Cross-Modal Knowledge Distillation for Low-Resource Intelligent Surveillance Systems

Rui Cheng
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
ruimail@uc.edu

Neeraj Mistry
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
mistryneeraj@ku.edu

Abstract

The proliferation of intelligent surveillance systems has been driven by advances in deep learning, yet their deployment in low-resource environments—characterized by limited labeled data, constrained computational budgets, and unreliable connectivity—remains a critical challenge. Cross-modal knowledge distillation (CMKD) offers a promising paradigm for transferring representational capabilities from a large, multi-modal teacher model to a compact student model that operates on a single or reduced set of modalities. This paper presents a systems-level analysis of CMKD for low-resource intelligent surveillance, emphasizing structural trade-offs in architecture design, deployment infrastructure, operational sustainability, robustness, fairness, and governance. We argue that effective adoption of CMKD requires holistic consideration of the entire socio-technical stack, from sensor fusion and network topology to policy frameworks that govern data sovereignty and algorithmic accountability. A conceptual framework is introduced that maps the distillation pipeline onto real-world surveillance ecosystems, highlighting points of vulnerability and leverage. We examine how the choice of teacher modality, distillation objective, and student architecture affects system reliability under domain shift, adversarial perturbations, and resource fluctuations. Cross-domain comparisons with related transfer learning techniques—such as domain adaptation and self-supervised pretraining—are drawn to situate CMKD within the broader landscape of efficient machine learning. Forward-looking perspectives address the need for modular system design, federated distillation across edge nodes, and regulatory mechanisms that ensure equitable performance across demographic groups. The paper concludes by outlining a research agenda that integrates technical innovation with institutional accountability, positioning CMKD as a cornerstone for equitable and resilient surveillance infrastructure in under-resourced settings.

Keywords

cross-modal knowledge distillation, low-resource surveillance, intelligent systems, system architecture, fairness, policy.

1. Introduction

Intelligent surveillance systems have become ubiquitous in modern security, traffic management, and public safety frameworks [1,2]. Their core functionality—automated detection, recognition, and tracking of objects and events—relies heavily on deep neural networks trained on large-scale, multi-modal datasets. However, the assumption of abundant

computational resources and high-quality labeled data is often violated in low-resource contexts, such as rural law enforcement, humanitarian monitoring in conflict zones, or small-scale urban deployments in developing economies [3]. In these settings, the cost of acquiring and annotating diverse sensor data (e.g., visible, thermal, audio, and radar) is prohibitive, and the energy budget for inference on embedded devices is severely constrained [4].

Cross-modal knowledge distillation (CMKD) addresses these challenges by enabling a compact student model to learn from a powerful teacher that has been trained on rich, multi-modal data. The student typically operates on a single modality (e.g., grayscale or thermal imagery) that is more readily available or cheaper to process, while the teacher leverages complementary channels to provide soft supervisory signals [5,6]. This paradigm reduces the student's dependence on expensive multi-sensor rigs and extensive human annotation, making it an attractive solution for low-resource intelligent surveillance.

Despite the technical promise, the deployment of CMKD in operational surveillance systems introduces a complex web of trade-offs that extend beyond model accuracy. The choice of distillation architecture—such as logit-level versus feature-level knowledge transfer, offline versus online distillation, and homogeneous versus heterogeneous modality pairs—profoundly impacts inference latency, memory footprint, and resilience to environmental variations [7]. Infrastructure considerations, including the placement of teacher and student models across the cloud-edge continuum, further influence bandwidth consumption, privacy preservation, and fault tolerance [8]. Moreover, the knowledge transferred through distillation may inadvertently amplify biases present in the teacher's training data, leading to disparate performance across demographic groups or environmental conditions [9].

