CrimNet: Two-Stage Crime Detection Networks Enhanced by Auxiliary Heads

Anonymous CVPR submission

Paper ID 22721

Abstract

*Detecting anomalous events in surveillance videos remains a challenging task due to the sparsity and subtlety of criminal activities. We present **CrimNet**, a novel two-stage framework for video anomaly detection that leverages auxiliary attention supervision to enhance spatiotemporal representation learning. CrimNet combines a 3D Convolutional Neural Network (C3D) with a lightweight Transformer Encoder to capture long-range temporal dependencies effectively. To further improve discriminability, we introduce an auxiliary attention entropy loss that encourages diverse and focused attention across heads. Trained on the standard UCF-Crime dataset, CrimNet exhibits strong cross-domain generalization when transferred to its event-based extension, **UCF-Crime-DVS**, maintaining robust performance despite sensing-modality shifts. Extensive experiments demonstrate that CrimNet achieves a new state-of-the-art frame-level AUC of **70.69%** on UCF-Crime-DVS, outperforming the previous best by **5.68%**. Our results highlight the effectiveness of structured auxiliary objectives and cross-modality transfer in boosting spatiotemporal modeling for weakly supervised video anomaly detection.*

1. Introduction

Anomaly detection plays a crucial role in high-stakes domains such as cybersecurity [5], fraud detection [11], and medical diagnostics [9], where rare but consequential events must be identified with high reliability. In the realm of video surveillance, anomaly detection presents even greater challenges: anomalies are temporally sparse, often visually subtle, and can occur unpredictably across long, untrimmed videos. These difficulties are further exacerbated under *weak supervision*, where only video-level labels are provided without precise temporal annotations. Benchmarks like UCF-Crime [14] and the recent event-based UCF-Crime-DVS [13] underscore the practical limitations of current methods in addressing these real-world constraints.

Most recent efforts approach weakly-supervised video anomaly detection using Multiple Instance Learning (MIL) [15] or ranking-based paradigms [16], assuming that at least one segment in a positively-labeled video contains an anomaly. However, such methods often falter in scenarios where anomalies are long-range, temporally diffuse, or visually indistinguishable from normal patterns. Their reliance on coarse labeling also makes them vulnerable to noisy supervision and poor generalization.

3D convolutional networks (e.g., C3D [18], R3D [8], and R(2+1)D [19]) are widely adopted to capture short-term spatiotemporal features. Yet, they lack the capacity to model long-range dependencies and often collapse under weak anomaly signals. Furthermore, existing models typically function as black boxes, offering little interpretability—an important consideration in security-critical applications.

To address these challenges, we propose **CrimNet**, a modular two-stage framework tailored for weakly supervised video anomaly detection. CrimNet first employs a 3D CNN backbone for localized spatiotemporal encoding, followed by a lightweight Transformer module that captures long-range temporal dependencies with minimal computational overhead. Crucially, we introduce an **auxiliary attention entropy loss** that explicitly regularizes attention distributions—encouraging diverse head specialization and improving both interpretability and generalization. Our method also exhibits faster convergence and enhanced training stability, making it suitable for deployment under real-world constraints.

We evaluate CrimNet on the challenging UCF-Crime-DVS benchmark [13], achieving a new state-of-the-art frame-level AUC of **70.69%**, surpassing the previous best by **5.68%**. Our contributions are summarized as follows:

- We present **CrimNet**, an efficient two-stage architecture for weakly supervised video anomaly detection. The first stage uses a lightweight 3D CNN feature selector to suppress background noise, while the second stage integrates a backbone encoder (e.g., C3D, R3D, R(2+1)D) with a Transformer module and auxiliary attention regularization to capture long-range temporal dependencies.
- We propose an **entropy-based auxiliary attention loss**

that prevents attention collapse by maximizing per-head entropy, ensuring balanced temporal coverage and stable optimization under weak supervision.

- CrimNet achieves a state-of-the-art frame-level AUC of **70.69%** on UCF-Crime-DVS, demonstrating strong performance and generalization. Extensive ablation and qualitative analyses verify the effectiveness of the auxiliary loss and the two-stage temporal reasoning framework.

2. Related Work

Weakly Supervised VAD: Challenges And Paradigms: Video anomaly detection (VAD) under weak supervision—where only video-level labels are available—has garnered significant attention due to its scalability and relevance to real-world surveillance [10, 15]. Most approaches adopt the multiple instance learning (MIL) paradigm, assuming that at least one snippet in an anomalous video contains the anomaly. Extensions such as MIL ranking [16] and contrastive instance learning [12] have attempted to improve discriminability, but suffer when anomalies are temporally diffuse or visually ambiguous. Datasets like UCF-Crime and UCF-Crime-DVS emphasize these difficulties, with long videos, complex scenes, and subtle anomaly manifestations.

