Article

A Space Information Extraction Method Based on Multi-modal Social Media Data: A Case Study on Urban Inundation

Yilong Wu[1], Yingjie Chen[1], Rongyu Zhang[2], Zhenfei Cui[1], Xinyi Liu[1], Jiayi Zhang[1], Yong Wu[1,3,*]

[1] School of Geographical Sciences,School of Carbon Neutrality Future Technology,Fujian Normal University,Fuzhou 350117,China;
[2] School of Software Engineering,Xiamen University of Technology,Xiamen 361000,China;
[3] Institute of Geography, Fujian Normal University, Fuzhou 350000, China
* Correspondence: wuyong3216@163.com;

Abstract: With the proliferation and development of social media platforms, social media data has become an important source for acquiring spatio-temporal information on various urban events. Providing accurate spatio-temporal information for events contributes to enhancing the capabilities of urban management and emergency response. However, existing research on mining spatio-temporal information of events often focuses solely on textual content, neglecting data from other modalities such as images and videos. Therefore, this study proposes an innovative spatio-temporal information extraction method for multi-modal social media data (MIST-SMMD), which extracts the spatio-temporal information of events from multi-modal data on Weibo at coarse and fine-grained hierarchical levels, serving as a beneficial supplement to existing urban event monitoring methods. This paper takes the "July 20th Zhengzhou Heavy Rainfall" incident as an example, to evaluate and analyze the effectiveness of the proposed method. The results indicate that in the coarse-grained spatial information extraction using only textual data, our method achieves a Spatial Precision of 87.54% within a 60m range, and reaches 100% Spatial Precision for ranges beyond 200m. For fine-grained spatial information extraction, the introduction of other modal data such as images and videos results in a significant improvement in Spatial Error. These results demonstrate the ability of the MIST-SMMD method to extract spatio-temporal information from urban events at both coarse and fine levels, and confirms the significant advantages of multi-modal data in enhancing the precision of spatial information extraction.

Keywords: multi-modal; social media; spatio-temporal information extraction; inundation

1. Introduction

Precise spatiotemporal localization plays a pivotal role in understanding and forecasting urban events. Spatiotemporal information extraction, a subfield of spatiotemporal data analysis and data mining, operationalizes this crucial function. Urban event spatial pattern recognition, urban planning and management, and urban science research all rely on accurate spatial information and effective spatiotemporal information extraction. In practice, the diversity of data sources is a key prerequisite. With the rapid development of Internet technology, social media platforms have become the primary channel for people to obtain and share information[1]. In December 2022, the number of monthly active Weibo users increased by 13 million year-on-year, reaching 586 million, setting a historical record[2]. With the popularity of social media, a large number of events with potential spatiotemporal information are widely spread, and this information has high application value in urban management and event response. For example, transportation planning, disaster management public opinion spatial analysis, and disaster analysis based on geographical tags[3-6] etc., timely and accurate acquisition and analysis of spatiotemporal information are of great significance for urban event spatial pattern recognition, urban planning and management, and urban science research[7,8]. Therefore, extracting precise spatiotemporal information can not only improve the efficiency and timeliness of urban management, command and disaster reduction work, but also contribute to understanding the inherent laws of the city and promoting the development of urban science.

Due to the convenience of text information processing, current research often focuses on using single-modal data (especially text data) for spatiotemporal information extraction, including rule-based methods and Named Entity Recognition (NER) methods. Rule-based methods rely on manually defined rules and patterns to extract information, usually requiring domain knowledge and language resources[9]. These methods, due to not requiring a large amount of annotated data or complex computing resources, have the advantage of quickly and efficiently implementing system prototypes and are therefore widely used by many researchers[10,11]. However, the diversity, informality, and ambiguity of social media text make it difficult for rule-based methods to cover all possible situations. It also requires a large amount of manual participation and maintenance, making it difficult to adapt to the rapid changes, frequent updates, and huge data volume of social media text. Methods based on Named Entity Recognition (NER) extract spatiotemporal information by detecting spatiotemporal-related entities in text data. In recent years, thanks to the rapid development of theory and application research in natural language processing in the field of machine learning[12,13], many studies have begun to use this method to extract spatiotemporal information[14,15]. The advantage of this method is that it can automatically recognize entities in the text, thereby reducing manual participation and maintenance. However, issues such as entity ambiguity, expression diversity, and nested entities in the text may affect its extraction effect.

Despite the fact that utilizing single-modal data can effectively extract potential spatiotemporal information to a certain degree, existing research has paid insufficient attention to the potential of other modes of data such as images and videos. Images and videos carry potential spatial information[16], and multi-modal data formed in conjunction with text data can provide more precise spatial information, thus helping to further enhance the accuracy and comprehensiveness of spatiotemporal information extraction. Previous studies have performed information mining and classification based on multi-modal social media data[17,18], but research on spatiotemporal information extraction from multi-modal data in social media is relatively scarce. Simultaneously, extracting spatiotemporal information from multi-modal social media data poses significant challenges due to issues such as noise, heterogeneity, and sparsity that commonly exist in publicly participated social media data[19,20].

In the past twenty years, the frequency and intensity of flood disasters in major cities worldwide have increased, posing serious threats to economic development and social stability[21]. In this context, investigating how to effectively extract spatiotemporal information of urban flood disasters has become an important subject. Predicting areas where urban inundation might occur in the future through the spatiotemporal information monitoring of urban flood disasters has become a critical means of managing urban inundation[22]. Currently, Internet of Things (IoT) sensing technology and remote sensing technology are commonly used for monitoring and analyzing urban flood disasters[23-25]. At a smaller spatial scale, IoT sensors can more accurately and swiftly respond to urban inundation issues, providing real-time warnings and monitoring[26]. On a larger spatial scale, although optical and radar satellite remote sensing can provide more effective continuous coverage of weather and inundation events compared to IoT sensing[27], flood disasters have short impacts on cities and the surface water coverage is small and concentrated. Due to influences such as cloud layers and vegetation canopy, microwave remote sensing is subject to the total reflection effect and cannot monitor or extract surface water information, causing the original access cycle to become even longer[28]. Therefore, the current methods still have numerous shortcomings in extracting spatiotemporal information from flood disasters, and they struggle to meet the high spatiotemporal resolution requirements of urban flood disaster monitoring. However, when flood disasters occur, people often share information on social media, which may include the time, location, degree, impact range, and duration of the disaster[29]. This information is significant for urban inundation management and prediction[30].

The objective of this research is to further explore the potential of other modal data (such as images and videos) for high-precision correction of extracted spatial information, based on the capability to extract spatiotemporal information from social media text. To this end, we propose a

MIST-SMMD method to address challenges such as multi-modal data fusion and heterogeneity processing, hoping to provide strong support for the early warning and management of urban events and disasters. At the same time, to evaluate and verify this method, we take urban floodwater accumulation events as an example for evaluation. We make publicly available the code, models, and datasets used in this study for researchers to reproduce and conduct further research. These are located at: https://github.com/orgs/MIST-SMMD/repositories (accessed on 19 May 2023).

The main contributions of MIST-SMMD are as follows：

- In terms of data preprocessing, compared with predecessors, we use a text classification model to filter related information and remove similar blog posts within the same day. This is beneficial for cleansing the noise in social media data and standardizing the dataset as much as possible.
- For the extraction of coarse-grained spatiotemporal information, we propose a set of strict standardization rules for spatiotemporal information, allowing the maximum structuring of potential spatiotemporal information.
- For the extraction of fine-grained spatial information, we propose an LSGL method. This leverages cascading computer vision models to further improve the accuracy of spatial information extracted from coarse-grained data, thus enhancing the utilization of image and video modal data from social media.

The structure of this paper is as follows: Section 2 introduces our innovative multi-modal social media data spatiotemporal information extraction method (MIST-SMMD); Section 3 uses the urban inundation event of the "July 20 Zhengzhou Torrential Rain" as an experiment to evaluate and verify this method; Section 4 discusses and analyzes the effectiveness of the method based on Section 3; Section 5 summarizes the entire research and proposes prospects.

2. Methods

2.1. Technical process

We introduce a method for extracting spatio-temporal information from multi-modal social media data, known as MIST-SMMD (Method of Identifying Spatio-temporal Information of Social Media Multimodal Data). The MIST-SMMD process comprises three steps:

Step One: Crawling and pre-processing of social media data.

Step Two: Coarse-grained extraction of spatio-temporal information.

Step Three: Fine-grained extraction of spatial information.

The Normative Dataset for Step Two is derived from the crawling and Pre-Processing of social media data performed in Step One. The Street View Image Dataset refers to all street view data from Baidu Maps. However, as Step Two only involves coarse-grained spatio-temporal extraction from microblog text, image data (including segmented video images) is not required. This data is instead used in Step Three for the fine-grained extraction of spatial information.

MIST-SMMD leverages the complementarity of multi-modal data and the flexibility and generalizability of model cascading, sequentially processing the text and images from social media. The overall flow of the method is shown in Figure 1.

