Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Low data image based urban noise estimation

Priyanshu Sarma-Sarkar

priyanshusarmasarkar@gmail.com

Maulana Abul Kalam Azad University of Technology, West Bengal

Rajkumar Saini

Luleå University of Technology

Research Article

Keywords: Urban noise prediction, Low-data learning, Semantic segmentation , Machine learning, Image analysis

Posted Date: August 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8807523/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Low data image based urban noise estimation

Priyanshu Sarma-Sarkar[1] and Rajkumar Saini[2]

[1] Maulana Abul Kalam Azad University of Technology, West Bengal, India
[2] Luleå University of Technology, Luleå, Sweden

Abstract. Urban noise pollution represents a critical environmental challenge with significant public health implications. Traditional noise monitoring requires expensive sensor networks, limiting scalability. We introduce a machine learning framework predicting background noise from casual smartphone imagery with only 400 samples. Our pipeline combines transformer-based semantic segmentation (SegFormer) with interpretable hand-crafted features (natural/manmade proportions, movable-object metrics, visual complexity) fed into Support Vector Regression. Critically, we introduce perceptually-grounded evaluation based on the ± 3 dB just-noticeable difference threshold established in psychoacoustic research. Our framework requires dramatically fewer resources than large-scale deep learning approaches, enabling scalable citizen-science driven urban acoustic monitoring.

Keywords: Urban noise prediction, Low-data learning, Semantic segmentation, Machine learning, Image analysis

1 Introduction

Environmental noise pollution has become one of the most pervasive forms of urban pollution, yet remains significantly underestimated compared to air and water pollution [13]. The WHO classifies it as a major public health concern, linking chronic exposure to cardiovascular disease, sleep disturbance, cognitive impairment in children, and reduced quality of life [13]. Unlike visible pollutants, noise is intangible and transient, making exposure difficult for residents to assess [8]. Current monitoring relies on two paradigms: fixed sensor networks of calibrated sound-level meters, and computational simulation models incorporating traffic data, meteorological variables, and 3D building geometries [8]. Both provide accurate measurements but face substantial limitations in scalability, cost, and deployment complexity in resource-constrained or rapidly urbanizing regions [8].

An emerging alternative leverages the strong correlation between urban acoustic and visual environments: a busy intersection filled with buses and pedestrians exhibits both distinctive visual patterns and elevated noise, while a tree-lined residential street appears calm and produces lower ambient sound, with vegetation reducing activity and absorbing sound [8]. If machine learning models can

reliably map visual scene characteristics to acoustic intensity, ubiquitous smartphone cameras could enable democratized large-scale noise mapping without specialized hardware. Huang et al. [8] demonstrated this feasibility using ResNet-34 on 46,000+ Baidu Street View panoramas (MAE≈2.01 dB, R^2 ≈0.64), but required 116,000+ paired samples collected via vehicle-mounted equipment, structured panoramic imagery, and substantial compute— barriers that preclude low-resource or citizen-driven deployment.

Our primary contributions are as follows:

- **Low-Data Framework:** Effective noise prediction from ∼400 casual smartphone images, approximately 300× fewer samples than prior deep learning approaches [8].
- **Unconstrained Image Capture:** Operates on everyday photographs with diverse angles, lighting, and framing, eliminating the need for structured panoramic imagery [8].
- **Interpretable Features:** Compact representation combining transformer-based segmentation with hand-crafted descriptors (natural/manmade proportions, movable-object metrics, gradient strength).
- **Perceptual Evaluation:** Adoption of the ±3 dB just-noticeable difference (JND) threshold [9], ensuring reported accuracy reflects human perceptual relevance.
- **Comparative ML Analysis:** Systematic comparison of SVR, GMM, XGBoost, and Random Forest under identical low-data constraints [8].

2 Related Work

Urban acoustic environments correlate strongly with visual characteristics, motivating vision-based noise estimation. Huang et al. [8] pioneered this direction using ResNet-34 on 116,000+ Baidu Street View panoramas coupled with Random Forest regression, achieving strong accuracy (MAE≈2.01 dB, R^2 ≈0.64). While demonstrating technical feasibility, their methodology imposes substantial barriers—specialized vehicle-mounted infrastructure, massive paired datasets, and significant computational resources—that preclude deployment in resource-constrained or citizen-science settings.