Temporal Modeling For VAD: CNNs vs Transformers: Capturing temporal dependencies is central to anomaly detection. Early methods utilize 3D convolutional networks (C3D [18], I3D [4], R(2+1)D [19]) to jointly model short-range spatial-temporal patterns. However, their local receptive fields limit the ability to capture long-term context, leading to degraded performance in dispersed or gradual anomalies. To address this, recent methods have incorporated Transformers [20] due to their ability to model long-range temporal dependencies. Architectures such as TimeSformer [3], ViViT [2], and VideoMAE [17] achieve state-of-the-art results on action recognition, but their reliance on full attention incurs significant computational cost, making them less suitable for real-time or resource-limited surveillance scenarios.

Hybrid Architectures And Auxiliary Supervision: To balance efficiency and expressivity, hybrid models combine CNN backbones with lightweight Transformer modules [3, 6, 7]. These designs enable coarse-to-fine temporal reasoning without excessive overhead. For instance, ActionFormer [21] uses auxiliary attention supervision to regularize attention maps for temporal localization, improving training stability and interpretability. However, most existing auxiliary losses focus on supervised settings with frame-level annotations.

In contrast, our method introduces a lightweight Transformer with an **entropy-based auxiliary attention loss** that explicitly promotes diversity among heads in a weakly labeled setting. This enhances interpretability and temporal sensitivity, enabling our model to capture both abrupt and prolonged anomalies. Compared to existing hybrid methods, our framework is tailored for the weakly supervised regime and delivers strong results on UCF-Crime-DVS without incurring the cost of full attention or fine-grained supervision.

3. Methods

We propose **CrimNet**, a two-stage framework designed for temporal anomaly detection under weak supervision. By introducing a lightweight feature selector and entropy-based attention regularization, CrimNet addresses both computational inefficiency and optimization instability in long, noisy surveillance videos. The overall architecture is shown in Figure 1.

Specifically, Stage 1 uses a shallow 3D CNN followed by upsampling to preserve spatial dimensions while suppressing irrelevant noise. Stage 2 employs a backbone classifier, where features are flattened and passed through a Transformer encoder for temporal modeling. The auxiliary attention loss stabilizes optimization by preventing attention collapse and maintaining balanced temporal coverage.

In weakly supervised video anomaly detection, models are trained with only video-level labels but must produce fine-grained frame-level anomaly scores. This discrepancy causes two major challenges: (1) excessive computation on irrelevant content, and (2) unstable training due to label ambiguity and long-range dependencies.

3.1. Two-stage Inference Flow

To tackle these issues, we adopt a principled two-stage design:

- **Stage 1: Feature Selection via Learnable Spatio-Temporal Filtering.** Unlike a reconstruction autoencoder, the first stage of **CrimNet** is *not* trained to reproduce RGB appearance. Its objective is to act as a *learnable denoising/filtering front-end* that suppresses background motion, compression artifacts, and sensor noise while preserving task-relevant cues for anomaly discrimination.

  Operator stack and shapes. Given $x \in \mathbb{R}^{B\times 3\times T\times H\times W}$, Stage 1 applies a shallow stack of 3D convolutions:

$$3 \xrightarrow[\text{stride/pool}]{\text{Conv3D}(k_t\times k\times k)} 16 \xrightarrow[\text{stride/pool}]{\text{Conv3D}} 32 \xrightarrow{\text{Conv3D}(1\times 1\times 1)} 3,$$

  interleaved with normalization and nonlinearity (e.g., BN/ReLU). Temporal/spatial pooling reduces resolution to attenuate high-frequency noise, followed by trilinear upsampling that restores the original (T, H, W) *resolution* for compatibility with the downstream backbone. The final output $\tilde{x} \in \mathbb{R}^{B\times 3\times T\times H\times W}$ is thus *shape-aligned* with the input but *content-filtered*.

