

Urban Facility Risk Assessment and Response Priority Decision-Making Based on Vision-Language Transformer

Xiaojun Li^a, Chengjin Jiang^b, Jinlu Sun^c, Jiaqin Mo^d, Jianjuan Fan^{*}

The Smart City Research Institute of China Electronics Technology Group Corporation, Shenzhen City 518000, China

^alixiaojun10@cetc.com.cn, ^bjiangchj@yeah.net, ^csunjinlu@mail.ustc.edu.cn, ^dmojiaqin@cetc.com.cn,

^{*} Corresponding author: Jianjuan Fan (Email: fanjianjuan@cetc.com.cn)

Abstract: This study addresses issues in urban management information systems to improve the intelligence level of urban facility risk identification and disposal priority decision. These issues include fragmented multi-source data, risk assessment relying on manual experience, and a lack of a quantitative basis for work order sorting. The study introduces Vision-Language Transformer (VLT) to construct an integrated multimodal framework for urban facility risk assessment and response priority decision-making. The framework takes road and municipal facility images, inspection records, and complaint work order texts as inputs. Through VLT, it fuses visual and linguistic features in a unified semantic space to obtain joint representations that can characterize both appearance defects and semantic hidden dangers simultaneously. Combined with business features such as location importance and historical failure frequency, it outputs risk scores and response priority rankings. Experiments select a public road damage dataset, dividing it into three subsets (A, B, C) corresponding to main road business districts, residential and public service facilities, and old areas. The optimized model of this study is compared with the Single-modal Vision Transformer-based Risk Assessment Framework (SVT-Risk) and the Early-Fusion Multimodal Transformer-based Risk Assessment Framework (EFMT-Risk). Results show that the accuracy of the optimized model in subsets A, B, and C is 0.91, 0.89, and 0.88; the mean average precision (MAP) in subsets A, B, and C reaches 0.87, 0.86, and 0.89, respectively. The overall performance is superior to the two comparison frameworks, indicating that the method can more accurately identify high-risk facilities and reasonably sort the disposal order; meanwhile, it can effectively alleviate the problems of inaccurate risk identification and unreasonable work order priority in traditional urban management information systems. Therefore, this study makes certain contributions to the field of urban facility risk assessment and intelligent urban management information system construction.

Keywords: Vision-Language Transformer; Urban facility risk assessment; Response priority decision-making; Multimodal deep learning; Urban management information system

1. Introduction

With the acceleration of urbanization, a large number of facilities such as roads, bridges, drainage pipelines, and lighting facilities have entered a period of concentrated aging. Coupled with extreme climates, emergencies, and high-intensity use, the safety risks associated with urban operations have risen significantly [1,2]. Existing urban management information systems have initially enabled work order circulation, event reporting, and visual monitoring. However, they still rely heavily on manual experience and post-event disposal in risk identification and judgment of disposal priority. It is difficult to timely extract key information from massive surveillance videos, inspection photos, and complaint texts [3]. In particular, multimodal data such as images and texts generally suffer from the problem of "operating independently", failing to form a unified semantic understanding and risk characterization. This leads to high-risk hazards being easily concealed, and limited urban management resources struggling to achieve refined and proactive risk prevention and scheduling [4].

In previous studies, Jana et al. (2024) found that the municipal road defect image recognition method based on deep convolutional neural networks could automatically identify surface defects such as potholes and cracks in large-scale inspection photos. The recognition results were connected to the urban management information platform for

generating maintenance work orders and statistical reports. Their study achieved certain results in visual recognition accuracy. However, it mainly focused on a single image modality, with limited use of multi-source information such as complaint texts and historical work orders, making it difficult to form a unified evaluation framework for comprehensive risks [5]. Ahmad and Sharoon (2025) demonstrated that applying text mining and topic modeling to public complaint documents and urban maintenance records could effectively identify high-frequency risk areas and common urban facility issues. Their method could provide municipal authorities with data-driven support for developing targeted maintenance strategies. This could extract risk topics and spatio-temporal distribution features from unstructured texts. However, it failed to fully consider visual information such as facility images and videos; it also lacked sufficient description of specific facility states, and did not establish a complete response priority decision-making mechanism [6].

Vision-Language Transformer (VLT) performs excellently in cross-modal representation learning, providing a new technical path for unified modeling of urban facility images and text descriptions. Based on this, on the one hand, it is necessary to use VLT to construct multimodal risk representations; facility appearance defects and on-site environments can be organically integrated with text descriptions and historical work orders to improve the accuracy and robustness of risk assessment. On the other hand,

under the premise of limited resources, urban management departments are more concerned about the decision-making problem of "which one to handle first". Therefore, there is an urgent need to further convert risk assessment results into an operable response priority sorting mechanism; meanwhile, it is embedded into existing urban management information systems to provide intelligent assistance for scheduling command and on-site disposal. This practical demand and technical opportunity together constitute the main motivation for this study.

