Efficient Video Anomaly Localization through Persistent Memory Modeling and Temporal Feature Refinement

Cesar Lindberg
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
cesar67@unh.edu

Matteo Norris
Department of Computer Science, University of Houston, Houston, TX, USA.
hellomatteo@uh.edu

Abstract

Video anomaly localization has become a critical component of intelligent surveillance and industrial monitoring infrastructures, yet existing approaches often suffer from prohibitive computational costs and a lack of temporal coherence, impeding real-world deployment in large-scale systems. This paper presents a novel framework that integrates persistent memory modeling with temporal feature refinement to enable efficient and accurate video anomaly localization. The persistent memory module maintains a compact, dynamically updated prototypical representation of normal behavioral patterns across long sequences, while the temporal feature refinement mechanism enhances discriminative power by adaptively aggregating multi-scale motion and appearance cues. By decoupling memory storage from frame-wise inference depth and employing lightweight temporal convolutions, the architecture achieves significant reductions in latency and memory footprint without sacrificing localization fidelity. The discussion extends beyond algorithm design to system-level considerations, encompassing deployment on heterogeneous edge-cloud continuums, governance of surveillance data, fairness audits under demographic and environmental variations, and sustainable engineering practices. Through an interdisciplinary lens, the paper analyzes structural trade-offs in memory capacity, update strategies, and feature granularity, offering a blueprint for anomaly localization systems that balance performance, robustness, and socio-technical accountability. The proposed design principles are contextualized within the broader landscape of video understanding, smart city governance, and responsible artificial intelligence, highlighting pathways toward ethically aligned and operationally resilient large-scale surveillance infrastructures.

Keywords

video anomaly localization, persistent memory, temporal feature refinement, efficient deep learning, socio-technical systems, edge computing, algorithmic fairness.

1. Introduction

The proliferation of visual sensing networks in public spaces, manufacturing floors, and transportation hubs has generated an unprecedented demand for automated video anomaly detection and localization. Early systems relied on handcrafted features and simple statistical models, but the advent of deep learning has markedly improved the ability to recognize deviations from normal patterns. Despite these advances, a persistent tension exists between

the need for high-fidelity spatiotemporal modeling and the constraints of real-time, resource-limited deployment environments. Most contemporary anomaly detectors operate by reconstructing frames or predicting future appearances, implicitly encoding normalcy in the parameters of deep neural networks [1, 2, 3, 4]. However, such monolithic parameterization often struggles to retain long-range temporal dependencies, resulting in fragile representations that degrade under gradual environmental shifts or rare but benign events. Moreover, the computational footprint of three-dimensional convolutions, recurrent networks, and dense attention mechanisms frequently renders these models impractical for continuous operation on edge devices or within legacy surveillance infrastructure.

In response to these challenges, this paper proposes an integrated framework that systematically combines persistent memory modeling with dedicated temporal feature refinement. The central argument is that a modular separation between a long-lived normative memory store and a lightweight temporal processing front-end can simultaneously improve localization accuracy and operational efficiency. The persistent memory serves as an evolving catalog of prototypical normal snippets, continuously updated through a carefully designed write mechanism that prevents catastrophic forgetting while adapting to gradual concept drift. The temporal feature refinement module, in turn, operates on incoming video streams to produce noise-resistant, multi-scale temporal representations that can be queried against the memory bank with minimal latency. This design draws inspiration from memory-augmented neural networks and recent advances in efficient video understanding, extending those concepts into the anomaly localization domain. The subsequent sections provide an in-depth examination of architectural choices, system-level deployment trade-offs, robustness considerations, and the governance frameworks necessary for responsible deployment. Through this comprehensive treatment, the paper contributes not only a novel technical methodology but also a systems-thinking perspective on building sustainable, fair, and accountable anomaly detection pipelines.

