% of the pixels have values below the threshold, it considers the video to have a low-light level. These are short-term abnormal human behaviors referring to unusual or atypical actions or reactions displayed by individuals over a short period of time. They are considered crucial for identifying potential dangers in public settings. Regarding temporal synchronization, both cameras (thermal and visible) were operated simultaneously during the same recording session. To ensure precise temporal alignment, we employed Adobe Premiere Pro to synchronize the two video streams using audio waveform alignment — a standard technique in professional video production that enables frame-accurate alignment at the millisecond level. Regarding spatial synchronization, the two cameras were physically co-mounted and oriented in the same direction throughout the entire data collection process, ensuring highly similar Fields of View between the two modalities. This setup minimizes angular misalignment between the thermal and visible streams, facilitating accurate cross-modal feature mapping within the fusion module.. We fused the visible videos and thermal videos using multimodal image fusion methods such as SeAFusion [4], SwinFusion [5], and YDTR [6]. Given the fused videos from each abnormal human behavior category, we annotated human object bounding boxes on the screen using the

Table 1. Overview of the newly collected abnormal behavior dataset

Category	Description	Outdoor				Indoor			
		# video	# frame	Duration (s)	Low-light Level (%)	# video	# frame	Duration (s)	Low-light Level (%)
Fighting	First Fighting	17	7128	285	86.0	30	14227	523	65.4
Running	Moving unusually fast	16	5159	206	96.2	40	18320	693	90.0
Biker	A person drives a bike	16	7130	264	79.4	-	-	-	-
Carrier	A person carries a bag/ backpack	32	11882	475	91.6	-	-	-	-
Bag Left Unattended	A person leaves a bag unattended	28	8631	330	86.9	46	19895	676	56.4
Total		109	39930	1562	88.39	116	52442	1893	70.32

CVAT annotation tool [35] and employed them for training YOLOv5n [36] to detect human targets in video segments. The predicted human bounding boxes for each fused video are then employed to localize the human location in the corresponding thermal and visible videos. The human bounding boxes extracted from video segments of both thermal and visible videos are resized to 112×112. The default temporal length is set to 16 out of 32 consecutive frames with a stride of 2, resulting in a final input with a 3D shape of $1 \times 16 \times 112 \times 112$. Figure 3 shows sample images for each of the five categories. Table 2 presents the number of human video segments in each category, where "Normal" behavior refers to expected or typical behavior within a given scenario, like standing or moving normally. The dataset follows a ratio 70:30 split between the training and testing sets.

Table 2. Number of human video segments for each category

Category	# video segments	
	Training Set	Testing Set
Normal	5686	2480
Fighting	565	243
Running	629	270
Biker	184	80
Carrier	332	143
Bags Left Unattended	704	302

4.2. Framework Implementation

We implemented our deep learning models using the PyTorch framework [37]. Our objective is to develop a lightweight surveillance system utilizing resource-efficient architectures for real-time video processing. To this end, we employ three distinctive architectures as our backbone feature extractors: 3D-ResNet18 [38], the Multilevel Features Cascade Fusion network (MFCF) [21], and the Global-Local Convolutional Autoencoder (AMFCFB) [39].

Unlike standard transfer learning approaches, we trained these feature extractors from scratch without relying on pre-trained weights from large-scale datasets (e.g., Kinetics). We initialized the network parameters using standard initialization techniques to ensure effective convergence. For all architectures, we exclude the final Average Pooling layer and Linear classification heads. This strategy forces the networks to learn descriptive and abstract 3D features directly from the target surveillance data, ensuring that the extracted representations are domain-specific and unbiased by external

Figure 3. Sample images in the Thermal-LLAB dataset. Top row: visible samples, Middle row: thermal samples, Bottom row: fused samples. Left to Right: (a) Fighting, (b) Running, (c) Bag Left Unattended, (d) Carrier, (e) Biker.

datasets. This approach facilitates robust multimodal feature fusion, thereby improving accuracy in detection and recognition tasks.

4.2.1. *Pre-Training Model for Anomaly Detection Task*

In the first stage of pre-training the anomaly detection model, the visible and thermal segments (X_v, X_t) are fed into the 3D feature extractors to generate individual 3D feature vectors (F_v, F_t) for each modality, resulting in an output feature size of $c \times 1 \times 1 \times 1$. Here, c signifies the number of output channels, specifically **512 for both 3D-ResNet18 and MFCF, and 1536 for AMFCFB**. Subsequently, our proposed 3D multimodal feature fusion module combines the extracted features to generate a 3D fused feature (F_{fused}) of the same size. The fused feature is then flattened and input into the projection head P. This module is structured as a Multilayer Perceptron (MLP) where the first linear layer maps the input dimension c to a hidden dimension of 256 units, followed by a ReLU activation function. A second linear layer then projects the hidden features to a 128-dimensional output vector. Crucially, **L2 normalization** is applied to the final output to project the embedding onto a unit hypersphere, which is essential for stabilizing the contrastive loss calculation. We employ Kaiming Normal initialization for the network weights to facilitate effective convergence. The resulting normalized latent feature e is utilized in the training phase as detailed in Section 3.3.1. We applied various image augmentations, including noise injection, rotation, horizontal flipping, and cropping, into the training process to enhance model robustness.

