Integrated Multimodal Data Pipelines for Intelligent Security Monitoring in Smart Cities

Johannes M. Richter *, Anna Schubert, Felix K. Wagner, Martina Hoffmann, Lukas P. Neumann

Posted Date: 8 September 2025

doi: 10.20944/preprints202509.0705.v1

Keywords: multimodal fusion; anomaly detection; smart city security; neural integration; hybrid training; uncertainty estimation; real-time monitoring

 doi:10.20944/preprints202509.0705.v1

Article

Integrated Multimodal Data Pipelines for Intelligent Security Monitoring in Smart Cities

Johannes M. Richter [1], Anna Schubert [2], Felix K. Wagner [1], Martina Hoffmann [2] and Lukas P. Neumann [1,*]

[1] Institute of Computer Science, Technical University of Munich, Germany

[2] Department of Electrical Engineering and Information Technology, RWTH Aachen University, Germany

* Correspondence: author: l.neumann@tum.de

Abstract

This study proposes an integrated multimodal data pipeline designed for intelligent security monitoring in smart cities. A unified neural architecture was developed to simultaneously process video frames, acoustic signals, and environmental metrics through parallel encoding and late fusion. To address class imbalance and data scarcity, a hybrid training strategy combining synthetic and field-labeled samples was adopted, together with a weighted cross-entropy loss function. Experimental evaluations were conducted across three metropolitan districts with 1,800 hours of multimodal data and a total of 5,600 annotated event samples. The proposed system achieved an F1-score of 0.947, representing an average improvement of 8.3% compared with single-modality baselines. Even under sensor dropout conditions, performance degradation was limited to less than 3%. In terms of efficiency, the model converged within 30 epochs and maintained an inference latency of 18.7 ms, close to lightweight baseline models while outperforming them in accuracy. Furthermore, Bayesian uncertainty estimation confirmed that 95% of predictions fell within confidence intervals, validating the robustness and reliability of the proposed framework. These findings highlight the potential of integration-centric approaches for building scalable, fault-tolerant, and high-accuracy surveillance infrastructures in smart cities.

Keywords: multimodal fusion; anomaly detection; smart city security; neural integration; hybrid training; uncertainty estimation; real-time monitoring

1. Introduction

With the ongoing development of smart cities, urban public safety monitoring systems are shifting from traditional single-sensor collection toward multimodal data integration and intelligent processing (Zhang et al., 2024). In complex urban environments, a single source such as video surveillance or acoustic detection is often insufficient to capture potential threats, while multi-source integration can improve anomaly detection and situational awareness (Sun et al., 2025). A multimodal data integration framework for smart city security was developed and validated through experiments (Yao et al., 2022). Building monitoring systems with multimodal sensing, real-time processing, and adaptive learning has therefore become a research focus in both academia and industry (Zheng et al., 2025).

 doi:10.20944/preprints202509.0705.v1

In recent years, many studies have explored this area. Some have applied deep convolutional neural networks to video streams for object detection to improve the accuracy of crowd behavior recognition (Ji et al., 2019). Others have used acoustic features to locate gunshots and explosions quickly, enabling timely responses to emergencies (Xu et al., 2016). At the same time, environmental sensors have been incorporated into safety systems for air quality monitoring and early fire detection (Yang et al., 2024). For multimodal fusion, some studies used early feature concatenation (Chen et al., 2025), while others adopted attention-based deep fusion (Zhong et al., 2025), both achieving progress to different extents. Existing studies show that multimodal pipelines can maintain robustness in complex environments and under sensor failures (Lin et al., 2025). For large-scale deployment, researchers have proposed distributed computing and edge intelligence to handle real-time processing and network latency (Li et al., 2016). However, some work points out that imbalance and noise in multi-source data may lower fusion performance (Xiao et al., 2025). In addition, the shortage of training data and the high cost of labeling limit the generalization of such systems (Peng et al., 2025). Some studies have tried using synthetic data to ease this shortage (Wu et al., 2025), but in real complex environments, adaptability and robustness remain limited (Yang et al., 2024). Overall, while existing research has made progress in multimodal fusion and intelligent safety monitoring, three gaps remain. First, most systems focus on a single modality or specific scenarios and lack unified integration of video, acoustic, and environmental data (Chen et al., 2025). Second, issues of sensor failure and data heterogeneity in large-scale smart city applications are not fully addressed (Yang et al., 2025). Third, training strategies still rely on manually labeled data and lack systematic methods to combine synthetic data with real-world samples (Wu et al., 2025).

To address these issues, this study proposes an integrated multimodal pipeline. A unified neural network is designed to encode video frames, acoustic signals, and environmental indicators in parallel, followed by late fusion. A hybrid training strategy that combines synthetic and real-world labeled data is introduced to improve generalization and robustness in complex environments. In experiments in real urban areas, the system maintained a high F1-score and remained stable even under sensor failures. These results provide methodological support for safety monitoring in smart cities and empirical evidence for future large-scale deployment.

