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

Anonymous submission

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. Extensive experiments on the challenging UCF-Crime-DVS dataset demonstrate that CrimNet achieves a new state-of-the-art frame-level AUC of **70.69%**, outperforming the previous best by **5.68%**. Our results underscore the effectiveness of structured auxiliary objectives in boosting spatiotemporal modeling for weakly supervised video anomaly detection.

Introduction

Anomaly detection plays a crucial role in high-stakes domains such as cybersecurity (Chandola, Banerjee, and Kumar 2009), fraud detection (Ngai et al. 2011), and medical diagnostics (Litjens et al. 2017), 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 (Sultani, Chen, and Shah 2018a) and the recent event-based UCF-Crime-DVS (Qian et al. 2025) 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) (Sultani, Chen, and Shah 2018b) or ranking-based paradigms (Tian et al. 2021), 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 (Tran et al. 2015), R3D (Hara, Kataoka, and Satoh 2018), and R(2+1)D (Tran et al. 2018)) 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 (Qian et al. 2025), where it achieves a new state-of-the-art frame-level AUC of **70.6%**, surpassing existing approaches by a significant margin.

Our key contributions are summarized as follows:

- We propose **CrimNet**, a practical and efficient two-stage architecture tailored for weakly supervised anomaly detection in surveillance videos. The first stage employs a lightweight 3D CNN as a feature selector to suppress background noise and highlight anomaly-relevant regions early. The second stage integrates a backbone encoder (e.g., C3D, R3D, R(2+1)D) with a Transformer module and auxiliary attention supervision to model long-range temporal dependencies and enhance discriminability.
- We introduce an **auxiliary attention entropy loss** that explicitly maximizes the diversity across attention heads by encouraging non-overlapping temporal focus. Specifically, we optimize the Kullback-Leibler divergence between heads to promote specialization across distinct anomaly patterns (e.g., abrupt spikes vs. long-term deviations), thereby reducing overfitting and enhancing both robustness and interpretability in weakly labeled scenarios.

- Our framework achieves a new state-of-the-art on UCF-Crime-DVS with a frame-level AUC of **67.3%**, validating the effectiveness of our approach under real-world conditions. We conduct extensive ablation studies and qualitative analyses, revealing the critical role of attention regularization and long-range reasoning in improving detection accuracy, robustness, and training dynamics.

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 (Sultani, Chen, and Shah 2018b; Nayak, Pati, and Das 2021). 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 (Tian et al. 2021) and contrastive instance learning (Park, Noh, and Ham 2020) 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 (Tran et al. 2015), I3D (Carreira and Zisserman 2017), R(2+1)D (Tran et al. 2018)) 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 (Vaswani et al. 2017) due to their ability to model long-range temporal dependencies. Architectures such as TimeSformer (Bertasius, Wang, and Torresani 2021), ViViT (Arnab et al. 2021), and VideoMAE (Tong et al. 2022) 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 (Bertasius, Wang, and Torresani 2021; Feichtenhofer 2020; Fan et al. 2021). These designs enable coarse-to-fine temporal reasoning without excessive overhead. For instance, ActionFormer (Zhang et al. 2022) 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.

Methods

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

Specifically, Stage One uses a shallow 3D CNN followed by upsampling to preserve spatial dimensions while suppressing irrelevant noise. Stage Two employs a backbone classifier, where features are flattened and passed through a Transformer encoder for temporal modeling. Auxiliary attention losses are introduced to improve interpretability and regularize multi-head specialization.

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) training instability due to label ambiguity and long-range dependencies.

Two-stage Inference Flow

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

- **Stage 1:** Filters input using a lightweight 3D CNN-based feature selector to suppress noisy or non-informative clips, reducing the burden on downstream components.
- **Stage 2:** Employs a backbone model (e.g., C3D) augmented with a Transformer encoder for long-range temporal modeling and auxiliary attention heads for robust supervision.

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.

