Identification of Illegal Urban Construction Based on VLMs and Remote Sensing Imagery

Shuhe Zhao [1], Zhixin Zheng [2, *]

[1] Ninghai Hailiang Senior High School, Ningbo, China

[2] School of Geosciences and Info-Physics, Central South University, Changsha, China

* Corresponding Author Email: 8215240113@csu.edu.cn

Abstract. Urbanization has led to a continuous increase in the demand for investigating illegal buildings, and manual inspection and drone patrol have shortcomings such as efficiency and airspace limitations. This paper researches urban illegal construction identification based on high-resolution remote sensing imagery and visual language models (Vision-Language Models, VLM). The LEVIR-CD public remote sensing dataset and the Qwen3.6-PLUS model were adopted. Multi-temporal remote sensing data and urban planning data are integrated for building a Visual Question Answering (VQA) judgment system. The experiment consisted of 10 training samples and 8 test samples. The model achieved a recognition accuracy of 62.5%, making it difficult to identify illegal buildings. Analysis revealed that insufficient sample size, poor model fit, and an unreasonable guided learning scheme were the contributing factors to insufficient accuracy. This study confirms the feasibility and long-term development prospects of this technical approach. Subsequent data expansion and algorithm optimization aim to improve the accuracy of urban building identification and illegal construction determination by training the Visual Language Model (VLM) to better adapt to high-resolution remote sensing images. At the same time, it aims to improve the stability of artificial intelligence technology in the field of urban illegal construction identification and processing, and reduce the proportion of manual identification.

Keywords: Remote sensing, illegal construction in cities, visual language model.

1. Introduction

Safety accidents caused by illegal construction of buildings are still common [1, 2]. The widespread development of remote sensing technology has made it possible to apply remote sensing to municipal construction. Furthermore, with the widespread acquisition of high-resolution remote sensing images, artificial intelligence technology has become an important technical approach for remote sensing analysis. The cross-modal reasoning capability of Vision-Language Models (VLM) provides new possibilities for solving the technical bottlenecks of traditional deep learning methods.

While existing technologies for identifying illegal construction have achieved a certain degree of automation, they still have significant shortcomings. To start with, the identification of most illegal constructions still relies on manual inspection, which is inefficient. For example, while using drones to identify illegal constructions improves efficiency, it still suffers from drawbacks such as drone bans, technological limitations, and legal risks [3, 4].

With the development of remote sensing technology and multimodal artificial intelligence, urban illegal construction identification based on VLM and remote sensing images offers a new possibility for solving the above problems. This paper, based on the high compatibility and wide application scenarios of the VLM visual language model, artificial intelligence technology, and remote sensing technology, studies the fundamental issues of accuracy and efficiency in illegal construction identification [5]. The results show that the recognition accuracy is controlled at about 60% with a small number of samples. Further research can be conducted to improve the accuracy and efficiency of illegal construction recognition by increasing the number of samples or training visual language models and artificial intelligence. Based on the shortcomings and adaptation optimization of remote sensing and visual language models, an automated illegal construction identification system that does not require manual identification is constructed. This study departs from the conventional review

model, combining the advantages of visual language models in image data analysis and processing with remote sensing, and proposes a targeted solution to address the shortcomings of the current illegal construction identification, which relies heavily on manual labor and is inefficient. This project uses high-resolution remote sensing imagery as the data source and large-scale visual language models such as Qianwen as the core technology, and combines multi-temporal remote sensing images to carry out research. First of all, the paper searched for and constructed a high-quality remote sensing dataset of illegal building construction, integrating publicly available datasets from multiple sources. Second, it built a professional visual question answering (VQA) system for illegal building identification and designed multi-dimensional professional question-answer pairs. Finally, it carried out scene adaptation optimization of the VLM model to achieve efficient detection and judgment, and to assist inspectors in carrying out their review work. The research findings of this project provide an innovative reference for the application of visual language models in remote sensing and urban construction.