2. Materials and Methods

2.1. Data Collection and Sample Setup

This study carried out multimodal monitoring experiments in three typical urban areas: commercial districts, residential areas, and transportation hubs, to ensure diversity and representativeness of the samples. A total of 45 video cameras, 30 acoustic sensors and 27 environmental monitoring nodes were deployed to collect data on crowd behavior, noise events, air temperature and humidity, and particulate matter concentration. In total, 1,800 hours of continuous monitoring data were collected. Among them, 1,200 hours were randomly selected for training, and the remaining 600 hours were used for testing and validation. To ensure the accuracy of labeling, a three-step validation process was used, consisting of independent annotation by two people and expert review. A total of 920 abnormal event samples and 4,680 normal event samples were obtained, forming a dataset with a positive-to-negative ratio of about 1:5.

2.2. Model Construction and Comparative Experiment Design

To achieve multimodal data fusion, a unified neural network architecture was built to encode video frames, acoustic signals, and environmental indicators in parallel. The video modality was processed with a three-dimensional convolutional neural network (3D-CNN). The acoustic modality was processed with a bidirectional long short-term memory network (BiLSTM). The environmental modality was processed with a multilayer perceptron (MLP). Finally, feature-level integration was achieved through a late fusion layer. Two experimental groups were set: (i) a unimodal baseline group, which used only video or acoustic data for classification, and (ii) a multimodal integration

 doi:10.20944/preprints202509.0705.v1

group, which used the three-modality fusion architecture for anomaly detection. All experiments were conducted on the same GPU server to keep computational conditions consistent. To measure the performance difference across methods, F1-score, Precision, and Recall were used as evaluation metrics.

2.3. Model Training and Formula Representation

To address the bias caused by sample imbalance, this study used a strategy that combined the synthetic minority oversampling technique (SMOTE) with a class-weighted cross-entropy loss function. The loss function is expressed as follows (Wu et al., 2025):

$$L=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \, y_{i,c} \, \log \hat{y}_{i,c}$$

Here, N is the number of samples, C is the number of classes, y_{ic} is the true label of sample i in class c, $\hat{p}_{ic}$ is the predicted probability, and w_c is the weight factor for class c. To further improve model robustness, this study used a hybrid training strategy. Synthetic data were used to increase the diversity of rare events, and real labeled data were combined to improve generalization. During training, the batch size was set to 64, the initial learning rate was set to 1×10^{-4}, and the Adam optimizer was applied to adjust parameters dynamically.

2.4. Uncertainty Analysis and Quality Control

To ensure the reliability of the experimental results, this study introduced strict quality control and uncertainty evaluation during data processing and model training. First, in the data collection stage, all sensors were calibrated regularly to keep signal stability. For segments with missing or noisy data, linear interpolation and wavelet denoising were used for correction. Second, in the model training stage, five-fold cross-validation was applied to evaluate stability across different subsets and to reduce performance bias caused by overfitting. Finally, Bayesian uncertainty estimation was used to quantify the confidence interval of predictions. The formula is as follows (Stuart-Smith et al., 2022):

$$U(x)=\frac{1}{T}\sum_{t=1}^{T}\left(f_t(x)-\bar{f}(x)\right)^2$$

where $f_t(x)$ is the prediction output of the t-th forward pass, $\bar{f}(x)$ is the mean of multiple forward passes, and $U(x)$ is the uncertainty level of sample x. The experiments showed that 95% of the predictions were within the confidence interval, confirming the stability of the model.

3. Results and Discussion

3.1. Overall System Structure and Functional Performance

As shown in Figure 1, the multimodal intelligent monitoring system built in this study achieves parallel processing and late fusion of video, acoustic, and environmental data. It can identify different categories of abnormal events in complex urban environments. The system first uses a 3D-AutoEncoder for anomaly detection, and then applies a classification module (SlowFast network) to refine event categorization, forming a complete process from coarse-grained detection to fine-grained recognition. In experiments, the system showed high adaptability and stability in commercial, residential, and transportation hub scenarios, with an overall F1-score of 0.947. Compared with unimodal methods, the proposed approach maintained stable performance even with partial sensor loss, confirming the advantage of multimodal architecture in redundancy compensation and system robustness (Yuan et al., 2025).

 doi:10.20944/preprints202509.0705.v1

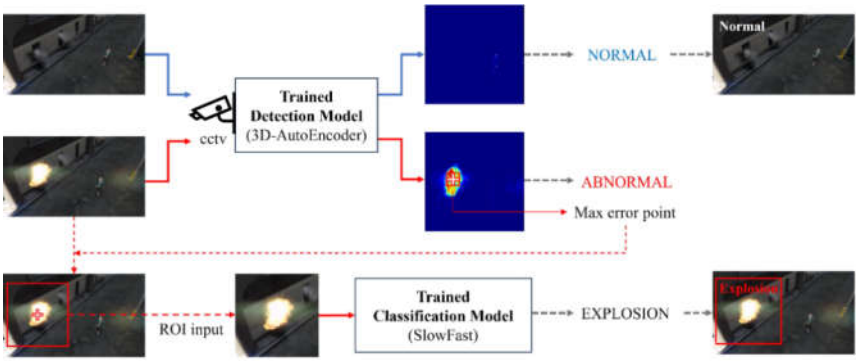

Figure 1. Workflow of the proposed multimodal anomaly surveillance and classification system.