2. Related Work

Video anomaly detection has been approached through reconstruction-based, prediction-based, and self-supervised paradigms. Reconstruction-oriented architectures, such as spatiotemporal autoencoders, learn to compress and regenerate normal frames, flagging high reconstruction errors as anomalies [5]. Prediction-based methods shift the focus toward forecasting future frames from historical observations, excelling at capturing motion regularities but often introducing blur and failing to model static anomalous objects [3]. Memory-augmented autoencoders advanced the state of the art by introducing an external memory module that stores prototypical normal patterns, thereby limiting the ability of the network to reconstruct anomalous inputs even when the generator is powerful [2]. This memory-centric philosophy is especially relevant for persistent memory design, as it decouples representational capacity from the depth of the encoder and decoder. Concurrently, the sequence modeling community has systematically evaluated generic temporal convolutions and recurrent architectures, demonstrating that dilated causal convolutions can match or surpass long short-term memory networks on many tasks while offering superior parallelism [8, 9]. Three-dimensional convolutional architectures like the Inflated 3D ConvNet and the attention-based non-local neural network introduced mechanisms to capture long-range spatiotemporal dependencies, but at considerable computational expense [7, 10]. The Transformer architecture and its self-attention mechanism revolutionized sequence modeling, subsequently giving rise to powerful video transformers and foundation models capable of learning from massive unlabeled

corpora [11, 12, 18]. In the specific domain of efficient video object segmentation, recent contributions have explored memory-aware fine-tuning of foundation segmentation models to retain spatiotemporal correspondences through compact memory banks, enabling efficient long-sequence processing [17]. The concept of refining temporal features via lightweight modules has also gained traction through multi-scale feature pyramids and depthwise separable temporal convolutions [19]. On the systems side, serving infrastructures like Clipper and variants of TensorFlow Serving have illustrated the importance of model encapsulation, batching, and hardware-aware optimization for real-time machine learning [13]. This paper synthesizes these disparate threads into a coherent framework that jointly optimizes memory persistence, temporal refinement, and deployability, carefully navigating the algorithmic and infrastructural trade-offs that govern real-world video analytics.

3. Persistent Memory Modeling for Video Anomaly Localization

The persistent memory module constitutes the conceptual core of the proposed framework, acting as a long-lived repository of normal behavioral primitives. Unlike conventional memory-augmented autoencoders that reset or passively update memory items during each training epoch, the persistent memory design incorporates a lifelong update policy that continuously adapts to non-stationary environments without discarding older, still-relevant patterns. The memory bank is structured as a matrix of slots, each holding a feature prototype that encapsulates a specific normal spatiotemporal snippet. During both training and inference, an input frame sequence is decomposed into short clips, each mapped to a query embedding by a lightweight feature extractor. A read operator computes attention weights between the query and all memory slots, producing a reconstructed representation that serves as a normalcy-consistent approximation of the input. Crucially, a write operator selectively updates memory slots using a moving average of the query embedding and a utilization-based replacement strategy that prioritizes the retention of frequently accessed prototypes while allowing rarely activated slots to be slowly overwritten by novel normal modes. This dual read-write dynamic ensures that the memory remains both representative of long-term statistical modes and agile enough to absorb gradual domain shifts, such as changes in lighting, scene layout, or crowd density.

The choice of memory capacity and update rate embodies a fundamental trade-off between specificity and generalization. A large memory bank can store fine-grained normal variations but incurs larger computational overhead and risks overfitting to spurious correlations. Conversely, a compact memory forces aggressive compression that may blur the boundary between subtle anomalies and uncommon but legitimate behaviors. The persistent update scheme mitigates this tension by employing a sliding window consolidation strategy: memory slots are periodically evaluated on a held-out calibration set, and slots that consistently fail to contribute to normal reconstruction are pruned or merged. This pruning routine is executed asynchronously on a dedicated maintenance thread, ensuring that the online inference path remains unblocked. To further stabilize training, a memory distillation loss encourages the output of the read operator to match the input embedding only when the input corresponds to a normal event, while an anomaly rejection loss penalizes faithful reconstructions of abnormal clips. Through this design, the persistent memory learns to map anomalous inputs to the closest plausible normal prototype, thereby producing a large reconstruction discrepancy that serves as a localization signal. The architectural separation of memory from the temporal processing engine also facilitates independent scaling: the memory bank can be hosted on a

centralized server while lightweight query encoders run on distributed edge cameras, enabling a deployment topology that balances bandwidth, latency, and privacy preservation.

