Computers, Environment and Urban Systems 126 (2026) 102401

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Multi-frequency street-level urban noise modeling and mapping through street view and remote sensing image fusion

Yan Zhang [a], Entong Ke [b], Mei-Po Kwan [a,c,*], Libo Fang [d], Mingxiao Li [e,f]

[a] Institute of Space and Earth Information Science, Fok Ying Tung Remote Sensing Science Building, The Chinese University of Hong Kong, Hong Kong SAR, China
[b] School of Resource and Environmental Sciences, Wuhan University, Wuhan, China
[c] Department of Geography and Resource Management, Wong Foo Yuan Building, The Chinese University of Hong Kong, Hong Kong SAR, China
[d] Hunan Architectural Design Institute Group Co. Ltd, Changsha, China
[e] Guangdong Key Laboratory of Urban Informatics, College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, China
[f] Key Laboratory for Geo-Environmental Monitoring of Great Bay Area MNR, Shenzhen, China

ARTICLE INFO

Keywords:
Noise exposure
Information fusion
Street view image
Noise modeling
Sensing bias

ABSTRACT

Urban noise pollution has become the third most significant environmental health threat following air and water pollution, while traditional noise modeling methods suffer from limitations including high costs, limited coverage, and an exclusive focus on total decibel values while neglecting frequency characteristics. This study proposes a method that combines street view imagery (SVI) and remote sensing imagery (RSI) to achieve precise modeling and mapping of multi-frequency noise exposure at the urban street scale. Using Xiangzhou District, Zhuhai City as a case study, we utilized approximately 6000 street view images and corresponding remote sensing images, and recorded 35,276 street noise audios containing 23 frequency bands (100 Hz-16,000 Hz) through volunteer cycling surveys. A multi-source fusion model was constructed based on a pre-trained vision transformer architecture, with 923 valid street noise-image paired samples used for training and validation. The sensitivity results demonstrate that: (1) the proposed multimodal fusion model achieves high predictive accuracy, with R^2 values for dBA prediction ranging from 0.417 to 0.649, with particularly higher accuracy observed for mid-frequency noise prediction; (2) 50-m resolution street-scale multi-frequency soundscape maps were successfully generated, providing scientific evidence for refined urban noise management; (3) explainable machine learning models revealed that buildings, roads, sidewalks, and terrain visual elements are the four most important factors affecting noise prediction, with road width showing a positive association with street noise levels. This study not only fills the gap in urban noise frequency characteristics research but also provides new methodological support for precise street-level noise pollution modeling and health-oriented urban planning. The source code is available at https://github.com/giserzy/NoisePrediction.

1. Introduction

Noise pollution represents the third most significant health threat in modern urban environments, following air and water pollution, making low-cost, large-scale noise modeling of paramount importance in environmental health research (Basner et al., 2014; Chen et al., 2024). The rapid urbanization process has rendered noise one of the most inescapable environmental stressors in daily urban exposure. Extensive epidemiological studies demonstrate that prolonged exposure to high-noise environments significantly increases the risk of cardiovascular diseases, sleep disorders, and psychological stress (Chang et al., 2025; Münzel et al., 2018; Song et al., 2025). Fine-scale noise exposure modeling constitutes a critical scientific challenge for precision governance in healthy city research.

Given the paramount importance of noise exposure modeling, three primary approaches have emerged for large-scale noise environment assessment. The first approach involves deploying fixed sensors (DY Wu et al., 2025) or mobile sensors (Song et al., 2024) for noise modeling. For instance, Cai et al. conducted extensive field surveys using professional GPS equipment to track participants (Cai et al., 2023; Wang et al., 2025).

* Corresponding author at: Institute of Space and Earth Information Science, Fok Ying Tung Remote Sensing Science Building, The Chinese University of Hong Kong, Hong Kong SAR, China

E-mail addresses: yanzhang@cuhk.edu.hk (Y. Zhang), entongke@whu.edu.cn (E. Ke), mpkwan@cuhk.edu.hk (M.-P. Kwan), fanglb@hnadi.com.cn (L. Fang), limx@szu.edu.cn (M. Li).

https://doi.org/10.1016/j.compenvurbsys.2026.102401
Received 13 August 2025; Received in revised form 4 December 2025; Accepted 12 January 2026
Available online 23 January 2026

However, this approach requires substantial costs and presents significant challenges for large-scale implementation. The second approach inputs static variables such as land use and built environment data, employing machine learning methods to learn and infer noise environments (Alvares-Sanches et al., 2021; Lan et al., 2024; Liu et al., 2021). The limitation of this modeling method lies in its reliance on relatively fixed land use types and similar static data (Chew & Wu, 2016; Zhou et al., 2024), which inadequately capture dynamic urban elements and struggle to generalize across complex and diverse urban environments. The third approach employs spatiotemporal big data for indirect estimation, including location-based noise complaints (Zhang et al., 2024; Jiao et al., 2025; H Wu et al., 2025) and street view imagery (Huang et al., 2024; Zhao et al., 2023). These methods effectively address both accuracy and coverage challenges. However, the former primarily reflects residents' subjective perceptions and awareness of rights protection, with complaint data often exhibiting uneven spatial coverage and potential data sparsity in areas with fewer residential units.

Street view imagery, while possessing advantages of widespread public accessibility, rich image content, and high spatial coverage (Biljecki & Ito, 2021; Zhang et al., 2021; Zhao et al., 2025), has been extensively used in large amount of city studies. Nevertheless, temporal variations in street view sampling can be affected by random traffic and pedestrian flow states, making noise inference susceptible to "temporal bias" interference and potentially yielding biased conclusions (Hou et al., 2025; Zhang & Kwan, 2025). For example, identical streets sampled during peak hours versus off-peak periods exhibit significant differences, with vehicle and pedestrian pixel ratios being particularly elevated during the morning and evening rush hours. This substantial variation in road features (Huang et al., 2025; Ma et al., 2023) can prevent model prediction convergence and even produce contradictory results. In contrast, remote sensing imagery serves as a complementary mapping to street view imagery, providing a stable bird's-eye perspective of built environments. This approach emphasizes overall environmental features such as buildings and vegetation rather than dynamic elements like pedestrians and vehicles (Zhang et al., 2025), offering crucial supplementation to the human-centered perspective of street view imagery. However, remote sensing imagery encounters limitations in detecting socioeconomic environments, where shopping centers, office buildings, and residential areas may appear morphologically similar while exhibiting different noise levels.

In addition, although existing research has proposed lots of methods for noise modeling (Huang et al., 2024; Xue et al., 2024; Zhang et al., 2023). However, comparing to professional equipment measurements, current modellings can only assess noise intensity (decibel values). But urban soundscapes are considerably more complex and diverse, extending beyond the single dimension of "loudness," as sound waves of different frequencies exert varying physiological and behavioral effects on humans. To illustrate with an intuitive cycling experience: when a heavy truck passes on the road, its impact on us far exceeds that of equally loud ambient noise. Despite identical decibel values, heavy vehicle noise concentrates primarily in low-frequency (20-200 Hz) and mid-frequency (200-2000 Hz) ranges. This type of noise possesses extraordinary penetrating power, capable of traveling considerable distances and remaining perceptible even indoors. Truck noise induces pronounced physiological discomfort, including chest tightness, palpitations, and even building vibrations. Conversely, passenger cars and pedestrian conversations primarily occupy mid-to-high frequency ranges (2000-20,000 Hz). Additionally, streets frequently contain high-frequency noise sources such as sudden braking, tire skidding, electric vehicle high-frequency motor sounds, and various warning signals. These diverse frequency components collectively constitute the complex urban soundscape, making accurate identification of multi-spectral acoustic environments critically important.

Confronting two key challenges in urban noise modeling (as shown in Fig. 1)- first, the inherent limitations of different-source data in noise perception—satellite imagery provides macro-level information on urban morphology, land use, and vegetation coverage, while street view imagery captures micro-level details such as street design, building conditions, and social order indicators. Second, traditional methods primarily focus on noise intensity (decibel values, A-weighted Sound Pressure Level, A-Level) while neglecting frequency characteristics.

This study proposes an innovative multisource fusion approach that combines street view imagery (SVI) and remote sensing imagery, employing machine learning technology to model and map urban multi-frequency real noise exposure conditions. We collected 6000 street view images within the study area (Xiangzhou District, Zhuhai City, China) along with corresponding 250 × 250-m resolution remote sensing imagery (Cao et al., 2023; Xing et al., 2023), complemented by field noise

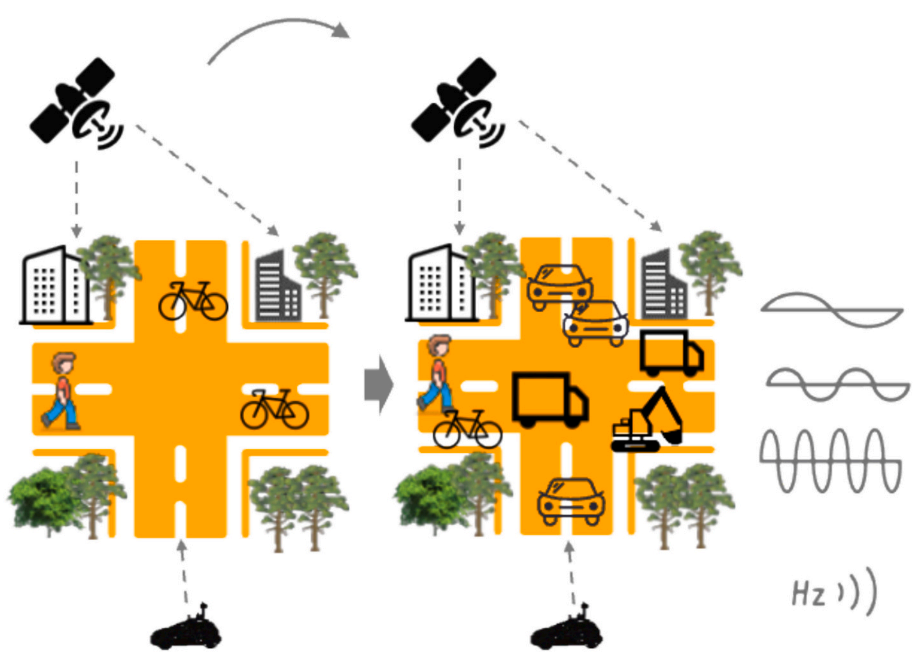


Fig. 1. Diagram illustrating the research gap. It highlights that remote sensing captures static built features, street view focuses on dynamic street-level characteristics, and different noise frequencies represent distinct types of sound sources.

exposure data (including low-frequency, mid-frequency, and high-frequency components) collected through volunteer cycling surveys. Based on a pre-trained vision transformer model and machine learning model, we designed a hybrid model based on multisource deep features to achieve street-level soundscape map prediction.