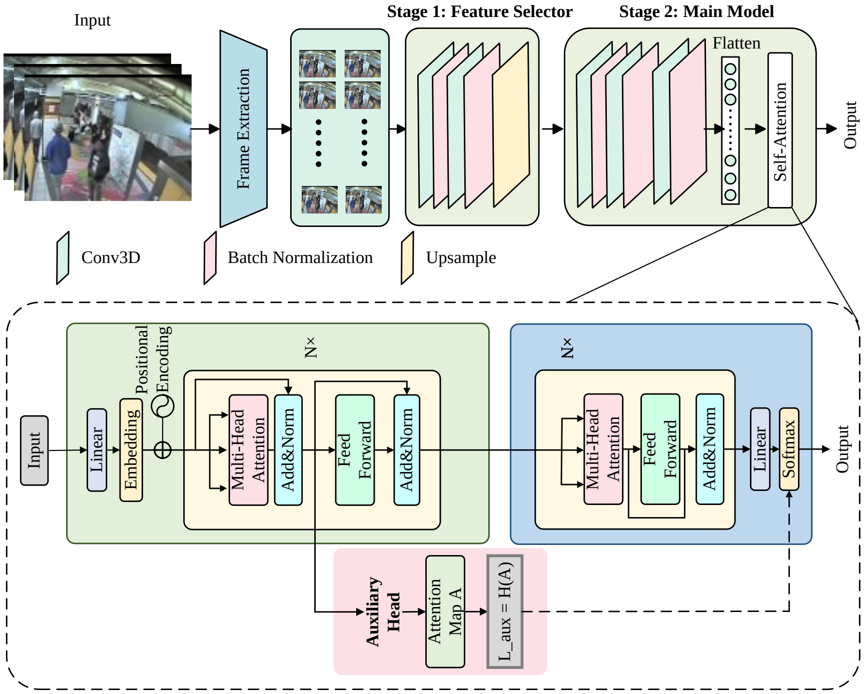


Figure 1. Overall architecture of **CrimNet**. The model consists of a two-stage pipeline: Stage 1 applies a lightweight 3D CNN-based FeatureSelector to reduce spatio-temporal noise; Stage 2 processes filtered features through a 3D backbone (e.g., C3D), followed by a Transformer encoder with auxiliary attention supervision. An auxiliary head introduces an entropy-based regularization term that prevents attention collapse and maintains balanced temporal coverage during training.

Why "3" channels again? The last $1\times1\times1$ convolution projects the 32-channel latent to 3 channels to: (i) keep the FLOPs/VRAM of the downstream 3D backbone identical to a standard 3-channel input (e.g., C3D); (ii) enforce a *bottleneck* that compacts spatio-temporal evidence into three learnable activation maps. These three channels *are not RGB*; they are *task-adapted feature maps* whose dynamic range is controlled by normalization (optionally clipped/squashed by $\tanh$/sigmoid if needed).

Why this suppresses noise instead of reconstructing appearance. Stage 1 is *never* optimized with a pixel-wise reconstruction loss. Its parameters are updated *only* through the downstream anomaly objective (and the optional information-bottleneck regularizer; see Eq. (2)). Consequently, gradients encourage the front-end to *discard* redundancy that does not help the anomaly classifier and to *retain* discriminative motion/appearance cues. Empirically, this yields smoothed backgrounds and clearer object/interaction traces, i.e., a task-aligned denoised tensor.

Information bottleneck view. The channel compression $32 \to 3$ implements a deterministic bottleneck $q_\phi(\tilde{x} \mid x)$ that discourages copying x and promotes retaining only $I(\tilde{x}; y)$-relevant information while reducing $I(\tilde{x}; x)$ (cf. Eq. (2)). In ablations, removing the $1\times1\times1$ projection (i.e., keeping many channels) weakens this effect and hurts generalization.

Compute/latency. Stage 1 has $\sim$1.2M parameters ($\approx$5% of total FLOPs) and adds $<$1 ms per clip on an RTX 4090 at T=16, H=W=112, while substantially stabilizing training (see Sec.4).

- **Stage 2: Backbone + Transformer with Entropy-Regularized Attention.** The filtered tensor $\tilde{x}$ is fed to a 3D CNN backbone (default: C3D) producing temporal embeddings $\mathbf{F} \in \mathbb{R}^{B\times T\times d}$ after global spatial pooling (d=512 by default). We add positional encodings and pass $\hat{\mathbf{F}}$ to a lightweight Transformer encoder to model long-range temporal dependencies while preserving clip granularity:

$$\begin{aligned} \mathbf{H} &= \mathcal{T}(\hat{\mathbf{F}}) \in \mathbb{R}^{B\times T\times d}, \\ \hat{y}_i &= \sigma(w^\top h_i), \quad i = 1, \dots, T. \end{aligned} \tag{1}$$

Why entropy regularization (Route A). Under weak supervision, attention heads tend to *collapse* onto a few dominant frames, harming temporal coverage and causing unstable gradients. We apply a per-head entropy term (Eq. (8)) to *prevent attention collapse* and maintain *balanced* temporal exploration within each head. This stabilizes optimization and improves robustness without enforcing inter-head dissimilarity (we do not use head-to-head KL), keeping computation minimal and the objective well-conditioned.

Supervision pathway. Stage 2 is trained end-to-end with the main anomaly loss on $\hat{y}$ plus the entropy regularizer on attention logits $Z^{(h)}$, and Stage 1 is updated solely through these downstream signals. Thus, the entire system aligns the front-end filtering and temporal reasoning with the final detection objective, rather than reconstructing appearance.

This coarse-to-fine approach enables more stable optimization, improves interpretability, and enhances localization accuracy.

Table 1 summarizes the layer-wise dimensionality transformations across the two stages of CrimNet.

Stage 1 progressively increases channel capacity while reducing the spatio-temporal resolution, yielding a receptive field large enough to suppress background motion while remaining lightweight (˜1.2M parameters, $\approx$ 5% of total FLOPs). The final trilinear upsampling restores the original temporal length T and image resolution 112×112, ensuring compatibility with off-the-shelf backbones.