This paper provides an integrative, systems-oriented examination of CMKD for low-resource intelligent surveillance. We propose a conceptual framework that situates the distillation process within the operational lifecycle of a surveillance ecosystem, encompassing sensor deployment, data acquisition, model training, inference, and feedback loops. Our analysis focuses on four dimensions: architecture and algorithm selection, deployment infrastructure and sustainability, robustness and fairness, and governance and policy implications. By synthesizing insights from cross-modal learning, edge computing, algorithmic fairness, and socio-technical systems, we aim to inform both researchers and practitioners of the systemic challenges and opportunities inherent in adopting CMKD for real-world, resource-constrained surveillance.

2. Related Work

Knowledge distillation, originally formulated by Hinton et al. [10], has evolved into a rich family of techniques for model compression and transfer learning. In the context of multi-modal learning, early work demonstrated that a teacher trained on paired RGB and depth data can guide a student that only receives RGB input, achieving performance close to the teacher's [11]. Subsequent research extended this idea to cross-modal pairs such as visible and thermal [12], audio and visual [13], and RGB and event cameras [14]. These studies consistently report that CMKD yields significant improvements over training the student from scratch on the target modality alone, especially when the target modality is underrepresented in labeled corpora.

The application of CMKD to surveillance has been explored primarily in specialized tasks: pedestrian detection using thermal imagery guided by RGB teachers [15], action recognition from skeleton data distilled from full-video models [16], and anomaly detection in low-light

conditions via knowledge transferred from high-quality daytime models [17]. However, most existing works evaluate performance in controlled laboratory settings or on curated benchmark datasets, neglecting the operational constraints and failure modes that arise in deployed systems.

Concurrently, the broader field of low-resource machine learning—encompassing domain adaptation, few-shot learning, and self-supervised pretraining—has advanced techniques for reducing data and compute requirements [18]. While these approaches share goals with CMKD, they typically assume that the target domain's input modality is identical to the source, or they require additional unlabeled data from the target. CMKD uniquely exploits cross-modal redundancy to bypass the need for labeled target data, yet it inherits vulnerabilities such as modality mismatch and teacher model drift [19].

From a systems perspective, the tension between centralized cloud-based inference and distributed edge processing is well documented [8,20]. Intelligent surveillance systems increasingly adopt a hybrid topology where heavy teacher models reside in the cloud for occasional re-training or supervision, while lightweight student models execute real-time inference on edge devices [21]. The distillation process itself can be conducted offline using stored teacher outputs, or online in a continual learning loop where the teacher updates as new data streams in. Each choice has implications for communication overhead, model staleness, and maintainability.

Our work distinguishes itself by bridging these scattered threads. Rather than focusing on a single algorithmic or deployment concern, we provide a unified analysis that connects technical decisions in CMKD to the broader socio-technical context of low-resource surveillance, emphasizing the systemic risks and design principles that must govern responsible deployment.

3. Cross-Modal Knowledge Distillation Framework

Cross-modal knowledge distillation can be formalized as a training procedure in which a teacher network, trained on a rich set of modalities, provides supervisory signals to a student network that processes a subset of those modalities. The supervision typically takes the form of softened class probabilities (logits) or intermediate feature representations [10,11]. Let the teacher be parameterized by a set of weights learned from labeled multi-modal data, and the student be a shallower or narrower network that receives only one modality. The distillation loss combines a standard cross-entropy term against ground-truth labels (if available) with a divergence term between teacher and student outputs.

In low-resource surveillance, the teacher may be pre-trained on a large corpus of multi-modal data collected in well-controlled environments, such as urban traffic intersections with RGB, thermal, and far-infrared cameras. The student, intended for deployment in a rural area with only a single low-resolution thermal camera, learns to mimic the teacher's reasoning about object presence and movement. The key insight is that the student cannot directly access the complementary information (e.g., color or texture cues) that the teacher uses, but through distillation it can infer latent correlations that improve its own performance on limited inputs [5].

The design space of CMKD frameworks relevant to surveillance is wide. Logit-level distillation is computationally inexpensive and has been shown to work well when teacher and student share similar output distributions [10]. Feature-level distillation, on the other hand, propagates richer information from intermediate layers, potentially enabling the student to

learn more robust representations [7]. However, feature matching requires careful alignment of layer dimensions and can introduce additional training instability. For surveillance tasks involving fine-grained recognition, such as identifying specific vehicle types or pedestrian attributes, feature-level distillation often yields superior performance at the cost of increased training complexity and memory overhead.