2. Method and research materials

The research methodology of this study is illustrated in Fig. 2. We collected real noise exposure data through field cycling along both sides of roads, trained models using corresponding visual data, including street view images and remote sensing images, conducted predictions and validation based on the trained noise models, and employed interpretable machine learning methods to explain model decisions.

2.1. Data collection and study area

The study area is located in Xiangzhou District, the main urban area of Zhuhai City in Guangdong Province in China. Considering that traffic noise constitutes the primary source of urban noise pollution in the study area (Chu et al., 2025), we recruited volunteers to cycle along traffic corridors during daytime hours and collected georeferenced noise measurements using calibrated smartphone applications. A total of 35,276 noise records with geographic coordinates (positional accuracy of within 15 m) were collected during June 26–27, 2025, covering different categories of urban districts. To maintain temporal consistency in noise environmental conditions, all field measurements were conducted during off-peak traffic periods. We employed uniform smartphone models (Redmi Note 11 Pro) that were calibrated against professional noise measurement equipment prior to deployment. The sampling rate was set at one measurement per second.

The collected noise measurements encompassed 23 frequency bands, as well as the aggregated average sound pressure level across all bands (unit: dBA). These bands ranged from 100 Hz to 16,000 Hz (Alvares-Sanches et al., 2021; Guo et al., 2022), which were categorized into three frequency ranges: low-frequency noise (20-200 Hz), mid-frequency noise (200-2000 Hz), and high-frequency noise (2000-20,000 Hz). All data collection was conducted during daytime hours to ensure consistency in urban activity patterns and traffic conditions. Street view imagery serves as an important representation of urban built environments (Zhang et al., 2023; Zhang et al., 2024). We systematically collected 5953 street view images (1024*384 pixels) along roadways within the study area at 50-m sampling intervals. Corresponding satellite imagery (30-m resolution visible light imagery) covering 250 × 250-m areas (Xing et al., 2023) was acquired to provide overhead perspectives of the urban built environment.

Street view imagery was acquired from Baidu Maps, representing the most recent available imagery (captured after 2020, with exact timing dependent on the service provider's update schedule). Since the study area of Zhuhai City is situated in a subtropical climate zone with evergreen vegetation throughout the year, street-level environmental characteristics such as green view index remain relatively constant across seasons. Consequently, the seasonal variation in image acquisition time does not compromise our analytical findings (Han et al., 2023). Remote sensing satellite imagery was obtained from Tianditu hybrid satellite imagery, with a spatial resolution and temporal coverage matching those reported in Xing et al. (2023).

Finally, we performed spatial matching between street view images and noise measurements, selecting street view images located within 15 m of noise recording points to establish paired datasets linking noise exposure with built environment characteristics. At last, we built one dataset including 923 training samples (923 SVI and matched noise records) with accurate correspondences between noise exposure records and geospatial images.

Fig. 3 shows two of our cycling trajectories (Route A and Route B) along with corresponding noise measurement variations (panels A and B on the right). To mitigate the influence of extreme values, we applied smoothing procedures to all collected noise records. As shown in panels A and B in Fig. 3, we plotted the smoothed variation curves for high-frequency, low-frequency, and mid-frequency noise components corresponding to both trajectories. It is important to note that these two trajectories represent only a subset of our complete data collection routes. During the data acquisition process, we deliberately covered a wide variety of street types to ensure sample diversity and representativeness.

2.2. Assessment of the noise sound pressure level

The physical basis of noise is pressure variations caused by sound waves. Sound pressure level is calculated using sound pressure (unit: Pascal, Pa) through a logarithmic formula, with the unit being decibels (dB).

$$L_p = 20 \cdot log_{10}\left(\frac{p}{p_{ref}}\right) \tag{1}$$

where p is the measured sound pressure, and pref is the reference sound pressure for human hearing threshold (20 μPa).

Ground truth noise data were collected through volunteer cycling surveys. This approach enables mobile data collection across diverse street environments. The measurement equipment captured noise levels across 23 different frequency bands (ranging from 100 Hz to 16,000 Hz). Sound pressure levels are calculated using third-octave band analysis. For instance, Leq100 denotes the equivalent continuous sound level within the third-octave band centered at 100 Hz, spanning approximately from 89.1 Hz (lower band edge) to 112.2 Hz (upper band edge). It should be noted that the frequency band divisions in this study follow a logarithmic scale rather than a linear scale. This logarithmic spacing aligns with the human auditory system's perceptual characteristics, which exhibit greater sensitivity to frequency changes at lower frequencies and progressively diminished sensitivity at higher frequencies. In our analysis, these bands were aggregated into three main categories based on their distinctive characteristics and health implications.

However, human ears exhibit varying sensitivity to different frequency sounds. For instance, we are most sensitive to mid-frequencies (such as human speech), while being less sensitive to very low frequencies (such as idling heavy trucks) and very high frequencies (such as electrical hum). A-weighting is an internationally standardized frequency weighting curve that simulates the auditory characteristics of human ears to low-to-moderate intensity sounds. When calculating final noise values, measurement instruments first analyze all frequency components contained in the noise (i.e., the “multi-frequency noise” mentioned), then apply gains or attenuations to different frequency sound pressure levels according to the A-weighting curve (weights of low-frequency and high-frequency components are reduced), and finally sum the energy across all frequencies to obtain a single dBA value. During the data collection process, we also took into account a series of interferences, such as wind noise. To avoid these interferences, we cycled at a relatively slow speed. Statistical data (Table 1) shows that the average and median speeds of our cycling were approximately 3 m/s, which is a rather low speed for normal cycling. At this speed, the impact of wind noise was reduced to an extremely low level. Moreover, during the two days of data collection, the weather was almost windless. The statistical summary of noise data is presented as follows.

2.3. Multi-source urban street-scale multi-spectral noise modeling method

This study proposes a multi-source data-based urban street-scale multi-spectral noise modeling method, aiming to achieve precise prediction and spatial mapping of urban low-frequency, mid-frequency, and high-frequency noise through the fusion of complementary geospatial image features. For 923 noise sampling records in Xiangzhou

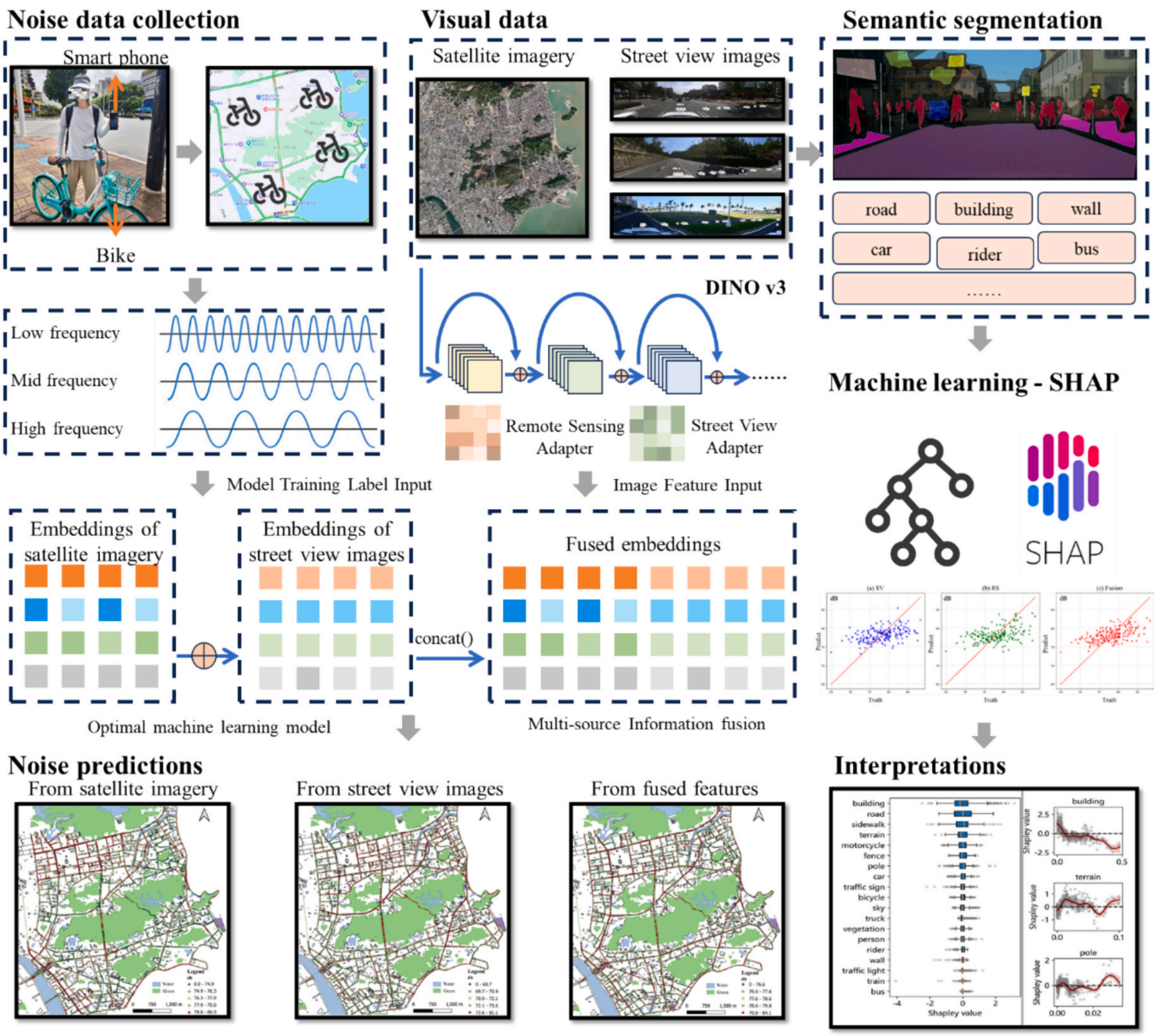


Fig. 2. Methodology flowchart of this study (Vision transformer and semantic segmentation were respectively applied to street view images for feature extraction, with the latter primarily used for interpretable machine learning due to its consistency with real-world scene elements).

District, Zhuhai City, we respectively processed high-resolution street view panoramic images (SVI, from Baidu Maps) and remote sensing images (RSI, from Google Maps). Street view panoramic images were acquired to provide a microscopic, human-centered perspective of urban landscapes. High-resolution satellite images were collected centered on each sampling point, providing a macroscopic bird's-eye view of urban morphology.