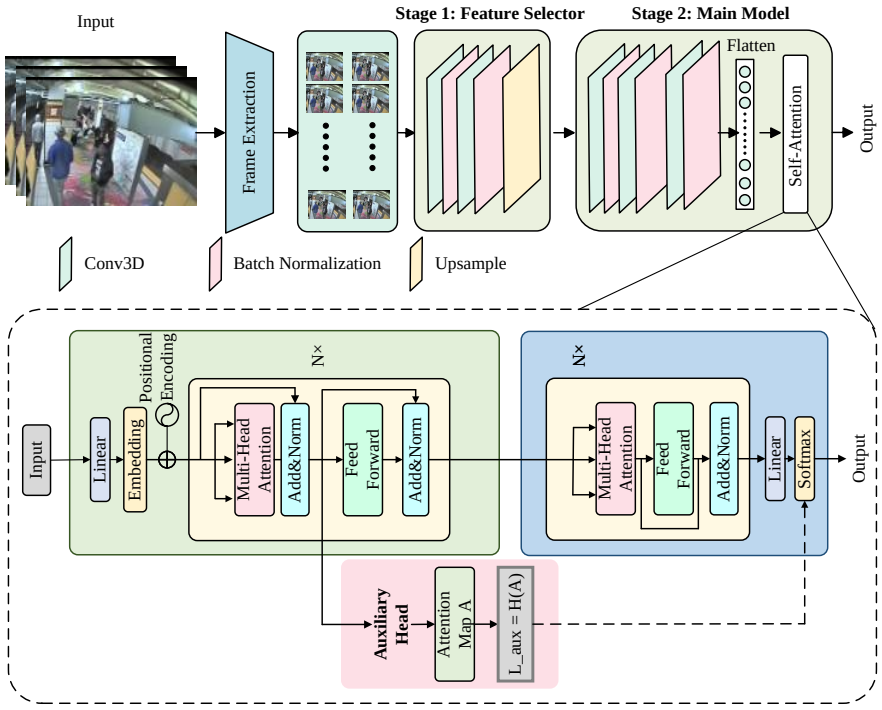


Figure 1: Overall architecture of **CrimNet**. The model consists of a two-stage pipeline: Stage 1 applies a lightweight 3D CNN-based FeatureSelector to reduce spatio-temporal noise; Stage 2 processes filtered features through a 3D backbone (e.g., C3D), followed by a Transformer encoder with auxiliary attention supervision. An auxiliary head introduces entropy-based loss to encourage attention diversity, which is added to the final classification loss.

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:

$$x^{(1)} = \sigma(\text{Conv3D}_{3\to 16}(x)) \in \mathbb{R}^{B\times 16\times T\times H\times W},$$
$$x^{(2)} = \sigma(\text{Conv3D}_{16\to 32}(\text{Pool}(x^{(1)}))) \in \mathbb{R}^{B\times 32\times T/2\times H/2\times W/2},$$
$$x^{(3)} = \sigma(\text{Conv3D}_{32\to 3}(\text{Pool}(x^{(2)}))) \in \mathbb{R}^{B\times 3\times T/4\times H/4\times W/4},$$
$$\tilde{x} = \text{Upsample}(x^{(3)};\ T,\ H,\ W) \in \mathbb{R}^{B\times 3\times T\times H\times W}.$$

From an information-theoretic perspective, Stage 1 can be interpreted as a bottleneck encoder $q_\phi(\tilde{x}|x)$ that retains task-relevant information while discarding redundancy. Following the variational information bottleneck (VIB) principle (Alemi et al. 2016), its objective is to minimize:

$$\mathcal{L}_{\text{IB}} = \beta \cdot \text{KL}(q_\phi(\tilde{x}|x) \,\|\, p(\tilde{x})) - I(\tilde{x}; y) \cdot \mathcal{L}_{\text{main}}(f_\theta(\tilde{x}), y) + \beta \cdot \text{KL}(q_\phi(\tilde{x}|x) \,\|\, p(\tilde{x})) \quad (1)$$