4.2.2. *Fine-Tuning for Recognition Task*

Upon completion of the pre-training with contrastive loss, the learned weights (W_v, W_t) of the 3D feature extractors are transferred to initialize the recognition model. To facilitate this transition, the projection head P is replaced with a classification head C comprising a Multilayer Perceptron (MLP). The structure of C adapts to the backbone: for the MFCF architecture, the first linear layer maps the input to a 256-dimensional hidden space, followed directly by a ReLU activation. Conversely, for ResNet and AMFCFB

architectures, a 1D Batch Normalization layer precedes the ReLU, and a Dropout layer (p = 0.4) is applied afterward to mitigate overfitting. Finally, the hidden features are projected to 6 output units (5 abnormal behaviors, 1 normal), and the model is fine-tuned using the Cross-Entropy loss.

5. Experiment Results

We conducted the training process for both models on an Ubuntu 22.04.5 LTS system, powered by an Intel(R) Xeon(R) W-3235 CPU @ 3.30GHz, 62GB of RAM, and two NVIDIA RTX A5000 GPUs, each equipped with 24GB of memory.

For the detection model (Stage 1), we trained it for 250 epochs using the SGD optimizer with an initial learning rate of 0.01, which was decayed by a factor of 10 every 20 epochs. The training batch size was set to 13. The temperature parameter (τ) for the contrastive loss was set to 0.1, and the weight decay was maintained at 1×10^{-4}. To address the data imbalance, the data loader was configured to sample a balanced mix of normal and abnormal segments during training.

For the recognition model (Stage 2), we initialized the backbone with the pre-trained weights from Stage 1 to leverage the learned spatiotemporal features. We tailored the fine-tuning strategy based on the backbone architecture to optimize feature adaptation. For the ResNet and AMFCFB models, we employed a 3-phase progressive unfreezing approach: initially training only the classification head, followed by partially unfreezing high-level layers to refine high-level features, and finally unfreezing the entire network for overall structural polishing. Conversely, for the MFCF backbone, the entire architecture was unfrozen from the start and trained end-to-end in a single phase.

5.1. Abnormal Human Behavior Detection

We evaluate the proposed framework on the Thermal-LLAB dataset under two experimental settings: unsupervised and supervised. In the unsupervised setting, we compare against STATE [14] and DSM-NET [15], both of which are reconstruction-based methods that model normal behavior patterns and flag deviations as anomalies. Both methods were trained on fused images generated by SwinFusion [5] to incorporate information from both visible and thermal modalities. STATE [14] extracts human bounding boxes using a pre-trained object detector and constructs spatiotemporal context cubes to train two reconstruction branches: STATE-raw for appearance reconstruction and STATE-motion for optical flow reconstruction. DSM-NET [15] employs a dual-stream memory module that separately reconstructs spatial appearance and temporal optical flow patches, identifying anomalies via high reconstruction error. Both methods were trained and evaluated following their original configurations.

In the supervised setting, we include MFF [9] as the primary multimodal baseline, as it is the most recent method evaluated on the MIHAB dataset. MFF uses two I3D [20] feature extractors to independently process thermal and visible streams, supplemented by optical flow features. All abnormal behavior categories are treated as a single positive class, with normal behavior as the negative class, consistent with standard anomaly detection protocols.

The quantitative results for abnormal event detection, summarized in Table 3, reveal several insights into the role of multimodal fusion and backbone architecture in this task.

Single-modality results highlight the advantage of thermal imaging under low-light conditions. The thermal-only MFCF configuration reaches an accuracy of 0.7892, compared to 0.7396 for the visible-only AMFCFB — a gap that directly reflects the signal degradation visible sensors suffer when

ambient lighting is poor, while thermal sensors continue to capture reliable heat signatures regardless of illumination.

Multimodal fusion consistently improves results across all backbones, with SeAFusion and SwinFusion both exceeding 0.90, confirming that thermal localization and visible texture are complementary rather than redundant.

A notable result is the strong detection performance of the AMFCFB backbone (0.9260) despite its relatively weaker recognition accuracy. This is consistent with its autoencoder design: the architecture learns to reconstruct normal patterns, so any deviation — regardless of anomaly type — produces a high reconstruction error that serves as a sensitive detection signal. The MFCF backbone, by contrast, uses R(2+1)D kernels optimized for classifying specific behavior labels, which makes it less sensitive to the subtle deviations that characterize anomaly detection.

MFF achieves the highest accuracy overall (0.9536), attributable to its use of Dual TV-L1 optical flow, which provides an explicit representation of motion magnitude and direction well suited to the high-intensity movements typical of anomalies. Our AMFCFB backbone reaches 0.9260 without computing dense optical flow, relying instead on a learned motion encoder — a trade-off that sacrifices a modest amount of accuracy in exchange for substantially lower computational overhead, which matters for real-time edge deployment.