In low-data regimes, Support Vector Machines excel through structural risk minimization and kernel-based nonlinear modeling [5], with Abu Zohair [1] demonstrating their superiority over neural networks when labels are scarce. SVR's ϵ-insensitive loss additionally tolerates small prediction errors without penalty [7,12]. Ensemble methods offer complementary strengths: Random Forest reduces variance through bootstrap aggregation [2], XGBoost achieves strong performance via regularized gradient boosting [3], and Gaussian Mixture Models capture natural clustering in heterogeneous urban scenes [11].

Traditional metrics (MAE, RMSE, R^2) quantify numerical deviation without regard to human perception. McShefferty et al. [9] established through psychoacoustic experiments that variations below ±3 dB fall beneath the just-noticeable difference (JND) threshold, meaning small numerical errors are perceptually

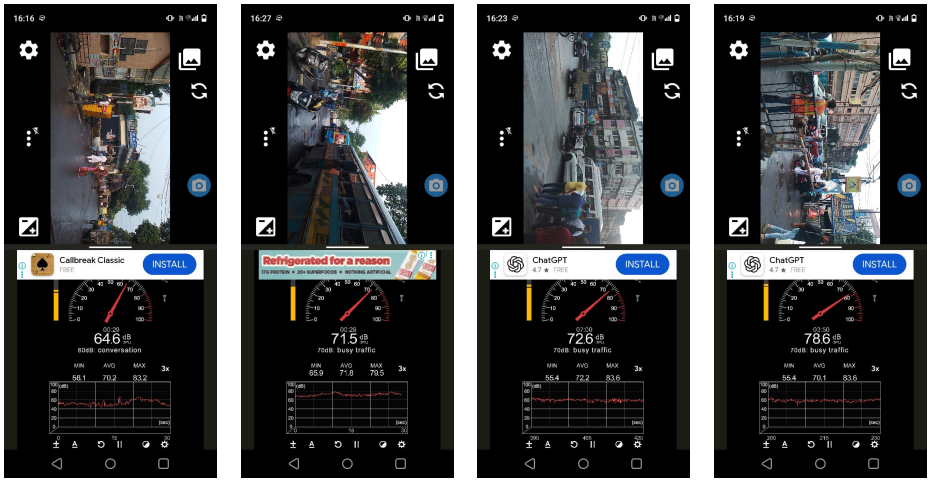


Fig. 1. Sample screenshots captured using a split-screen setup, where the upper section shows the camera view and the lower section displays the environmental noise level reading.

meaningless. Building on this, we adopt JND-grounded evaluation ensuring reported performance reflects genuine practical utility. Unlike Huang et al., our framework operates on ∼400 casually captured smartphone images without specialized infrastructure, combining transformer-based segmentation with interpretable features evaluated against psychoacoustically grounded criteria [8,9].

3 Data Collection and Preprocessing

Collecting paired visual-acoustic datasets traditionally requires synchronized sensor systems mounted on vehicles, imposing prohibitive costs for citizen-science initiatives [13]. We developed a lightweight protocol using commodity Android smartphones with a split-screen configuration: the upper half displays the camera viewfinder while the lower half shows a real-time sound meter with instantaneous decibel readings. A single screenshot simultaneously captures the visual scene and corresponding noise level, ensuring temporal synchronization.Screenshots are saved with timestamp-based filenames for traceability.

Raw screenshots underwent two-stage preprocessing. First, the visual scene (upper half) was cropped to eliminate the sound meter display and resized to a standardized 16:9 aspect ratio. Second, decibel readings were manually transcribed into a structured CSV and verified by independent annotators for accuracy.

