Remote Sensing Applications: Society and Environment 41 (2026) 101817

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

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

Automated detection and classification of bike lanes using multimodal imagery

Seung Jae Lieu [a,*], Bon Woo Koo [b], Uijeong Hwang [c], Subhrajit Guhathakurta [a]

[a] School of City and Regional Planning, Georgia Institute of Technology, 245 4th Street NW, Atlanta, GA, 30332, United States
[b] Department of Urban Planning and Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, Republic of Korea
[c] Atlanta Regional Commission, 229 Peachtree Street, Atlanta, GA, 30303, United States

ARTICLE INFO

Keywords:
Bicycle lanes
Computer vision
Multimodal
Street view imagery
Satellite imagery
Classification

ABSTRACT

Bike lanes are a critical element of urban infrastructure that promote cycling and support sustainable transportation goals. Effective planning and evaluation require comprehensive inventory datasets that both identify the locations of bike lanes and classify their types. However, existing data collection is limited by inconsistent municipal documentation practices and resource constraints. This paper introduces a computer vision–based approach for the automated detection and classification of bike lanes using publicly available multimodal imagery. Each data sample integrates two street view images, captured from opposite directions, with a corresponding satellite image, enabling complementary perspectives. This approach allows the model to reliably detect bike lane presence and distinguish between designated (marked lanes without physical barriers) and protected (lanes separated from traffic by physical barriers) types. To optimize performance, we conduct ablation experiments across three architectural dimensions: stage of modality concatenation, fusion strategy, and label structure. We also construct a training dataset using Google Street View and satellite imagery from 28 major U.S. cities to ensure broad applicability. Applying the model to over 1000 road segments in Atlanta, Georgia, we demonstrate its scalability and accuracy in a real-world urban setting. By providing an automated, transferable method for developing bike lane inventories, this research addresses a critical gap in infrastructure documentation and supports more effective planning of bicycle networks.

1. Introduction

Urban streets in most major U.S. cities have been designed to prioritize vehicular traffic, often at the expense of infrastructure for non-motorized transportation modes. Despite increasing awareness of the health, environmental, and mobility benefits of active transportation, and increasing efforts to promote bicycle use, cities continue to struggle with a low share of bicycle modes (Handy et al., 2014; Yang et al., 2021). This persistent challenge is largely driven by safety concerns and inadequate infrastructure (Dill, 2009; Hull and O'Holleran, 2014). In urban areas, these risks are amplified by high traffic volumes and limited bicycle facilities, many of which are poorly maintained or entirely absent (Buehler and Dill, 2016).

Transportation agencies and planners have invested in expanding bicycle infrastructure, but planning, prioritization, and evaluation efforts remain constrained by the lack of spatial data documenting where bike lanes are located and what types have been

* Corresponding author.
E-mail address: slieu3@gatech.edu (S.J. Lieu).

https://doi.org/10.1016/j.rsase.2025.101817
Received 6 October 2025; Received in revised form 18 November 2025; Accepted 1 December 2025
Available online 2 December 2025

implemented. Knowing the location of existing lanes is critical for creating continuous and connected networks, since gaps or discontinuities can discourage cycling and undermine safety (Caulfield et al., 2012; Pucher et al., 2010). Equally important is distinguishing between types of bike lanes, which differ considerably in their functionality and safety benefits. While designated lanes provide only painted separation from vehicles, protected lanes incorporate physical barriers that offer substantially greater security (Hwang and Guhathakurta, 2023). Comprehensive and reliable data on both the location and type of lanes is therefore fundamental to developing safe, connected, and effective cycling networks.

Recent advances in computer vision (CV) have enabled automated extraction of urban features from imagery, opening new opportunities for infrastructure inventory. However, bike lane detection and classification remain underdeveloped because of their small spatial footprint relative to roadways, variability in appearance, and frequent occlusion in imagery (Antwi et al., 2024; Ito and Biljecki, 2021). For instance, in street view images, bike lanes may be obstructed by parked cars, vegetation, or construction, while in satellite imagery they may be obscured by shadows or tree cover. Similar issues have been observed with other pedestrian facilities, such as sidewalks and crosswalks, where occlusion often results in omission or misclassification of features. Multimodal imagery analysis addresses these limitations by integrating complementary perspectives from satellite and street view imagery (Dong et al., 2020; Shen et al., 2018). Studies of pedestrian infrastructure have shown that multimodal approaches yield significantly higher accuracy than single-modality models, especially under challenging conditions such as dense tree cover or heavy shadowing (Mattyus et al., 2016; Ning et al., 2022; Zhang et al., 2025). Yet, despite these promising findings, bike lane detection and classification have not been