2. Data and Methods

2.1. Data Source

LEVIR-CD is a publicly available benchmark dataset released by the Beihang University team in 2020. It consists of RGB imagery of urban areas in Texas, USA, captured from Google Earth. It is specifically designed for building change detection. This dataset has high resolution and already labels the changed building portions, making it convenient to construct dedicated research datasets based on it [6]. Secondly, the VLM model selected is the Qwen 3.6-PLUS model, which is the latest big data model in the Qwen series, supports multimodal tasks, and is well-developed.

2.2. Model

The Qwen3.6-PLUS enhanced version model still adopts the underlying universal principle of the Qwen-VL series multimodal architecture and uses a three-stage sequential processing approach consisting of a visual encoder, a cross-modal adapter, and a large language model [7]. First, the features and spatial location information of the image regions are extracted. Then, the extracted image features are compressed into visual tokens through an adapter. Finally, the collected visual tokens and text tokens are input into the language model for processing. Through training, the model has been adapted to multimodal tasks in different scenarios and can also recognize buildings in aerial photography.

2.3. Method Introduction

The identification of illegal buildings in cities first requires finding and constructing a high-quality remote sensing dataset of illegal buildings. Simultaneously, Geographic Information System (GIS) technology is used to perform qualitative analysis and storage of land use information. In addition, Visual Question Answering (VQA) for illegal construction identification was used to optimize the model based on Few-shot prompts [8-10]. Then, satellite data processing is registered and integrated with the digital urban planning model, making it easier for the VLM model to integrate urban layout planning information into an understandable information pattern [11]. By comparing high-resolution remote sensing image information with urban land use planning information, the difference between urban built-up development land and ecologically restricted protection areas can be distinguished, and it can be clearly determined whether the construction planning information of the area where the building is located matches the actual use. After that, by comparing the remote sensing images from the two periods, the changes were analyzed, and the corresponding patches were extracted and labeled. Next, after comparing the actual and planned land use, the illegal construction was determined [12]. The detected illegal and irregular behaviors were marked, forming a methodology chain for urban illegal construction identification based on VLM and remote sensing images (Fig. 1).

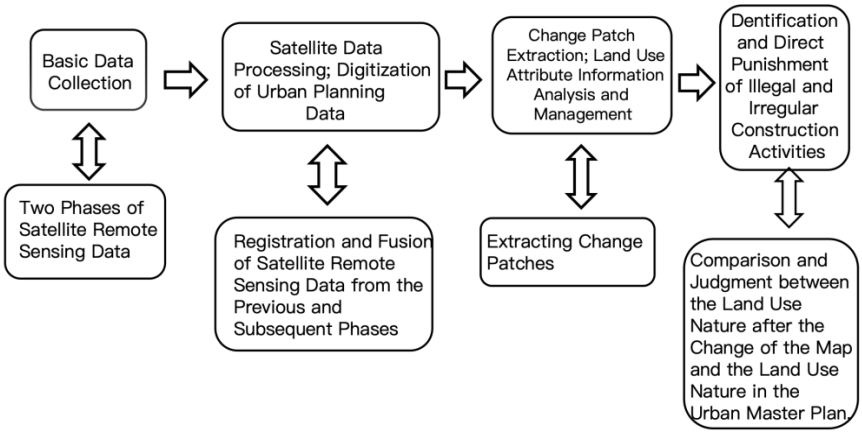


Fig. 1. Method chain for urban illegal construction identification based on VLM and remote sensing images (Photo/Picture credit: Original).

Ten sets of training data and eight sets of test data were constructed. After engineering optimization based on the prompt words in the few-shot dataset, the test was conducted, and the test results are as follows (Table 1).

Table 1. Statistical table of test results of urban illegal construction recognition based on VLM and remote sensing images