2. Integrated Framework for Urban Facility Risk Assessment and Response Priority

In this study, urban facility risk assessment and response

priority sorting are not two isolated tasks, but form an "integrated method framework" connected in series through VLT. The overall process can be summarized into "four stages": multi-source data collection and preprocessing, multimodal feature encoding and alignment, facility risk scoring, and response priority decision-making and system integration. The first two stages focus on using VLT to build a unified multimodal semantic space, deeply integrating the "visible" facility appearance with "understandable" text descriptions. The latter two stages then introduce urban management business rules and sorting strategies on this basis, further converting risk assessment results into operable scheduling instructions and work order priorities, as exhibited in Table 1:

Table 1: Core modules and key technologies of the method framework

Module name	Input data	Key technologies/models	Main function description
Multi-source data collection and preprocessing module	Facility images, inspection photos, complaint texts, work order records	Image enhancement, text cleaning, and annotation standards	Unified collection and cleaning of multimodal data related to urban facilities, constructing standardized samples and label systems.
Multimodal feature encoding and alignment module	Image tensors, text sequences	VLT	Images and text are mapped into a unified semantic space to learn multimodal joint representations of facility risks.
Facility risk scoring module	Multimodal joint feature vectors	Multilayer perceptron, classification/regression heads, task loss functions	Facility risk levels or continuous risk scores are output to provide foundational quantitative results for subsequent priority calculations.
Response priority decision-making module	Business features such as risk scores, location weights, and historical failures	Learning to rank models, weighted scoring mechanisms	Response priority scores are generated by integrating multiple factors to achieve sorting recommendations for different work orders/facility tasks.
Urban management information system integration and visualization display module	Priority lists, risk maps, and work order information	Web service interfaces, visualization, and dashboard design	Model outputs are seamlessly embedded into existing urban management information systems to provide decision support for dispatch command and on-site handling.

Re

garding specific implementation paths, the integrated framework first extracts image and text information related to urban facilities from surveillance cameras, the 12345 complaint platform, and historical work order systems through the "multi-source data collection and preprocessing module"; meanwhile, it establishes sample pairs and risk labels in accordance with unified standards. Subsequently, the "multimodal feature encoding and alignment module" is based on the VLT structure, adopting a visual Transformer for images and a linguistic Transformer for texts. It achieves semantic alignment through a cross-modal attention mechanism, generating joint representation vectors that can characterize both facility appearance defects and semantic descriptions simultaneously.

Next, the "facility risk scoring module" takes multimodal joint features as input and outputs the risk level or continuous risk score of facilities through a classification or regression head; it realizes quantitative evaluation of different facilities and different hidden dangers. On this basis, the "response priority decision-making module" further integrates business features such as location importance, service object sensitivity, historical failure frequency, and pedestrian and vehicle flow

intensity; it generates comprehensive priority scores using learning-to-rank or weighted scoring methods, thereby answering the core management question of "which one to handle first". Finally, through the "urban management information system integration and visualization display module", the model outputs are mapped to risk heatmaps on the command screen; it priority ranks in the pending work order list, and dispatches suggestions on mobile terminals, forming a closed loop from multimodal perception to intelligent decision-making.

3. Experimental Design and Result Analysis

3.1. Experimental design

The dataset selected for the experiment is "A Multi-National Image Dataset for Automatic Road Damage Detection". This dataset contains a total of 47,420 road images from 6 countries: Japan, India, the Czech Republic, Norway, the United States, and China. The dataset is mainly used for automatic detection and classification of road damage and can serve as a typical "urban facility risk"

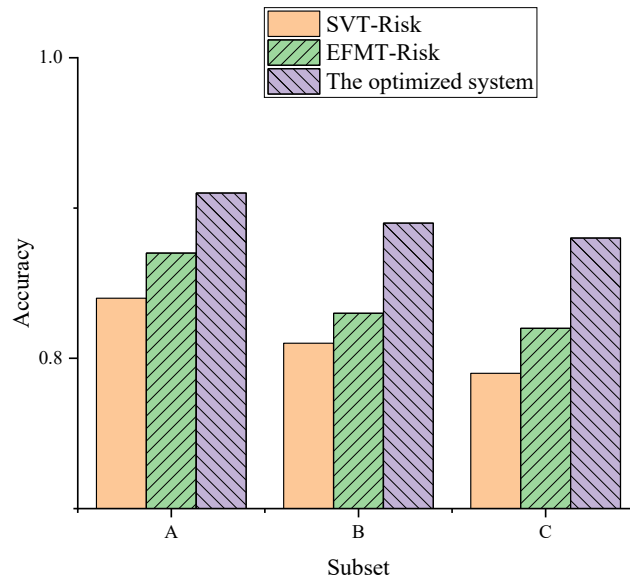
scenario. It can be downloaded via <https://arxiv.org/abs/2209.08538>.

