TLMA: Mitigating the Impact of Weakly Labeled Information for Video Anomaly Detection

Supplementary Material

7. Ablation Study Analysis

Table 8. Module ablation study on UCF-Crime dataset with detailed metrics. Best results are in **bold**.

Baseline	MA	$\mathcal{L}_{\text{Triplet}}$	AUC	AP	FAR_a	FAR_n
✓			85.69	25.79	75.79	3.30
✓	✓		86.84	27.09	67.29	2.62
✓		✓	88.37	40.60	30.49	1.93
✓	✓	✓	**89.47**	**44.09**	**29.49**	**1.56**

Table 9. Module ablation study on XD-Violence dataset with detailed metrics. Best results are in **bold**.

Baseline	MA	$\mathcal{L}_{\text{Triplet}}$	AUC	AP	FAR_a	FAR_n
✓			93.36	83.61	63.51	1.07
✓	✓		95.30	85.35	57.56	1.11
✓		✓	95.05	84.64	40.22	0.96
✓	✓	✓	**95.63**	**86.78**	**28.24**	**0.43**

We present the module ablation experimental results. FAR_a denotes the false alarm rate on abnormal videos, while FAR_n represents the false alarm rate on normal videos.

On UCF-Crime (Table 8), the triplet learning strategy provides the most substantial gains, improving AUC from 85.69% to 88.37% and AP to 40.60%, while reducing FAR_a from 75.79% to 30.49%. The complete TLMA framework achieves the best overall performance with 89.47% AUC and 44.09% AP.

On XD-Violence (Table 9), both components contribute to performance gains. The full framework achieves optimal results with 95.63% AUC and 86.78% AP, while attaining the lowest false alarm rates.

8. Fine-Grained Class-Wise Analysis

This section provides a comprehensive analysis of class-wise performance across all three datasets, combining detailed TLMA results with comparative analysis against baseline methods to offer complete insights into different anomaly types.

8.1. Overall Performance Patterns

As shown in Table 10, the detailed class-wise performance reveals important characteristics of different anomaly types.

On UCF-Crime, `Abuse` achieves the highest AUC (96.01%), while `Stealing` obtains the best AP (83.52%), indicating that certain social anomalies are more easily detectable than others. However, classes like `Explosion` (16.66% AP) and `Shoplifting` (15.58% AP) show significantly lower performance, suggesting challenges in detecting brief or visually subtle events.

On XD-Violence, `Riot` demonstrates exceptional performance with 97.85% AP, reflecting the effectiveness in capturing collective violent behaviors. The consistently high AP scores across all categories indicate robust violence detection capability.

MSAD exhibits the most varied performance pattern, with `Water_incident` achieving near-perfect scores while `People_falling` struggles significantly, highlighting diverse challenge levels.

8.2. Comparative Analysis with Baseline

As illustrated in Figures 8, 9 and 10, TLMA demonstrates consistent improvements over the baseline across most anomaly categories.

On **UCF-Crime** (Figure 8), TLMA achieves significant performance gains across 12 out of 13 categories, with particularly notable improvements in challenging classes such as `Explosion` (AUC: 38.77% → 63.69%) and `RoadAccidents` (AUC: 57.30% → 67.03%). The most substantial absolute improvement is observed in `Fighting` (AUC: 79.42% → 92.12%), indicating enhanced capability in detecting physical violence.

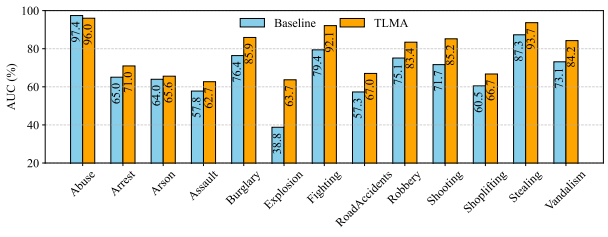

Figure 8. Performance comparison between baseline and TLMA across different anomaly categories on UCF-Crime dataset.

For **XD-Violence** (Figure 9), TLMA maintains strong performance across all violence categories while providing meaningful improvements in challenging scenarios. The most significant enhancement is observed in `Car accident` (AP: 45.78% → 51.78%), suggesting better handling of complex traffic incidents.