The final dataset comprises 400 image-noise pairs with deliberate diversity across four dimensions: spatial (high-traffic intersections, residential streets, urban parks, semi-rural zones), temporal (daytime 10:00–16:00, evening 17:00–19:00, nighttime 20:00–22:00), capture quality (variable framing, inconsistent

lighting, motion blur, occlusions), and acoustic range (42.9–82.6 dB(A), from quiet residential to busy commercial zones). Unlike curated street-view datasets, this variability provides a realistic testbed for low-data learning under unconstrained conditions [13].

4 Methodology

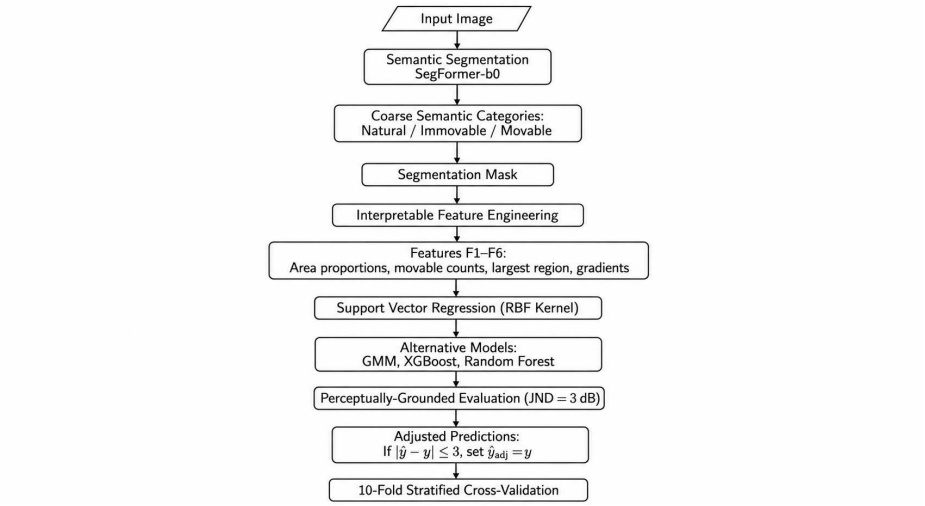


Fig. 2. Flowchart summarizing the proposed methodology, including segmentation, interpretable features, regression models, perceptual evaluation, and cross-validation.

The proposed framework as shown in Fig. 2 consists of three components: (i) semantic segmentation for scene understanding, (ii) interpretable feature engineering combining segmentation outputs with hand-crafted descriptors, and (iii) lightweight regression models optimized for small-data regimes.

4.1 Semantic Segmentation via SegFormer

Architecture Selection. We employ SegFormer-b0 [14], a transformer-based semantic segmentation model pre-trained on ADE20K. SegFormer offers advantages for low-data settings:

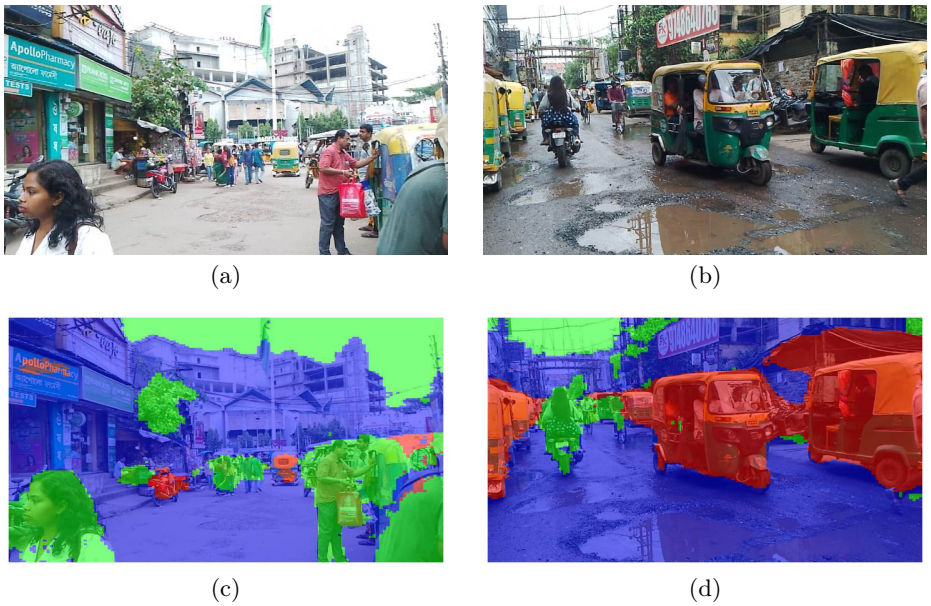
- **Hierarchical Encoder:** Multi-scale feature extraction (4 stages with progressively increasing receptive fields) captures fine details and global context without convolutional inductive biases [14]
- **Lightweight Decoder:** All-MLP design reduces parameters while maintaining performance, critical for computational efficiency when processing 400 images multiple times during cross-validation [14]