To transform this rich visual information into quantifiable noise modeling data, we employed the latest Vision-Transformer model as the backbone network for high-level feature extraction. We choose the pre-trained DINOv3[1] models provided by Meta to deal with the street-view and remote sensing imagery. Because it was not only pretrained on the aerial images but also on the Open World Image dataset. For street-view images, we employed the *dinov3-vitb16-pretrain-lvd1689m* model, which was trained on the Web dataset (LVD-1689 M)-a curated collection of 1.689 billion images extracted from a pool of 17 billion web images sourced from public Instagram posts. For remote sensing imagery, we utilized the *dinov3-vitl16-pretrain-sat493m* model, trained on the Satellite dataset (SAT-493 M) comprising 493 million 512 × 512 pixel images randomly sampled from Maxar RGB orthorectified imagery at 0.6-m resolution. These models generated 768-dimensional and 1024-dimensional feature embeddings, respectively, where each embedding corresponds to the Classification (CLS) token output, which serves as a global patch tokens representation of the input image.

This pre-training process endows the model with a rich, generalized understanding of visual patterns, from basic elements such as edges, textures, and colors in initial layers to more complex object parts and shapes in intermediate layers. Through transfer learning, we leverage this pre-existing knowledge and adapt it to our specific task of urban noise modeling. The image CLS token captures high-level semantic information about the input image while maintaining spatial invariance.

We input urban images into the network and forward-propagate through its vision transformer layers. For each input image, this process generates a high-dimensional vector ($\mathbf{F} \in \mathbb{R}^{768}; \in \mathbb{R}^{1024}$) as a dense semantic representation of the image content. The feature extraction process can be formalized as:

$$F_{SVI} = VIT(I_{SVI}) \tag{2}$$

$$F_{RSI} = VIT(I_{RSI}) \tag{3}$$

whereI_{SVI} *and* I_{RSI} represent the street view and remote sensing images, respectively.

The multi-source feature fusion is performed through concatenation:

$$F_{combined} = [F_{SVI}; F_{RSI}] \in \mathbb{R}^{1792} \tag{4}$$

[1] https://huggingface.co/facebook/dinov3-vit7b16-pretrain-lvd1689m

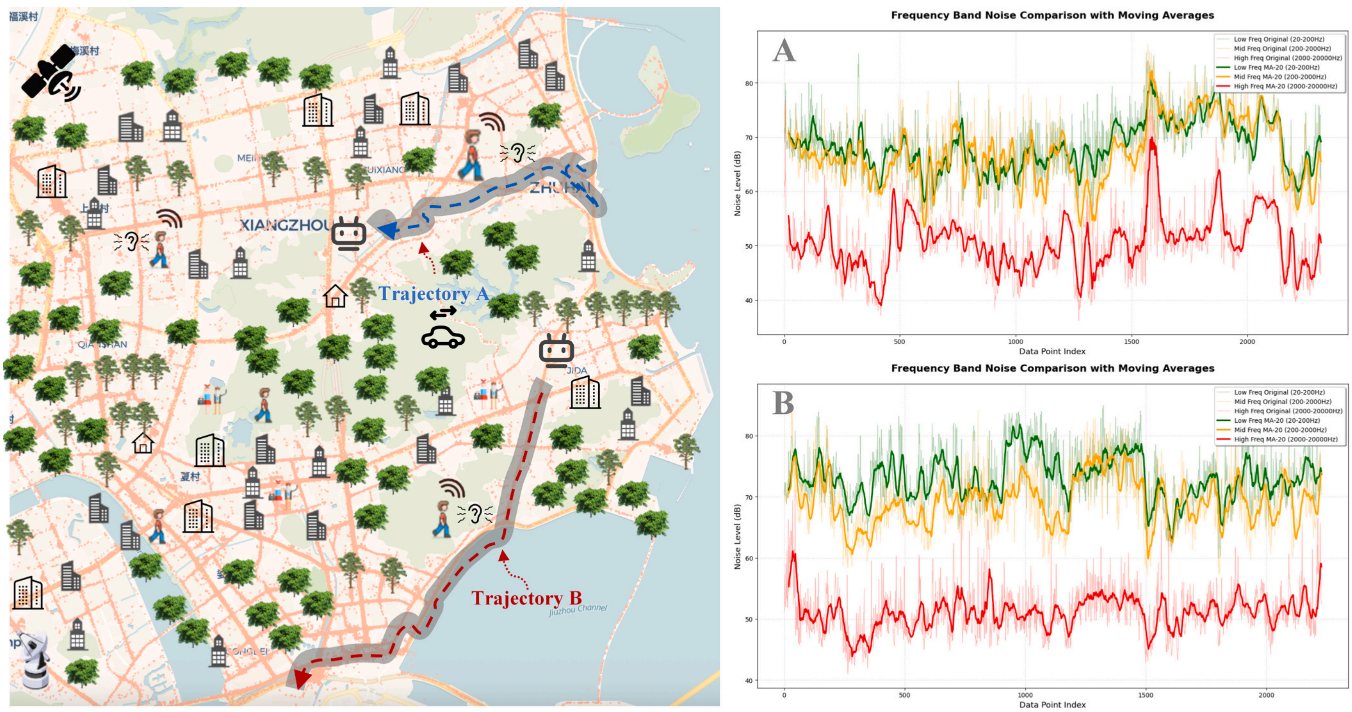


Fig. 3. Diagram of the study area and noise data collection results. In the line chart on the right, the y-axis depicts the average noise sound pressure level, and the x-axis denotes the sample ID. Sample IDs are sequential in both temporal and spatial dimensions, meaning each successive ID corresponds to the subsequent time step and reflects the volunteer's progression from one location to the next along the cycling route.

Table 1
Descriptive statistics of the collected noise exposure data.

Variable	Mean	Std. Dev.	Min	Max	Median	Skewness	Kurtosis
dBA	78.178	4.382	63.902	93.084	78.009	0.0857	0.2986
Accuracy	4.294	5.783	3.79	108.413	3.79	16.183	272.122
speed(m/s)	2.988	1.027	0	7.33	3.102	−0.4116	0.6029
leq100	71.643	5.005	58.879	86.215	71.913	−0.0817	−0.4823
leq125	71.879	4.603	58.019	83.804	71.943	−0.0729	−0.2755
leq160	72.381	4.209	60.484	83.255	72.377	−0.0732	−0.2362
leq200	72.824	4.154	60.137	85.032	72.892	−0.1231	−0.1796
leq250	73.093	4.108	59.4	85.894	73.371	−0.2299	0.0387
leq315	71.458	3.952	59.528	86.404	71.463	−0.0609	0.0734
leq400	69.823	4.094	56.936	83.391	69.853	−0.104	0.0179
leq500	69.238	4.155	56.89	82.605	69.169	0.0064	−0.0307
leq630	69.047	4.228	56.123	82.776	68.806	0.1022	0.0989
leq800	68.887	4.989	53.687	87.503	68.558	0.3626	0.4229
leq1000	68.659	5.667	49.898	86.473	68.201	0.3267	0.0647
leq1250	67.222	5.394	50.514	85.598	66.771	0.3552	0.186
leq1600	65.947	5.111	49.671	83.593	65.439	0.312	0.1545
leq2000	63.127	4.557	48.466	79.031	62.993	0.213	0.3225
leq2500	60.421	4.283	45.508	75.853	60.513	−0.0105	0.5627
leq3150	58.264	4.265	44.026	74.801	58.352	0.0914	0.6575
leq4000	56.944	4.172	42.729	80.157	56.982	0.2207	1.873
leq5000	56.627	5.317	42.572	83.111	56.035	0.9973	2.071
leq6300	54.551	7.161	41.012	83.869	52.621	1.501	2.538
leq8000	49.685	5.937	36.71	77.389	48.653	1.237	2.314
leq10000	44.765	4.923	35.263	70.026	43.659	1.172	1.826
leq12500	43.122	4.563	34.949	64.909	42.121	1.413	2.416
leq16000	42.174	3.385	35.262	58.539	41.568	1.29	2.338

Notes: leq denotes a frequency band, where f represents the center frequency. Sound pressure levels are calculated using third-octave band analysis. Std. Dev. = standard deviation. Skewness and kurtosis are calculated using the sample formulas. Speed represents the driving speed during the collection of the noise data. The mean value of GPS positioning accuracy is 4.294 m, which means our noise recording locations are relatively accurate.

The final noise prediction model can be expressed as:

$$\hat{y} = f(F_{combined}; \theta) \tag{5}$$

where f(·) represents the regression function, θ denotes the model training parameters, and $\hat{y}$ is the predicted noise level. For multi-spectral noise modeling, we extend this to predict noise levels across different frequency bands:

$$\hat{y} = [\hat{y}_{low}, \hat{y}_{mid}, \hat{y}_{high}]^T \tag{6}$$

where each component corresponds to the predicted noise level in low, mid, and high-frequency ranges, respectively.

The loss function for training is defined as:

$$L = (1/N)\sum_{i=1}^{N}\sum_{j\in\{low,mid,high\}}\alpha_j\left(\hat{y}_{i,j} - y_{i,j}\right)^2 \quad (7)$$

where N is the number of samples, α_j represents the weight for frequency band j, and $y_{i,j}$ is the ground truth noise level for sample i in frequency band j.

Building upon this foundation, we utilized the extracted multi-dimensional acoustic features as input to implement various classical machine learning baseline methods for comparative experiments, including XGBoost, LightGBM, Random Forest, Gradient Boosting, Support Vector Regression (SVR), k-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), and Lasso Regression. In addition, we also introduced the ensemble learning method, including a voting model and a stacking model, which could integrate different models to achieve better generalization and performance.

Specifically, the visual embeddings extracted from SVI and RSI formed the input feature vectors for the machine learning models, which learned the mapping relationship between these multi-scale built environment features and noise frequency levels. All models were evaluated with 5-fold cross-validation, using the coefficient of determination (R (Basner et al., 2014)), mean absolute error (MAE) and root mean square error (RMSE) as assessment metrics.

2.4. Semantic segmentation-based street element extraction and SHAP explainability analysis method

To enhance the interpretability of our noise prediction model and understand the relationship between specific urban environmental features (rather than meaningless high-dimensional values) and noise levels, we employed a semantic segmentation approach to extract detailed street-level features from the street view imagery. We utilized the pre-trained SegFormer model[2] (Xie et al., 2021), which was fine-tuned on the Cityscapes dataset. The Cityscapes dataset contains 5000 high-quality pixel-level annotated urban street scene images covering 19 common urban object categories, including roads, sidewalks, buildings, vegetation, sky, pedestrians, vehicles, traffic signs, and poles.