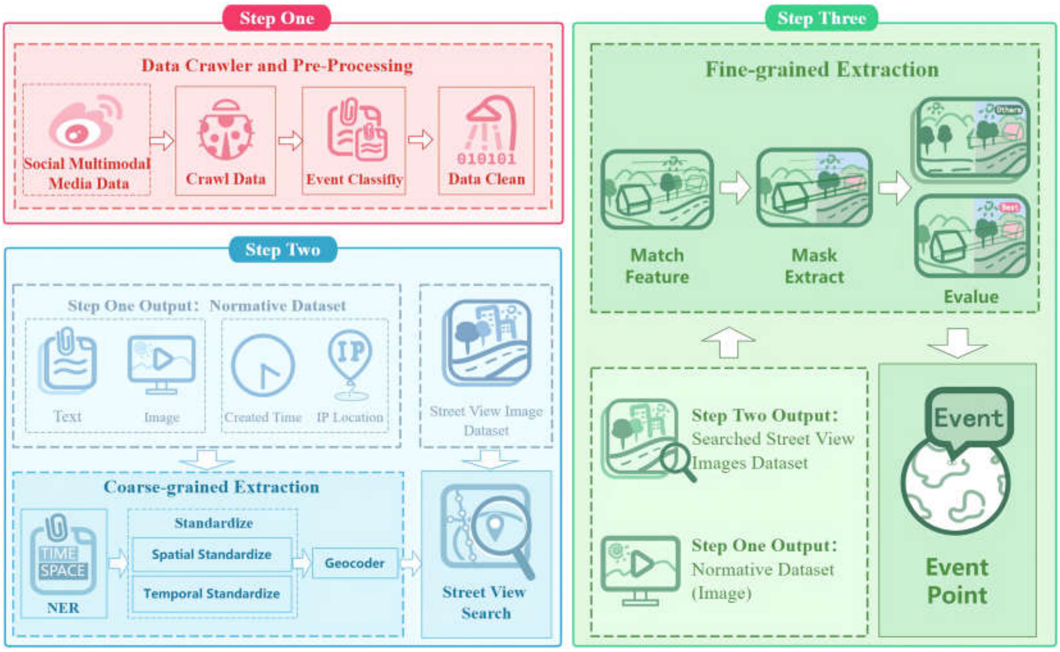


Figure 1. The overall structure of MIST-SMMD process

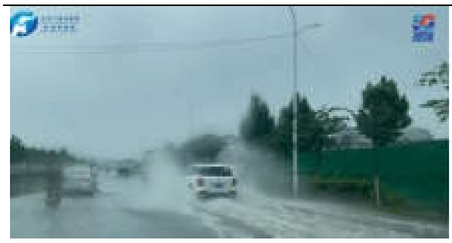
2.2. Data craw and Pre-Process

To obtain Weibo data, we utilize the public interface of Sina Weibo. The Weibo multi-modal data is crawled using Python programming language, by setting the time range and keywords related to city events. This data includes Weibo creation time, text content, images (if any), videos (if any), IP owned provinces (starting from August 1, 2022), etc. For videos, we extract stable frame images using the optical flow method.

Despite initial keyword filtering, not all Weibo posts containing event-related keywords are actually related to the event. Therefore, a text classification model can be used to classify the relevant text for the specified city event. Subsequently, to efficiently process text data, we need to clean the noise in the data. Character-level cleaning includes removing topic tags, zero-width spaces (ZWSP), @other users, emoji, HTML tags, etc. However, as an event often receives coverage from multiple media sources, overly similar report posts may lead to data redundancy. Therefore, we vectorize all text and use an efficient cosine similarity matrix to calculate the similarity of each text with all other texts, eventually removing highly similar Weibo posts (similarity>=0.9).

After the above three steps of data preprocessing, we obtain a normative city event Weibo dataset that is largely noise-free and relevant to the required event. An example of a processed dataset is shown in Table 1.

Table 1. Normative City Event Weibo Dataset Example.

Blog post	Blog Information	Blog Information Values
 7 月 19 日，记者在郑州金岱路距离南四环一公里处发现，金岱路的车道上积水严重，南北双向六车道有近 1 公里的积水带，最深处能淹没半个车轮，道路双向的	Created time	2021/7/19 14:28:17
	IP Location	Nodata
	Is relevant	True

外侧车道水更深，机动车速度稍快行驶，就会激起高于车身两倍的水花。目前这一积水情况还在持续，现场记者没有看到抽水作业，这一路段的积水为何如此严重？为何没有排水作业？河南交通广播的记者也会持续关注。（5G现场记者：靖一、雷静）	Mid	4660679711922369

* **Mid:**Unique identification code for each Weibo post.

2.3 Coarse-grained spatio-temporal information extraction

Due to the high degree of spontaneity and diversity in social media narratives and the lack of a unified text format, we designed a set of rigorous spatio-temporal information standardization rules to efficiently extract key spatio-temporal information from a large amount of Weibo data and lay a foundation for subsequent detailed research. These rules aim to ensure that during the standardization process of spatio-temporal narratives, different levels of potential spatio-temporal information are maximally utilized.

Before the standardization of spatio-temporal information, we first need to extract spatio-temporal information from the text. For the normative city event dataset processed by data preprocessing, we use NER (Named Entity Recognition) technology to identify entities related to spatio-temporal information. To improve the efficiency of subsequent spatio-temporal information standardization, we merge similar labels. Specifically, we combine DATE and TIME labels into the TIME category, as they can both be used as materials for time standardization. The GPE label is kept as a separate category, as it provides the basis for administrative divisions for spatial standardization. We integrate LOC and FAC labels into the FAC category because they can identify specific facilities or locations, which can serve as specific place names for spatial standardization. Table 2 shows the built-in labels required for extracting spatio-temporal information and the reclassified label types.

Table 2. Description of spaCy named entity labels and label classes classified in the present study.

Label type	Named entity labels	Description
TIME	DATE	Absolute or relative dates or periods
	TIME	Times smaller than a day
GPE	GPE	Geopolitical entity, i.e. countries, cities, states.
FAC	LOC	Non-GPE locations, mountain ranges, bodies of water
	FAC	Buildings, airports, highways, bridges, etc.

In terms of temporal-spatial standardization, specific attention is given to both temporal and spatial aspects. Hence, we utilized the JioNLP library, which currently provides the highest quality open-source temporal parsing tools and convenient location parsing tools [31]. In terms of temporal standardization, Weibo publication times are standardized to the format "Year-Month-Day", omitting the specific "Hour:Minute:Second". This is because events typically occur spontaneously and it is difficult to determine the exact time of the event based solely on the Weibo publication time and the implied time information in the text. Consequently, the lowest unit of time is retained only up to "Day", rather than the specific Weibo publication time or the detailed specifics implied in the text. With regard to spatial standardization, we transform the potential spatial information in Weibo posts into a "Province-City-District (County)-Specific Geographic Location" pattern for ease of comprehension during subsequent geocoding, and accurately convert it into the WGS1984 latitude and longitude coordinates for that address.

In this study, further precise handling of spatial information is of paramount importance. Firstly, data that does not include FAC entities must be excluded to ensure the progress of subsequent research. Building upon this, for temporal standardization, it is necessary to ascertain whether there are TIME class labels in the text. If not, the Weibo publication date is directly used as the final standardized time. If there are, some keywords are selected through forward screening, such as: "today", "yesterday", "day", etc. Utilizing the temporal parsing function provided by the JioNLP library and taking the Weibo publication time as a reference, entities with a named entity type of TIME are identified and used as correction times for temporal standardization. In the end, only meaningful point-in-time types are retained. If there are none, the Weibo publication date is selected as the final time.

During the process of spatial information standardization, a larger number of scenarios need to be handled. Firstly, it needs to be determined whether there are GPE labels in the text. Similar to temporal standardization, address standardization also requires a reference point, making the GPE labels critically important. Notably, as of August 1, 2022, the Office of the National Internet Information Office required internet information service providers to display user IP address ownership information, providing a new possibility for texts with only FAC labels and no GPE labels. However, cases involving overseas countries or regions must be excluded. Upon successfully obtaining a GPE label or IP address ownership location and FAC label, the address recognition function provided by JioNLP is used to standardize the content of the GPE label to the "district (county)" unit.

Different statuses of standardization results returned by the above temporal-spatial standardization are categorized, mainly into three types: 0, 1, and 2. Here, 0 represents a failure of standardization parsing, 1 represents incomplete standardization parsing, and 2 signifies successful standardization parsing. Based on the different types of standardization parsing, we only geocode the spatial information after standardization of types 1 and 2 using the Baidu Maps Geocoding API, converting the standardized addresses into Wgs1984 coordinates.

Through these series of steps, we effectively extract coarse-grained temporal-spatial information, laying a foundation for further research. The overall approach for the standardization of temporal-spatial information in Weibo text is visualized in Figure 2, demonstrating the program's assignment of different status types based on different standardization results. Additionally, three common examples of standardization rules are shown in Figure 3.