– **Pre-trained Representations.** ADE20K pre-training provides robust initialization for urban scenes, reducing overfitting risk [14]

Segmentation Processing For each input image, SegFormer produces pixel-wise class probability map. We apply argmax classification to obtain segmentation mask. To simplify feature extraction and enhance interpretability, we group ADE20K's granular classes into three coarse semantic categories aligned with acoustic relevance as shown in Fig. 3

– **Natural:** Vegetation (trees, grass, plant), sky, water, people
– **Manmade Immovable:** Buildings, roads, sidewalks, walls, bridges
– **Manmade Movable:** Vehicles (cars, buses, motorcycles)

This grouping reduces noise from fine-grained errors while preserving semantically meaningful distinctions.



(a) (b)

(c) (d)

Fig. 3. (a) and (b) displays sample street scenes, while (c) and (d) below show their corresponding annotated versions. In the annotated images, natural elements are highlighted in green, manmade movable objects in red, and manmade immovable structures in blue.

4.2 Interpretable Feature Engineering

From coarse segmentation masks, we extract six interpretable features capturing visual characteristics hypothesized to correlate with urban noise:

- **F1. Natural Area Proportion**
  Hypothesis: Greater vegetation correlates with quieter environments due to reduced activity and physical sound absorption.
- **F2. Manmade Movable Proportion**
  Hypothesis: Higher vehicle/pedestrian proportions correspond to dynamic noise sources (engine rumble, tire friction, vocalization).
- **F3. Manmade Immovable Proportion**
  Hypothesis: Infrastructure density indicates commercial/residential zones with characteristic baseline noise
- **F4. Movable-Object Count Score**
  Connected component analysis yields distinct regions. Sigmoid normalization:
  Hypothesis: Multiple distinct movable objects contribute additively to acoustic intensity
- **F5. Largest Movable Object Area**
  Let A_i denote pixel areas of detected regions, sorted descending:
  Hypothesis: Dominance of single large vehicle (bus, truck) may disproportionately contribute to noise spikes
- **F6. Gradient Strength**
  Sobel operator computes gradients:

$$G_x = S_x * I, \quad G_y = S_y * I \tag{1}$$

  where S_x, S_y are Sobel kernels, $*$ denotes convolution, I is grayscale image. Gradient magnitude:

$$|G| = \sqrt{G_x^2 + G_y^2} \tag{2}$$

  Feature is average magnitude:
  Hypothesis: Visual complexity proxies environmental busyness, cluttered scenes with many edges tend toward higher acoustic activity.

4.3 Support Vector Regression Model

Model Formulation. SVR extends SVM to continuous prediction via ϵ-insensitive loss [5]. Given training samples (x_i, y_i) where x_i is feature vector and y_i is noise label, SVR solves:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum(\xi_i + \xi_i^*) \tag{3}$$

subject to:

$$y_i - \langle w, \phi(x_i)\rangle - b \leq \epsilon + \xi_i, \quad \langle w, \phi(x_i)\rangle + b - y_i \leq \epsilon + \xi_i^* \tag{4}$$

where ϕ is nonlinear mapping induced by kernel, C controls trade-off between complexity and error, ϵ defines tolerance margin [5,7].

Kernel Selection. We employ Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{5}$$

RBF is well-suited due to: (i) nonlinear modeling (urban noise exhibits complex nonlinear relationships), (ii) smooth boundaries (desirable for acoustic estimation with gradual transitions), (iii) robustness to feature scale (RBF operates in transformed space where relative distances matter) [7].