For each street view image in our experiment, the segmentation process can be formalized as:

$$S = SegFormer(I_{SVI}) \quad (8)$$

where $S \in R^{H\times W\times C}$ represents the segmentation mask, H and W are the height and width of the input image, and C is the number of semantic categories (19 classes in Cityscapes). We extracted urban features by calculating the proportion of pixels belonging to each feature class:

$$P_c = \frac{\sum_{i,j}\mathbb{1}\left[S_{i,j} = c\right]}{H\times W} \quad (9)$$

This process generated a 19-dimensional feature vector for each image, where each dimension represents the coverage proportion of a specific urban feature. To understand which urban environmental features drive noise predictions, we employed SHAP (SHapley Additive exPlanations) values (Gu et al., 2025; Ke et al., 2025) for model interpretation. The Shapley value for feature i is defined as:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}}\frac{|S|!(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right] \quad (10)$$

We trained an XGBoost regression model using the 19-dimensional semantic segmentation features and employed TreeExplainer to compute SHAP values. We analyzed SHAP values at two levels: global feature importance identifies key urban features by calculating mean absolute SHAP values, while partial dependence analysis reveals non-linear relationships between urban features and noise levels:

$$\text{Importance}_i = \frac{1}{N}\sum_{j=1}^{N}\left|\phi_i^{(j)}\right| \quad (11)$$

Although this model based on semantic segmentation could not achieve the best accuracy, it provides a transparent framework for understanding feature-noise relationships. It could provide some evidence about the urban environmental factors that drive noise variations, supporting urban planning and noise mitigation strategies.

3. Results

3.1. Experimental collection of noise frequency distribution analysis

Fig. 4 illustrates the frequency variation and decibel intensity at different Hz along a cycling trajectory. It is observable that noise at the urban street scale is predominantly concentrated in the low- and mid-frequency ranges, a typical characteristic of urban traffic noise. Analysis of the boxplot reveals that noise intensity in the low-frequency band (around 250 Hz) exhibits higher values and greater variability, with the median generally maintained between 70 and 80 dB. This is primarily attributed to factors such as low-frequency vibrations from vehicle engines, friction between tires and the road surface, and vehicle body resonance.

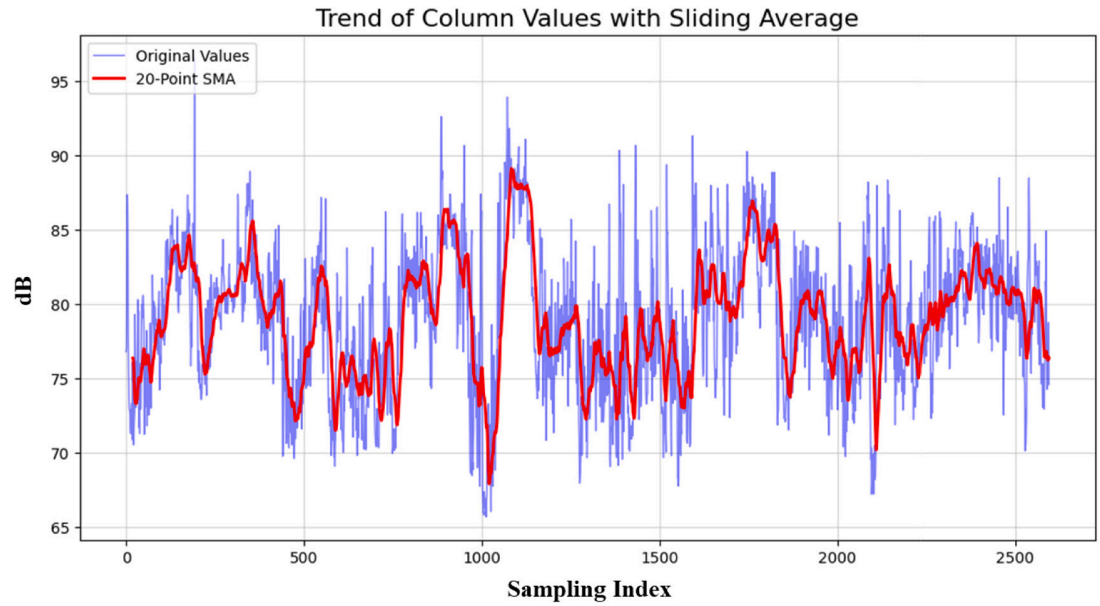
As the frequency increases, the noise intensity shows a distinct downward trend. In the high-frequency band (above 2000 Hz), the noise intensity is significantly reduced, with the median dropping to a range of 40–55 dB. This frequency distribution characteristic reflects the physical nature of urban traffic noise: low-frequency noise generated by large vehicles (e.g., buses, trucks) possesses greater penetrating power and can travel longer distances. In contrast, high-frequency noise (e.g., sharp braking, horn sounds), while potentially having a high instantaneous intensity, attenuates more rapidly during spatial propagation and has a relatively shorter duration. This non-uniform distribution of multi-spectral noise further underscores the necessity of frequency-band-specific modeling, as different frequency bands exhibit significant differences in their mechanisms of impact on human health and their sensitivity to environmental factors.

3.2. Model performance comparison and fusion effects

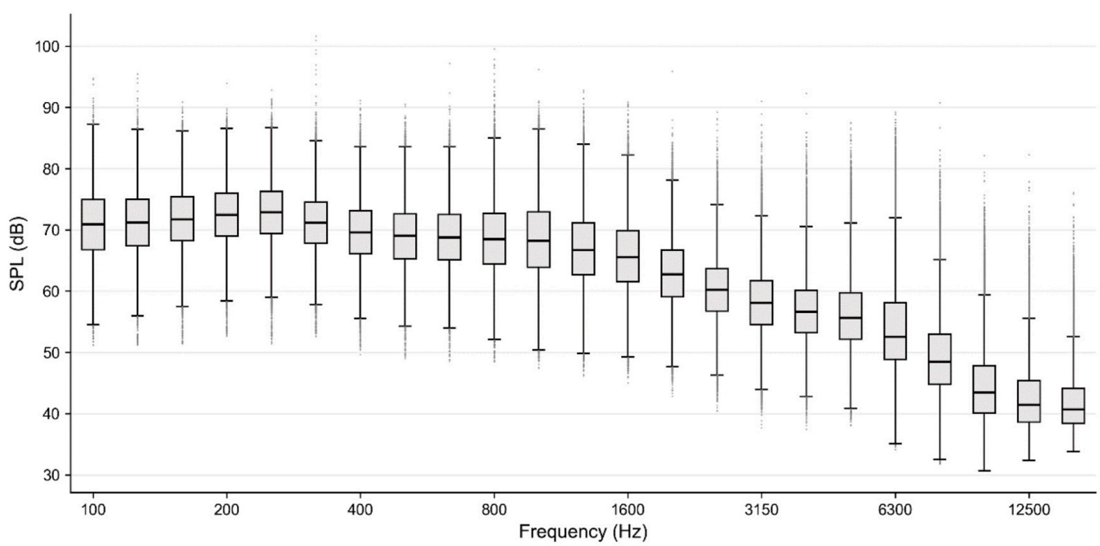
Given the difficulty of reconstructing multi-frequency noise exposure maps (Baclet et al., 2022), this study was based on the extracted built environment features combined with machine learning techniques (Baclet et al., 2023). Our goal was to predict four key acoustic metrics: the overall A-weighted Sound Pressure Level (dBA), low-frequency noise, mid-frequency noise, and high-frequency noise. For comparison, we evaluated various machine learning methods, including XGBoost, LightGBM, Random Forest, Gradient Boosting, SVR (Support Vector Regression), KNN (k-Nearest Neighbors), Lasso, and the ensemble learning Voting Model. Additionally, we conducted k-fold cross-validation on those models.

First, we constructed a minimum buffer zone of 15 m around each street view point. The 15-m buffer distance was determined by comprehensively considering GPS positioning accuracy and the distribution density of street view points. From approximately 6000 sampling points, we filtered 923 street views for which the corresponding buffer zones contained no fewer than one noise record. If there were more than one noise record within the buffer zone, the average value of these records was calculated as a replacement. Furthermore, we performed a sensitivity analysis to verify the rationality of the buffer distance selection and designed three comparative experiments with buffer zones of 25 m, 50 m, and 100 m respectively for calculation. The dataset comprised 923 multi-frequency noise-remote sensing-street view paired

[2] https://huggingface.co/nvidia/segformer-b0-finetuned-cityscapes-512-1024



(a)



(b)

Fig. 4. Noise exposure analysis along one cycling trajectory. (a) Noise exposure curve along a cycling route; (b) Frequency distribution histogram showing noise intensity across different frequency bands.

samples, and 5-fold cross-validation was conducted.

The sensitivity analysis presented in Tables 2–4 demonstrates the reliability of image-based noise prediction approaches. Taking dBA (mean noise) as an example, the R^2 values range from 0.417 to 0.649, indicating moderate to substantial predictive performance across different model configurations. Notably, the results reveal a positive relationship between buffer distance and prediction accuracy for remote sensing-based features. As buffer distance increases, the remote sensing approach achieves progressively higher accuracy, which can be attributed to two interconnected factors. First, larger buffer zones produce noise labels that better approximate the mean noise exposure within the area, reducing measurement uncertainty. Second, the inherently broader spatial coverage of remote sensing imagery becomes increasingly advantageous at larger scales, capturing contextual information that street-view imagery cannot adequately represent at such distances.

Among the evaluated models, SVR and KNN consistently demonstrate superior performance across all evaluation metrics (R (Basner et al., 2014), RMSE, and MAE). The SVR model, in particular, achieves the highest R^2 of 0.649 when combined with remote sensing features for mean noise prediction, suggesting its robust capability in capturing the complex non-linear relationships between environmental features and noise levels.

There are still some interesting findings: mid-frequency noise exhibits the highest predictability across all model configurations. This observation likely reflects the strong association between mid-frequency noise and prevalent human activity patterns, such as traffic flow and commercial operations, which are more readily captured by visual features in both street-view and remote sensing imagery. Conversely, low-

Table 2
Model Performance Across Different Frequency Bands with Multisource Features (25 m Buffer).