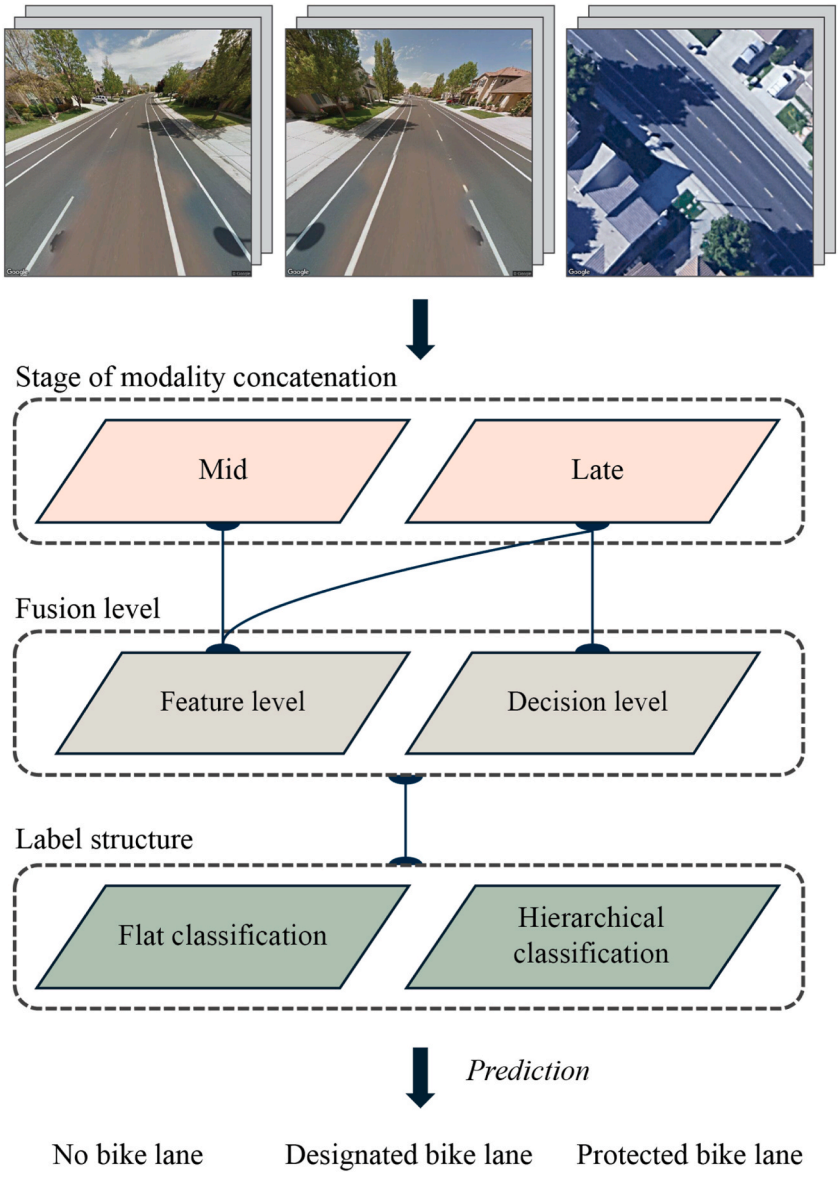

Fig. 1. Overview of model configuration dimensions for multimodal bike lane classification.

systematically examined using multimodal methods.

This study addresses this gap by developing a multimodal imagery analysis framework for classifying bike lane infrastructure into three categories: no bike lane, designated bike lane, and protected bike lane. Each data sample combines two street view images, captured from opposite directions, with a corresponding satellite image. Drawing on insights from the CV literature, we evaluate six alternative model architectures that vary across three dimensions: stage of modality concatenation, fusion strategy (feature-level versus decision-level), and label structure (flat versus hierarchical). Through systematic benchmarking these models (Fig. 1), we identify the most effective configuration and provide a framework that can be generalized for cities seeking to build comprehensive bike lane inventories. Another key contribution of this work is its reliance solely on publicly available imagery, making the approach scalable and transferable to cities lacking high-resolution or proprietary datasets. The findings offer practical insights into how multimodal image-based classification can support infrastructure monitoring and active transportation planning at scale.