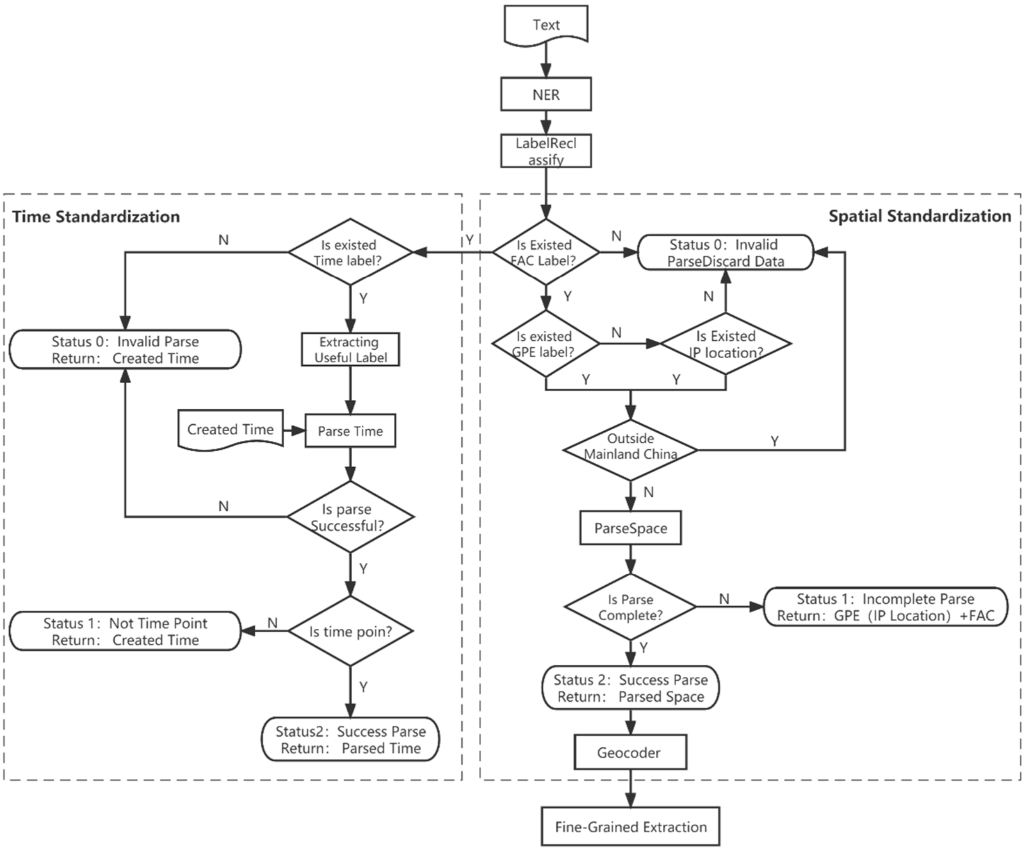
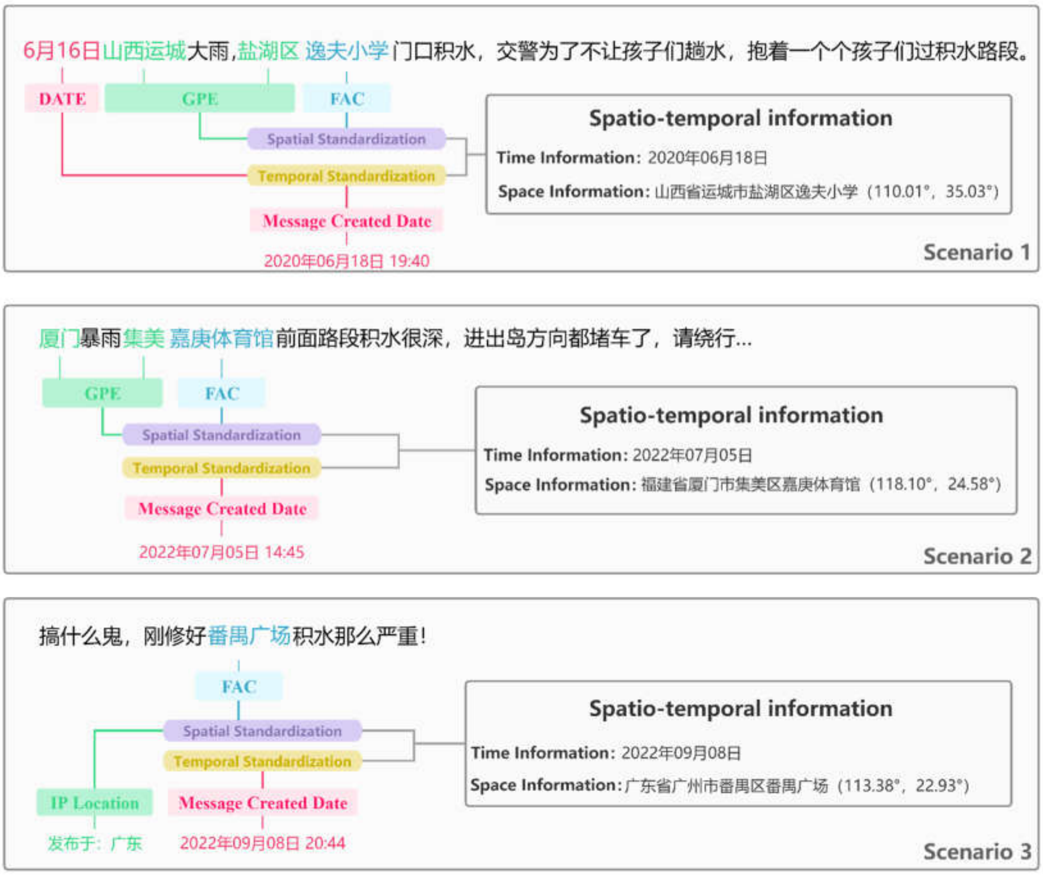


Figure 2. The flowchart of Spatio-temporal standardization.



6月16日山西运城大雨,盐湖区 逸夫小学门口积水，交警为了不让孩子们趟水，抱着一个个孩子们过积水路段。

Time Information：2020年06月18日

Space Information：山西省运城市盐湖区逸夫小学（110.01°，35.03°）

厦门暴雨集美 嘉庚体育馆前面路段积水很深，进出岛方向都堵车了，请绕行...

Time Information：2022年07月05日

Space Information：福建省厦门市集美区嘉庚体育馆（118.10°，24.58°）

搞什么鬼，刚修好番禺广场积水那么严重!

Time Information：2022年09月08日

Space Information：广东省广州市番禺区番禺广场（113.38°，22.93°）

Figure 3. Three common examples of standardization.

While coarse-grained spatial and temporal information has been effectively extracted via the steps described above, in social media data, users often express location and orientation based on their personal perception and cognition of the geographical environment. Thus, the spatial coordinates extracted through coarse-grained extraction may only reflect a representative building or place, while specific orientation descriptions, such as "nearby," "at the corner," "next to," etc., are often somewhat vague. One solution to this issue is to categorize the standardized addresses into two main classes, namely roads and non-roads, by referring to the categorization after Baidu Map geocoding. For standardized addresses of road type, streetscape sampling points are generated at 5-meter intervals along the road network vector in the Open Street Map (OSM) that corresponds to the road name. For non-road type standardized addresses, a buffer zone with a radius of 200 meters is created around the address, and streetscape sampling points are similarly generated at 5-meter intervals along the road network vector in the OSM that has been clipped within this zone.

However, the randomness of social media data raises another issue: in the same microblog post, the image may not be directly related to the text content. This implies that even if a space information is mentioned in the post, the image may not necessarily be related to this information. Additionally, given the varied quality of user-uploaded images or videos, there are not many clear, high-quality streetscape images that contain potential spatial information. To further explore these multimodal data, a semi-manual method can be adopted. First, based on the semantic segmentation of streetscapes, a simple algorithm determines whether each user-uploaded image is a streetscape image; that is, an image should contain at least three types of image semantics: road, sky, buildings, and poles. Then, through manual screening, high-quality, relevant images are selected from the microblogs and associated with the addresses standardized during the coarse-grained extraction phase. In this way, high-quality microblog image-text data can be screened out and categorized as

"Positive," while those coarse-grained standardized address points without high-quality, relevant images are categorized as "Negative."

2.4 Fine-grained extraction of spatial information

To extract fine-grained spatial information from the high-quality microblog image-text data above, a series of image processing techniques are required to compare it with streetscape images that already contain spatial information, thereby screening out the best match for spatial information migration. In this process, the matching degree between the social media images and streetscape images determines the reliability of the fine-grained spatial information. To maximize the reliability of this process as much as possible, we designed a cascade model LSGL(LoFTR-Seg Geo-Localization) based on match-extraction-evaluation.

Given the randomness of social media, most user-uploaded images are blurry, which greatly affects the selection of feature points. To solve this problem, the LSGL model adopts the LoFTR (Local Feature Transformer) [32] feature matching method in the matching stage. This method can not only extract feature points from blurry textures effectively but also maintain a certain relative positional relationship between feature point pairs through a self-attention mechanism, significantly improving the performance of streetscape image matching.

For the ideal streetscape image matching task, the feature matching degree between buildings generally represents the similarity of the shooting locations of the two scenes. However, in practical operations, the matching task is often affected by sky, road, vegetation, and other strong similarity feature information, resulting in a large number of feature points in the image that do not carry significant reference information. To reduce the influence of irrelevant information on the matching task, LSGL adopts the DETR model [33], which can efficiently segment images and label them at a relatively low performance overhead level, thereby extracting practical reference feature points from the images for further evaluation.

To select the best-matched streetscape image from all the matching results and extract its coordinates, a quantifiable indicator is required to assess the degree of image matching. With this goal in mind, we rely on the reference feature points of each scene to design this indicator from two dimensions: feature point feature vector matching degree and feature point spatial position difference.