Order	VLM identification results	Judgment reasons	Real results	test results
1	No (no illegal construction)	This area is all located within the scope of urban built-up development land in the overall land use planning map. Even if the road widening or building renovation is observed in the time series image, the plot does not belong to the restricted ecological control protection zone and does not meet the space judgment conditions for illegal development.	No (no illegal construction)	√
2	No (no illegal construction)	Although the plot is included in the restricted ecological control reserve, no new structures are observed by comparing the two time series remote sensing images; Only the farmland texture and the edge of the residential area are slightly extended. This part of the extended area belongs to the land for urban construction and development, which does not meet the core criteria, of "new structures within the time sequence".	No (no illegal construction)	√
3	Yes (illegal construction)	The new large-scale industrial plants and supporting roads appear in the upper left corner of the later image. The plot was idle woodland space in the early stage, and the plot is located in the restricted ecological protection zone; At the same time, it meets the two criteria of "no structures in the early stage, new construction in the late stage"+"located in the restricted development control zone".	Yes (illegal construction)	√
4	No (no illegal construction)	Part of the plot is included in the restricted ecological control protection zone, but the new residential buildings and roads in the later stage are distributed in the urban development land; Only a small amount of construction traces and idle space are left in the ecological control zone, no new structures are formed, and the double violation criteria are not met at the same time.	Yes (illegal construction)	×
5	No (no illegal construction)	Although the plot is located in the restricted ecological control protection zone, the time sequence image can only observe the land leveling, vehicle trajectory and early construction preparation traces, and there are no formed structures such as houses and workshops, which does not meet the core judgment conditions of "new structures in time sequence".	No (no illegal construction)	√
6	Yes (illegal construction)	The upper left corner of the later image is a piece of newly-built residential area and supporting roads. In the early stage, the plot was idle construction space of forest land and belonged to the restricted ecological control protection area; At the same time, it meets the two violation criteria of "no structure in the early stage, new construction in the late stage" and "located in the restricted regulatory zone of development"	Yes (illegal construction)	√
7	No (no illegal construction)	The plot is included in the restricted ecological control protection zone. Compared with the two phases of remote sensing images (B for the early stage and a for the late stage), the residential area and road facilities have all existed in the early stage, and there are no new structures within the time sequence, which does not meet the requirements for judging the illegal development in the time sequence dimension.	Yes (illegal construction)	×
8	No (no illegal construction)	Although the plot belongs to the restricted ecological control protection zone, there are no new structures in the comparison of the two phases of remote sensing images, only the farmland texture and road boundary are slightly fine tuned, which does not meet the core standard of illegal development of "new structures within the time sequence".	Yes (illegal construction)	×

3. Results and Discussion

Table 2. Test and analysis of urban illegal construction recognition based on VLM and remote sensing images

	Accuracy	Precision	F1-score	Recall
Model	0.625	1.000	0.571	0.400

After analyzing the experimental results (Table 2), it is found that the accuracy, F1 score, and recall rate are low, and the visual language model (VLM) has obvious omissions. The purpose of urban illegal construction recognition based on VLM and remote sensing images is to improve the efficiency and accuracy of illegal construction recognition. However, this experiment did not meet the corresponding expectations. The error sources and subsequent available solutions are as follows:

Firstly, too few training data sets may lead to the phenomenon of underfitting in model recognition, and the model can not well recognize the remote sensing image features of ground objects. The test data set is too small, and the random error of accuracy, F1 score, and recall rate calculation is relatively large. Subsequently, it is necessary to expand the dual temporal remote sensing image data, supplement the remote sensing images of different typical scenes, and weaken the prediction error[13].

Secondly, the general visual language model of Qwen3.6-plus may not be able to well recognize buildings in remote sensing images. The VLM model of the original Qwen3.6-plus may be mainly suitable for photographic images, and its perception ability for large-scale remote sensing images is insufficient. In view of the phenomenon that the recall is relatively small and qwen3.6-plus tends to conservatively miss the inspection, the reason may be that the existing VLM highly depends on the semantic environment and actively ignores the visual details. In the face of a large range of remote sensing images, to reduce the occurrence of texture interference illusion, the VLM model chooses to ignore the details of building changes that are not obvious [14]. In the future, it can improve the recognition ability of the model for buildings in remote sensing images by strengthening the guidance of prompt words, or improve the weight of remote sensing images based on the VLM model of the original Qwen3.6-plus, and rebuild the VLM model for remote sensing images [15].