2. Related work

2.1. Detection of street infrastructure in urban planning

The adoption of CV in urban planning has enabled cost-effective detection and classification of diverse street elements such as crosswalks, buildings, and vehicular lanes (Antwi et al., 2024; Hoang et al., 2019; Hoffmann et al., 2019; Kang et al., 2018; Lieu and Guhathakurta, 2025; Luttrell et al., 2024). In the context of bike infrastructure, however, the key task for planners extends beyond mere detection to classification, specifically distinguishing between designated (paint-only) and protected (physically separated) lanes. This distinction is well established in transportation research, as the type and quality of bicycle facilities strongly influence perceived safety and comfort (Hull and O'Holleran, 2014; Jones, 2012). Protected lanes, in particular, have been shown to significantly increase perceived safety and encourage cycling (Clark et al., 2019; Aldred and Dales, 2017). The distinction also carries practical implications for downstream applications such as low-stress bicycle network modeling. For instance, Hwang and Guhathakurta (2023) incorporated facility type into a simulation-based route choice model and found that protected bike lanes can reduce estimated traffic stress by up to 75 %, compared to 65 % for buffered lanes and 50 % for striped lanes.

Despite the importance of this distinction, most existing CV studies focus primarily on the detection of bicycle lanes rather than their type. These efforts have typically utilized single imagery sources, such as high-resolution aerial imagery to support transportation agencies (Antwi et al., 2025) or dashboard camera video for urban livability assessments (Agulto et al., 2023). A notable exception is Ding et al. (2021), who used crowdsourced street view imagery to classify bikeways into functional categories such as segregated and shared lanes for developing a bicycle routing service. While this work underscored the value of automated classification for practical applications, its reliance on a single modality (i.e., street view imagery) made it vulnerable to occlusions and incomplete visual information.

These challenges are not unique to bicycle infrastructure studies but persist in urban street element detection research, which often relies on a single imagery source, typically either satellite or street view imagery (Fang et al., 2022; Hosseini et al., 2022; Liu et al., 2023; Singh et al., 2024). This single-modality approach, however, introduces limitations stemming from the inherent constraints of each data type. Street view imagery provides fine-grained detail suitable for detecting markings, signage, and lane demarcations, but is frequently hindered by occlusions from parked vehicles or vegetation (Lieu and Guhathakurta, 2025). In contrast, satellite imagery offers broader spatial context and captures the continuity of road layouts, which is essential for network-level analysis. However, satellite images are often affected by occlusion from tree canopies, shadows, or other objects, as well as inconsistent lighting conditions that limit classification accuracy (Hosseini et al., 2022; Senlet and Elgammal, 2012).

To address these limitations, recent research has turned toward multimodal imagery fusion, which integrates satellite and street-level data to leverage their complementary strengths (Dong et al., 2020; Mattyus et al., 2016; Ning et al., 2022; Zhang et al., 2025). For example, Luttrell (2024) developed a "dual-perspective prediction model" for detecting crosswalks, combining aerial and street view images of the same location. This approach significantly improved performance under conditions of heavy occlusion, increasing accuracy by 49 % compared to the aerial-only model. Similarly, Cao et al. (2018) demonstrated the effectiveness of multimodal integration for urban land-use classification by combining aerial imagery with spatially interpolated features derived from street view data. Their study highlighted how ground-level perspectives can resolve ambiguities that overhead views alone often misclassify.

Despite these advances, no studies to date have applied multimodal imagery analysis specifically to the detection and classification of bike lanes. As a result, there is no established guidance on how best to integrate satellite and street view imagery for this task, nor on the relative effectiveness of different architectural strategies for distinguishing between designated and protected lanes. Furthermore, much of the existing research on related street elements relies on proprietary or high-resolution aerial imagery that is not publicly accessible, which limits scalability and real-world adoption. Consequently, urban planners currently lack robust and generalizable tools for evaluating not only the presence but also the types of cycling infrastructure, constraining their ability to develop continuous, safe, and effective bike networks.