Firstly, we consider the feature point feature vector matching degree. For the LoFTR feature matching method, it can output the feature point coordinates and the corresponding confidence. We first filter out feature points not within the target category based on their coordinates. Then, we use an exhaustive statistical method to calculate the number of remaining feature points. Subsequently, the confidence of each feature point is multiplied and added, and the average of the cumulative result is taken as the confidence of all feature points in the image. In mathematical terms, it is represented as:

$$R = \frac{\sum_{i=0}^{n} C_i}{n}, \tag{1}$$

In the formula,R represents the feature vector matching degree of the feature point,n represents the number of feature points, and $\mathrm{C_i}$ signifies the confidence of feature points.

Second, we consider the spatial position difference of feature points. As user images come from Weibo and are influenced by user devices, shooting level, etc., the features and objects in their images may be slightly offset compared to street view images. However, the spatial relationship between feature points should remain similar. Therefore, based on the coordinates of each pair of feature points in their respective images, we calculate their Euclidean distance and Euclidean direction as follows:

$$E_d = \sqrt{(x - x_0)^2 + (y - y_0)^2}, \tag{2}$$

$$E_a = tan^{-1}\left(\frac{y - y_0}{x - x_0}\right), \tag{3}$$

In equations (2) and (3), $\mathrm{E_d}$ and $\mathrm{E_a}$ respectively denote the Euclidean distance and direction of the feature points in the user image and the reference image. $\mathrm{x, y}$ represent the coordinates of the feature points in the user image,while $\mathrm{x_0}$,$\mathrm{y_0}$ signify the coordinates of the feature points in the reference image.

In order to assess the impact of changes in Euclidean distance and direction on the spatial position of feature points, we calculated the root mean square error for these two indices separately, resulting in RMSED and RMSEA. Multiplying these two values yields the spatial position discrepancy of the feature points, as shown in equation：

$$SM = RMSE_d \times RMSE_a, \tag{4}$$

Standardizing the indicators can more intuitively reflect the relative advantages of the evaluation results. Therefore, it is necessary to process the results of individual evaluations and round evaluations. The main methods are as follows：

$$StanR = \frac{R}{R_{max} - R_{min}}, \tag{5}$$

$$StanSM = \frac{SM}{SR_{max} - SR_{min}}, \tag{6}$$

In these equations, R and SM represent the matching degree and spatial position discrepancy of the feature vector in a single match, respectively. R_{max} and R_{min} are the optimal and worst feature vector matching degrees in a single round of matching, respectively.SR_{max} and SR_{min} are the optimal and worst spatial position discrepancies in a single round of matching, respectively.

Given the differing impacts of these two factors on the results of feature point matching, we have constructed the following final scoring method:

$$M = \frac{StanR}{StanSM}, \tag{7}$$

The more reliable the result of feature matching is, the higher the feature vector matching degree and the lower the spatial position matching degree.

Finally, we select the image with the optimal M value from all matching results and obtain its specific coordinates. We return this as the fine-grained spatial information. Through this series of processes, we have established a cascaded model that can better extract fine-grained spatio-temporal information. Figure 4 shows the impact of each level in this model on the image matching result.

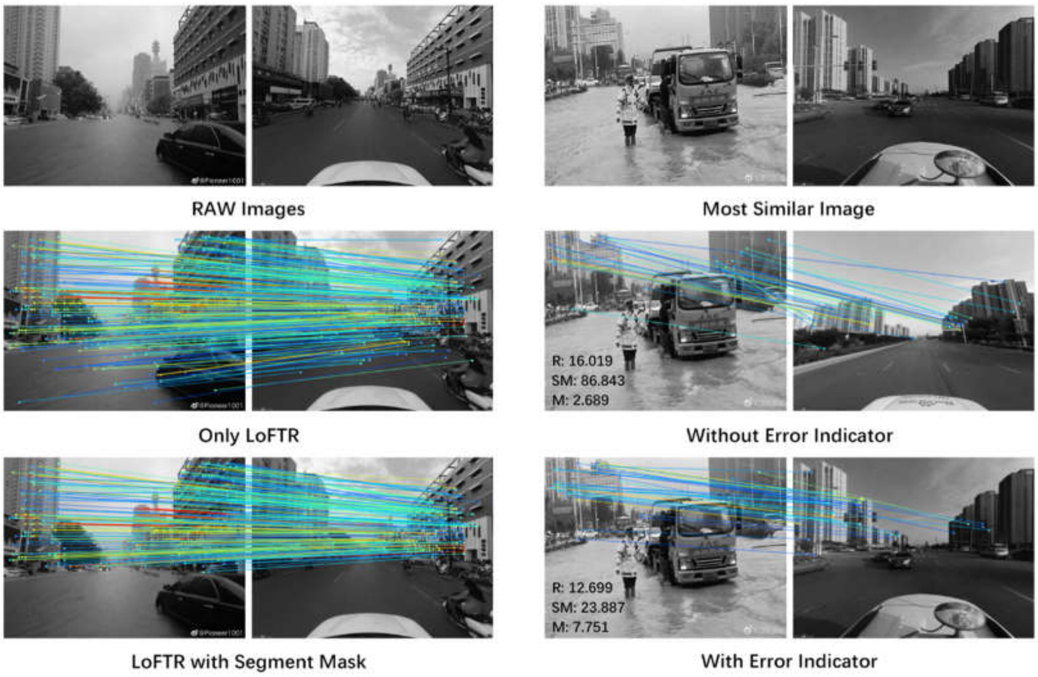


Figure 4. Effect of each level of model on matching results.

3. Experiments Setup

3.1 Research Event

This study has selected the "July 20 Heavy rainstorm in Zhengzhou" as the experimental event to validate the effectiveness of the MIST-SMMD method and conduct accuracy assessment. This event had a widespread impact, resulting in severe disaster losses and attracting significant societal attention. Between 2020 and 2023, this event had the highest volume of related urban inundation posts on Weibo, offering rich research value (see Figure 5).

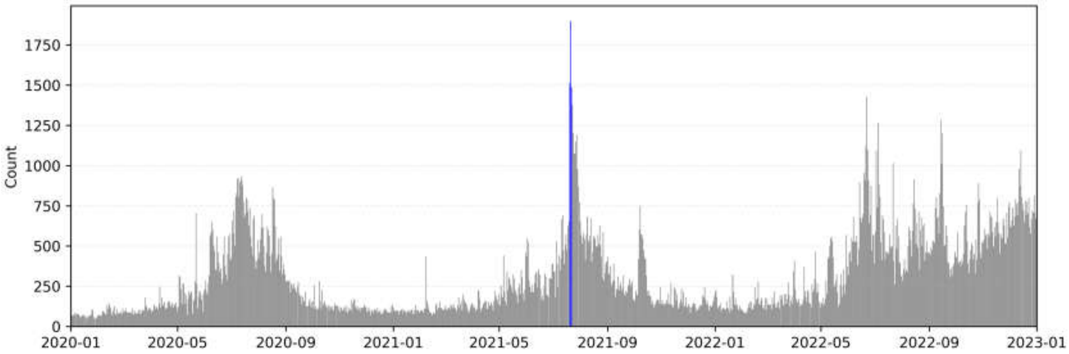


Figure 5. Distribution of Weibo Posts Related to Inundation Events (2020-2023).

Figure 6 displays the officially announced inundation points for this experimental event, along with the inundation spatial distribution (including water systems) extracted by the GF-3 radar satellite.

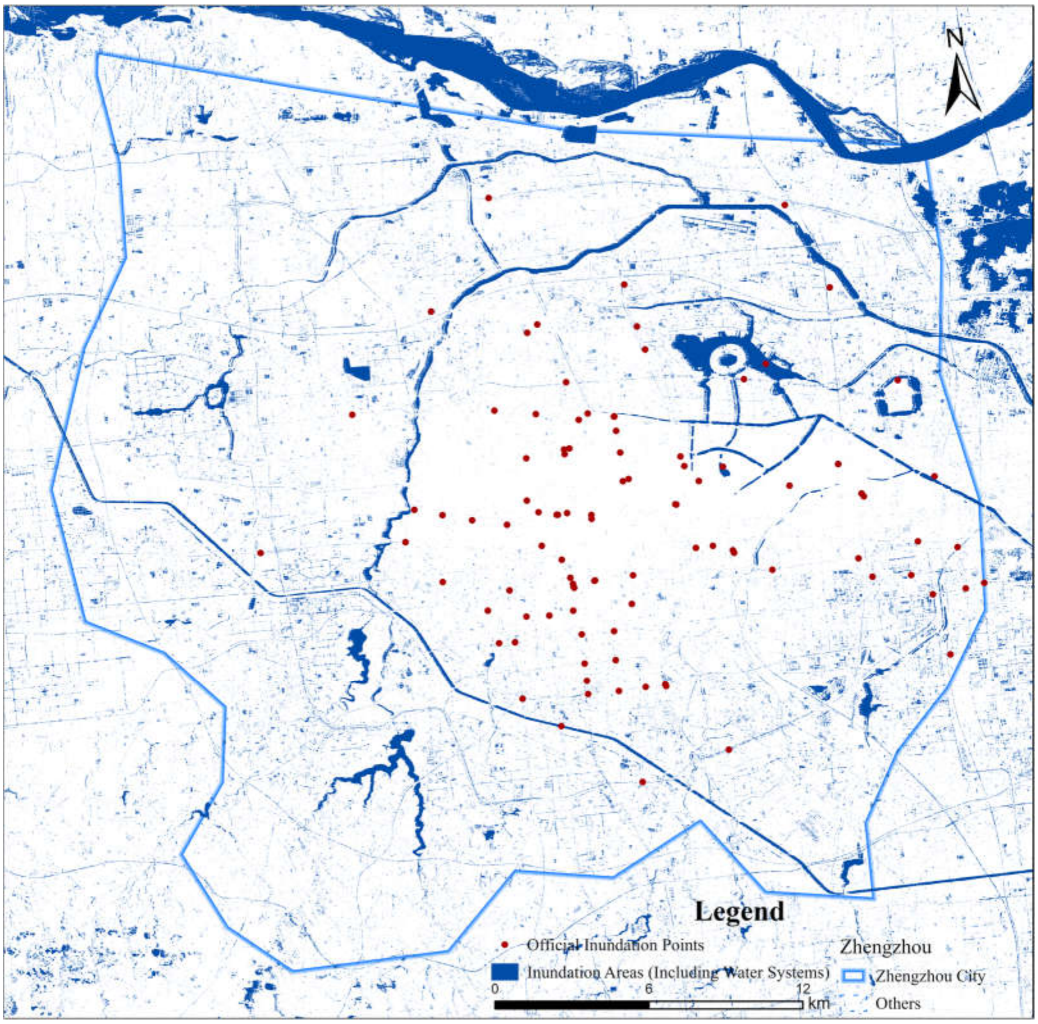

N
Legend
Official Inundation Points
Inundation Areas (Including Water Systems)
0
6
12
km
Zhengzhou
Zhengzhou City
Others