In addition, the method to guide VLM learning may not be good. In this experiment, the cue word engineering optimization based on few-shot was carried out by using the cue word guidance method. The learning efficiency of identifying building changes may be low due to the poor setting of cue words and other reasons. In the future, it should design specialized prompt phrases, standardize the judgment logic chain of illegal construction identification, and distinguish between normal land and illegal construction

To sum up, it can improve the recognition accuracy, precision, F1 score, and recall rate of the VLM model from the perspective of expanding the data set, enhancing the adaptation of the VLM model, and optimizing the guided learning method, to achieve the purpose of improving the efficiency and accuracy of identifying illegal constructions.

4. Conclusion

In this paper, the remote sensing image illegal construction data set is built based on the lEVIR-CD data set. The current Qwen3.6-plus visual language model is selected, and the VLM model is tried to be introduced into the field of remote sensing image recognition to carry out the automatic recognition of illegal urban buildings, improve the efficiency of illegal construction recognition, and then assist or even replace the conventional manual and UAV illegal construction recognition. According to the method chain, the basic data were collected and processed. After that, the VLM was used to extract the change spots, which were classified as illegal construction, and the final result was obtained. However, due to the influence of relatively few data sets, poor recognition of remote sensing images based on the universal visual language model, and guided learning strategies, the results obtained in this experiment did not reach the expected results, with an accuracy of 0.625, a precision of 1.000 F1-score0.571, Recall rate 0.400, nevertheless, after objective analysis, the technical route

and workflow are still feasible. In view of the high rate of missing detection of the VLM, the follow-up research work can improve the capture ability of the model for new features in remote sensing images by increasing data sets, enhancing the adaptation of the VLM model, and standardizing the guide words, and gradually improve the recognition ability to meet the accuracy requirements applicable to real life.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Hong K F. The reporter visited the site of the collapse of houses in Changsha, and the phenomenon of illegal construction was serious. Safety and Health, 2022, (7): 12-14.

[2] Zhengzhou Administration of Work Safety, Henan Province. Three people were killed and five injured in the collapse of illegal buildings. Labor Protection, 2020, (8): 55.

[3] Ouyang H L. The application of UAV urban building monitoring system in illegal construction investigation. New Urban Construction Technology, 2025, 34(9): 76-78.

[4] Onososen A O, Musonda I, Onatayo D, et al. Impediments to construction site digitalisation using unmanned aerial vehicles (UAVs). Drones, 2023, 7(1): 45.

[5] Kazanskiy N, Khabibullin R, Nikonorov A, et al. A comprehensive review of remote sensing and artificial intelligence integration: Advances, applications, and challenges. Sensors, 2025, 25(19): 5965.

[6] Chen H, Shi Z W. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing, 2020, 12(10): 1662.

[7] Bai J, Bai S, Yang S, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023.

[8] Zhi Y J, Jiang Y W, Yang Z, et al. The development status and prospect of the basic model for remote sensing image pre training. Chinese Journal of Image and Graphics, 2026, 31(4): 973-986.

[9] Zhao J D, Yu Y M, Tao Q C. Research on KB-VQA application of knowledge visual depth collaboration. Information Technology and Informatization, 2026, (2): 24-28.

[10] Leng J, Zhao X, Zhao L. Information gain and guided calibration with few-shot learning for prompt optimization in LLMs: Application in disease diagnosis. Proceedings of the 9th International Conference on Computational Biology and Bioinformatics, 2026: 169-174.

[11] Varol B, Yılmaz E Ö, Maktav D, et al. Detection of illegal constructions in urban cities: Comparing LIDAR data and stereo KOMPSAT-3 images with development plans. European Journal of Remote Sensing, 2019, 52(1): 335-344.

[12] He Y, Ma W, Ma Z, et al. Using unmanned aerial vehicle remote sensing and a monitoring information system to enhance the management of unauthorized structures. Applied Sciences, 2019, 9(22): 4954.

[13] Jiang Y C, Zhu B. Remote sensing image data enhancement based on generating countermeasure network under the condition of few samples. Laser & Optoelectronics Progress, 2021, 58(8): 0810022.

[14] Shahgir H S, Chen X, Fu Y, et al. VLMs need words: Vision language models ignore visual detail in favor of semantic anchors. arXiv, 2026.

[15] Tan X M, Xi B B, Xue C B, et al. A survey of remote sensing visual language models for zero sample classification. Journal of Remote Sensing, 2026, 30(2): 279.