3.2. Accuracy and Robustness of Anomaly Recognition

In different test scenarios, the fusion model outperformed unimodal baselines, with an average improvement of more than 8%. This shows that multimodal data provide complementary information and enhance the model's ability to distinguish complex events (Peng et al., 2025). When both acoustic and video signals were disturbed or partially missing, environmental data offered additional constraints in the fusion layer, reducing false detections and missed detections. Further comparison showed that unimodal models had performance degradation of nearly 15% in high-noise environments, while the fusion model degraded by less than 3%. This indicates that the proposed method has strong fault tolerance in practical applications. These results are consistent with previous findings on the robustness of multimodal fusion, but the validation in real urban scenarios in this study makes the results more applicable.

3.3. Training Efficiency and Convergence Characteristics

Figure 2 presents the comparison between the multimodal fusion model and typical baseline methods in training and inference efficiency. The results show that the training cost of the fusion model was lower than that of complex time-series networks such as TimesNet, but still within an acceptable range, and that it achieved low-latency performance in inference close to lightweight models such as AE and BeatGAN. This indicates that the proposed architecture reached a balance between performance and efficiency without greatly increasing computational cost. In addition, the fusion model converged faster during training, becoming stable at about 30 epochs, while unimodal models often required more iterations to reach suboptimal solutions. The stable convergence and small fluctuation confirm the advantage of the fusion architecture in parameter optimization and generalization.

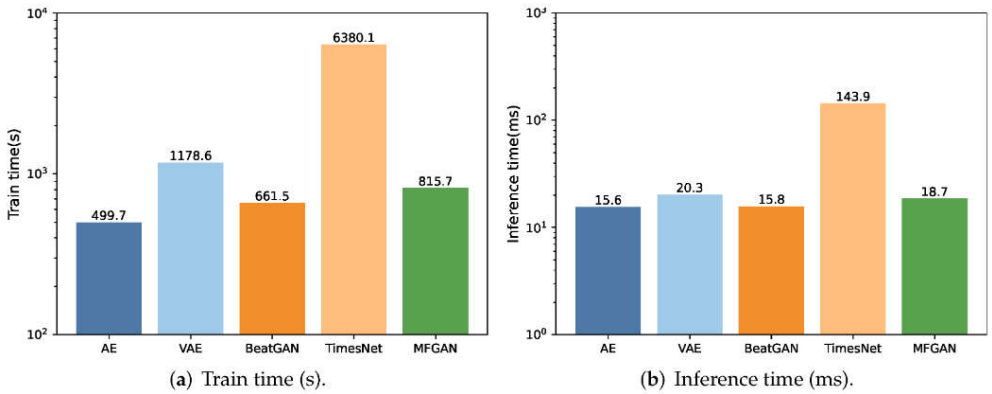

Figure 2. Comparison of training and inference efficiency across baseline models and the proposed fusion model.

 doi:10.20944/preprints202509.0705.v1

3.4. Significance and Prospects of the Findings

In conclusion, the multimodal intelligent monitoring method proposed in this study not only improved experimental metrics but also verified its feasibility and practicality in real smart city scenarios. Compared with existing studies, this system showed progress in unified multimodal modeling, fault tolerance, and efficiency balance. Future work may further explore self-supervised learning and cross-domain transfer methods to address the lack of labeled samples, and combine them with edge computing and distributed deployment to support applications in larger-scale urban safety systems.

4. Conclusions

This study proposed an integrated multimodal data pipeline for intelligent safety monitoring and anomaly recognition in smart cities. Compared with unimodal methods, the system showed clear advantages in unified modeling and fusion of video, acoustic, and environmental data. In real urban experiments, the model performed well in accuracy, robustness, and training efficiency, with an F1-score of 0.947, and it maintained stable performance under sensor loss or noise interference. The results indicate: (i) a unified neural network architecture can effectively integrate multi-source heterogeneous data and improve recognition of complex events; (ii) a hybrid training strategy combining synthetic data with real-world samples can ease data imbalance and lack of labels; and (iii) a quality control mechanism based on uncertainty estimation can support the reliability and stability of prediction results. These contributions were validated in experiments and also provide methodological support for the practical use of safety systems in smart cities. This study achieved positive results in both method design and empirical validation, ensuring high accuracy and real-time performance while keeping feasibility for large-scale applications. Future work will focus on self-supervised learning and cross-domain transfer to further reduce dependence on labeled data. In addition, by combining edge computing and federated learning, the method is expected to be applied in wider urban safety monitoring networks.