In Stage 2, the 3D backbone retains the canonical C3D stem; we report its output compactly as (Batch, T, d) because spatial dimensions are collapsed by global average pooling ($d = 512$ in our default setting). The Transformer encoder preserves temporal granularity, enabling clip-wise predictions without additional pooling. Its auxiliary head operates on the attention logits of each multi-head sub-layer (H=8), producing a (T, H) tensor on which the entropy loss is computed. Finally, a linear projection followed by a sigmoid delivers a scalar anomaly score per clip.

This design keeps the temporal index intact throughout Stage 2, so each clip maintains a unique representation path, facilitating precise localization at inference time.

Let $x \in \mathbb{R}^{B \times 3 \times T \times H \times W}$ be a mini-batch of RGB video clips, where B is batch size, T=16, H=W=112. Stage 1 is a shallow 3-layer 3D CNN with intermediate pooling and final upsampling:

$$\begin{aligned}
x^{(1)} &= \sigma(\mathrm{Conv3D}_{3\to16}(x)) \in \mathbb{R}^{B\times16\times T\times H\times W},\\
x^{(2)} &= \sigma(\mathrm{Conv3D}_{16\to32}(\mathrm{Pool}(x^{(1)}))) \in \mathbb{R}^{B\times32\times T/2\times H/2\times W/2},\\
x^{(3)} &= \sigma(\mathrm{Conv3D}_{32\to3}(\mathrm{Pool}(x^{(2)}))) \in \mathbb{R}^{B\times3\times T/4\times H/4\times W/4},\\
\tilde{x} &= \mathrm{Upsample}(x^{(3)};\ T,\ H,\ W) \in \mathbb{R}^{B\times3\times T\times H\times W}.
\end{aligned}$$

From an information-theoretic perspective, Stage 1 can be viewed as a bottleneck encoder $q_\phi(\tilde{x}|x)$ that retains task-relevant information while discarding redundancy. Following the variational information bottleneck (VIB) principle [1], its objective is to minimize:

$$\begin{aligned}
\mathcal{L}_{\mathrm{IB}} = \beta\,\mathrm{KL}(q_\phi(\tilde{x}|x)\,\|\,p(\tilde{x})) - I(\tilde{x};y)\,\mathcal{L}_{\mathrm{main}}(f_\theta(\tilde{x}),y)\\
+\beta\,\mathrm{KL}(q_\phi(\tilde{x}|x)\,\|\,p(\tilde{x}))
\end{aligned} \tag{2}$$

where $p(\tilde{x})$ is an isotropic Gaussian prior and β controls the compression-generalization trade-off.

3.2. Auxiliary Attention Head and Loss

In weakly supervised settings, Transformer attention often collapses onto a few salient frames, leading to unstable gradients and poor temporal coverage. To mitigate this issue, we introduce an entropy-based auxiliary loss that regularizes each attention head to maintain balanced focus over time.

Let $Z^{(h)} \in \mathbb{R}^{T\times T}$ denote the raw attention logits of head h, and $A^{(h)} = \mathrm{Softmax}(Z^{(h)}/\sqrt{d})$. The per-head attention entropy is computed as:

$$\mathcal{L}_{\mathrm{aux}} = \frac{1}{HT}\sum_{h=1}^{H}\sum_{i=1}^{T}\left(-\sum_{j=1}^{T} A_{ij}^{(h)} \log A_{ij}^{(h)}\right). \tag{3}$$

This term prevents attention collapse by penalizing overly peaked distributions and encouraging smoother temporal exploration within each head. In practice, $\mathcal{L}_{\mathrm{aux}}$ is applied to all attention layers and averaged across heads.

Gradient Perspective. The auxiliary entropy loss $\mathcal{L}_{\mathrm{aux}}$ acts directly on the attention logits $Z^{(h)}$ and provides intermediate supervision at every attention layer. Its derivative with respect to logits is:

$$\frac{\partial \mathcal{L}_{\mathrm{aux}}}{\partial Z_{ij}^{(h)}} = A_{ij}^{(h)}\left(1 + \log A_{ij}^{(h)} - \sum_{k} A_{ik}^{(h)} \log A_{ik}^{(h)}\right), \tag{4}$$

where $A_{ij}^{(h)}$ denotes the attention weight from token i to j. The term $\sum_k A_{ik}^{(h)} \log A_{ik}^{(h)}$ represents the negative Shannon entropy $H(A_{i,:}^{(h)})$. Hence, the gradient can be rewritten as:

$$\frac{\partial \mathcal{L}_{\mathrm{aux}}}{\partial Z_{ij}^{(h)}} = A_{ij}^{(h)}\left(1 + \log A_{ij}^{(h)} + H(A_{i,:}^{(h)})\right). \tag{5}$$

Intuitively:

- Larger $A_{ij}^{(h)}$ values produce stronger gradients, focusing updates on dominant regions.
- The $\log A_{ij}^{(h)}$ term penalizes overly confident (peaked) distributions, flattening the softmax output and preventing collapse.
- The entropy term $H(A_{i,:}^{(h)})$ moderates correction strength—high-entropy rows (well-distributed attention) indicate stable coverage and thus yield smaller updates.