5.2. Abnormal Human Behavior Recognition

Table 4 presents a comprehensive performance comparison between our proposed architectures and established baseline methods. A notable observation is the inconsistent performance of traditional 3D-CNN approaches across different anomaly categories. For instance, C3D struggles significantly with specific actions, achieving an F1-score of only 0.2500 for the ”Biker” class and 0.5882 for ”Running”. Similarly, while MFF shows improvements in some areas, it experiences a drastic performance drop in the ”Bag Left Unattended” category (0.4000). I3D delivers a strong baseline performance due to its pre-trained inflation from ImageNet, yet it still falls short of our proposed architectures in overall accuracy and struggles to dominate in highly dynamic classes. These fluctuations indicate that standard baselines often lack the robust spatiotemporal feature representation required to handle diverse and complex abnormal behaviors consistently.

In contrast, our proposed **AMFCFB Backbone** demonstrates superior robustness and generalization, achieving the highest overall accuracy of **0.9603**. Notably, it yields near-perfect F1-scores for the ”Normal” (0.9970) and ”Fighting” (0.9623) categories. This exceptional performance is a direct result of our two-stage training strategy, which effectively establishes a robust decision boundary between peaceful (”Normal”) and chaotic (”Abnormal”) behaviors. Furthermore, the AMFCFB architecture significantly outperforms all baselines in recognizing fast-paced anomalies, achieving remarkable F1-scores for ”Running” (0.9615) and ”Biker” (0.9500).

Interestingly, while the AMFCFB model excels globally, our **MFCF Backbone** proves highly effective in processing subtle human-object interactions. Specifically, for the ”Carrier” class, the MFCF Backbone achieves the highest F1-score of **0.8485** (compared to 0.6857 for AMFCFB and 0.7500 for I3D). This highlights the MFCF architecture’s specific capability to capture fine-grained local details, such as carried objects, which standard models frequently misclassify.

To further examine the class-wise behavior of the model, Figure 4 presents the confusion matrix of the AMFCFB backbone on the recognition task. The matrix confirms the patterns observed in the F1-score analysis: the Normal class achieves the highest accuracy with 496 out of 500 samples correctly classified, while dynamic

Table 3. The performance of the detection models on our dataset

Method			Input Type	Approach	Accuracy	Precision	Recall	F1-Score
Unsupervised			fused	STATE [14]	0.57	0.28	0.5	0.36
				DSM-Net [15]	0.72	0.56	0.51	0.46
Unsupervised			Visible	AMFCFB	0.7396	0.6965	0.6346	0.6447
Supervised			Thermal	MFCF	0.7892	0.824	0.6713	0.6906
Supervised	Single Feature Extractor		fused	SeAFusion [4]	0.9069	0.9091	0.8681	0.8849
				SwinFusion [5]	0.9000	0.9181	0.8465	0.8722
				YDTR [6]	0.8639	0.8568	0.8129	0.8299
	Dual I3D		Thermal + Visible	**MFF** [9]	**0.9536**	**0.9617**	**0.9372**	**0.9477**
	Proposed	Dual Feature Extractors	Thermal + Visible	AMFCFB backbone	0.9260	0.9377	0.8889	0.9085
				MFCF backbone	0.8167	0.8072	0.7417	0.7616
				3D-Resnet18	0.8931	0.9018	0.8434	0.8653

anomaly classes — Fighting (51/55 correct), Running (50/54 correct), and Biker (19/19) — consistently exhibit high recognition rates due to their distinctive spatiotemporal motion patterns. In line with the lower F1-score reported for the Carrier class, the confusion matrix reveals that 3 out of 31 samples were misclassified as Bag Left Unattended, highlighting the semantic similarity between these two classes. Similarly, the Bag Left Unattended class shows 14 out of 72 samples misclassified as Carrier, reflecting the temporal ambiguity of this anomaly type. These findings are consistent with the per-class F1-scores in Table 4 and reinforce the need for future improvements in distinguishing subtle, static anomaly classes.

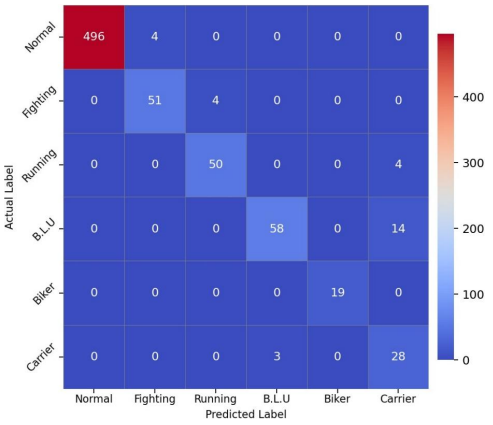

Figure 4. Confusion matrix of the AMFCFB model on the recognition task. The diagonal elements represent correctly classified samples, while off-diagonal elements indicate misclassification between behavior classes.

6. Ablation Studies

6.1. Transfer Learning

Transfer learning [40] is a pivotal technique in deep learning that initializes a model for a target task using weights pre-trained on a related source task. In our framework, during the pre-training phase of the detection model, the network learns to extract robust spatiotemporal features from the input data. By transferring these pre-trained weights to initialize the recognition model, we significantly accelerate convergence, mitigate overfitting, and enhance overall classification performance compared to training from scratch.

