From Frames to Events: Rethinking Evaluation in Human-Centric Video Anomaly Detection

Narges Rashvand, Shanle Yao, Armin Danesh Pazho, Babak Rahimi Ardabili, Hamed Tabkhi

Dept. of Electrical and Computer Engineering, UNC Charlotte

Problem Description

The current standard for evaluating Video Anomaly Detection (VAD) relies on frame-level metrics (e.g., AUC-ROC), which treat videos as collections of isolated snapshots. This paradigm is fundamentally misaligned with real-world operational requirements and the nature of human activities.

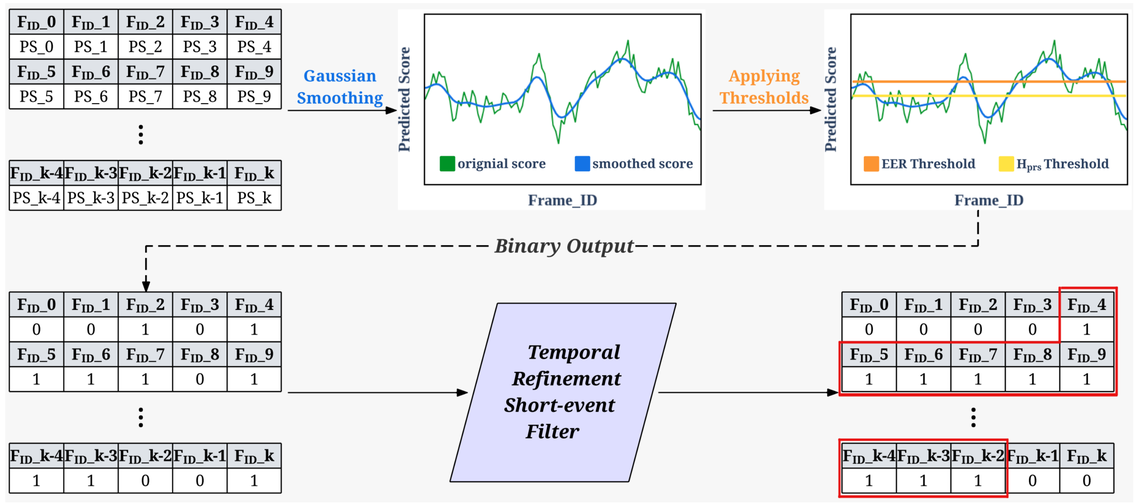
Comparison between frame-level and event-level evaluation paradigms for video anomaly detection (VAD)

Aspects	Frame-Level	Event-Level
Unit of analysis	Individual frames	Temporal events
Consideration of temporal continuity	No	Yes
Boundary localization	Cannot measure	Measured via tIoU
Operational trust	Low	High
Sensitivity to class imbalance	High	Lower
Suitable for real-world deployment	Limited	Yes

The Core Issues:

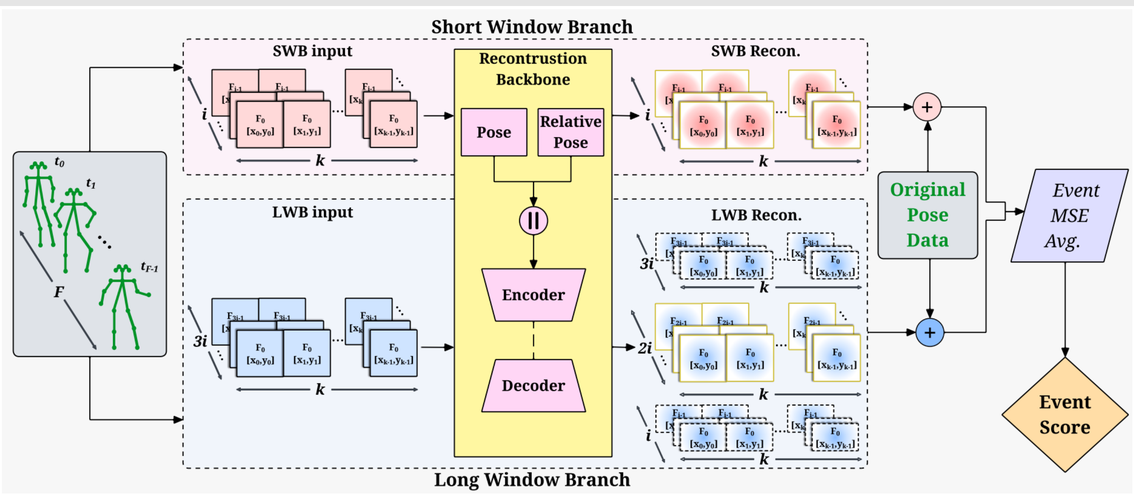
- Temporal Misalignment: Real-world anomalies are continuous temporal processes with a distinct onset, progression, and end. Frame-level metrics ignore the continuity of anomalies and cannot measure event boundaries or duration.
- Performance Overestimation: Models can achieve high frame-level scores by correctly identifying normal frames while failing to localize the actual boundaries of an anomalous event.
- Operational Irrelevance: Security personnel and surveillance operators respond to incidents, not individual frames. Frequent switching between normal and anomaly leads to many false alarms and poor real-world use.
- Incentivizing Poor Design: The reliance on frame-wise metrics encourages architectures optimized for snapshots rather than temporal reasoning and event coherence.