Table 1. Layer-wise output sizes in CrimNet. T denotes temporal length, H/W are spatial dimensions, and d is the feature embedding size.

Layer	Output Size
Stage 1: Feature Selector	
Input	(3, 16, 112, 112)
Conv3D (3→16)	(16, 16, 112, 112)
MaxPool3D	(16, 8, 56, 56)
Conv3D (16→32)	(32, 8, 56, 56)
MaxPool3D	(32, 4, 28, 28)
Conv3D (32→3)	(3, 4, 28, 28)
Upsample	(3, 16, 112, 112)
Stage 2: Backbone + Transformer	
3D CNN Backbone	(Batch, T, d)
+ Pos. Encoding	(T, d)
Transformer Encoder	(T, d)
Aux. Attention Head	(T, H)
Final Linear	(T, 1)

The proposed design offers multiple functional advantages. First, Stage 1 acts as an information bottleneck by selectively preserving task-relevant information $I(\tilde{x}; y)$ while suppressing redundant correlations $I(\tilde{x}; x)$, effectively reducing noise propagation. Second, the auxiliary loss enhances gradient flow by providing intermediate supervision, improving optimization efficiency. Lastly, the entropy regularization mitigates attention collapse in the Transformer, promoting balanced temporal coverage and improving both stability and interpretability.

To summarize the overall inference and training workflow, the step-by-step procedure of **CrimNet** is presented in Algorithm 1.

CrimNet outputs clip-level scores $\hat{y}_i$ and interpolates them to frame-level using linear upsampling, following prior protocols [15].

4. Experimental Results

4.1. Experimental Setup

We evaluate our method on two challenging large-scale benchmarks for video anomaly detection: UCF-Crime and its event-based extension UCF-Crime-DVS, both designed for real-world surveillance scenarios.

UCF-Crime comprises 1,900 untrimmed surveillance videos spanning 13 anomaly categories (e.g., burglary, robbery, assault) and a normal class. It is characterized by high intra-class variance, weak video-level annotations, and significant class imbalance. Due to its scale and realism, it has become a standard benchmark for weakly-supervised anomaly detection.

UCF-Crime-DVS builds upon the original dataset by incorporating event-based camera recordings, offering temporally precise event streams while preserving the same class taxonomy. The increased temporal resolution introduces additional challenges in modeling fine-grained motion and dynamics.

Algorithm 1: CrimNet Inference and Training Procedure

```
Input: Video clip batch $x \in \mathbb{R}^{B\times 3\times T\times H\times W}$, labels $y$
Output: Frame-level anomaly scores $\hat{y}$
Stage 1: Feature Selection
$\tilde{x} \leftarrow \text{FeatureSelector}(x)$   // 3D CNN filtering + upsampling
Stage 2: Backbone + Transformer
$\mathbf{F} \leftarrow f_{\text{C3D}}(\tilde{x})$   // Temporal embeddings
$\hat{\mathbf{F}} \leftarrow \mathbf{F} + \text{PosEnc}(\mathbf{F})$
$\mathbf{H} \leftarrow \text{TransformerEncoder}(\hat{\mathbf{F}})$   // Temporal modeling
$\hat{y} \leftarrow \sigma(W\mathbf{H})$   // Clip-level anomaly prediction
Auxiliary Attention Supervision
for each head $h = 1, \dots, H$ do
    $A^{(h)} \leftarrow \text{Softmax}(Z^{(h)}/\sqrt{d})$
    $\mathcal{L}_{\text{aux}} \mathrel{+}= \frac{1}{HT}\sum_i -\sum_j A_{ij}^{(h)} \log A_{ij}^{(h)}$
     // Prevent attention collapse via entropy regularization
end
Total Loss
$\mathcal{L} \leftarrow \mathcal{L}_{\text{main}}(\hat{y}, y) + \lambda\mathcal{L}_{\text{aux}}$
return $\hat{y}$ and $\mathcal{L}$
```

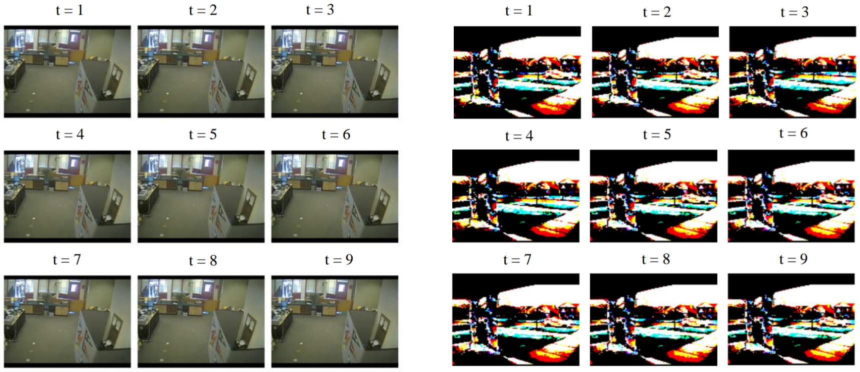
Figure 2 provides an illustrative example from the UCF-Crime dataset. Subfigure (a) visualizes the first nine frames of a raw *Arrest* video segment, while (b) presents its augmented version after applying our spatio-temporal data enhancement strategies, revealing improved variability and robustness to visual distortions.