We conducted an ablation study to evaluate the impact of this transfer learning strategy. As detailed in Table 5, initializing the recognition models with pre-trained detection weights consistently improves accuracy across all architectures. Notably, the **MFCF Backbone**

Table 4. Performance comparison of recognition models on our dataset. Note: The AMFCFB Backbone achieves near-perfect F1-scores on Normal and Fighting classes due to the effective two-stage training strategy

Approach	Accuracy	F1-Score					
		Normal	Fighting	Running	Biker	Carrier	Bag Left Unattended
C3D [19]	0.8944	0.9639	0.6667	0.5882	0.2500	0.6000	0.8696
I3D [20]	0.9391	0.9845	0.8667	0.8387	0.8000	0.7500	**0.8980**
MFF [9]	0.8528	0.9549	0.7586	0.6316	0.8000	0.7619	0.4000
AMFCFB Backbone (Ours)	**0.9603**	**0.9970**	**0.9623**	**0.9615**	**0.9500**	0.6857	0.8345
MFCF Backbone (Ours)	0.9425	0.9881	0.9111	0.7872	0.8571	**0.8485**	0.8228
3D-ResNet18 (Ours)	0.8988	0.9760	0.7706	0.7216	0.6512	0.6301	0.7857

Table 5. The accuracy of recognition models with and without using the pre-trained detection weights

Model	With pre-trained weights	Without pre-trained weights
AMFCFB Backbone	**0.9603**	0.9535
MFCF Backbone	0.9425	0.8728
3D-ResNet18	0.8988	0.8974

exhibited the most substantial benefit, achieving an accuracy increase of approximately **7%** (improving from 0.8728 to 0.9425). Similarly, our proposed **AMFCFB Backbone** reached its peak accuracy of **0.9603** when initialized with these weights, outperforming its from-scratch counterpart (0.9535). Even the standard 3D-ResNet18 showed marginal gains.

These results strongly validate the efficacy of the contrastive learning scheme applied to the 3D fused features during Stage 1. By contrasting different human behaviors beforehand, the model pre-learns rich, highly representative feature embeddings. This prior knowledge directly translates to more accurate and stable abnormal event recognition, especially when dealing with complex or imbalanced datasets.

6.2. Training Detection Model without the 3D Multimodal Feature Fusion

In this ablation study, we performed an experiment on the detection model utilizing 3D-ResNet18 as feature extractors. To assess the impact of different fusion modules, we replaced our proposed 3D fusion module with the fusion module introduced in MFF [9]. Furthermore, we examined the scenario where the proposed 3D fusion module is omitted from the detection model, assuming an equal contribution of thermal and visual information from each pixel. The resulting fused feature F_{fused}, was computed as the summation of the extracted visible feature map (F_v) and the extracted thermal feature map (F_t) via Eq. 11.

Table 6. The accuracy of the detection model in various experiments

Type	Accuracy
Without 3D fusion module	0.8617
With 3D fusion module in MFF [9]	0.8883
With proposed 3D fusion module	**0.9260**

$$F_{fused} = \frac{F_v + F_t}{2} \tag{11}$$

Table 6 shows the accuracy comparison in different setups. Notably, our detection model, enhanced by the proposed fusion module, exhibited a remarkable 5.1% increase in accuracy compared to the MFF [9] fusion method and an impressive 8% improvement over the model without a fusion module. The effectiveness of our proposed 3D fusion module lies in its ability to integrate information from both types of images, maximizing the benefits derived from each. This confirmed our assumption that thermal

and visible variations might play distinct roles in different regions of the frames, and the proposed 3D fusion module attempts to aggregate the most useful characteristics from both inputs.

6.3. Supervised Contrastive Loss versus Cross Entropy Loss and Focal Loss

Cross entropy loss [41] and focal loss [42] are commonly used loss functions in deep learning for classification tasks. In this experiment, cross-entropy loss and focal loss are used in place of contrastive loss when training the detection model. We replaced the projection head with a fully connected layer to train the detection model featuring 3D-ResNet18 as feature extractors using cross-entropy loss and focal loss. Due to the unbalanced distribution of normal and abnormal data used in the training process, we used the weighted cross-entropy loss, where weights are computed based on inverse class frequency, to focus on the minority class of abnormal behavior. Focal loss is a modification of cross-entropy loss that is designed to address the class imbalance in the data distribution. Focal loss encourages the model to focus more on abnormal samples during training, which can lead to improved performance on abnormal event detection. The comparative results are illustrated in Table 7. The contrastive loss outperformed other losses in terms of detection accuracy with the highest value of 0.912. The cross-entropy loss was outperformed by focal loss, which had a performance of 0.859 versus 0.880. This confirmed that the contrastive loss is appropriate for learning representations that are robust to variations in the data.