Table 10. Detailed class-wise performance across all datasets.

UCF-Crime			XD-Violence			MSAD		
Class	**AUC**	**AP**	**Class**	**AUC**	**AP**	**Class**	**AUC**	**AP**
Abuse	96.01	61.25	Fighting	90.13	89.84	Assault	73.09	79.48
Arrest	70.96	52.06	Car accident	69.02	51.78	Explosion	92.18	93.55
Arson	65.58	38.94	Shooting	84.93	76.46	Fighting	82.84	89.64
Assault	62.68	42.29	Explosion	84.99	73.64	Fire	92.97	96.40
Burglary	85.91	67.18	Riot	93.13	97.85	Object_falling	90.65	94.97
Explosion	63.69	16.66	Abuse	85.29	75.31	People_falling	63.11	55.54
Fighting	92.12	84.10				Robbery	70.85	87.61
RoadAccidents	67.03	15.99				Shooting	80.28	84.00
Robbery	83.43	78.45				Traffic_accident	71.08	60.16
Shooting	85.20	46.06				Vandalism	89.63	85.77
Shoplifting	66.73	15.58				Water_incident	99.99	100.00
Stealing	93.67	83.52						
Vandalism	84.25	70.45						
Abnormal	76.16	45.45	Abnormal	86.20	86.23	Abnormal	79.52	84.42
Overall	89.47	44.09	Overall	95.63	86.78	Overall	93.68	81.30

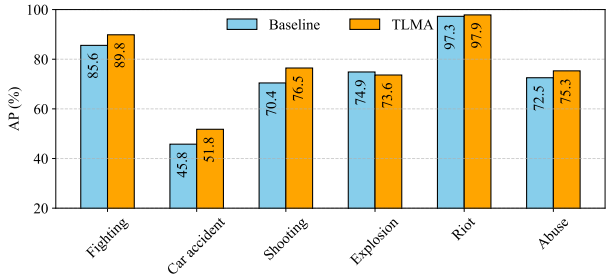

Figure 9. Performance comparison between baseline and TLMA across different anomaly categories on XD-Violence dataset.

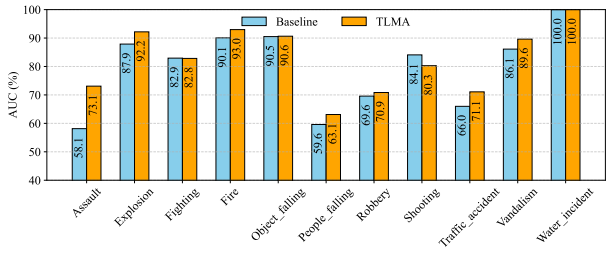

Figure 10. Performance comparison between baseline and TLMA across different anomaly categories on MSAD dataset.

On **MSAD** (Figure 10), TLMA shows particularly strong improvements in difficult categories such as `Assault` (AUC: 58.14% $\rightarrow$ 73.09%) and `Traffic_accident` (AUC: 65.99% $\rightarrow$ 71.08%), indicating better generalization to complex real-world scenarios.

Across all datasets, violence-related categories (`Fighting`, `Assault`, `Shooting`) generally show stable performance in both absolute terms and relative improvements, while transient events (`Explosion`, `RoadAccidents`) and subtle activities (`Shoplifting`, `People_falling`) present greater detection challenges but benefit most from our approach.

The consistent performance improvements demonstrate TLMA's effectiveness in enhancing anomaly detection capability, particularly for challenging categories that require sophisticated temporal and motion understanding. The framework shows robust generalization across different anomaly types and dataset characteristics, with the comparative analysis confirming its superiority over baseline methods in handling diverse anomaly detection scenarios.

9. Qualitative Results

9.1. Anomaly Score

As shown in Figure 11, the qualitative results on the XD-Violence dataset demonstrate the superior performance of our TLMA framework in accurately localizing anomalous events. The anomaly scores generated by TLMA (blue curve) exhibit closer alignment with the ground-truth anomaly intervals (pink regions) compared to the baseline method (orange curve). Notably, TLMA produces fewer false alarms in normal video segments while maintaining higher precision in detecting actual anomalous events. This visualization confirms that our approach effectively reduces misclassification caused by weakly labeled information and provides more precise anomaly detection.