Feature Type	Model	Mean Noise			Low Freq			Mid Freq			High Freq		
		R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Street View	XGBoost	0.290	3.176	2.502	0.227	3.262	2.569	0.276	3.090	2.461	0.245	3.686	2.769
	LightGBM	0.278	3.204	2.533	0.219	3.275	2.569	0.273	3.095	2.448	0.233	3.713	2.799
	Random Forest	0.193	3.386	2.665	0.132	3.458	2.725	0.182	3.282	2.588	0.152	3.906	2.881
	Gradient Boosting	0.291	3.173	2.509	0.212	3.292	2.589	0.240	3.164	2.500	0.209	3.771	2.852
	SVR	0.367	2.998	2.318	0.311	3.078	2.375	0.359	2.907	2.244	**0.316**	**3.504**	**2.642**
	KNN	**0.370**	**2.993**	**2.313**	**0.316**	**3.066**	**2.344**	**0.360**	**2.904**	**2.224**	0.314	3.503	2.634
	Lasso	0.275	3.208	2.525	0.191	3.336	2.634	0.248	3.147	2.496	0.230	3.721	2.827
	Voting	0.353	3.032	2.378	0.298	3.109	2.428	0.338	2.953	2.332	0.302	3.544	2.656
	Stacking	0.302	3.149	2.471	0.184	3.353	2.645	0.278	3.085	2.426	0.141	3.931	2.895
Remote Sensing	XGBoost	0.259	3.243	2.517	0.274	3.163	2.465	0.251	3.141	2.422	0.282	3.591	2.707
	LightGBM	0.289	3.176	2.480	0.270	3.169	2.456	0.266	3.106	2.407	0.270	3.622	2.760
	Random Forest	0.188	3.395	2.658	0.190	3.342	2.630	0.179	3.288	2.580	0.206	3.778	2.825
	Gradient Boosting	0.236	3.292	2.559	0.256	3.201	2.488	0.240	3.163	2.434	0.253	3.658	2.761
	SVR	**0.374**	**2.978**	**2.297**	**0.392**	**2.892**	**2.225**	**0.371**	**2.876**	**2.226**	**0.426**	**3.210**	**2.441**
	KNN	0.318	3.114	2.377	0.308	3.085	2.320	0.313	3.006	2.323	0.368	3.369	2.532
	Lasso	0.256	3.249	2.503	0.184	3.351	2.619	0.237	3.168	2.450	0.218	3.751	2.860
	Voting	0.337	3.068	2.377	0.335	3.025	2.338	0.327	2.977	2.307	0.363	3.382	2.555
	Stacking	0.190	3.393	2.658	0.152	3.419	2.709	0.205	3.237	2.538	0.213	3.765	2.783
Fusion	XGBoost	0.330	3.086	2.407	0.279	3.150	2.468	0.291	3.056	2.369	0.296	3.560	2.704
	LightGBM	0.329	3.086	2.400	0.265	3.176	2.477	0.295	3.046	2.321	0.300	3.546	2.738
	Random Forest	0.209	3.352	2.618	0.168	3.386	2.677	0.189	3.268	2.545	0.193	3.810	2.840
	Gradient Boosting	0.307	3.135	2.453	0.248	3.212	2.507	0.288	3.060	2.351	0.269	3.627	2.763
	SVR	**0.417**	**2.877**	**2.196**	0.382	2.916	2.266	0.405	2.801	2.145	**0.402**	**3.270**	**2.490**
	KNN	0.398	2.922	2.221	**0.390**	**2.897**	**2.212**	**0.413**	**2.780**	**2.110**	0.375	3.342	2.518
	Lasso	0.321	3.103	2.411	0.228	3.256	2.547	0.261	3.118	2.425	0.274	3.614	2.747
	Voting	0.386	2.953	2.282	0.344	3.005	2.342	0.365	2.892	2.223	0.369	3.368	2.563
	Stacking	0.300	3.151	2.456	0.212	3.293	2.594	0.270	3.100	2.417	0.229	3.724	2.756

Table 3
Model Performance Across Different Frequency Bands with Multisource Features (50 m Buffer).

Feature Type	Model	Mean Noise			Low Freq			Mid Freq			High Freq		
		R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Street View	XGBoost	0.335	2.738	2.057	0.240	2.851	2.224	0.293	2.733	2.088	0.288	3.150	2.340
	LightGBM	0.331	2.744	2.085	0.228	2.872	2.255	0.309	2.703	2.064	0.271	3.187	2.369
	Random Forest	0.232	2.943	2.244	0.143	3.028	2.402	0.198	2.911	2.230	0.181	3.376	2.455
	Gradient Boosting	0.325	2.756	2.077	0.227	2.874	2.240	0.302	2.716	2.073	0.281	3.164	2.357
	SVR	**0.453**	**2.480**	**1.868**	**0.356**	**2.623**	**1.978**	0.418	2.479	1.880	**0.378**	**2.940**	**2.184**
	KNN	0.452	2.485	1.851	0.339	2.655	1.960	**0.443**	**2.427**	**1.800**	0.368	2.963	2.181
	Lasso	0.295	2.818	2.143	0.194	2.934	2.301	0.253	2.810	2.160	0.240	3.253	2.429
	Voting	0.406	2.589	1.945	0.315	2.707	2.086	0.381	2.559	1.950	0.343	3.024	2.223
	Stacking	0.312	2.785	2.122	0.180	2.962	2.337	0.309	2.704	2.068	0.093	3.554	2.584
Remote Sensing	XGBoost	0.368	2.667	2.030	0.308	2.720	2.090	0.366	2.590	1.947	0.316	3.084	2.221
	LightGBM	0.381	2.639	2.005	0.323	2.691	2.082	0.399	2.522	1.904	0.321	3.071	2.219
	Random Forest	0.251	2.906	2.218	0.201	2.925	2.296	0.253	2.810	2.136	0.229	3.276	2.385
	Gradient Boosting	0.356	2.696	2.029	0.271	2.791	2.133	0.354	2.614	1.959	0.274	3.176	2.307
	SVR	**0.525**	**2.311**	**1.744**	**0.480**	**2.361**	**1.759**	**0.514**	**2.267**	**1.713**	**0.529**	**2.554**	**1.869**
	KNN	0.418	2.561	1.903	0.377	2.582	1.922	0.408	2.503	1.857	0.433	2.805	2.043
	Lasso	0.337	2.732	2.069	0.176	2.969	2.291	0.313	2.695	2.044	0.232	3.271	2.425
	Voting	0.438	2.516	1.899	0.369	2.599	1.982	0.431	2.454	1.846	0.406	2.874	2.085
	Stacking	0.362	2.684	2.038	0.119	3.070	2.440	0.355	2.612	1.991	0.162	3.402	2.467
Fusion	XGBoost	0.405	2.589	1.963	0.313	2.705	2.101	0.385	2.550	1.920	0.346	3.016	2.249
	LightGBM	0.417	2.564	1.939	0.316	2.702	2.093	0.386	2.547	1.931	0.360	2.984	2.257
	Random Forest	0.250	2.908	2.206	0.177	2.969	2.343	0.236	2.843	2.162	0.208	3.321	2.401
	Gradient Boosting	0.389	2.622	1.984	0.301	2.729	2.136	0.380	2.560	1.939	0.313	3.090	2.284
	SVR	**0.518**	**2.326**	**1.737**	**0.447**	**2.431**	**1.831**	**0.507**	**2.281**	**1.702**	0.490	2.659	1.991
	KNN	0.472	2.434	1.783	0.434	2.460	1.816	0.485	2.325	1.694	0.468	2.716	1.989
	Lasso	0.367	2.670	2.044	0.262	2.808	2.180	0.321	2.681	2.051	0.284	3.157	2.355
	Voting	0.459	2.470	1.850	0.385	2.564	1.974	0.444	2.425	1.820	0.420	2.840	2.107
	Stacking	0.355	2.695	2.039	0.231	2.867	2.243	0.339	2.641	1.999	0.236	3.252	2.376

frequency noise presents the greatest prediction challenge, despite remote sensing features showing better performance in this frequency band. This difficulty may stem from the propagation characteristics of low-frequency sound waves, which travel longer distances and are less influenced by static visual environmental factors, thereby weakening the correlation between image features and noise levels. In summary, these findings reveal the unique soundscape characteristics at different frequencies and emphasize the importance of multi-source data fusion.

Although our sensitivity analysis (Tables 2–4) demonstrates that the 100-m buffer zone achieves the highest R^2 value (0.649 for mean noise level prediction in dBA), we primarily employed the 25-m buffer zone-trained model for visualization and mapping applications in this study. This choice is grounded in several methodological and practical considerations. First, the 25-m buffer corresponds to our 50-m street view image sampling interval, ensuring that adjacent sampling points maintain spatial independence without overlapping coverage areas.

Table 4
Model Performance Across Different Frequency Bands with Multisource Features (100 m Buffer).

Feature Type	Model	Mean Noise			Low Freq			Mid Freq			High Freq		
		R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Street View	XGBoost	0.376	2.364	1.759	0.276	2.511	1.957	0.313	2.402	1.784	0.306	2.724	1.960
	LightGBM	0.371	2.372	1.786	0.280	2.505	1.945	0.326	2.380	1.787	0.297	2.739	1.988
	Random Forest	0.239	2.610	1.993	0.167	2.694	2.141	0.206	2.582	1.963	0.184	2.955	2.118
	Gradient Boosting	0.371	2.376	1.762	0.245	2.564	1.983	0.303	2.418	1.808	0.264	2.806	1.987
	SVR	0.504	2.107	1.550	**0.414**	**2.256**	**1.695**	0.483	2.085	1.558	0.420	2.491	1.797
	KNN	**0.514**	**2.087**	**1.509**	0.397	2.286	1.699	**0.512**	**2.026**	**1.499**	**0.430**	**2.469**	**1.758**
	Lasso	0.305	2.496	1.903	0.214	2.615	2.036	0.256	2.501	1.905	0.221	2.890	2.099
	Voting	0.443	2.235	1.664	0.351	2.376	1.828	0.409	2.230	1.668	0.362	2.614	1.863
	Stacking	0.326	2.459	1.872	0.134	2.748	2.204	0.311	2.407	1.831	0.035	3.215	2.301
Remote Sensing	XGBoost	0.439	2.244	1.664	0.379	2.326	1.790	0.411	2.227	1.611	0.390	2.548	1.796
	LightGBM	0.451	2.218	1.652	0.415	2.258	1.750	0.451	2.150	1.577	0.371	2.589	1.813
	Random Forest	0.301	2.504	1.880	0.249	2.559	2.023	0.292	2.441	1.813	0.268	2.798	1.969
	Gradient Boosting	0.416	2.290	1.700	0.370	2.344	1.816	0.425	2.200	1.605	0.376	2.578	1.830
	SVR	**0.649**	**1.774**	**1.262**	**0.594**	**1.881**	**1.383**	**0.623**	**1.778**	**1.265**	**0.628**	**1.985**	**1.396**
	KNN	0.550	2.010	1.434	0.456	2.176	1.579	0.518	2.013	1.406	0.536	2.223	1.556
	Lasso	0.410	2.297	1.716	0.237	2.580	2.010	0.375	2.290	1.704	0.246	2.840	2.046
	Voting	0.520	2.075	1.529	0.451	2.187	1.676	0.499	2.053	1.482	0.464	2.392	1.673
	Stacking	0.465	2.190	1.634	0.149	2.720	2.190	0.415	2.217	1.636	−0.023	3.309	2.364
Fusion	XGBoost	0.435	2.250	1.654	0.366	2.351	1.799	0.434	2.184	1.600	0.408	2.519	1.832
	LightGBM	0.462	2.196	1.627	0.380	2.324	1.781	0.440	2.170	1.614	0.442	2.444	1.776
	Random Forest	0.275	2.551	1.912	0.227	2.596	2.067	0.280	2.462	1.846	0.229	2.875	2.041
	Gradient Boosting	0.419	2.281	1.664	0.353	2.374	1.824	0.414	2.220	1.640	0.396	2.541	1.828
	SVR	**0.596**	**1.900**	**1.384**	**0.549**	**1.981**	**1.476**	**0.594**	**1.846**	**1.328**	**0.592**	**2.085**	**1.522**
	KNN	0.555	1.996	1.394	0.488	2.112	1.525	0.562	1.918	1.362	0.559	2.172	1.522
	Lasso	0.394	2.331	1.747	0.298	2.474	1.908	0.380	2.282	1.699	0.296	2.746	2.009
	Voting	0.505	2.108	1.538	0.443	2.204	1.679	0.501	2.048	1.495	0.478	2.365	1.692
	Stacking	0.419	2.281	1.704	0.242	2.570	2.026	0.385	2.274	1.691	0.049	3.190	2.285