Figure 6. Officially Reported Inundation Points and Inundation Areas Spatial Distribution.

To address urban inundation, we selected 11 keywords highly related to the inundation event, including "inundation, water accumulation, flooding, water immersion, water invasion, water rise, water disaster, wash away, drainage, wading, water ingress," and scraped and preprocessed Weibo data from July 18 to July 20, 2021. After undergoing preprocessing steps such as character cleaning, applying classification models, and removing similar articles, the quality and relevance of the dataset to the target event were enhanced. We thus obtained a regularized dataset regarding the "July 20 Heavy rainstorm in Zhengzhou" event, laying a solid foundation for our subsequent coarse-grained and fine-grained spatio-temporal information extraction. Table 3 shows the statistical results of preprocessed Weibo data scraped during these three days.

Table 3. Statistics of the Pre-processed Dataset for the July 20 Heavy Rainstorm in Zhengzhou.

Type	Only text	With Text + Images（Video）	Total
Origin	12338	14222	26560
Text classify	6750	7886	14636
Data clean	1096	1951	3047
Space Filter	623	942	1565

From Table 3, we can see that the original inundation-related Weibo data has been preprocessed from 26,560 items to 3,047 regularized data entries. Additionally, it can be observed that data containing both text and images (or videos) always exceeds data with text only, further proving the richness of other modal data in social media. However, as Weibo is a social media platform facing all of China and even globally, we first need to carry out coarse-grained spatio-temporal information extraction and then further filter spatially to the scope of Zhengzhou city. Figure 7 presents the spatial distribution of inundation event points extracted with coarse granularity during these three days across all of China (a), and within the range of Zhengzhou city (b, c, d) on a daily basis.

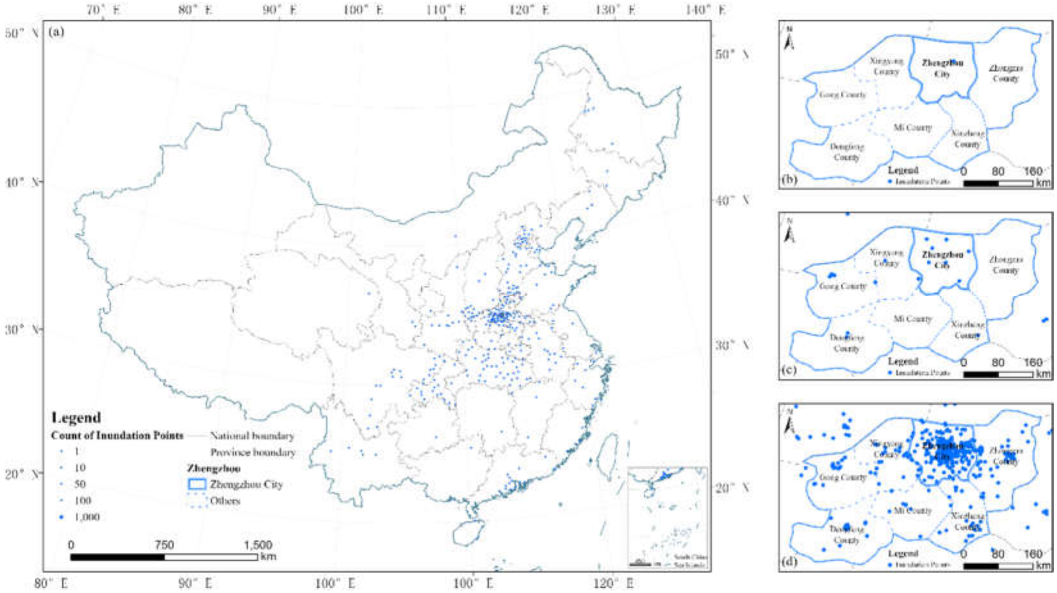


Figure 7. Spatial distribution of inundation events during the period from July 18 to July 20. (a)Coarse-grained inundation event points extracted from China during the period from July 18 to July 20. (b)Spatial distribution of inundation event points in Zhengzhou on July 18.(c) Spatial distribution of inundation event points in Zhengzhou on July 19.(d) Spatial distribution of inundation event points in Zhengzhou on July 20.

Subsequently, we have selected 23 pairs of high-quality Weibo image-text data, i.e., "Positive," from Weibo posts with coarse-grained spatial information extraction. The normalized address points with coarse-grained information but without high-quality, relevant images were deemed "Negative." Figure 8 illustrates the Positive and Negative points during the three-day period from July 18 to July 20, 2021, in the urban area of Zhengzhou city, along with several exemplary images corresponding to the Positive and Negative points.

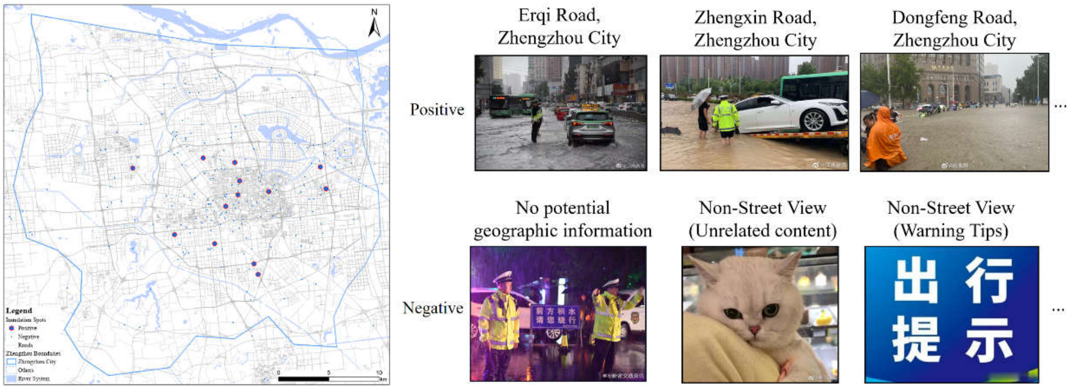

N
Legend
Inundation Spots
Positive
Negative
Roads
Zhengzhou Boundaries
Zhengzhou City
Others
River System
0
5
10
km
Erqi Road, Zhengzhou City
Zhengxin Road, Zhengzhou City
Dongfeng Road, Zhengzhou City
Positive
...
No potential geographic information
Non-Street View (Unrelated content)
Non-Street View (Warning Tips)
Negative
出 行
提 示
...

Figure 8. Spatial Distribution of Positive and Negative Points in Zhengzhou City and Corresponding Typical Images during July 18-20, 2021.

3.2 Experiments environment

The evaluation testing environment for MIST-SMMD runs on the deep learning frameworks of PyTorch 1.10.0 and TensorFlow 2.6, and performance assessment was completed on a Windows Server 2016 workstation equipped with a Tesla P40 GPU, a dual-route E5 2680 V2 CPU, and 64GB RAM.

3.3 Evaluation Metrics

To comprehensively evaluate the accuracy of the extracted event point spatial information and verify the advantages of multi-modal data, this study designed two methods for evaluation.

The accuracy of the standardized spatial information extracted at the coarse-grained level is based on the spatial distribution of flood inundation in the Zhengzhou city heavy rain and flood disaster as the benchmark dataset. When the spatial information exists in a submerged area within a specified nearby range, the spatial information is considered accurate. It should be noted that our method serves mainly as a supplement to traditional methods, so here we only evaluate the precision metric and do not involve recall.

The calculation formula for spatial precision is:

$$Spacial\ Precision = \frac{TP}{TP+FP}, \tag{8}$$

Herein, TP denotes the number of inundation areas present within the designated proximity of each coarse-grained inundation point extracted, while FP represents the number of cases in which no inundation areas exist within the specified proximity of each coarse-grained inundation point.