All experiments are conducted on a server with an **NVIDIA RTX 4090 (24GB)** GPU and an **Intel Xeon Gold 6430** CPU. We use the **Adam** optimizer (initial LR: `1e-5`, weight decay: `5e-4`) with cosine annealing or one-cycle LR scheduling, training for 60 epochs with a **batch size of 16**.

Evaluation Metrics: We use frame-level AUC as the primary metric, consistent with prior works [15, 16]. We also report false alarm rate (FAR) where applicable. Visualizations are performed using attention-weighted scores and detection heatmaps.

4.2. Quantitative Comparison

Table 2 presents a comprehensive comparison of frame-level AUC and false alarm rate (FAR) on the UCF-Crime-DVS dataset. **CrimNet** achieves a frame-level AUC of **70.69%**, establishing a new state-of-the-art among all



(a) Category examples (Arrest). (b) Data Enhancement (Arrest).

Figure 2. Overview of the UCF-Crime dataset: (a) sample video frames from various anomaly classes; (b) Sample of Data Enhancement

weakly supervised methods. Compared to the most competitive prior approach, MSF [13], which achieves 65.01%, our method offers a notable improvement of **+5.68%**, validating the effectiveness of integrating Transformer-based temporal modeling and auxiliary entropy-guided attention regularization.

Although MSF achieves the lowest FAR (3.27%), CrimNet maintains a low false alarm rate of **5.10%**, achieving a favorable trade-off between recall and precision—crucial in real-world surveillance applications where missed detections or over-alerting both carry high costs. Furthermore, CrimNet consistently outperforms both traditional ANN-based frameworks (e.g., AR-Net, RTFM) and event-driven SNN-based designs (e.g., PLIF, SEW-ResNet), underscoring the generalizability of our hybrid architecture across sensing modalities.

Interestingly, while most existing methods plateau around 60–62% AUC, our Transformer-enhanced framework breaks this ceiling, revealing the limitations of static local features and motivating the need for long-range temporal modeling. The entropy-regularized auxiliary head plays a crucial role in capturing diverse temporal dependencies, which is difficult for earlier MIL-based or recurrent architectures to achieve.

5. Analysis and Interpretability

5.1. Ablation Study

To dissect the contribution of each CrimNet component, Table 3 reports a comprehensive ablation study across three popular backbones: C3D, R3D, and R(2+1)D, evaluated on both UCF-Crime and UCF-Crime-DVS.

Across all architectures and datasets, the full configuration (Two-Stage + Transformer + Aux Head) delivers the best performance, indicating strong complementarity between spatial feature filtering and temporal attention modeling. On the UCF dataset with C3D, AUC rises from **68.23% (baseline)** to **85.12%**, an absolute improvement of **+16.89%**. A similar gain is observed in accuracy, which increases from 23.00% to 37.14%.

Analyzing the individual module effects:

- **Transformer + Auxiliary Head only:** Provides the largest individual gain, improving temporal reasoning by focusing attention on discriminative segments. For example, with C3D, AUC improves to 75.23%, outperforming the Two-Stage-only variant by 5.02%.
- **Two-Stage Only:** Acts as a spatial noise suppressor, improving early-stage feature quality. Gains are consistent, but slightly lower than the Transformer.
- **Combined Setup:** Synergistically enhances both spatial localization and temporal pattern modeling, with clear cumulative effects.

Notably, all three backbones exhibit the same trend, demonstrating that the architectural benefits are model-agnostic. Moreover, similar improvements are observed on UCF-Crime-DVS, suggesting strong transferability across input formats—from RGB videos to temporally dense event streams.

These findings underscore that: (1) The proposed architecture is robust and general across both CNN backbones and datasets; (2) Long-range attention with entropy regularization plays a crucial role in anomaly modeling; (3) The modular design of CrimNet allows progressive integration without breaking compatibility with legacy backbones.

Effect of Two-Stage Architecture. Introducing the lightweight feature selector (Stage 1) improves performance by filtering out irrelevant spatial noise and emphasizing anomaly-related cues early. For instance, the C3D backbone improves from 68.23% to 70.21% (UCF) in AUC when Stage 1 is included.

Effect of Transformer + Auxiliary Head. The auxiliary attention module enhances temporal reasoning. When used alone, it boosts C3D performance from 68.23% to 75.23% AUC. When combined with Stage 1, the full Crim-

Table 2. Comparison of AUC and FAR on UCF-Crime-DVS dataset with existing methods. CrimNet achieves the highest frame-level AUC.