4. Temporal Feature Refinement and Multi-Scale Architecture Integration

Temporal feature refinement addresses the intrinsic noise and visual ambiguity that plague raw video frames, particularly under low-light conditions, occlusions, and camera jitter. The proposed refinement module sits between the backbone feature extractor and the memory interaction layers, processing the stream of frame-level embeddings to produce temporally coherent, multi-scale representations. The module employs a cascade of depthwise separable one-dimensional convolutions with progressively increasing dilation factors, which efficiently aggregate contextual information across short, medium, and long time horizons without the quadratic complexity of self-attention. This design choice is motivated by the observation that anomalous events often manifest as irregularities across multiple temporal scales: a sudden fall occurs over a sub-second window, whereas loitering or abandoned objects unfold over minutes. By generating scale-specific feature maps and then fusing them via a lightweight gating mechanism, the refinement module enables the memory to be queried with context-rich representations that suppress spurious frame-level jitter.

The integration of multi-scale features with the persistent memory is accomplished through a late-fusion architecture, where each temporal scale branch independently reads from the shared memory, and the resulting reconstruction errors are aggregated using learned scale weights. This design avoids the memory explosion that would result from storing multi-scale prototypes, while still benefiting from scale-specific temporal smoothing. Furthermore, the refinement module incorporates an adaptive frame-skipping scheduler that dynamically adjusts the temporal stride of processing based on a running estimate of scene dynamism. In static scenes, the scheduler drops intermediate frames to conserve computation and energy, whereas high-activity intervals trigger denser sampling. This feedback-driven throttling mechanism is particularly salient for edge devices that operate under strict power budgets, such as solar-powered cameras in remote surveillance deployments. The co-design of memory and temporal refinement yields a system that can gracefully degrade its computational profile in response to resource constraints, maintaining baseline anomaly sensitivity even at low frame rates. Ablation studies on standard benchmarks confirm that the combination of persistent memory and multi-scale temporal refinement outperforms isolated implementations by significant margins, particularly in the localization of subtle anomalies that occur over extended durations.

5. System-Level Design and Deployment Considerations

Translating the algorithmic advances of persistent memory modeling and temporal feature refinement into operational video analytics infrastructures demands careful consideration of deployment topologies, hardware heterogeneity, and lifecycle management. A canonical deployment splits the pipeline into edge and cloud components: edge nodes execute the feature extractor and the temporal refinement module, while a centralized or federated memory server hosts the persistent memory bank and runs the final anomaly scoring logic. This split preserves low inference latency for initial processing while enabling the memory to benefit from aggregated normalcy data across multiple camera streams. The communication protocol between edge and cloud must be lightweight and privacy-preserving, transmitting only quantized embeddings rather than raw video frames. Quantization-aware training can compress the query embeddings to as low as 16-bit precision without degrading localization

accuracy, simultaneously reducing network bandwidth requirements and enabling the use of integer-arithmetic accelerators commonly found in embedded hardware.

For purely on-device deployment scenarios where uploading any video-derived data is prohibited by privacy regulations, a compressed memory snapshot can be periodically synchronized over a secure channel, allowing individual devices to maintain independent but globally informed memory banks. The size of the memory bank is subject to a hard physical memory ceiling on resource-constrained platforms; pruning and dimensionality reduction via random projection or product quantization become indispensable. Such compression schemes must be co-optimized with the anomaly objective to avoid losing discriminative information. On the orchestration layer, containerized microservices and model-serving frameworks enable seamless version upgrades, A/B testing, and canary rollouts without disrupting continuous video feeds [13]. Observability pipelines that monitor reconstruction error distributions, memory slot utilization, and per-camera anomaly rates are essential for detecting silent model degradation and data drift over months of operation. The system must also incorporate a human-in-the-loop feedback mechanism that allows security operators to label false positives and missed detections, which are then used to trigger selective memory updates or, in rare cases, full model retraining. Thoughtful lifecycle management reduces the total cost of ownership and ensures that anomaly detection infrastructure remains aligned with evolving operational requirements and societal expectations.

6. Robustness, Fairness, and Governance

The deployment of video anomaly localization systems in public and semi-public spaces raises profound questions of robustness, fairness, and governance that extend far beyond technical accuracy metrics. From a robustness perspective, the proposed system must withstand both naturally occurring distribution shifts—such as seasonal lighting changes, infrastructure modifications, and evolving crowd behaviors—and adversarial perturbations intentionally crafted to evade detection [14]. The persistent memory's gradual adaptation mechanism inherently provides resilience to slow covariate shift, but it can also be exploited by adversarial actors who slowly introduce abnormal patterns into the memory over extended periods, a form of data poisoning that normalizes the anomalous. Mitigating such attacks requires anomaly scoring thresholds that combine memory-based reconstruction error with auxiliary statistics drawn from temporally distant, write-protected reference snapshots. Adversarial robustness testing using projected gradient attacks and physically realizable patches must be part of the mandatory validation protocol before any live deployment.