This 50-m sampling resolution has been widely adopted in street-level urban research (Wang et al., 2025; Zhang et al., 2023), providing a standardized framework for comparative studies. Second, the relationship between buffer size and predictive performance is not simply monotonic. The coverage areas of buffer sizes differ substantially: 15-m (707 m^2), 25-m (1,964 m^2), and 50-m (7,854 m^2) buffer zones. Larger buffer zones inevitably incorporate more diverse environmental information, which can dilute localized noise variations. For instance, since typical urban road widths in our study area rarely exceed 50 m, a 50-m buffer zone simultaneously captures features from both sides of the street. This spatial averaging effect smooths acoustic differences between opposite street sides, reducing the model's ability to capture fine-grained soundscape variations. Our field measurements confirmed notable acoustic differences between opposite sides of the same street, which motivated our multimodal fusion approach despite remote sensing features showing slightly higher individual performance. Finally, different buffer sizes serve different geospatial contexts. The 15-m buffer-trained models are more suitable for narrow alleys and confined spaces where sound propagation is limited, while 50-m buffer models better capture noise patterns in open areas with broader acoustic propagation. While we acknowledge the superior R^2 performance of the 100-m buffer in sensitivity analysis, we advocate for the 25-m buffer as the optimal scale for street-level noise mapping model training, considering the built environment characteristics of our study area (primarily the urban core of Zhuhai City). Given our 50-m sampling interval, the 25-m buffer model strikes an optimal balance between prediction accuracy and spatial coverage, providing sufficient environmental context while preserving local acoustic characteristics.

Fig. 5. Comparison of model prediction performance using different feature inputs. The scatter plots show prediction results for models using (a) street view imagery features only, (b) remote sensing imagery features only, and (c) fused multi-source features. The x-axis represents the ground-truth noise values from the test set, while the y-axis represents the model's predicted values. The fused model (c) shows the tightest clustering of points around the ideal y = x line (black dashed line), indicating superior predictive accuracy.

3.3. Scatter plots and prediction spatial distribution analysis

Fig. 5 presents a quantitative comparison of the dBA prediction (mean noise exposure) results, using the SVR model at a 25 m buffer (50-m resolution) as an example. This model was chosen due to its optimally balanced prediction accuracy with spatial independence, avoiding overlap between adjacent prediction points while preserving fine-grained local acoustic variations (Song et al., 2026). The scatter plots illustrate the model predictions using features from (a) street view imagery only (orange points, left), (b) remote sensing imagery only (green points, center), and (c) fused features from both modalities (blue points, right). In each plot, the x-axis represents the ground-truth values from the test set, the y-axis represents the model's predictions, and the red dashed line indicates the ideal y = x prediction line.

As illustrated by the prediction performance, the multi-source fusion method (c) achieves the highest predictive accuracy. Its scatter points are the most tightly clustered around the ideal prediction line, indicating excellent agreement between predicted and true values. In contrast, the model using only SVI features (a) shows greater prediction variance, with notable biases in high-noise environments (above 80 dBA). This is likely attributable to the "temporal bias" of street view images, which can be heavily influenced by transient traffic and environmental conditions at the moment of capture. The RS-based model (b) yields a more stable overall distribution but lacks the sensitivity to capture fine-grained noise variations, reflecting the inherent spatial resolution limitations of satellite data. Our field validation data reveal acoustic variations could exceed 5 dBA between opposite street sides. This fine-grained difference detail may be smoothed systematically by the 30-m resolution of satellite imagery.

In addition to the quantitative plots, we also mapped the spatial prediction results from the three different feature extraction approaches across different frequencies in Fig. 6. The predictions from the RS-based model exhibit pronounced block-like patterns, a direct result of the resolution limits and characteristics of satellite imagery. While this model captures broad features, it struggles with fine-scale variations. For instance, since our sampling was conducted at 50-m intervals along roadways, consecutive points on the same street often correspond to very similar remote sensing imagery, leading to homogenized predicted

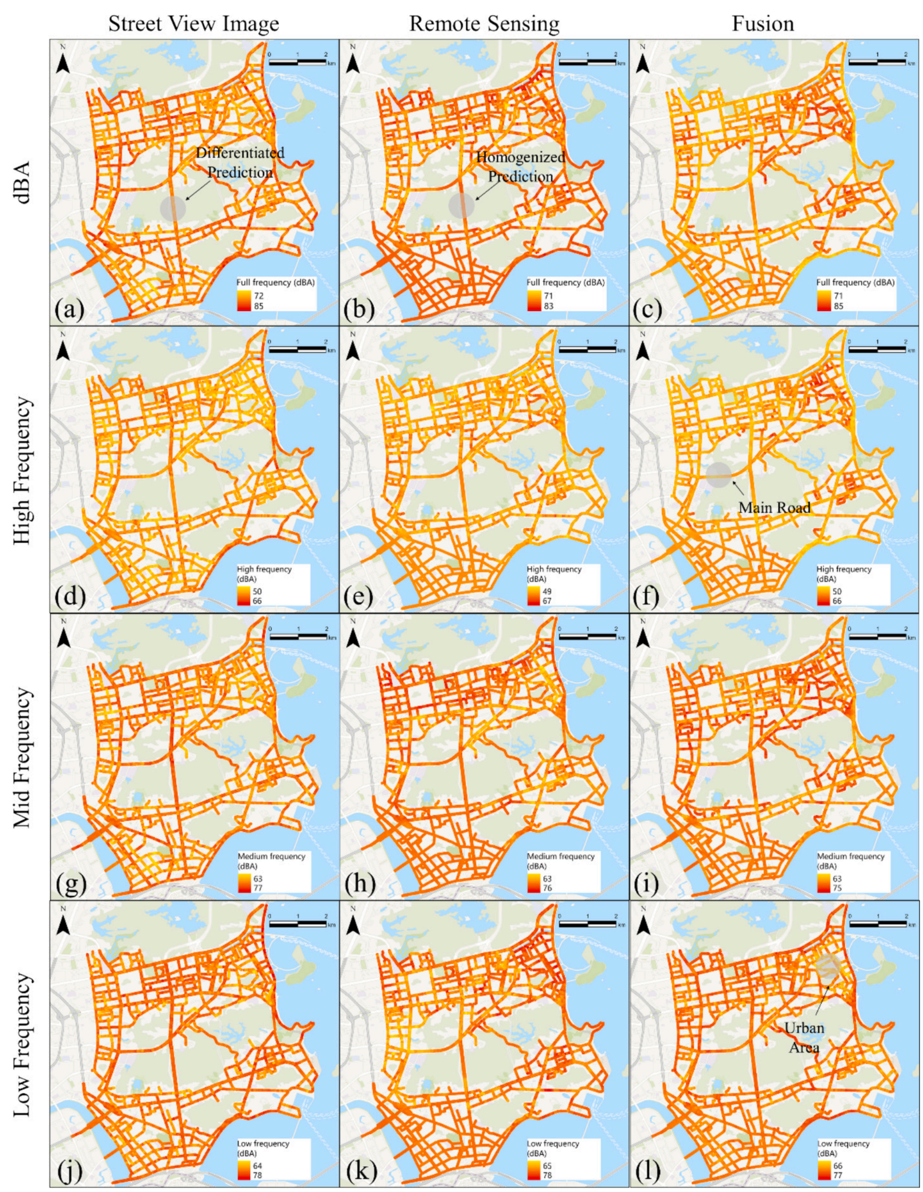


Fig. 6. Spatial distribution of street-level noise predictions at different frequencies using various feature inputs. Subplots a-l present fine-grained (50-m resolution) noise maps across different frequencies generated from different feature sets. The prediction results not only preserve the detailed perceptual capabilities of street view data but also leverage the spatial consistency of remote sensing data to reduce prediction uncertainty.

values and coarse prediction granularity.

In contrast, the SVI-based model demonstrates a superior capability in capturing intra-urban socioeconomic characteristics with richer spatial variation. However, its limitation lies in its susceptibility to the influence of immediate vehicle and pedestrian density, which can lead to inconsistencies such as substantial prediction discrepancies between opposite sides of the same street. Compared to individual data-driven models, the fusion model exhibits a more uniform and consistent pattern. Although its R^2 (0.417–0.596) is marginally lower than the RS-based model, the fusion approach preserves fine-grained spatial heterogeneity essential for practical noise mapping applications.

This is attributed to its ability to integrate the complementary perspectives of Street View Imagery at the micro-scale and Remote Sensing at the macro-scale. SVI captures dynamic street-level socioeconomic characteristics, such as traffic density, architectural styles, and commercial activities, while RS provides stable macro-level spatial context, including land use, building density, and vegetation coverage. Specifically, regarding different noise frequencies, main roads accommodate higher traffic volumes and demonstrate stronger correlations with low-frequency noise. In contrast, urban area roads tend to be narrower and exhibit positive correlations with mid- to high-frequency noise.

3.4. Explainable learning results

While the encoder-based hybrid approach achieves superior predictive performance, it operates as a 'black box' that provides limited insights into which urban elements drive noise variations. To complement our predictive model and provide interpretable insights into the relationship between built environment characteristics and noise levels, we employed an additional explainable modeling approach based on the semantic segmentation to explain the associations between street elements and noise (dBA). We first obtained the proportion of various street elements through semantic segmentation. It is based on the pretrained Segformer model trained on the cityscape dataset (Ma et al., 2025). Subsequently, we fitted noise levels using these street features based on XGBoost regression ($R^2 = 0.189$, RMSE = 3.88). Finally, based on Tree Explainer, we quantified the influence of all street elements on noise levels, namely their Shapley values. The interpretation results are shown in Fig. 7. It should be noted that the relatively modest R^2 value (0.189) of the segmentation-based XGBoost model is expected and acceptable in this context. Unlike the deep feature-based models that capture complex visual patterns through high-dimensional embeddings, the segmentation approach deliberately trades predictive accuracy for interpretability by using only 19 human-understandable urban element categories. This design choice enables transparent analysis of how specific street elements influence noise levels.