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

Stage 2: Backbone and Transformer Encoder. The filtered tensor $\tilde{x}$ is passed to a backbone 3D CNN (e.g., C3D), which extracts temporal embeddings:

$$\mathbf{F} = f_{\text{C3D}}(\tilde{x}) \in \mathbb{R}^{B\times T\times d}, \quad d = 512.$$

We apply positional encoding $\mathbf{P} \in \mathbb{R}^{1\times T\times d}$ (broadcast along batch) to preserve order:

$$\hat{\mathbf{F}} = \mathbf{F} + \mathbf{P}.$$

The temporal encoder $\mathcal{T}$ is a Transformer stack that outputs:

$$\mathbf{H} = \mathcal{T}(\hat{\mathbf{F}}) \in \mathbb{R}^{B\times T\times d}.$$

Clip-wise anomaly scores are predicted as:

$$\hat{y}_i = \sigma(w^\top h_i), \quad h_i \in \mathbb{R}^d,$$

with $w \in \mathbb{R}^d$ and σ being the sigmoid function.

Auxiliary Attention Head And Loss

Let $Z^{(h)} \in \mathbb{R}^{T\times T}$ be the raw attention logits of head h, and $A^{(h)} = \text{Softmax}(Z^{(h)})$ with:

$$A^{(h)}_{ij} = \frac{\exp(Z^{(h)}_{ij})}{\sum_{k=1}^{T}\exp(Z^{(h)}_{ik})}, \quad A^{(h)} = \text{Softmax}(Z^{(h)}/\sqrt{d})$$

We extract token-level attention marginal by row: $A^{(h)}_{i,:}$, and define entropy loss across heads:

$$\mathcal{L}_{\text{aux}} = \frac{1}{HT}\sum_{h=1}^{H}\sum_{i=1}^{T}\left(-\sum_{j=1}^{T} A^{(h)}_{ij}\log A^{(h)}_{ij}\right)$$

This entropy term encourages diverse and non-overlapping attention focus across heads, mitigating redundancy and improving temporal coverage. In practice, the auxiliary loss is

applied to all attention blocks in the Transformer encoder, with averaged contribution.

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

$$\frac{\partial \mathcal{L}_{\text{aux}}}{\partial Z_{ij}^{(h)}} = A_{ij}^{(h)} \left(1 + \log A_{ij}^{(h)} - \sum_k A_{ik}^{(h)} \log A_{ik}^{(h)}\right),$$

where $A^{(h)} = \text{Softmax}(Z^{(h)})$, and $A_{ij}^{(h)}$ denotes the attention weight from token i to j in head h. The term $\sum_k A_{ik}^{(h)} \log A_{ik}^{(h)}$ corresponds to the (negative) Shannon entropy $H(A_{i,:}^{(h)})$ of the attention row, thus the gradient can be rewritten as:

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

This gradient formulation has several intuitive implications:

- The presence of $A_{ij}^{(h)}$ scales the gradient proportionally to the attention itself. Positions receiving higher attention contribute more to the update.
- The term $\log A_{ij}^{(h)}$ penalizes overly confident (peaked) distributions. When $A_{ij} \approx 1$, $\log A_{ij} \approx 0$, and the gradient is positive, reducing the corresponding logit and flattening the softmax output.
- The entropy term $H(A_{i,:}^{(h)})$ modulates the total force of correction. High-entropy rows (i.e., diverse attention) reduce the gradient magnitude, effectively preserving already-distributed attention maps.

Table 1: Layer-wise output sizes in CrimNet. T denotes temporal length, H/W are spatial dimensions, and d is 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 input correlations $I(\tilde{x}; x)$, effectively reducing noise propagation to downstream components. Second, the introduction of an **auxiliary loss** facilitates gradient flow by providing intermediate supervision, which enhances optimization and accelerates convergence. Lastly, the attention entropy regularization mitigates head collapse in the Transformer, encouraging diverse temporal focus and improving both stability and interpretability of attention maps.