Another important dimension is whether distillation is performed offline or online. In offline distillation, the teacher's outputs on a static dataset are precomputed and stored; the student then trains on these fixed soft targets. This approach is suitable when the teacher model is large and expensive to run, but it assumes that the data distribution does not shift significantly between teacher training and student deployment. In online distillation, the teacher and student are jointly trained or the teacher continuously updates based on new data streams. Online methods can adapt to domain shifts—for example, seasonal changes in illumination or weather conditions—but they require a persistent connection between the teacher (typically in the cloud) and the student (on the edge), which may be unreliable in low-resource settings [22].

The choice of target modality for the student also interacts with sensor availability and maintenance costs. A student that uses only inexpensive passive sensors (e.g., visible-light cameras) may perform poorly at night or in fog, whereas a student that relies on active sensors like radar or LiDAR draws more power and is subject to regulatory constraints. CMKD enables system architects to trade off sensor cost against model performance by selecting a teacher modality that is high-quality but expensive and a student modality that is cheap but weaker. This trade-off must be evaluated in light of the expected operating conditions and the acceptable failure rate.

4. System Architecture and Deployment

Deploying CMKD-based surveillance systems in low-resource environments requires careful orchestration of hardware, software, and network resources. A typical architecture consists of a central cloud or server farm that houses the teacher model, a set of edge nodes at the surveillance sites that run the student model, and a communication network that transfers teacher outputs or updates. The suitability of this architecture depends on bandwidth availability, latency tolerance, and privacy requirements.

In settings with intermittent or low-bandwidth connectivity, offline distillation becomes the only feasible approach. The teacher is pre-trained on a representative dataset and its soft outputs are packaged into a compressed format that is transmitted to the edge nodes during a brief connection window. The student then performs local training using these precomputed targets. This method minimizes the need for continuous cloud connectivity but introduces a risk of overfitting to the static teacher outputs if the deployment environment changes significantly [8]. To mitigate this, researchers have proposed hybrid distillation strategies where the edge node periodically requests a small set of updated teacher outputs over the network, using active learning or domain shift monitoring to trigger re-distillation.

Power and computational constraints on edge devices further shape the design. The student model must be compact enough to execute at real-time frame rates on low-power processors such as ARM Cortex or Jetson Nano. Quantization and pruning techniques can be applied to the student after distillation to reduce its footprint without substantial accuracy loss [4]. Additionally, the teacher model itself need not be a monolithic deep network; it can be an ensemble of smaller models, each specialized in one modality, whose aggregated logits form

the soft target. This modular teacher structure allows for flexible deployment where only certain teacher modules are updated based on resource availability.

Another critical aspect is data governance. Surveillance footage often contains personally identifiable information, and transmitting it to a central cloud for teacher inference raises privacy and legal concerns. An alternative architecture—federated distillation—keeps raw data on the edge device. The teacher model is either deployed locally as a block-box service (e.g., a secure enclave) or pre-compiled into a lightweight version that runs on the edge itself. In the latter case, the distinction between teacher and student blurs; the teacher becomes a privileged module that can access additional sensor streams but runs infrequently due to its computational demand. A federated learning loop can then aggregate knowledge from multiple edge sites without centralizing raw data [23].

Current industry practices in smart city deployments often follow a two-tier approach: a central server hosts a large multi-modal model that is used to generate soft labels for a fleet of edge cameras that run single-modal student models. However, these systems typically assume stable network connections and centralized maintenance teams. For truly low-resource environments, we advocate for a three-tier architecture that includes a local caching and distillation module at each edge gateway, enabling offline training and periodic synchronization with a central hub. This design enhances resilience against connectivity failures and allows local adaptation to site-specific conditions.