Fig. 7(a) presents the global influence of street elements. Buildings, roads, sidewalks, and terrain are the four features with the greatest impact on model predictions. Their Shapley value distributions span the widest ranges, indicating that they can substantially increase or decrease noise prediction values. Fig. 7(b) provides an in-depth illustration of how numerical changes in the top 9 street elements by feature importance specifically affect their contributions to noise prediction. Although the curves and trends for buildings, roads, and sidewalks differ, they all demonstrate that wider roads are more prone to generating traffic noise. Spacious roads typically need to accommodate higher traffic volumes and faster vehicle speeds. Engine noise and tire friction sounds are the primary sources of traffic noise. Moreover, open road surfaces lack noise barriers, allowing sound to propagate unobstructed in all directions. Therefore, noise represents the superposition of nearby and distant noise sources, and this "broadcasting effect" contributes to volume amplification. The nonlinear influence of terrain reflects more complex urban environments. When terrain values increase from 0 to 0.02, noise levels gradually rise. At this range, the model primarily captures not large expanses of grassland or greenery, but rather curbs adjacent to roadways (as detailed in Cityscape). These hard boundaries locally and slightly increase noise levels. However, when terrain values exceed 0.02, street elements include not only curbs but also substantial areas of green belts or soil. These soft surfaces have sound-absorbing properties; therefore, when terrain proportions exceed 0.02, larger terrain coverage generally suppresses noise levels. Other street elements follow trends consistent with our understanding of the physical world. For instance, higher proportions of motorcycles and cars promote increased noise level predictions. Additionally, while fences, poles, and traffic signs may exhibit correlations in data distribution, they generally hover around the Shapley value baseline of zero, indicating minimal impact on prediction

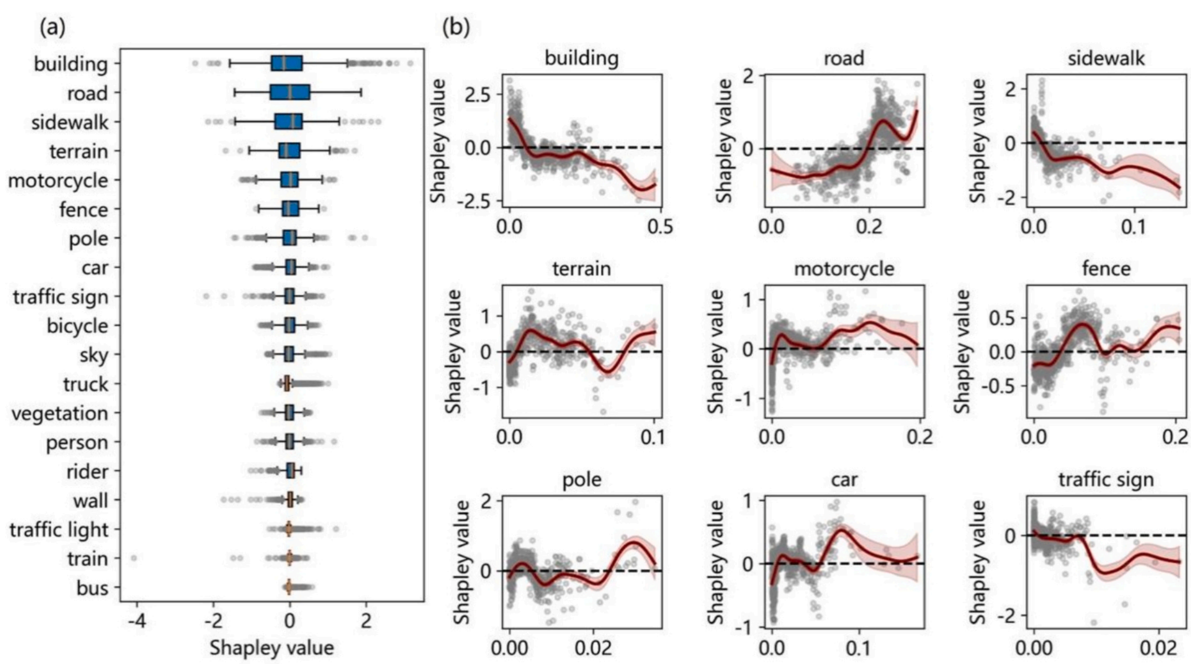


Fig. 7. SHAP-based Explainability Analysis of Street Visual Elements. (a) Global feature importance ranking from highest to lowest (top to bottom); (b) Partial dependence plots for the top 9 most important street element features.

results.

4. Conclusion and discussion

This study successfully constructed and validated a multimodal machine learning framework that integrates the microscopic perspective of street view imagery with the macroscopic perspective of remote sensing imagery, achieving significant breakthroughs in multi-spectral noise prediction and high-resolution mapping at the urban street scale. The research innovatively combines the street-level detail perception capability of street view imagery with the spatial macro-consistency of remote sensing imagery, effectively overcoming the inherent limitations of single data sources. This approach not only addresses the temporal instability issues of street view imagery but also compensates for the shortcomings of remote sensing imagery in perceiving microscopic socioeconomic activities, achieving complementary advantages.

This study overcomes the traditional limitation of focusing solely on A-weighted total decibel values, revealing that urban noise is predominantly composed of low- and mid-frequency components. It also achieves simultaneous and accurate prediction of noise across three distinct frequency bands: low, mid, and high. Across different soundscape scales, the model achieves high-precision predictions with R^2 values ranging from 0.417 to 0.649, with superior predictive accuracy particularly evident for mid-frequency noise. However, substantial variance remains unexplained, as unmeasured factors such as instantaneous traffic conditions, human behavior, and meteorological variations contribute to noise variability beyond what static visual features can capture. Furthermore, the study successfully generates street-level multi-frequency soundscape maps at an unprecedented spatial resolution of 50 m, achieving a new level of spatial granularity. Through SHAP explainable analysis, we quantified the contribution of urban visual elements such as buildings, roads, and sidewalks to noise prediction, providing scientific evidence for urban planning decisions. Compared to traditional field measurement methods, this study only requires street view imagery and remote sensing imagery corresponding to streets, providing a cost-effective and large-scale applicable new paradigm for urban acoustic environment monitoring that significantly reduces data collection costs while achieving broader spatial coverage.

Although this study still has room for improvement in model generalization capability and data timeliness. For instance, smartphones are highly convenient and can help us simultaneously record speed, location, and noise levels. Although we calibrated them against professional equipment prior to data collection, errors are unavoidable. The spatial resolution of the remote sensing data utilized is coarser compared to the pretrained dataset, potentially compromising the reliability of the transfer learning process. But our built methodological framework has good scalability. Future research can further integrate dynamic traffic data, meteorological conditions, and other factors to construct a more comprehensive urban acoustic environment spatial-temporal dynamic monitoring system. In conclusion, this study not only achieves important innovation in technical methods but also provides powerful scientific support tools and data foundations for precise urban noise pollution management, public health policy formulation, and healthy city construction.

CRediT authorship contribution statement

Yan Zhang: Writing – review & editing, Writing – original draft, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Entong Ke:** Writing – review & editing, Data curation. **Mei-Po Kwan:** Writing – original draft, Funding acquisition, Conceptualization. **Libo Fang:** Data curation. **Mingxiao Li:** Writing – review & editing, Data curation.

Funding

This research was supported by a grant from the Smart Traffic Fund (Grant no: PSRI/44/2208/PR). It was also supported by grants from the Hong Kong Research Grants Council (General Research Fund Grant no. 14605920, 14606922, 14603724; Collaborative Research Fund Grant no. C4023-20GF; Research Matching Grants RMG 8601219, 8601242, 3110151), a grant from the Research Committee on Research Sustainability of Major Research Grants Council Funding Schemes (3133235) of the Chinese University of Hong Kong (CUHK), a grant from the $1 + 1 + 1$ Platform Program of CUHK (4760974), and a grant from the Vice-Chancellor's One-off Discretionary Fund (Smart and Sustainable Cities: City of Commons) (4930787) of CUHK. Open Research Fund Program of Shenzhen Key Laboratory of Spatial Smart Sensing and Services (GEM-Lab-2024010) and National Natural Science Foundation of China (42501579). The funders had no role in the study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Declaration of competing interest

The authors declare that they have no financial and personal relationships with other people or organizations that can inappropriately influence our work.

Appendix A. Appendix

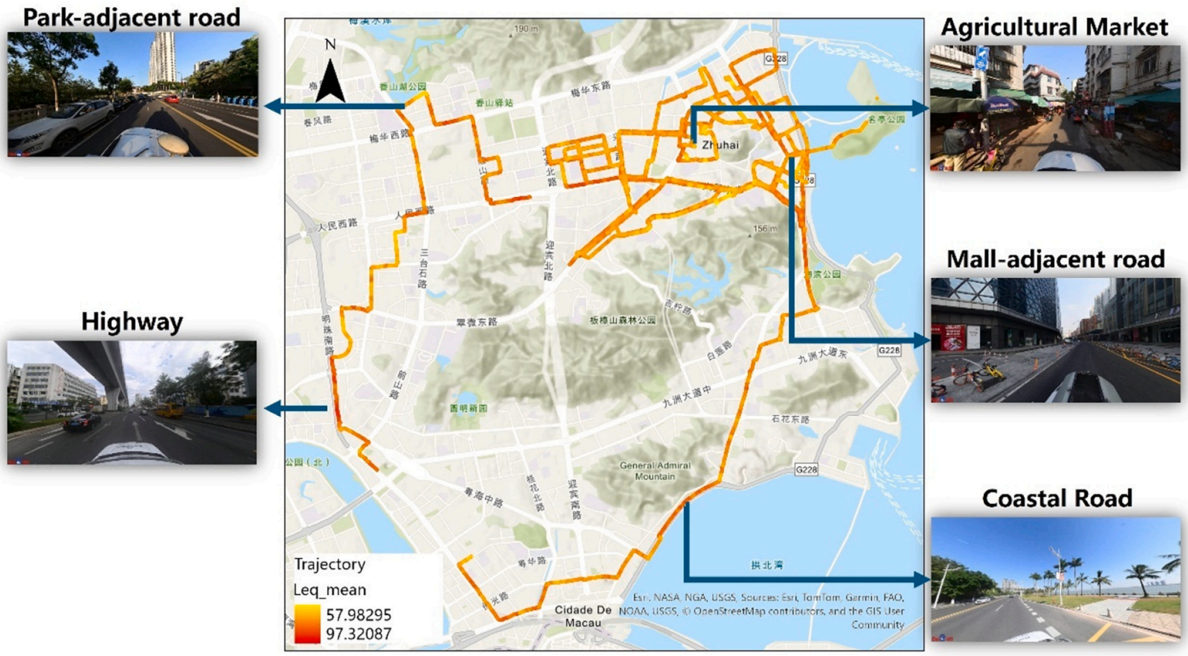


Fig. A.1. The spatial distribution of collected noise data points.