Method	Architecture	Supervision	AUC (%)	FAR (%)
Sultani et al. (2018)	ANN	Weakly-supervised	55.56	8.69
3C-Net (2019)	ANN	Weakly-supervised	59.22	9.50
AR-Net (2020)	ANN	Weakly-supervised	60.71	8.51
Wu et al. (2020)	ANN	Weakly-supervised	58.58	34.35
RTFM (2021)	ANN	Weakly-supervised	52.67	13.19
TSA (2023)	ANN	Weakly-supervised	51.86	22.36
SEW-ResNet (2021a)	SNN	Weakly-supervised	53.99	11.79
PLIF (2021b)	SNN	Weakly-supervised	54.74	9.17
Zhou et al. (2023)	SNN	Weakly-supervised	62.78	11.52
MSF (2025)	SNN	Weakly-supervised	65.01	**3.27**
CrimNet (Ours)	ANN	Weakly-supervised	**70.69**	5.10

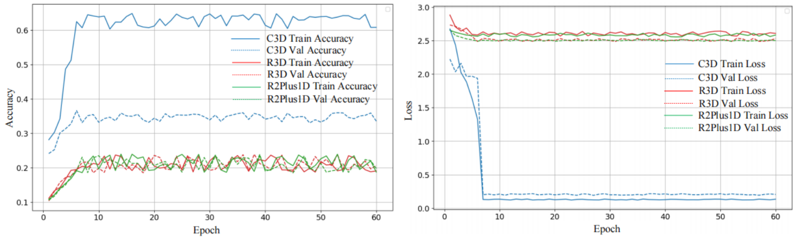


Figure 3. Training dynamics of CrimNet with C3D, R3D, and R(2+1)D under the full Two-Stage + Self-Attention setup. C3D shows better convergence and generalization.

Net model achieves the best result—**85.12%** on UCF and **70.69%** on DVS.

5.2. Training Dynamics and Generalization Behavior

To understand the impact of our architectural components on learning dynamics, we visualize the training and validation loss/accuracy curves across three backbones under the full configuration (Two-Stage + Transformer + Aux Head). As shown in Figure 3, the C3D-based CrimNet not only converges significantly faster but also exhibits higher stability and generalization.

- **Convergence Speed:** C3D reaches over 60% training accuracy and 40% validation accuracy by epoch 60, whereas R3D and R(2+1)D plateau around 20% with slower and noisier convergence. This validates that CrimNet benefits from compact spatial features, which are amplified by attention regularization.
- **Transformer Stabilization:** The entropy-guided auxiliary loss explicitly prevents attention collapse, as reflected by lower validation loss variance in C3D. This stabilizing effect is less prominent in deeper backbones due to their higher capacity and potential overfitting under weak labels.

5.3. Confusion Matrix Analysis: Temporal Focus vs. Class Discriminability

To further assess per-class behavior, we present confusion matrices under all backbones in Figure 4. Notably, the C3D-based variant demonstrates the strongest diagonal patterns—especially in high-frequency categories like Robbery (Class 9) and Road Accidents (Class 8)—suggesting stable localization and temporal focus.

- **Error Concentration in Long-tail Classes:** R3D and R(2+1)D exhibit higher confusion in rare classes (e.g., Abuse, Arson), highlighting that over-parameterized backbones may struggle to generalize from limited data under weak supervision.
- **Interpretation Via CrimNet Modules:** Stage 1 filters out spatial noise early, aiding low-level discrimination, while the auxiliary attention loss ensures long-range dependencies are effectively captured, which boosts recall in motion-heavy anomalies (e.g., Assault, Fighting).
- **Entropy-to-AUC Correlation:** Empirically, models with higher average attention entropy (measured on test-time attention maps) tend to yield higher AUC—indicating that diverse temporal focus correlates with better

These findings validate the design intuition of CrimNet: (1) A lightweight spatial filter reduces overfitting risk in deeper backbones; (2) Auxiliary entropy improves attention allo-

Table 3. Ablation study comparing the effect of Two-Stage and Transformer-Auxiliary modules across different backbones.