For the 23 coarse-grained spatial data points refined by fine-grained correction extracted in this case, we use the Space Error as an evaluation index to standardize different methods, due to the limited sample size. Furthermore, we extended two different indicators: the error between the spatial coordinates after fine-grained correction and the actual coordinates under limited sample evaluation, and the superiority of spatial information incorporating image modality compared to using textual unimodal spatial information alone.

$$\text{Space } Error = \sqrt{(Lon_{true} - Lon_{fin})^2 + (Lat_{true} - Lat_{fin})^2}\ , \tag{9}$$

$$MAE_{SE} = \frac{1}{n}\sum_{i=1}^{n} Space\ Error, \tag{10}$$

$$RMSE_{SE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Space\ Error)^2}, \tag{11}$$

In this, (Lon_{true}, Lat_{true}) 为 refers to the longitude and latitude coordinates of the official inundation points corresponding to the content of the Weibo posts mentioned above, (Lot_{fin}, Lat_{fin}) represents the longitude and latitude coordinates extracted by fine-grained methods, and n stands for the sample size.

4. Experimental Results and Analysis

4.1 Effectiveness Analysis

In the coarse-grained extraction stage of spatio-temporal information, we found that with the increase of the defined "nearby" range, the Spacial Precision of the extracted inundation event points also correspondingly improved (as shown in Figure 9). Particularly, we observed two Gradient Points (local maxima of the Spacial Precision curve, indicating rapid increases in Spacial Precision within a certain range). When the range expanded to 52m, the Spacial Precision reached 65.88%, and when it further expanded to 60m, the Spacial Precision rose to 87.54%. Ultimately, within a range of 201m, the Spacial Precision peaked at 100%. This signifies that our coarse-grained spatial information extraction method can cover most inundation areas within a relatively small range (for instance, 52m and 60m), demonstrating the effectiveness of coarse-grained spatial information extraction.

Additionally, during this stage of extraction, we employed a pre-trained Bert-base-Chinese model implemented by the spaCy library for Named Entity Recognition (NER). This model not only possesses the functionality required by our study (including event classification during the data preprocessing stage), but it also ranks first in terms of efficiency among common NLP tools[33], meeting our need to process massive amounts of Weibo data. Despite the efficiency of this model and technique, they may still be limited by inherent uncertainties, such as possible bias in extraction results for ambiguous or vaguely expressed text.

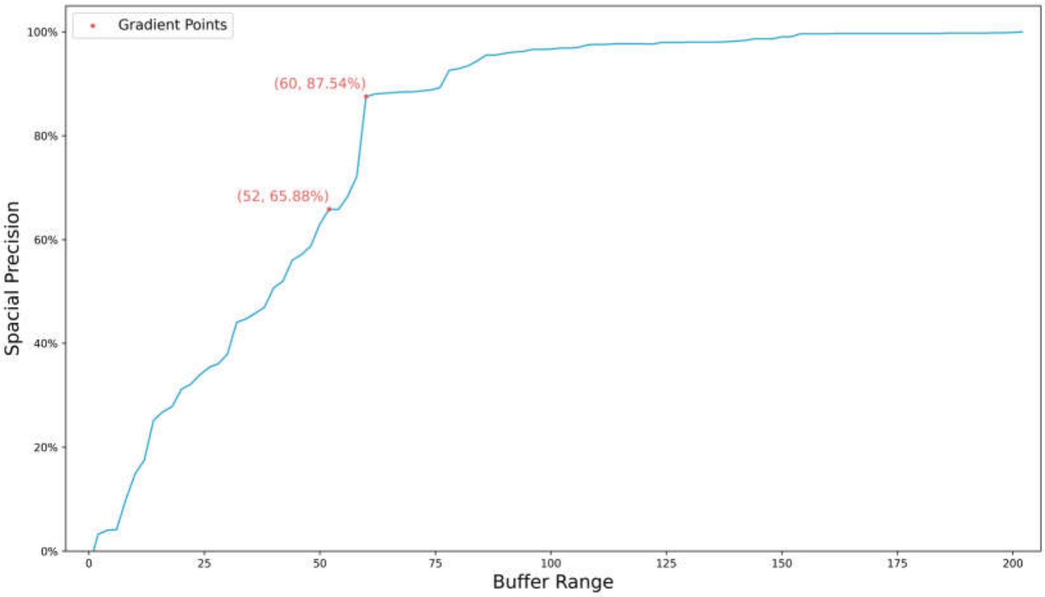


Figure 9. Spatial Precision of Coarse-grained Spatial Information Extraction within Different Buffer Ranges.

After conducting finer-grained spatial information extraction on 23 pairs of high-quality text and image data, it can be observed in Table 4 that MAE_{SE} and $RMSE_{SE}$ remain above 50, even in multi-modal scenarios (Text + Images). This is primarily influenced by a few instances of large error points. However, in actuality, it can be discerned from Figure 10 that the Space Error for the majority of the data points remains fairly low, around 20. This level of error is sufficient to meet the needs of many practical applications, such as in dealing with sudden urban disasters or events (such as floods, earthquakes, etc.), where real-time and precise spatial information is needed to guide rescue and disaster relief work. Against this backdrop, social media data provides a low-cost, wide-coverage data source that can effectively supplement traditional monitoring systems. Moreover, we also found that fine-grained extraction can significantly reduce spatial error compared to coarse-grained extraction methods based solely on text, improving the overall MAE_{SE} and $RMSE_{SE}$ by 95.53% and 93.62%, respectively, as shown in Table 4 (see Figure 10). These results validate that in the process of event point spatial information extraction, the use of images, videos, and other multi-modal data can effectively compensate for the spatial accuracy deficiency of a single modality, thereby enhancing the precision of spatial information. Although the number of text and image data pairs is relatively small, this does not imply that our method lacks value. Due to the spontaneity of social media data, the data points providing high-quality images are relatively few, but with the spread of social media and the development of the internet, we expect this situation to improve. Additionally, it is possible to reduce errors and omissions in manually filtering high-quality data by training a multi-modal fusion classification model for high-quality social media data and conducting large-scale collection on social media. However, there is still significant room for reducing the time cost of fine-grained spatial information extraction. Due to the significant variability in road lengths, the standardization of coarse-grained road types has led to considerable fluctuation in the number of street view sampling points, which further impacts the stability of the time cost. Conversely, for non-road type coarse-grained standardization addresses, setting a 200-meter buffer zone and generating sampling points every 5 meters results in an average of 635 sets of street view images per event point, totaling 2540 images. The total time to obtain and match each image averages 2.55 seconds, and thus, the total time averages 1 hour and 48 minutes. Furthermore, due to the API request restrictions of street view image data, the time to obtain images may increase if the model is used outside mainland China. It's worth noting that when dealing with multi-modal data, even though our method has categorized the text, there still remains the issue of determining whether the image is related to the standardized address. This could result in numerous irrelevant images being ineffectively matched with text, indirectly increasing the time cost. Additionally, from the data perspective, street view only covers fixed routes and not all places have street views. Meanwhile, the quality and shooting angle of user-uploaded images could also impact extraction results. Furthermore, during image matching analysis, accuracy might also be limited by the model's training data and its generalization capabilities.

Table 4. Comparison of Spatial Error between Coarse-grained Spatial Information Extraction with Text Modality Only and Multi-modal Data Integration for Refining Coarse-grained Spatial Information Extraction with Image Modality

Space Error	Only Text	Text + Images	Improvement
MAE_{SE}	1491.13	66.63	95.53%
$RMSE_{SE}$	2068.43	131.88	93.62%

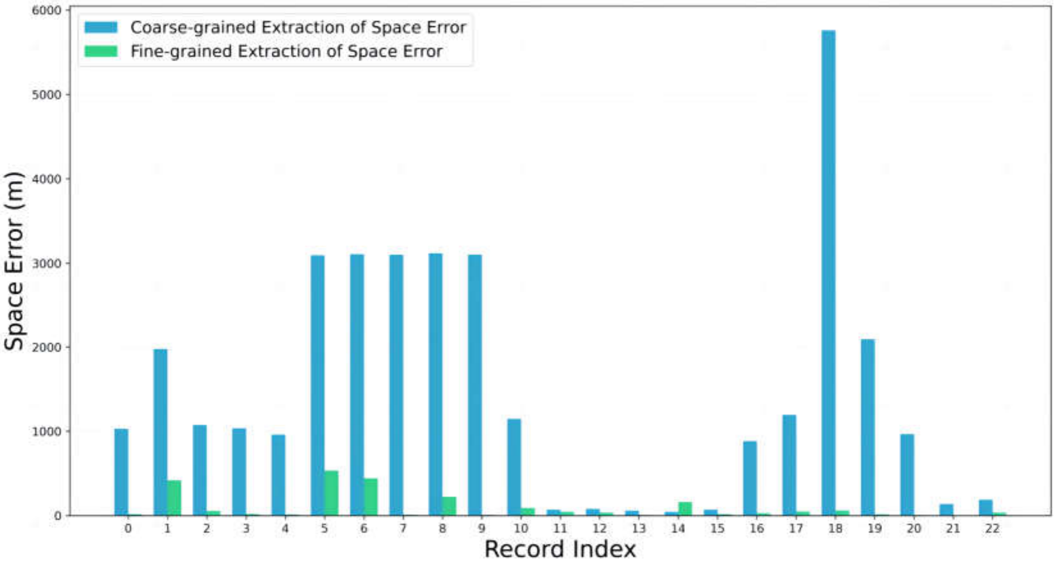


Figure 10. Comparison of Coarse and Fine Grained Extraction.