The experiments of this study are completed on a workstation. The Central Processing Unit (CPU) model is Intel Xeon Silver 4210; the Graphics Processing Unit (GPU) model is NVIDIA GeForce RTX 3090 (with 24 gigabytes (GB) of video memory). The memory capacity is 128GB; the storage hard disk type is 1-terabyte (TB) NVMe Solid State Drive (SSD); the operating system version is Ubuntu 20.04 Long Term Support (LTS); the Python version is 3.10, and the deep learning framework version is PyTorch 2.1. The image encoder adopted by the proposed optimized model is ViT-B/16, with a hidden dimension of 768 for joint features, 12 heads for multi-head self-attention, and 12 Transformer encoding layers; the input image resolution is 224×224, and the maximum text length is 128. During the training phase, the batch size is 32, the initial value of the base learning rate is 3×10^{-5} , the optimizer type is AdamW, and the weight decay coefficient is 0.01; the learning rate scheduler adopts linear warm-up plus cosine annealing with a warm-up ratio of the first 10% of training steps. The total number of training epochs is set to 50, and an early stopping strategy is used to prevent overfitting. The loss function type is a weighted combination of cross-entropy classification loss and ranking loss; class weights are balanced according to the training set distribution to improve the recognition and ranking effects of high-risk samples. The comparison frameworks selected for the experiment are Single-modal Vision Transformer-based Risk Assessment Framework (SVT-Risk) and Early-Fusion Multimodal Transformer-based Risk Assessment Framework (EFMT-Risk). The selection of SVT-Risk and EFMT-Risk as comparison models is mainly based on three considerations.

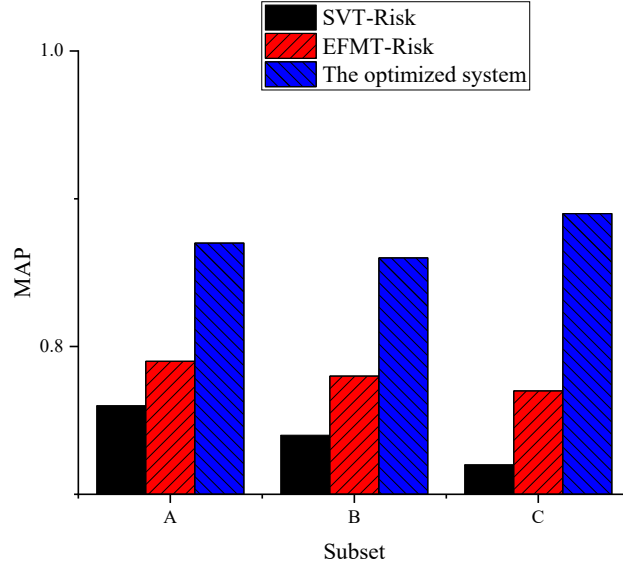
First, SVT-Risk represents the current mainstream single-modal Transformer visual framework, which only uses facility images for risk identification. It can serve as a baseline of "only seeing, not reading text" to test the upper limit of model performance when no linguistic information is used at all. Second, EFMT-Risk introduces both images and texts in structure, but only adopts an early fusion method of feature concatenation. It is a common multimodal enhancement scheme with low implementation cost in engineering; it can also be used as a "simple multimodal" baseline to compare whether the deep cross-modal interaction based on VLT in this study truly brings gains. Third, these two frameworks have high comparability with the proposed optimized model in terms of encoder selection and training settings. This is conducive to fairly evaluating the impact of different depths of multimodal interaction on the performance of urban facility risk assessment and response priority decision-making under the same data and experimental conditions.

3.2. Performance comparison experiment

To verify the comprehensive performance of different frameworks in urban facility risk assessment and response priority decision-making tasks, this section selects two indicators: accuracy and mean average precision (MAP). The dataset is divided into three subsets (A, B, C) according to diverse scenarios. These scenarios correspond to main road and core business district scenarios, residential and public service facility scenarios, and complex scenarios such as old areas and back streets; they aim to test the model's robustness and generalization ability across various urban contexts. The experimental results are presented in Figure 1:



(a)



(b)

Figure 1: Performance comparison experiment results ((a): Accuracy; (b): MAP)