As shown in Figure 12, the qualitative results on the MSAD dataset further validate the effectiveness of our

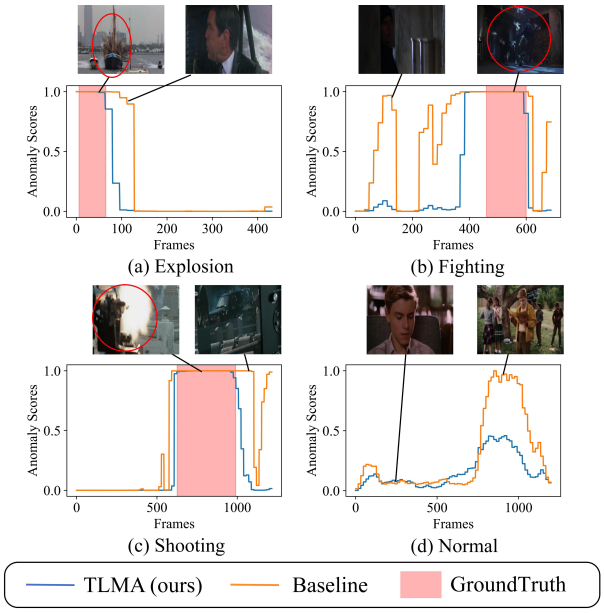

Figure 11. Qualitative results on XD-Violence dataset. The red circles highlight specific occurrences of anomalies.

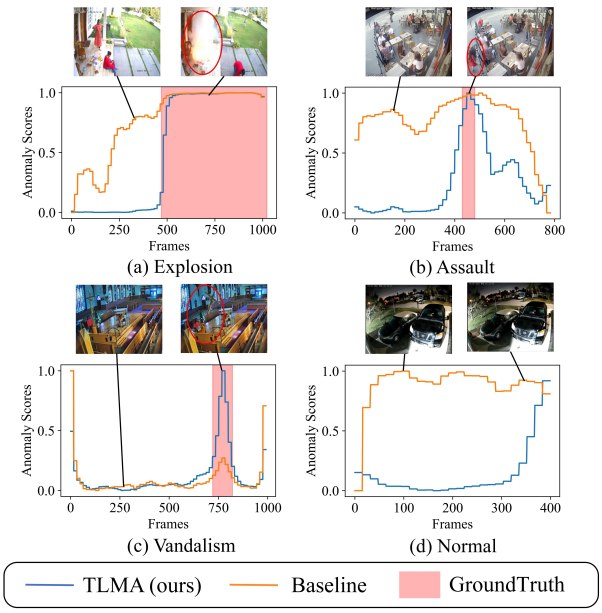

Figure 12. Qualitative results on MSAD dataset. The red circles highlight specific occurrences of anomalies.

TLMA framework. The anomaly scores produced by TLMA demonstrate significantly better alignment with the ground-truth annotations compared to the baseline method. Notably, as illustrated in subfigure (d), our approach substantially reduces false alarms in normal video segments where the baseline method generates nearly continuous false detections. This visual evidence confirms that TLMA

Figure 13. MA visualization in a tracking camera scenario. The subject (red box) is prioritized via relative motion saliency despite background interference. Per Eq. (7), the module adaptively converges to global pooling under intense motion to ensure robust feature aggregation.

effectively mitigates the misclassification issues prevalent in weakly-supervised settings, providing more reliable and precise anomaly localization in complex surveillance scenarios.

9.2. Case Study on Tracking Cameras

Our Motion Aware (MA) module demonstrates robust generalization to complex surveillance data within tracking camera scenarios, as visualized in Fig. 13. 1) Adaptive Convergence: In scenarios where rapid camera tracking fills the motion map with high-intensity gradients, Eq. (7) enables the module to adaptively revert to a global pooling strategy, thereby safeguarding the semantic integrity of the extracted features. 2) Relative Motion Discernment: Since Sobel filters capture local edge gradients rather than absolute displacement, the module prioritizes subjects as long as their motion differs from the global background translation. Even amidst background interference, the unique relative motion trajectory of the subject (red box) remains salient, ensuring effective feature aggregation.