It must be emphasized that our method should serve as a supplementary measure. As previously stated, in the multi-modal data of social media, the volume of data that can simultaneously realize spatio-temporal standardization parsing and accurate image matching is not substantial. Therefore, the spatio-temporal information extracted from social media data can only serve as an effective supplement to traditional urban event monitoring and cannot fully replace conventional methods.

4.2 Analysis of Fine-grained extraction

Furthermore, during the fine-grained extraction, we conducted ablation experiments to compare the advantages of the LSGL model that introduced masks and matching algorithms when extracting fine-grained spatial information. Overall, as seen in Table 5, whether it's the MAE_{SE} or $RMSE_{SE}$ metrics, FM+SS performs the best, and the performance using only FM is the worst. This is not entirely consistent with our initial expectation that FM+SS+QIFM would be optimal. Although the MAE_{SE} of

FM+SS+QIFM is higher than FM+QIFM, its $RMSE_{SE}$ is lower. This suggests that among our 23 pairs of high-quality text and image data, FM+SS has the strongest overall performance, while FM+SS+QIFM's precision can only rank second, and its robustness third.

Table 5. Spatial Error of Fine-grained Spatial Information Extraction with Different Combinations of Quantitative Indicators for Feature Matching, Semantic Segmentation, and Feature Matching Degree.

Space Error	FM	FM+SS	FM+QIFM	FM+SS+QIFM
MAE_{SE}	124.30	66.63	110.33	100.74
$RMSE_{SE}$	227.35	131.88	179.42	181.16

* **FM:**Feature Matching; **SS**:Semantic Segmentation; **QIFM**:Quantitative Indicators for Feature Matching

As shown in Figure 11, we noticed that in some instances, the results of the four ablation experiment methods are consistent, and the Space Error is relatively low. This often occurs when the image resolution uploaded by social media users is high, and the images have a large number of semantic feature points, such as buildings, which are recognizable. This consistency reflects our method's dependency on high-quality images. Specifically, as shown in Figure12(a), user-uploaded high-quality images provide clearer, more detailed visual information and features such as buildings, which aids the LSGL in more accurately identifying and extracting event point spatial information. However, this also implies that in cases where the image quality is too low or lacks key façade feature information of buildings (such as Figure12(b)), this method may face challenges.

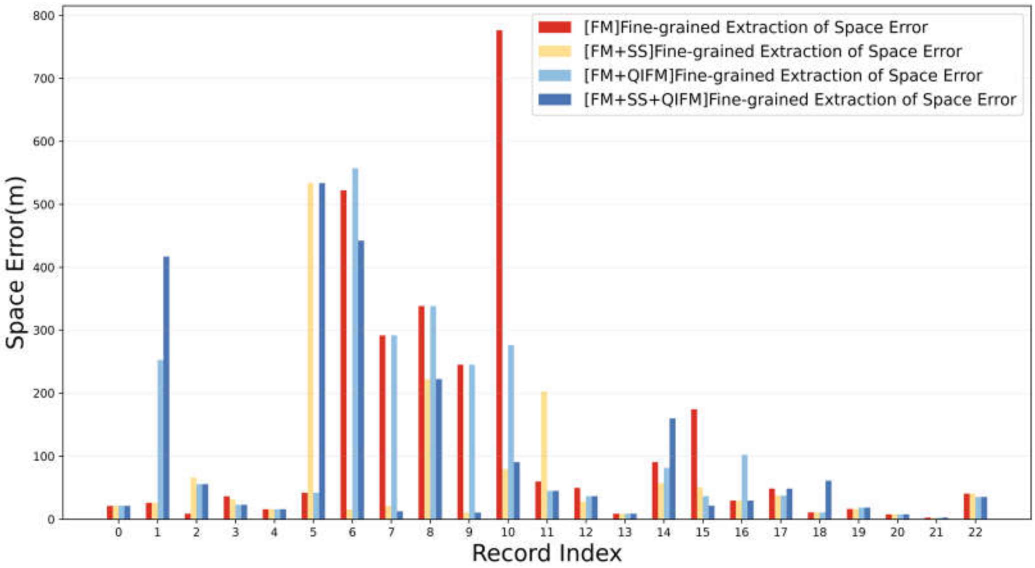


Figure 11. Ablation Experiments for Fine-grained Extraction.

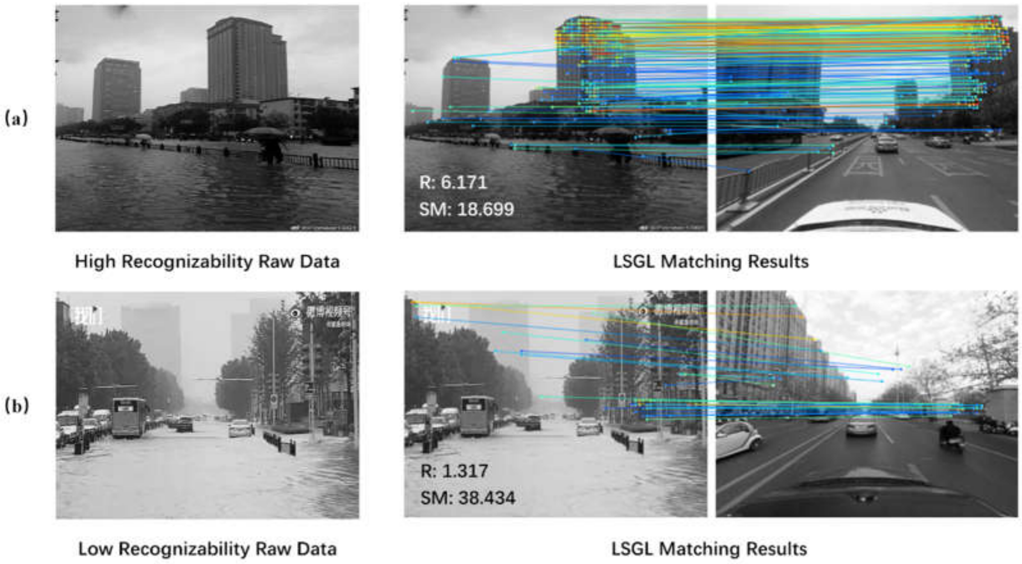
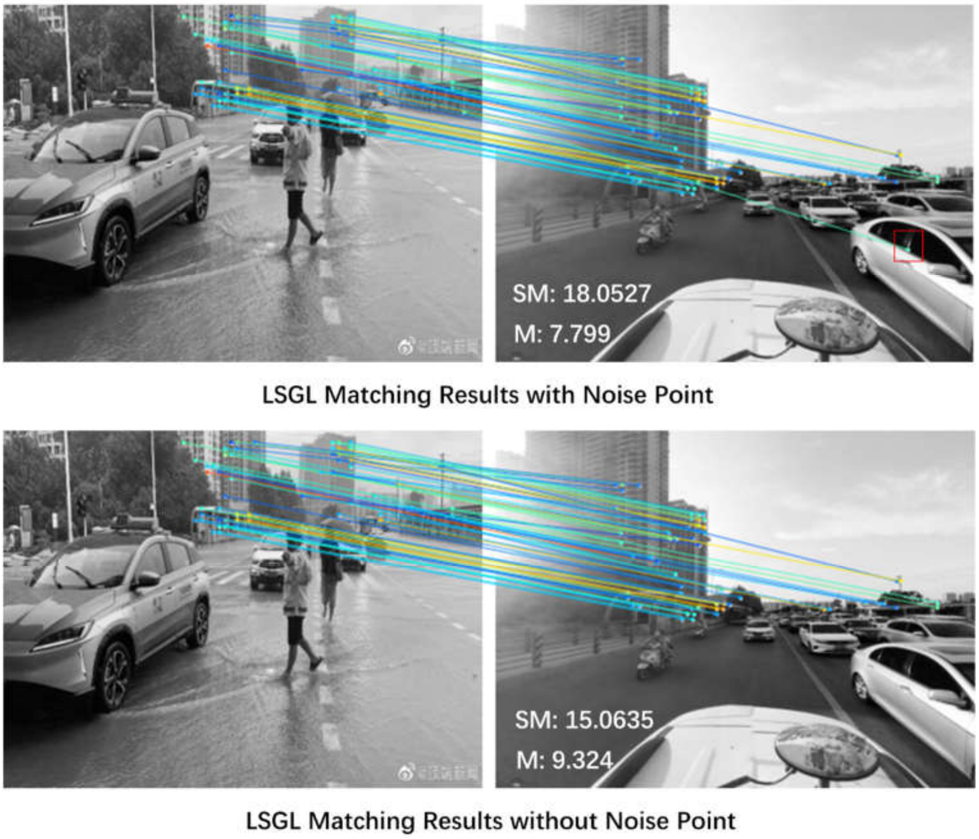


Figure 12. Matching results of LSGL for images with different levels of recognizability

We found that introducing SS and QIFM can significantly enhance the performance of LSGL in practice. This is mainly because SS can effectively filter out irrelevant background information, and QIFM can provide more intuitive and precise spatial distance metrics. The combination of these two methods enables LSGL to more accurately locate inundation event points, thereby enhancing overall spatial precision. However, sometimes we also observe that introducing QIFM makes the results worse. This is mainly due to the limited performance of the semantic segmentation model used for masking, leading to less accurate masks. In this situation, an imprecise mask may not completely eliminate features like cars and pedestrians that impact street view matching (such as in Figure13). The matching results of these noise points usually bring abnormal Euclidean distance and Euclidean direction, interfering with the computation of spatial location difference.