6.4. Running Performance in Different Devices

Table 8 summarizes the computational complexity and memory footprint of all evaluated models. Among the proposed architectures, AMFCFB is the most expressive, with 295.71M parameters and 234.35 GFLOPs per clip, reflecting its dual-backbone design that jointly processes thermal and visible modalities. Despite this higher complexity, AMFCFB achieves 96.03% accuracy (Table 4), representing a consistent 2.12% improvement over I3D (12.29M parameters, 111.51 GFLOPs), which demonstrates that the additional capacity is well justified by the performance gain. In contrast, MFCF offers a lightweight alternative with only 57.49M parameters, 19.27 GFLOPs per clip, and a memory footprint of 219.3 MB — the smallest among all compared models — making it particularly suitable for resource-constrained deployment.

Table 7. Accuracy of detection model with 3D-ResNet18 feature extractors using different loss functions

Loss Type	Overall Accuracy
Cross-Entropy Loss	0.859
Focal Loss	0.880
Contrastive Loss	**0.912**

Table 8. The complexity and efficiency of Models

Model	Complexity		Memory
	Parameters (M)	FLOPs / clip (G)	Memory (MB)
C3D	78.02	38.55	1323.2
I3D	12.29	111.51	546.6
MFF	32.79	164.69	738.2
AMFCFB	295.71	234.35	1400.9
MFCF	57.49	19.27	219.3

Table 9 presents inference speeds measured on three hardware platforms. AMFCFB achieves 10.06 FPS on the Jetson Orin Nano (4GB), 6.43 FPS on a laptop CPU(Intel Core i7-10750H), and 145.09 FPS on a desktop GPU (NVIDIA GTX 1080 Ti), confirming its practicality for real-time action recognition on both edge and server-grade hardware. MFCF, owing to its compact design, achieves the highest throughput across all platforms: 38.43 FPS on Jetson, 21.83 FPS on CPU, and 656.60 FPS on GPU, making it the

preferred choice when inference latency is the primary constraint.

Table 9. Inference Speed Across Different Devices

Device	FPS				
	C3D	I3D	MFF	**AMFCFB**	**MFCF**
Jetson	25.35	25.34	2.9	10.06	38.43
Laptop CPU	63.93	32.71	16.55	6.43	21.83
GTX 1080Ti	555.05	1241.4	166.89	145.09	656.6

7. Discussion

7.1. Performance Disparity Across Behavior Classes

The experimental results reveal a notable performance gap between dynamic and static anomaly classes. Dynamic behaviors such as *Fighting* (F1: 0.9623) and *Running* (F1: 0.9615) produce strong, consistent motion signals and the resulting embeddings are well separated in the latent space. In contrast, the *Carrier* class (F1: 0.6857 for AMFCFB) is visually almost identical to normal walking; the only distinguishing cue is the presence of a small carried object, which occupies a limited spatial region and produces no distinctive motion signal on its own. *Bag Left Unattended* (F1: 0.8345) was falling into the problem of defining moment — the person walking away and leaving the bag — this action may be outside the fixed 16-frame window of our experiments, so the segment the model actually sees may contain no clear anomalous signal at all.

7.2. Open-Set Detection Capability

An important property of the two-stage pipeline is that the detection stage has a degree of open-set capability that the recognition stage does not. The contrastive detection model learns what normal behavior looks like in the embedding space; at inference, any segment whose cosine similarity to the normal template vector v_n falls below threshold θ is flagged as abnormal, without needing to know what type of anomaly it is. As a result, previously unseen events such as fainting or seizures may still be caught at the detection stage even though the recognition stage — a closed-set classifier trained on five fixed categories — cannot assign them a meaningful label. This distinction is worth stating explicitly: the two stages offer different coverage, and neither alone represents the full capability of the framework.

7.3. Modality Contribution Analysis

Table 3 shows a consistent advantage for thermal imaging over visible-only input under low-light conditions: the thermal-only MFCF configuration (0.7892) outperforms the visible-only AMFCFB (0.7396) by nearly 5 percentage points. This gap reflects a straightforward physical constraint — visible sensors lose signal when ambient light drops, while thermal sensors capture heat signatures regardless of illumination. When both modalities are combined, performance improves further across all backbones, which confirms that the two sources carry complementary rather than redundant information.

8. Limitations

8.1. Dataset Scale and Diversity

The Thermal-LLAB dataset, while newly introduced, comprises only 225 videos across five behavioral categories. Significant class imbalance exists within the dataset — for instance, the *Biker* category contains only 184 training segments compared to 5,686 for the *Normal* class. Furthermore, certain categories such as *Biker* and *Carrier* were captured exclusively in outdoor environments, restricting scene diversity. The reported recognition accuracy of 96.03% was evaluated strictly on the test split defined by this benchmark and may not fully reflect generalization to more diverse real-world scenarios.

8.2. Closed-Set Recognition

The recognition stage of the proposed framework operates as a closed-set classifier trained on five predefined anomaly categories. In real-world deployment, abnormal behavior is an open-set and ill-defined concept — events such as fainting, strokes, or other unforeseen anomalies can readily occur beyond the scope of the current category set. While the contrastive detection stage may flag such unseen events as anomalous (see Section 7), the recognition stage will fail to assign them a meaningful category label.