2.2. Model architectural dimensions

To effectively utilize multimodal imagery for bike lane classification, three key architectural configurations require investigation: the stage at which modalities are concatenated, the fusion level, and the label structure. The computer vision literature remains inconclusive on the optimal approach for combining different data sources, largely due to the heterogeneity in perspective, geometry, and visual content across modalities. A systematic exploration of these dimensions is therefore critical to understand how best to

integrate complementary information for this task.

Concatenation stage refers to when representations from different modalities are combined in the network. Early-stage concatenates raw inputs before feature extraction, mid-stage merges latent embeddings from separate feature extractors at intermediate layers, and late-stage keeps modalities fully separate until their features or predictions are integrated. Recent studies show late-stage strategies often employ parallel network branches that process each modality independently before combining them using sophisticated feature fusion modules (Guo et al., 2024a; Zhao et al., 2023). This approach allows for the integration of distinct feature types, such as combining global context from satellite imagery with dynamic local details from street-view data to identify traffic accident hotspots or fusing complementary information from different views to improve remote sensing scene classification. Mid-stage is widely favored for its balance between modularity and joint representation learning (Guo et al., 2024b). Studies using mid-stage concatenation typically deploy separate feature extractors for each modality and integrate latent features at an intermediate layer for joint training (Fan et al., 2022; Luo et al., 2024). This approach has also demonstrated superior performance in fine-grained land use classification and urban scene understanding tasks due to its capacity to retain modality-specific strengths while enabling cross-modal synergy.

Fusion level refers to how modalities are integrated once their features are extracted. While concatenation timing determines when fusion occurs in the processing pipeline, fusion level specifies how deep in the representation hierarchy the integration takes place: feature-level fusion or decision-level fusion. Feature-level fusion combines latent representations (feature vectors) to learn a joint representation, which is effective for capturing cross-modal interactions. Studies have employed feature-level fusion to integrate complementary information by fusing image-based and auxiliary inputs such as digital surface models, thermal, or infrared data at the feature fusion stage (Cao et al., 2018; Filho et al., 2023; Li et al., 2024; Workman et al., 2017). However, this approach may struggle with highly dissimilar geometries like aerial and street views, where feature misalignment can degrade performance (Hoffmann et al., 2019). On the other hand, decision-level fusion combines outputs of independently processed modalities, such as their class probabilities or final predictions, to make a final decision. This approach often uses ensemble methods and can prioritize the more confident modality, offering robustness with noisy or incomplete data. Implementations include model blending, which averages the probability outputs of a diverse ensemble of models, and weighted voting, where the "vote" of each classifier is weighted by its accuracy on specific classes (Hoffmann et al., 2019; Shen et al., 2018). A practical application of this is found in urban feature detection, where a dual-perspective method was developed to identify occluded crosswalks by combining the class probabilities from two separate models (i.e., one trained on aerial-view imagery and the other on street-view imagery) using a soft voting function to produce a more robust final prediction (Zhang et al., 2025).

The label structure defines how target classes are organized, typically as either flat or hierarchical. A flat structure assigns each instance to a single, mutually exclusive category. It is well-suited for classification tasks with broad, clearly defined categories and has been used to distinguish building functions such as commercial, residential, public, and industrial types by leveraging decision-level fusion of aerial and street view images (Hoffmann et al., 2019), as well as for pixel-wise semantic segmentation using infrared and DSM imagery (Audebert et al., 2017). In contrast, hierarchical label structures organize categories across multiple levels, capturing nested or context-dependent relationships. Examples include using unmanned aerial vehicle imagery and DSM with fuzzy logic to refine urban land use and land cover classification (Gibril et al., 2020; Shackelford and Davis, 2003).

3. Methods and data

Building on the architectural dimensions identified in section. 2.2, we systematically evaluate six multimodal models that vary along three axes: (1) stage of modality concatenation, (2) fusion level, and (3) label structure. The objective of this methodological design is not to re-define these concepts but to operationalize them for the task of bike lane classification and to assess their relative effectiveness in an applied urban context.