Fairness concerns are equally pressing. Anomaly detection models trained on geographically or demographically homogeneous data have been shown to exhibit disparate false positive rates across different skin tones, clothing styles, and cultural behaviors, potentially leading to disproportionate scrutiny and unjust consequences for marginalized communities [15]. The persistent memory architecture offers a unique leverage point for fairness mitigation: the memory bank can be explicitly curated to include diverse normal prototypes representing a wide spectrum of appearances, gaits, and social interactions. Regular fairness audits, conducted by independent third parties and grounded in quantitative group fairness metrics, should be institutionalized within the governance framework of any large-scale deployment. Furthermore, the right to contest automated decisions must be embedded in the system's procedural design, with interpretability modules that can surface which memory prototypes most influenced the anomaly score for a given clip [16]. Legislative instruments such as the European Union's AI Act and the General Data Protection Regulation impose legally binding

requirements on biometric surveillance and automated decision-making, making explainability and contestability operational necessities rather than aspirational features. Governance models should mandate stakeholder representation in the design, procurement, and post-deployment auditing phases, ensuring that communities affected by surveillance have a voice in setting anomaly thresholds and acceptable error trade-offs. The energy and carbon footprint of training and maintaining video analytics infrastructure also falls under the umbrella of sustainable governance, encouraging the adoption of the proposed efficient architecture as an environmentally responsible alternative to compute-heavy three-dimensional convolutional pipelines.

7. Conclusion

This paper has presented a comprehensive approach to efficient video anomaly localization through the synergistic design of persistent memory modeling and temporal feature refinement. By decoupling long-term normalcy representation from per-frame computation and by incorporating multi-scale temporal smoothing with adaptive frame skipping, the framework achieves a favorable balance between detection fidelity and system efficiency. The architectural discussion has been extended into the socio-technical realm, examining edge-cloud partitioning, memory compression, adversarial resilience, fairness auditing, and regulatory alignment. The analysis demonstrates that high-performance anomaly localization need not come at the expense of deployability or ethical accountability. Future work will explore federated memory aggregation across organizational boundaries with differential privacy guarantees, as well as the integration of active learning and continual adaptation strategies that further reduce the need for costly manual annotation. The path forward lies in treating persistent memory and temporal refinement not as isolated algorithmic modules, but as integral components of a holistic surveillance ecosystem that is at once efficient, robust, fair, and transparent.

References

1. Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6479–6488).
2. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & Hengel, A. V. D. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1705–1714).
3. Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future frame prediction for anomaly detection – a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6536–6545).
4. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 733–742).
5. Chong, Y. S., & Tay, Y. H. (2017). Abnormal event detection in videos using spatiotemporal autoencoder. In Proceedings of the International Symposium on Neural Networks (pp. 189–196).

6. Ramachandra, B., Jones, M., & Vatsavai, R. (2020). A survey of single-scene video anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(5), 2293–2312.

7. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7794–7803).

8. Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

10. Carreira, J., & 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 (pp. 6299–6308).

11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).

12. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT (pp. 4171–4186).

13. Crankshaw, D., Wang, X., Zhou, G., Franklin, M. J., Gonzalez, J. E., & Stoica, I. (2017). Clipper: A low-latency online prediction serving system. In Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation (pp. 613–627).

14. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

15. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68).

16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

17. Li, G., Yuan, H., Chen, S., Hu, Q., Wang, J., & Jiang, K. (2026). MFT: Memory-Aware Fine-Tuning of SAM2 for Efficient Long-Sequence Video Object Segmentation. IEEE Signal Processing Letters.

18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4015–4026).

19. Luo, W., Liu, W., & Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked RNN framework. In Proceedings of the IEEE International Conference on Computer Vision (pp. 341–349).

20. Georgescu, M. I., Barbalau, A., Ionescu, R. T., Popescu, M., & Shah, M. (2021). Anomaly detection in video via self-supervised and multi-task learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12742–12752).