8.3. Multimodal Synchronization in Large-Scale Deployment

Although the Thermal-LLAB data collection ensured tight temporal and spatial synchronization between visible and thermal cameras — using audio waveform alignment in Adobe Premiere Pro and fixed physical co-mounting — maintaining such synchronization in large-scale real-world deployment may be challenging due to network latency or frame-rate mismatches between heterogeneous camera systems.

8.4. Thermal Sensitivity to Ambient Temperature

Thermal cameras are inherently sensitive to ambient temperature conditions. In environments near heat sources, under intense direct sunlight, or in industrial settings with elevated surface temperatures, thermal features can become significantly noisy, making human heat signatures difficult to distinguish from environmental thermal backgrounds. The Thermal-LLAB dataset was collected under controlled low-light conditions, and extreme ambient temperature scenarios were not included in our evaluation.

9. Future Work

Several promising directions emerge from the limitations identified above.

9.1. Dataset Expansion

We are going focus on expanding the Thermal-LLAB dataset to include a broader range of behavioral categories (e.g., fainting, vandalism, loitering), diverse indoor and outdoor scenes, and varied ambient temperature conditions. Evaluation on established international benchmarks such as UCF-Crime [43] would further validate the generalizability of the framework.

9.2. Improving Performance on Subtle Anomaly Classes.

To address the lower F1-scores on the *Carrier* and *Bag Left Unattended* classes, further research will investigate: (i) integrating lightweight object detection modules to explicitly encode carried-object cues, and (ii) incorporating long-range temporal modeling via LSTM or Transformer-based sequence encoders to capture cause-and-effect patterns that span beyond fixed 16-frame segments.

9.3. Open-Set Recognition.

The recognition stage will be extended with open-set recognition or out-of-distribution (OOD) detection techniques — such as OpenMax or energy-based OOD scoring — to enable the system to handle previously unseen anomaly types beyond the predefined category set.

9.4. Backbone Modernization.

While the current framework employs 3D CNN feature extractors, lightweight Video Transformer architectures (e.g., Video Swin Transformer) are a natural extension that may offer improved global context modeling with competitive computational efficiency. A direct comparison with such architectures is identified as a valuable direction for future research.

9.5. Robust Synchronization for Deployment.

Our future work will investigate automatic temporal calibration and geometric alignment techniques — such as homography estimation and adaptive frame-rate matching — to improve the practical deployability of the framework under real-world heterogeneous camera configurations.

9.6. Thermal Robustness.

To address thermal sensitivity under high ambient temperatures, the next work will explore background thermal compensation techniques and adaptive thermal contrast normalization to improve framework robustness in challenging environments such as industrial settings or extreme outdoor conditions.

10. Conclusion

This paper introduced a 3D multimodal feature fusion framework for abnormal human behavior recognition in low-light surveillance, combining lightweight 3D architectures with a weighted assembling strategy to integrate thermal and visible spatiotemporal features.

Extensive experiments demonstrate the superiority of our approach over state-of-the-art methods. The AMFCFB model achieves 96.03% recognition accuracy, while the contrastive detection stage attains 92.60% AUC. Our two-stage training strategy with contrastive learning establishes a clear decision boundary between normal and abnormal behaviors, yielding near-perfect F1-scores for "Normal" (0.9970) and "Fighting" (0.9623). The approach effectively handles class imbalance, maintaining strong performance on challenging dynamic classes such as "Running" (0.9615) and "Biker" (0.9500).

The primary remaining bottleneck is performance on subtle, static anomaly classes — particularly *Carrier* (F1: 0.6857) and *Bag Left Unattended* (F1: 0.8345) — where short fixed-length segments and the absence of strong motion cues limit the current framework. Addressing these cases through longer temporal modeling and explicit object-level reasoning is the most direct path to improving the system further. On the deployment side, the AMFCFB backbone achieves 10.06 FPS on the Jetson Orin Nano and 145.09 FPS on a GTX 1080 Ti, confirming practical viability across a range of hardware configurations.

Acknowledgments

The work is partially supported by the Tan Tao University Foundation for Science and Technology Development under Grant No. TTU.RS.24.102.004.

The work of Cuong Tran was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under the Grant No. 2019-0-00231 funded by MSIT of Korea, and has been done when he was with Department of Computer Engineering, Sejong University, Seoul, 05006 Korea.

References

[1] V. Chandola, A. Banerjee, V. Kumar, Anomaly Detection: A survey, ACM Computing Surveys, Vol. 41, No. 3, 2009, pp. 1–58.
URL https://doi.org/10.1145/1541880.1541882

[2] S. A. Jebur, K. A. Hussein, H. K. Hoomod, L. Alzubaidi, J. Santamaría, Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance, Electronics, Vol. 12, No. 1, 2022, pp. 29.
URL https://doi.org/10.3390/electronics12010029

[3] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, A. Kumar, A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets, Applied Artificial Intelligence, Vol. 36, No. 1, 2022, pp. 2093705.
URL https://doi.org/10.1080/08839514.2022.2093705

[4] L. Tang, J. Yuan, J. Ma, Image Fusion in the Toop of High-Level Vision Tasks: A Semantic-aware Real-Time Infrared and Visible Image Fusion Network, Information Fusion, Vol. 82, 2022, pp. 28–42.
URL https://doi.org/10.1016/j.inffus.2021.12.004