LSGL Matching Results with Noise Point

LSGL Matching Results without Noise Point

Figure 13. The effect of noise points on LSGL matching results

Also, the situation with a record index of 14 in Figure 11 is somewhat unique, and it is for the above reasons that introducing QIFM makes the result worse. Simultaneously, the coarse-grained spatial error is relatively low, making further fine-grained extraction decrease the Space Error. This indicates that while SS and QIFM can improve results in most cases, we also need to be aware of their possible limitations and challenges.

Finally, we also encountered situations where the performance of all methods was not satisfactory. This usually occurs when distant buildings are affected by fog, and distinct feature points cannot be extracted. In fact, the previously mentioned lack of high-quality coarse-grained standardized addresses is why this study needs to semi-automatically screen high-quality coarse-grained standardized addresses. This situation highlights the challenges of LSGL fine-grained spatial information extraction when dealing with non-high-quality coarse-grained standardized addresses.

5. Conclusions

This study presents an innovative method - the MIST-SMMD method, which can extract spatio-temporal information of urban events from coarse to fine-grained through hierarchical processing. Leveraging the advantages of multi-modal data, our research reveals the enormous potential of social media data (especially Weibo) as a source for acquiring dynamic, high-precision information on urban events.

Our method can be broadly applied to the field of urban disaster management, and it also has potential in other areas where real-time and precise spatial information is required. For example, in the monitoring and management of traffic congestion and accidents, since not all road sections are equipped with monitoring equipment, our method can provide on-site spatio-tempinal information about traffic congestion or real-time situations based on real-time information on social media. This could help traffic management departments adjust signal light settings in a timely manner or dispatch rescue vehicles promptly. Moreover, picture and video data on social media have potential utility value, for example, to extract the severity of events, or for data archiving, temporal tracking, and further in-depth analysis of the same event at different time points.

Future research could explore more potential directions and improvement strategies, including adopting more advanced models to enhance the accuracy of urban event classification and named entity extraction, more comprehensively integrating unutilized information in social media, and introducing more other types of data sources to enhance the robustness of data extraction and analysis. Furthermore, we believe that the real-time extraction and processing of event information from multi-modal social media data has significant potential for urban emergency systems. It could contribute to more efficient and timely urban management, command, and disaster reduction work.

Author Contributions: Conceptualization, Yilong Wu and Yingjie Chen;methodology, Yilong Wu,Yingjie Chen and Rongyu Zhang;software,Rongyu Zhang,Yilong Wu,Yingjie Chen and Zhenfei Cui;evaluation,Yilong Wu and Yingjie Chen;data annotation,Xinyi Liu;writing—original draft preparation,Yilong Wu,Jiayi Zhang and Xinyi Liu;writing—review & editing,Yong Wu and Yilong Wu.All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fujian Normal University College Students' Innovative Entrepreneurial Training Projects Funded in 2023, grant number cxxl-2023292" .

Data citation: [dataset] Lu el. 2022. Spatial distribution dataset of flood inundation in Zhengzhou City, Henan Province, July 2021 by heavy rainfall and flooding; ChinaGEOSS Data Sharing Network; 2017YFB0504100.

Acknowledgments: Thanks to Xiaochen Qin for his constructive suggestions and Shuying Luo for her art support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ryan, T.; Allen, K.A.; Gray, D.L.; McInerney, D.M. How social are social media? A review of online social behaviour and connectedness. Journal of Relationships Research 2017, 8, e8.
2. Weibo Reports Fourth Quarter and Fiscal Year 2022 Unaudited Financial Results. Available online: http://ir.weibo.com/node/8856/pdf (accessed on 15 May 15, 2023).
3. Zhang Z.Spatial analysis of Internet sensation based on social meadia—Taking the Jiuzhaigou earthquake as an example.NanJing University, 2019.
4. Li S.,Zhao F.,Zhou Y.Analysis of public opinion and disaster loss estimates from typhoonsbased on Microblog data. Ch'ing-hua Ta Hsueh Hsueh Pao, Tzu Jan K'o Hsueh Pan J. Tsinghua Univ., Sci. Technol, 2022, 62(01),pp:43-51.
5. Wu Q.,Qiu Y. Effectiveness Analysis of Typhoon Disaster Reflected by Microblog Data Location Information. J. Geomat.Sci. Technol. 2019,36(04),pp:406-411.
6. Liang C.,Lin G.,Zhang M.,Assessing the Effectiveness of Social Media Data in Mapping the Distribution of Typhoon Disasters. J Geogr Inf Sci,2018,20(06),pp:807-816.
7. Yu, M.; Bambacus, M.; Cervone, G.; Clarke, K.; Duffy, D.; Huang, Q.; Li, J.; Li, W.; Li, Z.; Liu, Q. spatio-temporal event detection: A review. International Journal of Digital Earth 2020, 13, 1339-1365.
8. Etzioni, O.; Cafarella, M.; Downey, D.; Kok, S.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D.S.; Yates, A. Web-scale information extraction in knowitall: (preliminary results). In Proceedings of the Proceedings of the 13th international conference on World Wide Web, 2004; pp. 100-110.
9. Ritter, A.; Etzioni, O.; Clark, S. Open domain event extraction from twitter. In Proceedings of the Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012; pp. 1104-1112.
10. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 2015.
11. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018
12. Ma, K.; Tan, Y.; Tian, M.; Xie, X.; Qiu, Q.; Li, S.; Wang, X. Extraction of temporal information from social media messages using the BERT model. Earth Science Informatics 2022, 15, 573-584.
13. Yuan, W.; Yang, L.; Yang, Q.; Sheng, Y.; Wang, Z. Extracting Spatio-Temporal Information from Chinese Archaeological Site Text. ISPRS International Journal of Geo-Information 2022, 11, 175.
14. MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., & Blanford, J. (2011). SensePlace2: GeoTwitter analytics support for situational awareness. In VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings ,pp:181-190.

15. Zou, Z.; Gan, H.; Huang, Q.; Cai, T.; Cao, K. Disaster image classification by fusing multimodal social media data. ISPRS International Journal of Geo-Information 2021, 10, 636.
16. Ofli, F.; Alam, F.; Imran, M. Analysis of social media data using multimodal deep learning for disaster response. arXiv preprint arXiv:2004.11838 2020.
17. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 2018, 41, 423-443.
18. Shuai X.,Hu S.,Liu Ql. Internet media-based acquisition and processing model of earthquake disaster situation.J. Nat. Disasters,2013,22(3),pp:178-184.
19. Zhang S.,Yang Z.,Wang Y. Simulation on Flood Disaster in Urban Building Complex System Based on LBM. J Simul,2022,34(12),pp:2584-2594.11. Yuan, F.; Xu, Y.; Li, Q.; Mostafavi, A. Spatio-temporal graph convolutional networks for road network inundation status prediction during urban flooding. Computers, Environment and Urban Systems 2022, 97, 101870.
20. Faxi Yuan, Yuanchang Xu, Qingchun Li, Ali Mostafavi.Spatio-Temporal Graph Convolutional Networks for Road Network Inundation Status Prediction during Urban Flooding.Comput Environ Urban Syst,2022,Volume 97, Article 102289.
21. Xu, L.; Ma, A. Coarse-to-fine waterlogging probability assessment based on remote sensing image and social media data. Geo-spatial Information Science 2021, 24, 279-301.
22. Panteras, G.; Cervone, G. Enhancing the temporal resolution of satellite-based flood extent generation using crowdsourced data for disaster monitoring. International journal of remote sensing 2018, 39, 1459-1474.
23. Zhang Z.,Wang Z.,Fang D.Optimal Design of Urban Waterlogging Monitoring and WarningSystem in Wuhan Based on Internet of Things and GPRS Technology. Saf Environ Eng,2018,25(02),pp:37-43.
24. Zeng Z.,Xv J.,Wang Y.Advances in flood risk identification and dynamic modelling based on remote sensing spatial information. Adv Water Sci, 2020,31(03),pp:463-472.[27]Wang, R.-Q.; Mao, H.; Wang, Y.; Rae, C.; Shaw, W. Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. Computers & Geosciences 2018, 111, 139-147.
25. Songchon, C.; Wright, G.; Beevers, L. Quality assessment of crowdsourced social media data for urban flood management. Computers, Environment and Urban Systems 2021, 90, 101690.
26. BLE, SOCIAL MEDIA & FLOOD RISK AWARENESS . Available online: https://www.fema.gov/sites/default/files/documents/fema_ble-social-media-flood-risk-awareness.pdf (accessed on 15 May 15, 2023).
27. Songchon Chanin,Wright Grant,Beevers Lindsay. Quality assessment of crowdsourced social media data for urban flood management. Comput Environ Urban Syst,2021,Volume 90, pp: 101690.
28. Wang, X.; Kondratyuk, D.; Christiansen, E.; Kitani, K.M.; Alon, Y.; Eban, E. Wisdom of committees: An overlooked approach to faster and more accurate models. arXiv preprint arXiv:2012.01988 2020.
29. JioNLP. Available online: https://github.com/dongrixinyu/JioNLP (accessed on 15 May 15, 2023).
30. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-free local feature matching with transformers. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021; pp. 8922-8931.
31. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, 2020; pp. 213-229.
32. Schmitt, X.; Kubler, S.; Robert, J.; Papadakis, M.; LeTraon, Y. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2019; pp. 338-343.