Backbone	Two-Stage	Transformer + Aux Head	AUC (%) - UCF	Acc (%) - UCF	AUC (%) - DVS	Acc (%) - DVS
C3D	55	55	68.23	23.00	60.23	21.56
C3D	51	55	70.21	30.22	61.23	24.23
C3D	55	51	75.23	32.14	62.15	24.11
C3D	51	51	**85.12**	**37.14**	**70.69**	**29.76**
R3D	55	55	67.42	22.10	59.73	20.50
R3D	51	55	69.33	29.20	60.91	23.42
R3D	55	51	74.10	31.05	61.40	23.98
R3D	51	51	77.80	36.22	66.30	28.50
R(2+1)D	55	55	67.95	22.56	59.94	20.83
R(2+1)D	51	55	69.80	29.80	61.01	23.87
R(2+1)D	55	51	74.65	31.55	61.73	24.20
R(2+1)D	51	51	76.30	36.75	66.70	29.10

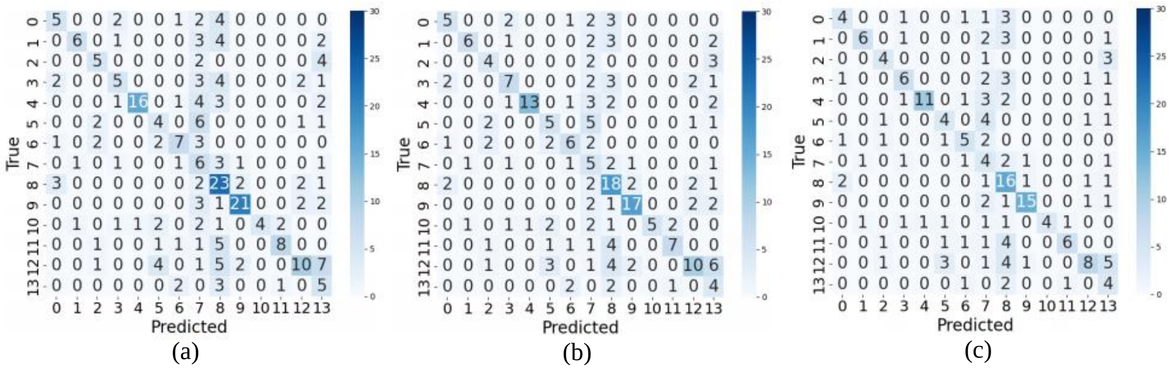


Figure 4. Confusion matrices of CrimNet on C3D (a), R3D (b), and R(2+1)D (c) backbones. Diagonal strength indicates classification quality; C3D demonstrates higher consistency.

cation and generalization; (3) CrimNet's modularity allows consistent gain across all backbones and anomaly types.

5.4. Attention Visualization and Entropy Effect

To further illustrate the impact of the proposed auxiliary entropy loss, we visualize the average self-attention maps of the Transformer encoder under two settings: (a) without and (b) with entropy regularization. As shown in Figure 5, the model trained *without* the auxiliary loss exhibits highly concentrated diagonal attention patterns—indicating head collapse and limited temporal coverage. In contrast, when the entropy loss is applied, attention becomes more distributed across time steps, capturing longer-range dependencies and improving robustness to label noise. This qualitative result directly aligns with our quantitative findings, confirming that entropy-guided supervision effectively prevents over-concentration and stabilizes training.

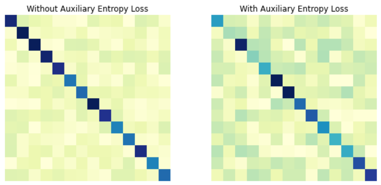


Figure 5. Visualization of average attention weights without (left) and with (right) the auxiliary entropy loss. Entropy regularization mitigates attention collapse, yielding more diverse and temporally aware attention maps.

6. Conclusion

We introduced **CrimNet**, a modular two-stage framework for weakly supervised video anomaly detection that integrates a lightweight 3D CNN-based feature selector with a Transformer enhanced by an entropy-regularized auxiliary head. This design effectively mitigates temporal ambiguity and label sparsity, enabling robust long-range modeling under weak supervision. Experiments on UCF-Crime and UCF-Crime-DVS demonstrate state-of-the-art performance with up to **+5.68%** improvement in frame-level AUC, while ablation studies validate the complementary roles of spatial filtering and attention regularization. Overall, CrimNet provides a simple, interpretable, and deployable solution for real-world surveillance anomaly detection.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 4

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021. 2

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8134–8148, 2021. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 2

[5] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009. 1

[6] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Dai, and Jitendra Malik. Multiscale vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6824–6835, 2021. 2

[7] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213, 2020. 2

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 1

[9] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud AA Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AW van der Laak, Bram van Ginneken, and Clara I Sanchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[10] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 2021. 2

[11] Eric WT Ngai, Yao Hu, YH Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011. 1

[12] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020. 2

[13] Yuanbin Qian, Shuhan Ye, Chong Wang, Xiaojie Cai, Jiangbo Qian, and Jiafei Wu. Ucf-crime-dvs: A novel event-based dataset for video anomaly detection with spiking neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39, 2025. 1, 6

[14] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488, 2018. 1

[15] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488, 2018. 1, 2, 5

[16] Yuxin Tian et al. Weakly-supervised action localization with multi-instance learning and self-attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 5

[17] Zhan Tong, Yibing Song, Jue Wang, and Jie Shen. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[18] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 1, 2

[19] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018. 1, 2

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[21] Yue Zhang, Yuxuan Yang, Zhiwu Ma, Bernt Schiele, and Siyu Wang. Actionformer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 492–510, 2022. 2