CrimNet outputs clip-level scores $\hat{y}_i$ and interpolates them to frame-level using linear upsampling, following prior protocols (Sultani, Chen, and Shah 2018b).

Experimental Results

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.

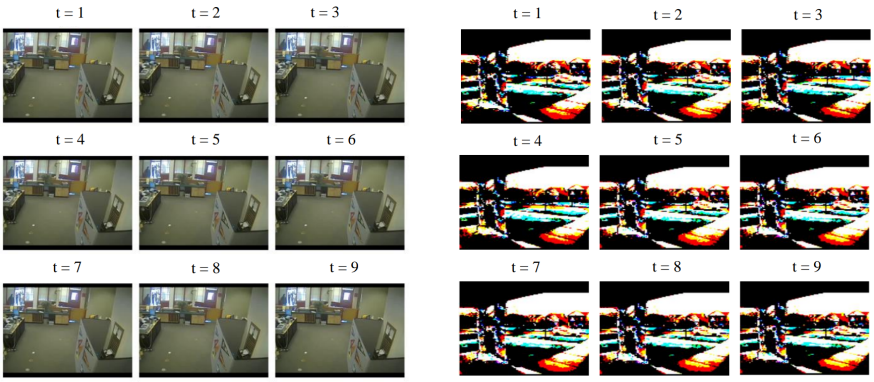
Figure 2 presents a visual overview of the dataset. Subfigure (a) shows representative frames sampled from various anomaly classes, highlighting the diversity in scene context and motion. Subfigure (b) illustrates the distribution of video samples across all 14 classes, with pronounced imbalance in categories such as *Robbery* and *Abuse*.

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 (Sultani, Chen, and Shah 2018b; Tian et al. 2021). We also report false alarm rate (FAR) where applicable. Visualizations are performed using attention-weighted scores and detection heatmaps.

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 weakly supervised methods. Compared to the most competitive prior approach, MSF (Qian et al. 2025), 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.



(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

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.

Analysis and Interpretability

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 CrimNet model achieves the best result—**85.12%** on UCF and **67.23%** on DVS.

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

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

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	✗	✗	68.23	23.00	60.23	21.56
C3D	✓	✗	70.21	30.22	61.23	24.23
C3D	✗	✓	75.23	32.14	62.15	24.11
C3D	✓	✓	**85.12**	**37.14**	**70.69**	**29.76**
R3D	✗	✗	67.42	22.10	59.73	20.50
R3D	✓	✗	69.33	29.20	60.91	23.42
R3D	✗	✓	74.10	31.05	61.40	23.98
R3D	✓	✓	77.80	36.22	66.30	28.50
R(2+1)D	✗	✗	67.95	22.56	59.94	20.83
R(2+1)D	✓	✗	69.80	29.80	61.01	23.87
R(2+1)D	✗	✓	74.65	31.55	61.73	24.20
R(2+1)D	✓	✓	76.30	36.75	66.70	29.10

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.

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

Conclusion

We presented **CrimNet**, a modular two-stage architecture for weakly supervised video anomaly detection that addresses the core challenges of temporal ambiguity and label sparsity. Our method combines a lightweight 3D CNN-based feature selector for early-stage spatial noise filtering, with a Transformer encoder augmented by an entropy-regularized auxiliary head to improve long-range temporal modeling under weak supervision.