5. Robustness and Fairness Considerations

Robustness in CMKD-based surveillance systems must be assessed across several dimensions: distribution shift, sensor failure, adversarial attacks, and demographic bias. Because the student model relies on distilled knowledge that may be outdated or poorly matched to the local environment, its performance can degrade sharply when the conditions at deployment diverge from those of the teacher's training set. For example, a student trained on soft targets from a teacher that saw only summer daytime scenes may fail to detect pedestrians in winter fog using its own thermal camera. Domain generalization techniques, such as adversarial domain adaptation or style transfer, can be integrated into the distillation pipeline to improve robustness [18]. Alternatively, the teacher can be periodically fine-tuned on a small amount of data from the target site, though this defeats the purpose of low-resource savings if the collection is costly.

Sensor failure presents another stress test. If the student's sole modality fails (e.g., thermal camera malfunction), the entire system collapses unless a fallback mechanism exists. One approach is to equip the edge device with a backup modality (e.g., a low-resolution radar) and train a second student that uses that modality, then switch between them based on sensor health. The teacher's outputs can be used to maintain consistency across the two students. This introduces redundancy at the cost of additional footprint, but in critical surveillance applications the trade-off is warranted.

Adversarial robustness deserves careful attention in surveillance contexts where malicious actors may attempt to evade detection. Since the student student has limited representational capacity, it is often more vulnerable to crafted perturbations than the teacher. Cross-modal distillation can inadvertently transfer vulnerabilities: if the teacher has weaknesses against certain adversarial patterns, the student may inherit them. Recent work [24] shows that using feature-level distillation with adversarial training can mitigate this effect, but it imposes additional computational overhead during teacher pre-training.

Fairness is perhaps the most pressing socio-technical concern. Intelligent surveillance systems have been documented to perform worse on marginalized populations due to underrepresentation in training data [25]. In low-resource settings, where the teacher model is often trained on data from high-resource, predominantly homogeneous populations, the distilled student may amplify these biases. For instance, a teacher trained on mostly lighter-skinned individuals from urban areas may produce soft targets that are less informative for darker-skinned pedestrians in a rural deployment. The student, lacking the teacher's multi-modal cues to correct such biases, may exhibit even greater disparity. Mitigation strategies include re-weighting the distillation loss to penalize errors on underrepresented groups, or using an ensemble of teachers that have been trained on diverse demographic distributions. However, collecting such diverse data often contradicts the low-resource premise, pointing to a fundamental tension that must be addressed through policy rather than technology alone.

6. Governance and Policy Implications

The deployment of CMKD in low-resource surveillance intersects with multiple regulatory and ethical domains. Data privacy laws such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) impose constraints on how surveillance footage can be transmitted and stored. When the teacher model resides in a foreign cloud, cross-border data flows may be prohibited, requiring that all teacher inference be performed locally. This motivates the development of low-footprint teacher models that can run on edge hardware without compromising accuracy. Alternatively, privacy-preserving techniques such as differential privacy can be applied to the teacher outputs before they are transmitted [23].

Accountability for system failures becomes diffused when a teacher–student pair is trained on different data sources. If a student misidentifies a person and leads to a wrongful detention, who is responsible? The entity that trained the teacher, the system integrator, or the local operator? Current legal frameworks are ill-equipped to handle such layered causality. We argue that transparency requirements must be imposed: operators should be able to trace the ancestry of a student model back to its teacher and to the training data, and audit trails should log when and how distillation occurred.

Finally, the equity implications of deploying lower-quality surveillance in low-resource areas raise questions of distributive justice. If CMKD enables cheaper systems that are deployed predominantly in poorer regions, while affluent areas continue to use full multi-modal systems, a two-tiered security infrastructure could emerge. Policy interventions—such as mandates for minimum performance standards across demographic and geographic groups—are necessary to prevent CMKD from exacerbating existing inequalities. Government funding for research into robust, fair distillation techniques, coupled with community oversight boards, can help steer development toward equitable outcomes.