To maintain comparability across all experiments, we employ the Swin Transformer (Swin-S) as the backbone feature extractor for every model. The Swin-S is pretrained on ImageNet and selected for its ability to capture hierarchical feature representations and to model long-range spatial dependencies more effectively than conventional convolutional networks. We adopt a transfer learning approach, freezing the early layers of the backbone and fine-tuning only the last two stages together with a newly added classification head. This strategy leverages pretrained knowledge while allowing the models to adapt to the specific task of bike lane classification, ensuring that differences in performance can be attributed to architectural variations rather than discrepancies in feature extraction capacity.

The inputs to the models consist of three co-located RGB images: two street view images captured from opposite directions and one

Table 1
Model configurations across concatenation stage, fusion level, and label structure.

Model	Concatenation stage	Fusion level	Label structure
1	Late	Feature-level	Flat
2	Mid	Feature-level	Flat
3	Late	Feature-level	Hierarchical
4	Mid	Feature-level	Hierarchical
5	Late	Decision-level	Flat
6	Late	Decision-level	Hierarchical

satellite image of the same location. All images are resized to 384 × 384 pixels and geographically paired to ensure consistency across modalities. Additional details on the training dataset are provided in Section 3.2, while the training configuration details are described in Appendix A. Moreover, because model performance can vary across runs due to the stochastic nature of weight initialization, data shuffling, and optimization dynamics, we repeat training with five different random seeds for each model to obtain more robust and reliable performance estimates.

3.1. Model configurations across architectural dimensions

There are six different models spanning three architectural dimensions (Table 1). The first dimension concerns the stage of concatenation, where we implement both mid-stage and late-stage strategies. In the mid-stage concatenation setting, the two street view images are each passed through independent Swin-S backbones to generate latent feature embeddings (Fig. 2). These embeddings are concatenated and processed through a projection block composed of fully connected layers, producing a unified ground-level representation. The satellite image is also encoded through its own Swin-S backbone to generate a satellite feature embedding. The ground-level representation and the satellite embedding are then concatenated and passed through an additional fusion layer and the final classification head. This design ensures early interaction between the two street-level perspectives while deferring integration with aerial context until after each modality has been independently processed through its backbone.

In the late-stage concatenation setting, all three images are independently processed by parallel Swin backbones. The resulting features remain separate until the final stage of the model. At this point, we investigate two fusion levels: feature-level and decision-level (Fig. 3). For feature-level fusion, we extract latent features from each modality after their respective Swin backbones, concatenate these vectors along the feature dimension, and pass the combined embedding through a shared fully connected projection head. Regarding the decision-level fusion setting, each modality's extracted features are independently passed through a classification head to produce logits. These logits are then aggregated to produce a final prediction. We implemented and tested two ensemble strategies for this aggregation: (1) an element-wise max pooling, where the maximum value across the three modality logits is selected; and (2) a learnable weighted average, where modality-specific weights are optimized during training using softmax normalization. The