Through comprehensive experiments on UCF-Crime and the event-driven UCF-Crime-DVS dataset, CrimNet achieves state-of-the-art frame-level AUC scores, surpassing previous methods by up to **+5.68%**. Ablation studies confirm the complementary roles of spatial filtering and temporal attention regularization, with consistent gains across backbones (C3D,

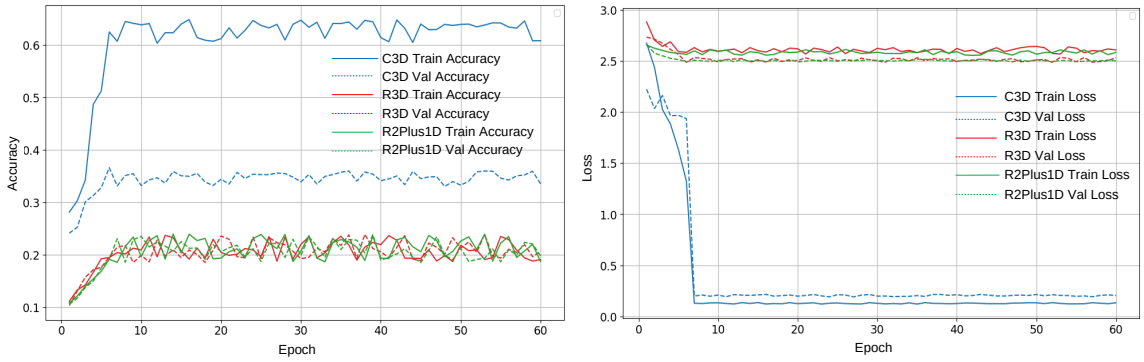


Figure 3: Training dynamics of CrimNet with C3D, R3D, and R(2+1)D under the full Two-Stage + Self-Attention setup. C3D shows better convergence and generalization.

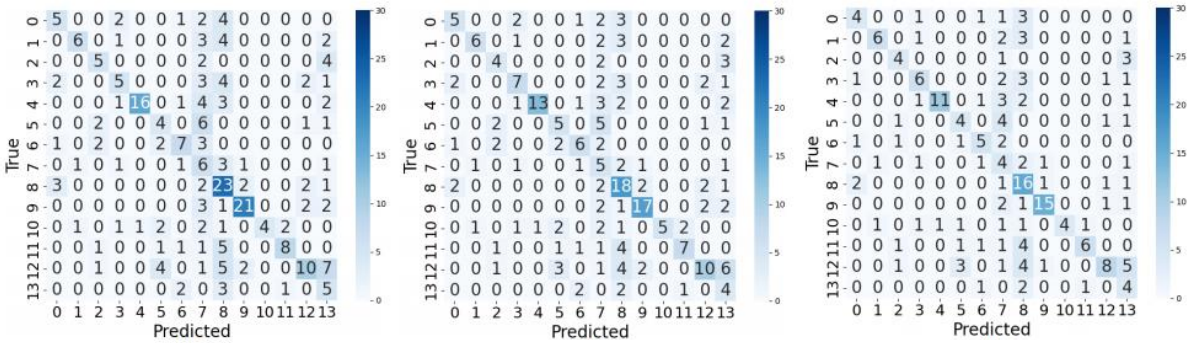


Figure 4: Confusion matrices of CrimNet on C3D (a), R3D (b), and R(2+1)D (c) backbones. Diagonal strength indicates classification quality; C3D demonstrates higher consistency.

R3D, R(2+1)D) and input modalities.

From a learning dynamics perspective, our auxiliary attention entropy loss serves as an effective inductive bias that stabilizes training and encourages attention head diversity. Visualization results further reveal interpretable temporal specialization among heads, supporting the theoretical motivation behind our entropy-based regularization.

CrimNet offers a simple yet effective framework that is broadly applicable to real-world surveillance scenarios. Its plug-and-play nature allows seamless integration with existing models, making it well-suited for deployment in weakly labeled or resource-constrained environments.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. ViViT: A video vision transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 6836–6846.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 8134–8148.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6299–6308.

Chandola, V.; Banerjee, A.; and Kumar, V. 2009. Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3): 1–58.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Dai, J.; and Malik, J. 2021. Multiscale vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 6824–6835.