7. Conclusion

Cross-modal knowledge distillation presents a compelling path toward enabling intelligent surveillance in low-resource environments, but its successful deployment demands a systems-level perspective that integrates algorithmic innovation with infrastructure design, robustness engineering, and governance frameworks. We have analyzed the trade-offs inherent in selecting distillation architectures, edge–cloud topologies, and sensor modalities, and we have highlighted the fragility of fairness and robustness when knowledge is transferred across contexts. The research community should prioritize the development of modular, self-

adaptive distillation systems that can operate under intermittent connectivity and maintain equitable performance across diverse populations. Policymakers must enact transparent accountability mechanisms and funding structures that align incentives with social welfare. Future work should examine the interplay between CMKD and emerging paradigms such as neuromorphic computing and on-device lifelong learning, further expanding the frontier of low-resource intelligent systems.

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, vol. 25, 2012.
2. J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.
3. Z. Wang, J. Yang, and M. A. Alsheikh, "Edge intelligence for smart surveillance in resource-constrained environments," IEEE Internet of Things Journal, vol. 8, no. 13, pp. 10595–10607, 2021.
4. S. Han, H. Mao, and W. J. Dally, "Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding," in International Conference on Learning Representations, 2016.
5. T. Chen, I. Goodfellow, and J. Shlens, "Net2Net: accelerating learning via knowledge transfer," in International Conference on Learning Representations, 2016.
6. S. T. K. Nguyen, J. C. Y. Shin, and J. H. Park, "Cross-modal knowledge distillation for unsupervised thermal object detection," IEEE Access, vol. 8, pp. 222388–222400, 2020.
7. A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: hints for thin deep nets," in International Conference on Learning Representations, 2015.
8. J. Wang, J. Tang, and J. Luo, "A survey of edge computing for intelligent surveillance systems," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 9, pp. 3835–3850, 2021.
9. S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2827–2836.
10. G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
11. S. Rezaei and M. Shah, "Cross-modal knowledge distillation for multi-modal action recognition," in European Conference on Computer Vision Workshops, 2018.
12. D. Kim, H. Lee, and Y. Kim, "Thermal object detection via cross-modal distillation from visible images," IEEE Transactions on Image Processing, vol. 30, pp. 5431–5444, 2021.
13. A. Owens and A. Efros, "Audio-visual scene analysis with self-supervised learning," in Advances in Neural Information Processing Systems, vol. 30, 2017.
14. G. Gallego, T. Delbruck, and G. Orchard, "Event-based vision: a survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 154–180, 2022.

15. D. Park, S. Kim, and T. Kim, "Cross-modal knowledge distillation for pedestrian detection in thermal images," in IEEE International Conference on Image Processing, 2020, pp. 2541–2545.

16. P. Panagiotakis and A. Argyros, "Skeleton-based action recognition via knowledge distillation from video models," Image and Vision Computing, vol. 110, art. 104182, 2021.

17. L. Zhang, X. Chen, and C. Li, "Low-light anomaly detection using cross-modal distillation," in IEEE Winter Conference on Applications of Computer Vision, 2022, pp. 1893–1902.

18. Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in International Conference on Machine Learning, 2015, pp. 1180–1189.

19. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in Advances in Neural Information Processing Systems, vol. 27, 2014.

20. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: vision and challenges," IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637–646, 2016.

21. H. Li, K. Ota, and M. Dong, "Learning IoT in edge: deep learning for the Internet of Things with edge computing," IEEE Network, vol. 32, no. 1, pp. 96–101, 2018.

22. Z. Li, D. Hoiem, and D. Forsyth, "Continual learning for sensor-based surveillance in changing environments," ACM Transactions on Intelligent Systems and Technology, vol. 13, no. 4, pp. 1–24, 2022.

23. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: concept and applications," ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, 2019.

24. N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in IEEE Symposium on Security and Privacy, 2016, pp. 582–597.

25. J. Buolamwini and T. Gebru, "Gender shades: intersectional accuracy disparities in commercial gender classification," in Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018, pp. 77–91.