The Score-Refinement Pipeline



The three-stage Frame-to-Event Transformation framework smooths raw anomaly scores, applies adaptive thresholds, and refines detections to produce coherent anomalous events aligned with human motion dynamics.

Dual-Branch Reconstruction Event VAD



Dual-branch event-level anomaly detection framework. An input pose sequence is processed by Short- and Long-Window branches sharing a transformer-based reconstruction model for absolute and relative pose modeling. During inference, reconstruction errors from both branches are aligned, fused, and temporally pooled to generate an event-level anomaly score.

Results

Frame-level evaluation performance of pose-based VAD models

Metric	STG-NF[15] SHT[19]	STG-NF[15] CHAD[6]	STG-NF[15] NWPUC[4]	STG-NF[15] HuVAD[25]	SPARTA[24] SHT[19]	SPARTA[24] CHAD[6]	SPARTA[24] NWPUC[4]	SPARTA[24] HuVAD[25]	TS-GAD[23] SHT[19]	TS-GAD[23] CHAD[6]	TS-GAD[23] NWPUC[4]	TS-GAD[23] HuVAD[25]
AUC-ROC	0.866	0.570	0.617	0.520	0.861	0.569	0.635	0.649	0.807	0.553	0.618	0.623
AUC-PR	0.838	0.569	0.276	0.252	0.831	0.517	0.258	0.327	0.740	0.499	0.300	0.305
EER	0.218	0.458	0.408	0.478	0.225	0.444	0.408	0.409	0.256	0.461	0.418	0.417
$F1@\tau_{EER}$	0.746	0.52	0.330	0.347	0.738	0.538	0.329	0.413	0.703	0.521	0.322	0.406
$F1@\tau_{H_{prs}}$	0.744	0.518	0.323	0.377	0.729	0.536	0.332	0.405	0.692	0.521	0.322	0.421

Results

Frame-level and event-level statistics of VAD benchmarks

Granularity	Characteristic	SHT[19]	CHAD[6]	HuVAD[25]	NWPUC[4]
Frame	Normal Frames	24,077	67,303	694,415	318,793
	Anomalous Frames	16,714	59,172	225,075	65,266
Event	Anomalous Events	121	190	1,691	137
	Avg. Duration (f)	138.13	311.43	133.10	476.39

Event-level performance of pose-based VAD models, and our proposed framework, across the SHT, CHAD, NWPUC, and HuVAD datasets, using EER Threshold. Results are compared across three configurations: Baselines, Baseline+Post-processing, and our natively event-aware model.