Feichtenhofer, C. 2020. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 203–213.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. 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)*, 6546–6555.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J. A.; van Ginneken, B.; and Sanchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88.

Nayak, R.; Pati, U. C.; and Das, S. K. 2021. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106: 104078.

Ngai, E. W.; Hu, Y.; Wong, Y.; Chen, Y.; and Sun, X. 2011. 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.

Park, H.; Noh, J.; and Ham, B. 2020. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14372–14381.

Qian, Y.; Ye, S.; Wang, C.; Cai, X.; Qian, J.; and Wu, J. 2025. 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.

Sultani, W.; Chen, C.; and Shah, M. 2018a. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6479–6488.

Sultani, W.; Chen, C.; and Shah, M. 2018b. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6479–6488.

Tian, Y.; et al. 2021. Weakly-supervised Action Localization with Multi-Instance Learning and Self-Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tong, Z.; Song, Y.; Wang, J.; and Shen, J. 2022. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4489–4497.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6450–6459.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Zhang, Y.; Yang, Y.; Ma, Z.; Schiele, B.; and Wang, S. 2022. ActionFormer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 492–510.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) NA

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) NA

2.4. Proofs of all novel claims are included (yes/partial/no) NA

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) NA

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) NA

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) NA

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) NA

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the pa-

per with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) partial

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) partial

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) partial

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) NA

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) no

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) no

Appendix

A. Experimental Settings

We summarize the key experimental configurations used in CrimNet training and evaluation. All experiments were conducted using PyTorch on an NVIDIA RTX 4090 GPU.

Table 4: Training Configuration Details

Parameter	**Value**
GPU	NVIDIA RTX 4090
Batch Size	16
Optimizer	Adam
Learning Rate	1×10^{-5}
Weight Decay	5×10^{-4}
Scheduler	CosineAnnealingLR ($T_{max} = 10$)
Epochs	50
Backbone Network	C3D / R3D / R(2+1)D
Transformer Heads	8
Transformer Embedding Size	512
Auxiliary Loss	Attention Entropy Loss
Training Time	$\sim$3 hours / model
Framework	PyTorch 2.0
Dataset	UCF-Crime-DVS

B. Lightweight Deployment Analysis

We explored a variant of CrimNet by replacing the Transformer encoder with a standard Vision Transformer (ViT) block. However, we observed that under the same GPU environment (NVIDIA RTX 4090) and hyperparameter settings (e.g., batch size = 16, 8 attention heads), the ViT-based variant encountered out-of-memory issues during training. To make it fit within GPU memory constraints, we had to reduce the number of attention heads and shrink intermediate embedding dimensions.

Such reductions, however, significantly weakened the model's temporal modeling capability, leading to a noticeable drop in performance. This supports our design choice: the current CrimNet architecture, which incorporates a compact and stable Transformer encoder with auxiliary attention heads, achieves a better trade-off between accuracy and deployability.

C. Attention Map Visualization with and without Entropy Loss

To further illustrate the effect of our auxiliary entropy loss, we visualize the attention maps from a representative attention head under two settings: with and without the proposed auxiliary supervision. As shown in Figure 5, the model without entropy loss tends to focus heavily on the diagonal, indicating a collapse of attention and limited diversity. In contrast, the entropy-supervised model displays a more distributed attention pattern, demonstrating its ability to attend to a broader range of classes. This validates the effectiveness of the auxiliary loss in promoting attention diversity and enhancing interpretability.

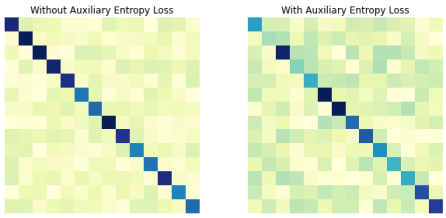


Figure 5: Attention heatmaps comparison of a single head with/without auxiliary entropy loss. The auxiliary objective helps mitigate attention collapse and encourages diverse focus across class regions.