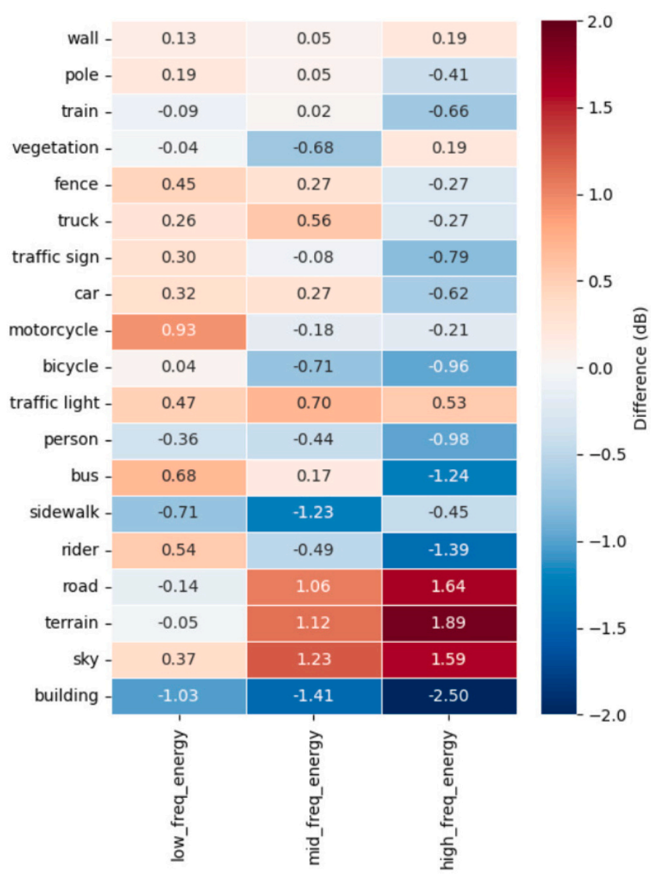


Fig. A.2. The comprehensive correlation analysis between visual elements from street-view imagery and noise frequency characteristics.

Fig. A3. The Representative optical remote sensing imagery samples from the study area. The remote sensing imagery used for our inference is shown in Fig. A3. Despite lower resolution compared to the pretraining data (0.6-m resolution), semantic features relevant to noise prediction, including building patterns, road networks, vegetation, and street configurations, remain clearly identifiable.

Data availability

Data will be made available on request.

References

Alvares-Sanches, T., Osborne, P. E., & White, P. R. (2021). Mobile surveys and machine learning can improve urban noise mapping: Beyond A-weighted measurements of exposure. *Science of the Total Environment, 775*, Article 145600.

Baclet, S., Venkataraman, S., Rumpler, R., Billsjö, R., Horvath, J., & Österlund, P. E. (2022). From strategic noise maps to receiver-centric noise exposure sensitivity mapping. *Transportation Research Part D: Transport and Environment, 102*, Article 103114.

Baclet, S., et al. (2023). Near-real-time dynamic noise mapping and exposure assessment using calibrated microscopic traffic simulations. *Transportation Research Part D: Transport and Environment, 124*, Article 103922.

Basner, M., Babisch, W., Davis, A., et al. (2014). Auditory and non-auditory effects of noise on health. *The Lancet, 383*(9925), 1325–1332.

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning, 215*, Article 104217.

Cai, J., Kwan, M. P., Kan, Z., et al. (2023). Perceiving noise in daily life: How real-time sound characteristics affect personal momentary noise annoyance in various activity microenvironments and times of day. *Health & Place, 83*, Article 103053.

Cao, R., Liao, C., Li, Q., et al. (2023). Integrating satellite and street-level images for local climate zone mapping. *International Journal of Applied Earth Observation and Geoinformation, 119*, Article 103323.

Chang, D., Chen, X., Bai, Y., et al. (2025). Soundscape co-governance for a healthy megacity. *NPJ Urban Sustainability, 5*(1), 30.

Chen, S., He, P., Yu, B., et al. (2024). The challenge of noise pollution in high-density urban areas: Relationship between 2D/3D urban morphology and noise perception. *Building and Environment, 253*, Article 111313.

Chew, Y. R., & Wu, B. S. (2016). A soundscape approach to analyze traffic noise in the city of Taipei, Taiwan. *Computers, Environment and Urban Systems, 59*, 78–85.

Chu, S., Wang, Y., Liu, J., et al. (2025). Pandemic-induced shifts in urban noise perception: A spatial-temporal analysis in a Chinese megacity. *Journal of Environmental Management, 391*, Article 126584.

Gu, T., Zhao, H., Yue, L., et al. (2025). Attribution analysis of urban social resilience differences under rainstorm disaster impact: Insights from interpretable spatial machine learning framework. *Sustainable Cities and Society, 118*, Article 106029.

Guo, L. H., Cheng, S., Liu, J., et al. (2022). Does social perception data express the spatio-temporal pattern of perceived urban noise? A case study based on 3,137 noise complaints in Fuzhou, China. *Applied Acoustics, 201*, Article 109129.

Han, Y., Zhong, T., Yeh, A. G. O., et al. (2023). Mapping seasonal changes of street greenery using multi-temporal street-view images. *Sustainable Cities and Society, 92*, Article 104498.

Hou, C., Zhang, F., Kang, Y., et al. (2025). Transferred bias uncovers the balance between the development of physical and socioeconomic environments of cities. *Annals of the American Association of Geographers, 115*(1), 148–166.

Huang, J., Fei, T., Kang, Y., et al. (2024). Estimating urban noise along road network from street view imagery. *International Journal of Geographical Information Science, 38*(1), 128–155.

Huang, Y., Zhang, F., Wu, L., et al. (2025). Measuring urban physical environments using image deep features. *Cities, 166*, Article 106196.

Jiao, Y., Wang, Z., Li, C., et al. (2025). How to enhance urban noise management: Exploring the influencing factors of noise complaints at multiple scales integrating citizen perception. *Environmental Impact Assessment Review, 112*, Article 107783.

Ke, E., Zhao, J., & Zhao, Y. (2025). Investigating the influence of nonlinear spatial heterogeneity in urban flooding factors using geographic explainable artificial intelligence. *Journal of Hydrology, 648*, Article 132398.

Lan, Z., Li, F., & Cai, M. (2024). Road traffic noise exposure assessment based on spatiotemporal data fusion. *Transportation Research Part D: Transport and Environment, 127*, Article 104044.

Liu, Y., Oiamo, T., Rainham, D., et al. (2021). Integrating random forests and propagation models for high-resolution noise mapping. *Environmental Research, 195*, Article 110905.

Ma, H., Xu, Q., & Zhang, Y. (2023). High or low? Exploring the restorative effects of visual levels on campus spaces using machine learning and street view imagery. *Urban Forestry & Urban Greening, 88*, Article 128087.

Ma, H., Zhang, Y., Cui, Q., Kwan, M.-P., Cai, J., et al. (2025). Perceived distance to greenery affects psychological restoration. *Environment and Planning B: Urban Analytics and City Science, 0*(0). https://doi.org/10.1177/23998083251362610

Münzel, T., Schmidt, F. P., Steven, S., et al. (2018). Environmental noise and the cardiovascular system. *Journal of the American College of Cardiology, 71*(6), 688–697.

Song, J., Zhou, S., Kwan, M. P., et al. (2025). Association between real-time noise exposure in broader activity contexts and job satisfaction: Evidence from Guangzhou, China. *Cities, 161*, Article 105912.

Song, L., Liu, D., Kwan, M. P., et al. (2024). Machine-based understanding of noise perception in urban environments using mobility-based sensing data. *Computers, Environment and Urban Systems, 114*, Article 102204.

Song, L., Zhang, A., & Kwan, M. P. (2026). Exploring the influencing factors of noise complaints in New York City based on an interpretable machine learning model. *Environmental Impact Assessment Review, 116*, Article 108100.

Wang, L., Zhang, T., He, J., et al. (2025). Cross-platform complementarity: Assessing the data quality and availability of Google street view and Baidu street view. *Transactions in Urban Data, Science, and Technology, 4*(1), 22–47.

Wu, H., Wang, C., Liu, Y., et al. (2025). Spatiotemporal analysis of neighborhood noise complaints and related factors: A case study in Beijing, China. *Building and Environment, 113332*.

Xie, E., Wang, W., Yu, Z., et al. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems, 34*, 12077–12090.

Xing, Z., Yang, S., Zan, X., Dong, X., Yao, Y., Liu, Z., & Zhang, X. (2023). Flood vulnerability assessment of urban buildings based on integrating high-resolution remote sensing and street view images. *Sustainable Cities and Society, 92*, Article 104467.

Xue, W., Liang, C., Hao, M., et al. (2024). Rapid method for binary image-based road traffic noise mapping. *Transportation Research Part D: Transport and Environment, 135*, Article 104392.

Zhang, Y., Chen, N., Du, W., et al. (2021). Multi-source sensor based urban habitat and resident health sensing: A case study of Wuhan, China. *Building and Environment, 198*, Article 107883.

Zhang, Y., Kwan, M. P., Cai, J., et al. (2025). CrossGraphNet: A cross-spatiotemporal graph-based method for traffic speed reconstruction using remote sensing vehicle detection. *International Journal of Digital Earth, 18*(1), Article 2528640.

Zhang, Y., Kwan, M. P., & Ma, H. (2024). Sensing noise exposure and its inequality based on noise complaint data through vision-language hybrid method. *Applied Geography, 171*, Article 103369.

Zhang, Y., Liu, P., & Biljecki, F. (2023). Knowledge and topology: A two layer spatially dependent graph neural networks to identify urban functions with time-series street view image. *ISPRS Journal of Photogrammetry and Remote Sensing, 198*, 153–168.

Zhang, Yan, & Kwan, Mei-Po. (2025). Considering human mobility is essential for accurate environmental exposure assessment. *Sci. Bull., 70*(21), 3457–3460. https://doi.org/10.1016/j.scib.2025.07.016. ISSN 2095-9273, https://www.sciencedirect.com/science/article/pii/S2095927325007364.

Zhao, T., Liang, X., Biljecki, F., et al. (2025). Quantifying seasonal bias in street view imagery for urban form assessment: A global analysis of 40 cities. *Computers, Environment and Urban Systems, 120*, Article 102302.

Zhao, T., Liang, X., Tu, W., et al. (2023). Sensing urban soundscapes from street view imagery. *Computers, Environment and Urban Systems, 99*, Article 101915.

Zhou, Z., Zhang, M., Gao, X., et al. (2024). Analysis of traffic noise spatial distribution characteristics and influencing factors in high-density cities. *Applied Acoustics, 217*, Article 109838.