Results in Figure 1 indicate that on the three subsets, the proposed optimized model's accuracy attains 0.91, 0.89, and 0.88, respectively. This represents a stable improvement compared to SVT-Risk (0.84 on subset A, 0.79 on subset C) and EFMT-Risk (0.87 on subset A, 0.83 on subset B). It shows that in diverse scenarios, the optimized model can more accurately distinguish facility risk levels and reduce misjudgments and omissions through more sufficient vision-language interaction. In terms of ranking performance, the MAP of the proposed optimized model on subsets A, B, and C is 0.87, 0.86, and 0.89, respectively, which is significantly higher than EFMT-Risk (0.78 and 0.77 on subsets B and C) and SVT-Risk (0.76 on subset A). This illustrates that the optimized model has better relevance and discrimination in ranking high-risk facilities at the top; it also facilitates the priority identification of severe damage on main roads, hidden dangers in people's livelihood-sensitive areas, and potential high-risk points in old areas. Overall, both in classification accuracy and priority ranking quality, the optimized model proposed here demonstrates leading performance with small fluctuations across the three scenario subsets, reflecting good cross-scenario generalization ability.

3.3. Discussion

The comparison results demonstrate that SVT-Risk, which relies solely on image information, is more susceptible to the influence of lighting conditions, shooting angles, occlusions, and pavement texture complexity. Thus, it has limited ability to capture subtle risk signs in complex urban scenarios. Although EFMT-Risk with early fusion uses both images and texts, it only completes multimodal fusion through simple feature concatenation; this problem makes it difficult to fully model the fine-grained correspondence among hidden danger descriptions, historical work order semantics, and visual details. In contrast, the integrated framework of this study, based on VLT, achieves deeper cross-modal interaction at the feature level and unifies risk assessment and response priority decision-making at the task level. Therefore, it can maintain high recognition ability and ranking quality across different regions and facility types.

4. Conclusion

This study constructs a multimodal integrated method framework for urban management scenarios. By introducing VLT, it deeply aligns and interacts facility images with text information such as inspection records and complaint work orders in a unified semantic space, realizing refined characterization of urban facility risk levels. On this basis, the study integrates business features such as location importance and historical failure frequency to form a response priority sorting mechanism oriented to urban management business. Overall, this study offers a feasible technical path for urban management information systems to transform from "passive order receiving" to "proactive early warning + intelligent scheduling". At the same time, it has a certain reference value for improving the safety of urban facility operation and maintenance and the efficiency of resource allocation.

Although this study has achieved relatively ideal experimental results, it still has limitations. Firstly, the experiments are mainly based on a single public dataset and limited scenario division, lacking coverage of other types of urban facilities (such as bridges, lighting, pipelines, etc.) and diverse environments of different cities. The cross-city and cross-facility transfer capability of the model has not been systematically verified. Secondly, the current multimodal data mainly consists of images and structured/semi-structured texts. Meanwhile, more abundant information sources, such as social media, meteorological, and Internet of Things sensor data, have not been fully introduced, resulting in deficiencies in risk characterization in temporal-spatial dimensions and behavioral aspects. In addition, the VLT-based framework also has certain thresholds in terms of computational overhead, model complexity, and online deployment. Future research can be carried out in three directions. First, conduct large-scale verification on data from more cities and multiple types of facilities to improve the model's generalization and transferability. Second, introduce multi-source data, such as sensors, trajectories, and meteorology, to construct a more comprehensive temporal-spatial risk representation; additionally, methods such as causal inference and

reinforcement learning are combined to expand priority decision-making into dynamic scheduling and resource optimization. Third, explore lightweight, multi-task, and incremental update mechanisms to enable the model to be efficiently deployed on the edge side or existing urban management infrastructure; this can better support long-term operation and iterative upgrading in actual urban governance scenarios.

References

- [1] Sanusi B O. Smart Infrastructure: Leveraging IoT and AI for Predictive Maintenance in Urban Facilities[J]. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 2025, 17(02): 26-37.
- [2] Gu G L I, Liu J, Shao W. Urban flood risk assessment under rapid urbanization in Zhengzhou City, China[J]. Regional Sustainability, 2023, 4(3): 332-348.
- [3] Tang H, Zhu Q, Qin B, et al. UAV path planning based on third-party risk modeling[J]. Scientific reports, 2023, 13(1): 22259.
- [4] Ghani J, Nawab J, Ullah Z, et al. Multivariate statistical methods and GIS-based evaluation of potable water in urban children's parks due to potentially toxic elements contamination: a children's health risk assessment study in a developing country[J]. Sustainability, 2023, 15(17): 13177.
- [5] Jana S, Middya A I, Roy S. Participatory sensing based urban road condition classification using transfer learning[J]. Mobile Networks and Applications, 2024, 29(1): 42-58.
- [6] Ahmad R E, Sharoon O. Harnessing Data Science for Sectoral Efficiency in the Governance Framework of the Punjab[J]. Pakistan Social Sciences Review, 2025, 9(1): 157-171.