Hyperparameter Optimization. Grid search over C, ϵ, γ via 10-fold cross-validation. Optimal: $\epsilon = 5.56$, $C = 40$, $\gamma = 0.01$. The large ϵ reflects design philosophy: prioritize avoiding large errors over perfect predictions, acknowledging small deviations are perceptually inconsequential [12].

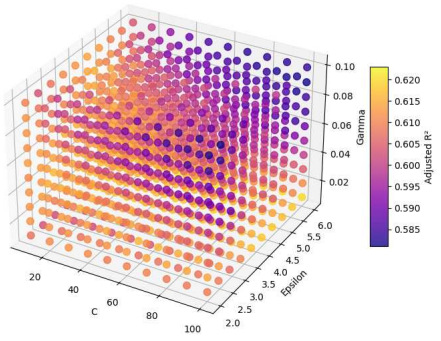


Fig. 4. 3D Scatter for C, ϵ, and γ showing Adjusted R^2 values as colours.

4.4 Alternative Machine Learning Models

Gaussian Mixture Model Regression. GMMs model joint distribution as weighted Gaussian sum [11]:

$$p(x, y) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x, y | \mu_k, \Sigma_k) \tag{6}$$

Prediction via conditional expectation. Advantageous when data clusters into subgroups (distinct urban contexts). Optimized via cross-validation, finding optimal $K = 5$.

XGBoost Regression. Employs gradient boosting with regularized objectives building tree ensembles iteratively [3]:

$$\mathcal{L} = \sum l(y_i, \hat{y}_i) + \sum \Omega(f_k) \tag{7}$$

Second-order gradient statistics enable accurate optimization. Optimal: 50 estimators, depth 3, learning rate 0.05.

Random Forest Regression. Constructs ensemble of trees, each trained on bootstrap sample with random feature subsets [2]: Bootstrap aggregation reduces variance while random selection decreases correlation. Optimal: 600 trees, depth 2.

4.5 Perceptually-Grounded Evaluation Framework

Motivation. Traditional metrics quantify numerical deviations without considering human perceptual capabilities [9]. This creates fundamental disconnect: 2 dB error is numerically measurable but perceptually meaningless if humans cannot detect such variations. McShefferty et al. [9] established through psychoacoustic experiments that just-noticeable difference (JND) for environmental noise is ±3 dB, variations below this threshold fall below perceptual resolution of human auditory systems.

Adjusted Predictions. For each prediction $\hat{y}$ and ground truth y, we compute adjusted predictions accounting for perceptual indistinguishability:

$$\hat{y}_{\text{adj}} = \begin{cases} y & \text{if } |\hat{y} - y| \leq 3 \\ \hat{y} & \text{otherwise} \end{cases} \tag{8}$$

Predictions within ±3 dB treated as perceptually correct (reset to true value); only errors exceeding JND penalized [9]. Standard metrics computed using adjusted predictions.

Advantages: (i) alignment with human experience, (ii) focus on consequential errors, (iii) practical relevance for public health applications [9].

4.6 Cross-Validation Strategy

Given limited dataset (400 samples), we employ 10-fold stratified cross-validation maximizing data efficiency: (i) samples divided into 10 folds maintaining equal noise level distribution, (ii) 9 subsets (360 samples) for training, 1 subset (40 samples) for validation, (iii) hyperparameter optimization via nested grid search within training set, (iv) performance metrics averaged across folds with standard deviations [1].

5 Experimental Results

5.1 Support Vector Regression Results

Table 1 summarizes SVR performance across 10 folds. Raw metrics show moderate accuracy (MAE=5.33 dB, RMSE=6.65 dB, $R^2 = 0.16$), while perceptually-adjusted metrics reveal substantially stronger performance ($R^2 = 0.62$), indicating the majority of raw errors fall within ±3 dB perceptual tolerance [9]. The 287% R^2 improvement underscores the importance of perceptually-grounded evaluation, and the low adjusted R^2 standard deviation (0.06) indicates stable, generalizable performance across folds [1].