[5] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, SwinFusion: Cross-Domain Long-Range Learning for General Image Fusion via Swin Transformer, IEEE/CAA Journal of Automatica Sinica, Vol. 9, No. 7, 2022, pp. 1200–1217.
URL https://doi.org/10.1109/JAS.2022.105686

[6] W. Tang, F. He, Y. Liu, YDTR: Infrared and Visible Image Fusion via Y-shape Dynamic Transformer, IEEE Transactions on Multimedia (2022).
URL https://doi.org/10.1109/TMM.2022.3192661

[7] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, S. Yogamani, Spatio-Contextual Deep Network-based Multimodal Pedestrian Detection for Autonomous Driving, IEEE transactions on intelligent transportation systems, Vol. 23, No. 9, 2022, pp. 15940–15950.
URL https://doi.org/10.1109/TITS.2022.3146575

[8] J. Qiu, R. Yao, Y. Zhou, P. Wang, Y. Zhang, H. Zhu, Visible and Infrared Object Tracking via Convolution-Transformer Network with Joint Multimodal Feature Learning, IEEE Geoscience and Remote Sensing Letters, Vol. 20, 2023, pp. 1–5.
URL https://doi.org/10.1109/LGRS.2023.3259583

[9] V.-A. Nguyen, S. G. Kong, Multimodal Feature Fusion for Illumination-Invariant Recognition of Abnormal Human Behaviors, Information Fusion, Vol. 100, 2023, pp. 101949.
URL https://doi.org/10.1016/j.inffus.2023.101949

[10] O. Köpüklü, N. Kose, A. Gunduz, G. Rigoll, Resource Efficient 3D Convolutional Neural Retworks, in: The IEEE/CVF International Conference on Computer Vision Workshop, 2019, pp. 1910–1919.
URL https://arxiv.org/pdf/1904.02422

[11] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised Contrastive Learning, Advances in neural information processing systems, Vol. 33, 2020, pp. 18661–18673.
URL https://dl.acm.org/doi/abs/10.5555/3495724.3497291

[12] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, L. S. Davis, Learning Temporal Regularity in Video Sequences, in: The IEEE conference on computer vision and pattern recognition, 2016, pp. 733–742.
URL https://doi.org/10.48550/arXiv.1604.04574

[13] W. Liu, W. Luo, D. Lian, S. Gao, Future frame prediction for anomaly detection–a new baseline, in: The IEEE conference on computer vision and pattern recognition, 2018, pp. 6536–6545.
URL https://doi.org/10.48550/arXiv.1712.09867

[14] Y. Wang, C. Qin, Y. Bai, Y. Xu, X. Ma, Y. Fu, Making Reconstruction-based Method Great Again for Video Anomaly Detection, in: 2022 IEEE International Conference on Data Mining, 2022, pp. 1215–1220.
URL https://doi.org/10.48550/arXiv.2301.12048

[15] Z. Wang, Y. Chen, Anomaly Detection with Dual-Stream Memory Network, Journal of Visual Communication and Image Representation, Vol. 90, 2023, pp. 103739.
URL https://doi.org/10.1016/j.jvcir.2022.103739

[16] W. Chen, Z. Yu, C. Yang, Y. Lu, Abnormal Behavior Recognition Based on 3D Dense Connections, International Journal of Neural Systems, Vol. 34, 2024, pp. 2450049.
URL https://doi.org/10.1142/s0129065724500497

[17] D. Ahn, S. Kim, H. Hong, B. C. Ko, Star-Transformer: a Spatio-Temporal Cross Attention Transformer for Human Action Recognition, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, pp. 3330–3339.
URL https://doi.org/10.1109/WACV56688.2023.00333

[18] M. Munsif, S. U. Khan, N. Khan, S. W. Baik, Attention-based Deep Learning Framework for Action Recognition in a Dark Environment, Hum. Centric Comput. Inf. Sci, Vol. 14, 2024, pp. 1–22.
URL https://doi.org/10.22967/HCIS.2024.14.004

[19] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, in: The IEEE international conference on computer vision, 2015, pp. 4489–4497.
URL https://doi.org/10.1109/ICCV.2015.510

[20] J. Carreira, A. Zisserman, Quo Vadis, Action Recognition? a New Model and the Kinetics Dataset, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
URL https://doi.org/10.48550/arXiv.1705.07750

[21] H. Han, N. Ouyang, X. Kuang, S. Wang, F. Xiong, Multilevel Features Cascade Fusion Network for Infrared Video Human Behavior Recognition, Displays, Vol. 87, 2025, pp. 102921.
URL https://doi.org/10.1016/j.displa.2024.102921

[22] D. Tsiktsiris, A. Lalas, M. Dasygenis, K. Votis, Multimodal Abnormal Event Detection in Public Transportation, IEEE Access, Vol. 12, 2024, pp. 133469–133480.
URL https://doi.org/10.1109/ACCESS.2024.3425308