Data	Model	$tIOU=0.5$ Prec.	$tIOU=0.5$ Rec.	$tIOU=0.5$ F_1	$tIOU=0.4$ Prec.	$tIOU=0.4$ Rec.	$tIOU=0.4$ F_1	$tIOU=0.3$ Prec.	$tIOU=0.3$ Rec.	$tIOU=0.3$ F_1	$tIOU=0.2$ Prec.	$tIOU=0.2$ Rec.	$tIOU=0.2$ F_1	Average F_1
SHT[19]	**Baselines**													
	STG-NF[15]	50.59%	71.07%	0.591	60.00%	84.30%	0.700	61.76%	86.78%	0.721	64.71%	90.91%	0.756	0.692
	TS-GAD[23]	36.79%	58.68%	0.452	45.08%	71.90%	0.554	54.92%	87.60%	0.675	57.51%	91.74%	0.707	0.597
	SPARTA[24]	44.62%	68.60%	0.540	52.15%	80.17%	0.631	56.45%	86.78%	0.684	0.596%	91.74%	0.723	0.644
	Baselines + Proposed Post-Processing													
	STG-NF[15] w/ ES	55.41%	71.90%	0.625	63.06%	81.82%	0.712	66.88%	86.78%	0.755	69.43%	90.08%	0.784	0.7193
	TS-GAD[23] w/ ES	41.38%	59.50%	0.488	50.00%	71.90%	0.589	59.70%	85.95%	0.705	63.79%	91.74%	0.752	0.6335
	SPARTA[24] w/ ES	48.19%	66.12%	0.557	56.63%	77.69%	0.655	62.65%	85.95%	0.724	63.86%	87.60%	0.738	0.6685
	Dual-Branch Event Detection													
	Dual (Ours)	46.27%	59.59%	0.521	55.26%	76.94%	0.643	63.09%	87.69%	0.734	66.45%	87.12%	0.754	0.662
CHAD[6]	**Baselines**													
	STG-NF[15]	7.31%	23.68%	0.111	11.69%	37.89%	0.178	17.05%	55.26%	0.260	22.08%	71.58%	0.337	0.222
	TS-GAD[23]	3.92%	13.68%	0.061	8.14%	28.42%	0.126	13.57%	47.37%	0.211	18.40%	64.21%	0.286	0.171
	SPARTA[24]	10.00%	25.26%	0.143	14.58%	36.84%	0.209	21.25%	53.68%	0.304	26.25%	66.32%	0.376	0.258
	Baselines + Proposed Post-Processing													
	STG-NF[15] w/ ES	8.21%	24.21%	0.122	11.96%	35.26%	0.178	18.39%	54.21%	0.274	23.57%	69.47%	0.352	0.231
	TS-GAD[23] w/ ES	5.38%	16.32%	0.080	10.24%	31.05%	0.154	16.32%	49.47%	0.245	21.53%	65.26%	0.323	0.200
	SPARTA[24] w/ ES	10.91%	25.26%	0.152	16.14%	37.37%	0.225	23.18%	53.68%	0.323	28.41%	65.79%	0.396	0.274
	Dual-Branch Event Detection													
	Dual (Ours)	16.71%	30.53%	0.216	25.07%	45.79%	0.324	34.87%	63.68%	0.451	39.77%	72.63%	0.514	0.376
NWPUC[4]	**Baselines**													
	STG-NF[15]	3.92%	23.36%	0.067	5.63%	33.58%	0.096	6.98%	41.61%	0.119	9.55%	56.93%	0.163	0.111
	TS-GAD[23]	3.89%	21.32%	0.065	5.23%	28.68%	0.088	7.65%	41.91%	0.129	9.93%	54.41%	0.168	0.112
	SPARTA[24]	3.76%	23.36%	0.064	5.28%	32.85%	0.091	7.28%	45.26%	0.125	9.98%	62.04%	0.171	0.113
	Baselines + Proposed Post-Processing													
	STG-NF[15] w/ ES	4.39%	24.09%	0.074	5.99%	32.85%	0.101	7.86%	43.07%	0.132	10.12%	55.47%	0.171	0.119
	TS-GAD[23] w/ ES	4.26%	21.32%	0.071	5.73%	28.68%	0.095	8.22%	41.18%	0.137	10.72%	53.68%	0.178	0.120
	SPARTA[24] w/ ES	4.13%	23.36%	0.070	5.68%	32.12%	0.096	7.62%	43.07%	0.129	10.98%	62.04%	0.186	0.120
	Dual-Branch Event Detection													
	Dual (Ours)	4.00%	21.90%	0.068	5.39%	33.58%	0.093	8.42%	57.66%	0.147	9.50%	64.96%	0.166	0.118
HuVAD[25]	**Baselines**													
	STG-NF[15]	15.12%	24.48%	0.186	22.78%	36.90%	0.281	30.60%	49.56%	0.378	39.43%	63.87%	0.487	0.333
	TS-GAD[23]	13.83%	26.91%	0.182	20.45%	39.80%	0.270	28.14%	54.76%	0.371	34.55%	67.24%	0.456	0.320
	SPARTA[24]	17.19%	31.99%	0.223	23.86%	44.41%	0.310	32.02%	59.61%	0.416	39.99%	74.45%	0.520	0.367
	Baselines + Proposed Post-Processing													
	STG-NF[15] w/ ES	15.85%	24.31%	0.191	23.91%	36.66%	0.289	31.78%	48.73%	0.384	41.50%	63.63%	0.502	0.341
	TS-GAD[23] w/ ES	15.18%	26.97%	0.194	22.38%	39.74%	0.286	30.00%	53.28%	0.383	37.26%	66.17%	0.476	0.334
	SPARTA[24] w/ ES	17.80%	31.4%	0.227	24.71%	43.58%	0.315	33.46%	59.02%	0.427	41.80%	73.74%	0.533	0.375
	Dual-Branch Event Detection													
	Dual (Ours)	22.20%	33.18%	0.266	30.30%	45.29%	0.363	39.31%	58.75%	0.471	47.41%	70.86%	0.568	0.417

Acknowledgment

This research is supported by the National Science Foundation (NSF) under Award Number 2329816.

GitHub Repository

References

Noghre, G. A., Pazho, A. D., & Tabkhi, H. (2024). Human-centric video anomaly detection through spatio-temporal pose tokenization and transformer. *arXiv preprint arXiv:2408.15185*.

Hirschorn, O., & Avidan, S. (2023). Normalizing flows for human pose anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13545 -13554).

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection – a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.