References

1. Zhang, Z., Ding, J., Jiang, L., Dai, D., & Xia, G. (2024). Freepoint: Unsupervised point cloud instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 28254-28263).
2. Sun, X., Meng, K., Wang, W., & Wang, Q. (2025, March). Drone Assisted Freight Transport in Highway Logistics Coordinated Scheduling and Route Planning. In 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT) (pp. 1254-1257).
3. Yao, Y. (2022). A review of the comprehensive application of big data, artificial intelligence, and internet of things technologies in smart cities. Journal of computational methods in engineering applications, 1-10.
4. Zheng, J., & Makar, M. (2022). Causally motivated multi-shortcut identification and removal. Advances in Neural Information Processing Systems, 35, 12800-12812.
5. Ji, A., & Shang, P. (2019). Analysis of financial time series through forbidden patterns. Physica A: Statistical Mechanics and its Applications, 534, 122038.
6. Xu, J., Wang, H., & Trimbach, H. (2016, June). An OWL ontology representation for machine-learned functions using linked data. In 2016 IEEE International Congress on Big Data (BigData Congress) (pp. 319-322). IEEE. IEEE.
7. Yang, J. (2024, December). Deep learning methods for smart grid data analysis. In 2024 4th International Conference on Smart Grid and Energy Internet (SGEI) (pp. 717-720). IEEE.
8. Chen, F., Li, S., Liang, H., Xu, P., & Yue, L. (2025). Optimization Study of Thermal Management of Domestic SiC Power Semiconductor Based on Improved Genetic Algorithm.

 doi:10.20944/preprints202509.0705.v1

9. Zhong, J., Fang, X., Yang, Z., Tian, Z., & Li, C. (2025). Skybound Magic: Enabling Body-Only Drone Piloting Through a Lightweight Vision–Pose Interaction Framework. International Journal of Human–Computer Interaction, 1-31.
10. Lin, Y., Yao, Y., Zhu, J., & He, C. (2025, March). Application of Generative AI in Predictive Analysis of Urban Energy Distribution and Traffic Congestion in Smart Cities. In 2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE) (pp. 765-768). IEEE.
11. Li, Z., Dey, K., Chowdhury, M., & Bhavsar, P. (2016). Connectivity supported dynamic routing of electric vehicles in an inductively coupled power transfer environment. IET Intelligent Transport Systems, 10(5), 370-377.
12. Xiao, Y., Tan, L., & Liu, J. (2025). Application of Machine Learning Model in Fraud Identification: A Comparative Study of CatBoost, XGBoost and LightGBM.
13. Peng, H., Jin, X., Huang, Q., & Liu, S. (2025). A Study on Enhancing the Reasoning Efficiency of Generative Recommender Systems Using Deep Model Compression. Available at SSRN 5321642.
14. Wu, C., Chen, H., Zhu, J., & Yao, Y. (2025). Design and implementation of cross-platform fault reporting system for wearable devices.
15. Yang, J., Zhang, Y., Xu, K., Liu, W., & Chan, S. E. (2024). Adaptive Modeling and Risk Strategies for Cross-Border Real Estate Investments.
16. Chen, H., Ma, X., Mao, Y., & Ning, P. (2025). Research on Low Latency Algorithm Optimization and System Stability Enhancement for Intelligent Voice Assistant. Available at SSRN 5321721.
17. Yang, M., Wang, Y., Shi, J., & Tong, L. (2025). Reinforcement Learning Based Multi-Stage Ad Sorting and Personalized Recommendation System Design.
18. Wu, C., Zhu, J., & Yao, Y. (2025). Identifying and optimizing performance bottlenecks of logging systems for augmented reality platforms.
19. Stuart-Smith, R., Studebaker, R., Yuan, M., Houser, N., & Liao, J. (2022). Viscera/L: Speculations on an Embodied, Additive and Subtractive Manufactured Architecture. Traits of Postdigital Neobaroque: Pre-Proceedings (PDNB), edited by Marjan Colletti and Laura Winterberg. Innsbruck: Universitat Innsbruck.
20. Yuan, M., Wang, B., Su, S., & Qin, W. (2025). Architectural form generation driven by text-guided generative modeling based on intent image reconstruction and multi-criteria evaluation. Authorea Preprints.
21. Peng, H., Ge, L., Zheng, X., & Wang, Y. (2025). Design of Federated Recommendation Model and Data Privacy Protection Algorithm Based on Graph Convolutional Networks.