[23] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, Y. Yu, Current Advances and Future Perspectives of Image Fusion: A Comprehensive Review, Information Fusion, Vol. 90, 2023, pp. 185–217.
URL https://doi.org/10.1016/j.inffus.2022.09.019

[24] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image Fusion Meets Deep Learning: A Survey and Perspective, Information Fusion, Vol. 76, 2021, pp. 323–336.
URL https://doi.org/10.1016/j.inffus.2021.06.008

[25] X. Zhang, Y. Demiris, Visible and Infrared Image

Fusion Using Deep Learning, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).
URL https://dl.acm.org/doi/10.1109/TPAMI.2023.3261282

[26] D. Wei, Y. Liu, X. Zhu, J. Liu, X. Zeng, MSAF: Multimodal Supervise-Attention Enhanced fusion for Video Anomaly Detection, IEEE Signal Processing Letters, Vol. 29, 2022, pp. 2178–2182.
URL https://doi.org/10.1109/LSP.2022.3216500

[27] X. Liu, Z. Wang, H. Gao, X. Li, L. Wang, Q. Miao, HATF: Multi-Modal Feature Learning for Infrared and Visible Image Fusion via Hybrid Attention Transformer, Remote Sensing, Vol. 16, No. 5, (2024).
URL https://www.mdpi.com/2072-4292/16/5/803

[28] C. Jiang, H. Ren, H. Yang, H. Huo, P. Zhu, Z. Yao, J. Li, M. Sun, S. Yang, M2FNet: Multi-modal Fusion Network for Object Detection from Visible and Thermal Infrared Images, International Journal of Applied Earth Observation and Geoinformation, Vol. 130, 2024, pp. 103918.
URL https://doi.org/10.1016/j.jag.2024.103918

[29] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality Reduction by Learning an Invariant Mapping, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 1735–1742.
URL https://doi.org/10.1109/CVPR.2006.100

[30] Y. Zheng, M. Jin, Y. Liu, L. Chi, K. T. Phan, Y.-P. P. Chen, Generative and Contrastive Self-Supervised Learning for Graph Anomaly Detection, IEEE Transactions on Knowledge and Data Engineering (2021).
URL https://doi.org/10.1109/TKDE.2021.3119326

[31] O. Köpüklü, J. Zheng, H. Xu, G. Rigoll, Driver Anomaly Detection: A Dataset and Contrastive Learning Approach, in: The IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 91–100.
URL https://doi.org/10.48550/arXiv.2009.14660

[32] S. Sun, X. Gong, Hierarchical Semantic Contrast for Scene-Aware Video Anomaly Detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22846–22856.
URL https://doi.org/10.48550/arXiv.2303.13051

[33] H. N. Ngoc, N. N. Xuan, T. H. Bui, D. H. Hung, S. Q. H. Truong, V. Hoang, An Efficient Approach for Real-Time Abnormal Human Behavior Recognition on Surveillance Cameras, in: 17th International Conference on Automatic Face and Gesture Recognition, 2023, pp. 1–6.
URL https://doi.org/10.1109/FG57933.2023.10042648

[34] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, F. Makedon, A Survey on Contrastive Self-Supervised Learning, Technologies, Vol. 9, No. 1, 2020, pp. 2.
URL https://doi.org/10.3390/technologies9010002

[35] CVAT. ai Corporation, Computer Vision Annotation Tool (2022).
URL http://https://www.cvat.ai

[36] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, Z. Yifu, C. Wong, D. Montes, et al., ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation, Zenodo (2022).
URL https://zenodo.org/records/7347926

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., Pytorch: An Imperative Style, High-Performance Deep Learning Library, Advances in neural information processing systems, Vol. 32, (2019).
URL https://dl.acm.org/doi/10.5555/3454287.3455008

[38] K. Hara, H. Kataoka, Y. Satoh, Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition, in: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), 2017, pp. 3154–3160.
URL https://doi.org/10.48550/arXiv.1708.07632

[39] F. Sun, J. Zhang, X. Wu, Z. Zheng, X. Yang, Video Anomaly Detection Based on Global-Local Convolutional Autoencoder, Electronics, Vol. 13, No. 22, (2024).
URL https://doi.org/10.3390/electronics13224415

[40] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How Transferable Are Features in Deep Neural Networks?, in: The 27th International Conference on Neural Information Processing Systems - Volume 2, 2014, pp. 3320–3328.
URL https://doi.org/10.48550/arXiv.1411.1792

[41] I. J. Good, Rational Decisions, Journal of the Royal Statistical Society: Series B (Methodological), Vol. 14, No. 1, 2018, pp. 107–114.
URL https://doi.org/10.1111/j.2517-6161.1952.tb00104.x

[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal Loss for Dense Object Detection, IEEE Trans on Pattern Analysis and Machine Intelligence, Vol. 42, No. 2, 2020, pp. 318–327.
URL https://doi.org/10.1109/TPAMI.2018.2858826

[43] W. Sultani, C. Chen, M. Shah, Real-world Anomaly Detection in Surveillance Videos (2019). arXiv: 1801.04264.
URL https://arxiv.org/abs/1801.04264