Table 1. SVR 10-Fold Cross-Validation Results

Metric	Raw	Perceptually-Adjusted
MAE (dB)	5.33 ± 0.38	$\mathbf{2.93 \pm 0.31}$
RMSE (dB)	6.65 ± 0.43	$\mathbf{4.47 \pm 0.49}$
R^2	0.16 ± 0.13	$\mathbf{0.62 \pm 0.06}$

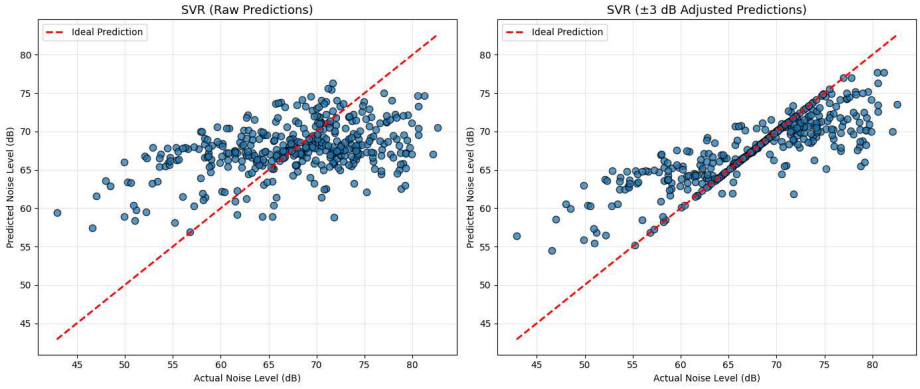


Fig. 5. Predicted vs. actual background noise levels (dB) for the SVR model using 10-fold cross-validation. The left plot shows raw SVR predictions, while the right plot shows predictions.

Table 2. Comparative Performance (Mean ± Std across 10 folds)

Model	Raw R^2	Adjusted R^2
SVR	0.16 ± 0.13	$\mathbf{0.62 \pm 0.06}$
XGBoost	0.15 ± 0.12	$\mathbf{0.61 \pm 0.06}$
GMM	0.14 ± 0.13	0.61 ± 0.07
Random Forest	0.04 ± 0.11	0.54 ± 0.09

5.2 Comparative Model Performance

Table 2 compares all four models. SVR leads with adjusted R^2=0.62, closely followed by XGBoost at R^2=0.61, suggesting that kernel-based methods and gradient-boosted approaches with explicit regularization are well suited to this low-data setting [1,3]. GMM performs competitively at 0.61, suggesting natural clustering structure in the data [11]. Random Forest underperforms (adjusted $R^2 = 0.54$), likely due to overfitting with the shallow trees required in this low-data setting [2]. Across all models, the dramatic raw-to-adjusted gap validates that prediction errors predominantly fall within the ±3 dB JND threshold [9].

Table 3. Comparison with Huang et al.

Method	Data	MAE (dB)	R^2	Note
Huang et al. [8]	116,000+ panoramas	2.01	0.64	Raw
Ours (SVR)	400 images	5.33	0.16	Raw
Ours (SVR)	400 images	2.93	0.62	Perceptually-adjusted

6 Discussion

6.1 Key Findings and Implications

Our results establish that useful urban noise predictions can be achieved with ∼400 loosely captured images—300× fewer samples than state-of-the-art deep learning approaches [8]. This data efficiency stems from semantic feature engineering that encodes domain knowledge directly, reducing the burden on learning algorithms to discover patterns from raw pixels, combined with small-data models (SVR, XGBoost) that avoid overfitting through explicit regularization [1]. The finding has direct implications for scalable urban sensing: citizen scientists and municipal agencies in resource-limited settings can deploy our framework without massive labeled datasets or specialized infrastructure [13].

Unlike end-to-end deep learning, our pipeline explicitly engineers interpretable features from semantic categories, object counts, and visual complexity. This sacrifices representational flexibility but offers transparency—predictions can be explained in human-understandable terms (e.g., "high noise due to multiple vehicles and visual clutter"), facilitating trust in public policy contexts [13]. Hand-crafted features grounded in stable semantic concepts may also generalize better across cities compared to low-level CNN texture patterns [4,6], and feature importance analysis enables targeted debugging [1]. The trade-off is a limited performance ceiling: an end-to-end model trained on 100,000+ images would likely capture subtle patterns (vehicle types, road textures) our coarse features miss [8].

The dramatic gap between raw ($R^2 = 0.16$) and perceptually-adjusted ($R^2 = 0.62$) metrics underscores a critical methodological point: evaluation criteria must align with application requirements [9]. For public health noise mapping, the relevant question is not sub-1 dB precision but whether the system distinguishes perceptually meaningful categories (quiet vs. moderate vs. loud) [13,9]. Crucially, our adjusted MAE of 2.93 dB falls below the ±3 dB JND threshold, meaning the perceptual adjustment reflects genuine accuracy rather than post-hoc rationalization [9].

6.2 Comparison with Prior Work

For applications tolerant of perceptual-level accuracy, our lightweight interpretable approach approximates deep learning performance while dramatically reducing resource requirements [1].

6.3 Robustness and Limitations

Our framework inherits a vulnerability from its dependence on SegFormer: test images exhibiting degradation (nighttime sensor noise, rain occlusion) not well-represented in training data will suffer reduced segmentation accuracy, cascading into prediction errors [4,10]. However, our interpretable design partially mitigates this—coarse semantic features like "proportion of natural area" are more robust to low-level corruption than raw pixel values (a blurry tree is still segmented as vegetation), grouping ADE20K's 1,038 classes into three coarse categories reduces sensitivity to fine-grained misclassification, and gradient strength may capture visual complexity even when segmentation degrades [6,4]. Systematic evaluation on corrupted test sets remains critical future work. Beyond robustness, three further limitations warrant attention. First, worst-performing predictions (errors >8 dB) predominantly involve nighttime or poorly-lit images; mitigation strategies include denoising preprocessing and low-light robust segmentation [10]. Second, our single-frame approach cannot capture temporal dynamics—an empty intersection photographed moments before a truck passes will yield low predictions against a high measured value; incorporating optical flow or time-of-day features could address this. Third, perspective bias means a nearby parked car occupies far more pixels than a distant highway with dozens of vehicles, despite the highway contributing more noise; monocular depth estimation could weight features by estimated distance. Additionally, stationary vehicles with running engines contribute acoustic energy that remains entirely invisible in static imagery.

Future directions include multi-modal fusion of sparse acoustic measurements with dense visual predictions [13], foundation models (DINOv2, CLIP) for improved distribution-shift robustness [4], active learning for intelligent sample selection [1], and spatiotemporal modeling via RNNs or graph neural networks to capture dynamic noise patterns [13].

7 Conclusion

This work demonstrates that urban background noise levels can be predicted from loosely captured street-level images even under severe data constraints. By combining semantic information extracted via transformer-based segmentation with interpretable hand-crafted visual features, the proposed framework achieves performance comparable to data-intensive deep learning approaches while requiring orders of magnitude fewer training samples. The results indicate that meaningful relationships between visual urban context and ambient sound levels can be captured without relying on large-scale curated datasets or end-to-end black-box models.

Evaluation using perceptually grounded criteria shows that the majority of predictions fall within the just-noticeable difference for sound level perception, supporting the practical relevance of the proposed approach. Overall, the framework provides a lightweight and accessible alternative for large-scale urban noise assessment, particularly in resource-limited or citizen-science scenarios.Future

work will focus on extending the approach to more diverse urban environments, improving generalization across cities, and incorporating temporal and multi-modal cues to better capture dynamic soundscape characteristics.

While our approach cannot match raw numerical precision of large-scale deep learning systems trained on 116,483+ samples, perceptually-adjusted results demonstrate that majority of prediction errors fall below human auditory thresholds, meaning the system achieves practical accuracy for urban noise mapping applications [9]. The dramatic performance gap between raw ($R^2 = 0.16$) and adjusted ($R^2 = 0.62$) metrics underscores a critical methodological insight: *evaluation criteria must align with application requirements and human perceptual capabilities* [9].

Looking forward, promising research directions include: (i) multi-modal fusion combining sparse acoustic measurements with dense visual predictions, (ii) active learning strategies for intelligent sample selection under labeling constraints [1], and (iii) spatiotemporal modeling to capture dynamic noise patterns across time and space. Our framework's low-resource requirements position it as an enabler for environmental justice initiatives, empowering communities to document and advocate against disproportionate noise pollution exposure [13].

In conclusion, this work establishes the feasibility of image-only, low-data urban noise prediction and opens pathways toward scalable, democratized environmental sensing. By combining domain-informed feature engineering, small-data optimized machine learning, and perceptually-meaningful evaluation, we demonstrate that practical acoustic monitoring can be achieved without the massive datasets and specialized infrastructure traditionally required.

8 Data Availability Statement

The source code and the dataset can be made available on request

References

1. Abu Zohair, L.M.: Prediction of student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education **16**(1) (Aug 2019). https://doi.org/10.1186/s41239-019-0160-3, http://dx.doi.org/10.1186/s41239-019-0160-3
2. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (Oct 2001). https://doi.org/10.1023/a:1010933404324, http://dx.doi.org/10.1023/a:1010933404324
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. ACM (Aug 2016). https://doi.org/10.1145/2939672.2939785, http://dx.doi.org/10.1145/2939672.2939785
4. Chhipa, P.C., Rodahl Holmgren, J., De, K., Saini, R., Liwicki, M.: Can self-supervised representation learning methods withstand distribution shifts and corruptions? In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). p. 4469–4478. IEEE (Oct 2023). https://doi.org/10.1109/iccvw60793.2023.00481, http://dx.doi.org/10.1109/iccvw60793.2023.00481

5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995). https://doi.org/10.1007/bf00994018, http://dx.doi.org/10.1007/bf00994018
6. Ericsson, L., Gouk, H., Hospedales, T.M.: Why do self-supervised models transfer? investigating the impact of invariance on downstream tasks (2022), https://arxiv.org/abs/2111.11398
7. Girma, H.G.: A tutorial on support vector machine. Tech. Rep. arXiv:0904.1276, arXiv (2009)
8. Huang, J., Fei, T., Kang, Y., Li, J., Liu, Z., Wu, G.: Estimating urban noise along road network from street view imagery. International Journal of Geographical Information Science **38**(1), 128–155 (Nov 2023). https://doi.org/10.1080/13658816.2023.2274475, http://dx.doi.org/10.1080/13658816.2023.2274475
9. McShefferty, D., Whitmer, W.M., Akeroyd, M.A.: The just-meaningful difference in speech-to-noise ratio. Trends in Hearing **20** (Jan 2016). https://doi.org/10.1177/2331216515626570, http://dx.doi.org/10.1177/2331216515626570
10. Rajnoha, M., Burget, R., Povoda, L.: Image background noise impact on convolutional neural network training. In: 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). p. 1–4. IEEE (Nov 2018). https://doi.org/10.1109/icumt.2018.8631242, http://dx.doi.org/10.1109/icumt.2018.8631242
11. Reynolds, D.: Gaussian mixture models. In: Li, S.Z., Jain, A. (eds.) Encyclopedia of Biometrics, pp. 659–663. Springer US, Boston, MA (2009). https://doi.org/10.1007/978-0-387-73003-5_196, https://doi.org/10.1007/978-0-387-73003-5_196
12. Wei, J., He, X.: Support vector regression model with variant tolerance. Measurement and Control **56**(9-10), 1705–1719 (2023). https://doi.org/10.1177/00202940231180620, http://dx.doi.org/10.1177/00202940231180620
13. World Health Organization: Environmental noise guidelines for the european region. Tech. rep. (January 2018), https://www.who.int/publications/i/item/9789289053563
14. Xie, E., et al.: Segformer: Simple and efficient design for semantic segmentation with transformers (2021), https://arxiv.org/abs/2105.15203