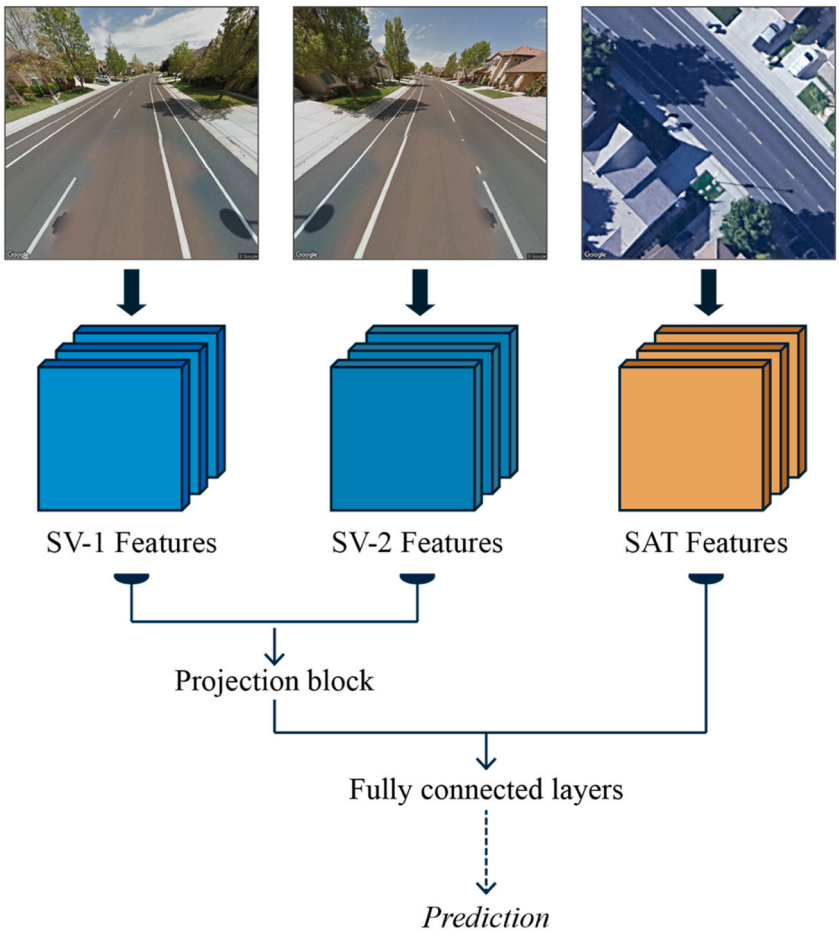

Fig. 2. Mid-stage concatenation: Feature-level fusion only. (SV = street view imagery; SAT = satellite imagery).

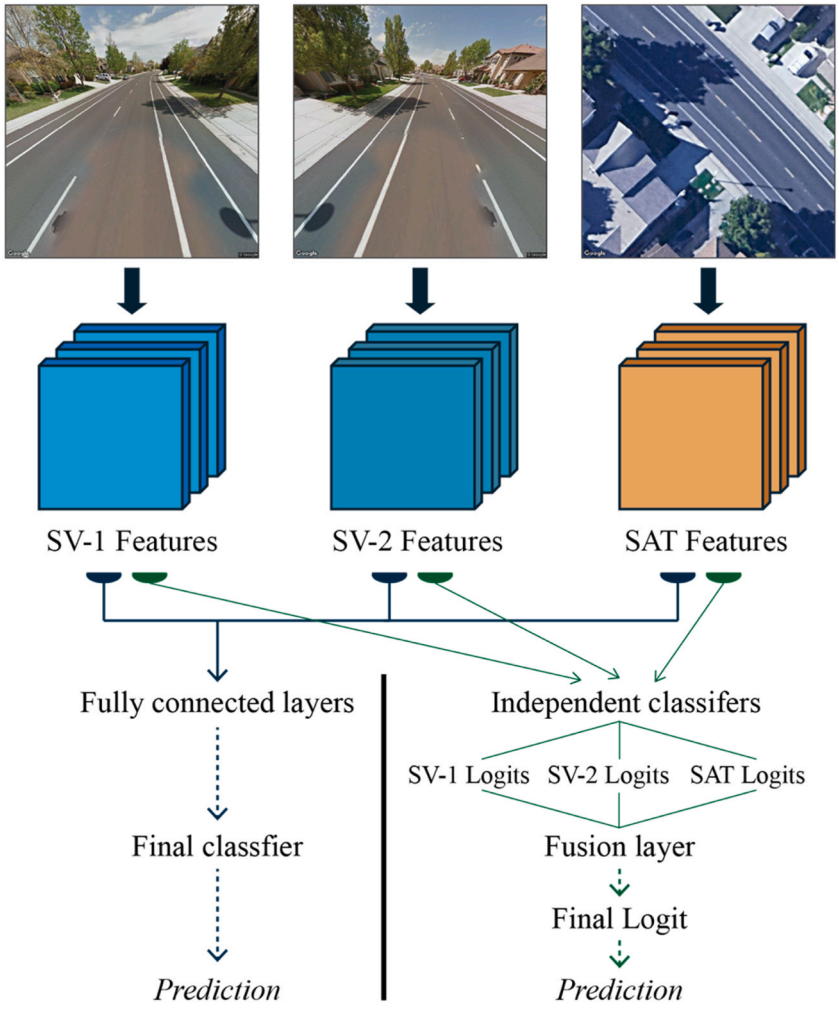

Fig. 3. Late-stage concatenation: Feature-level (left) and decision-level (right) fusion.

weighted strategy consistently outperformed max pooling, and we therefore adopt it for the decision-level fusion models. We do not consider combinations of mid-stage concatenation with decision-level fusion, as concatenation of features at the mid-stage inherently prevents independent classification pathways. Overall, the distinction between mid- and late-stage concatenation reflects alternative assumptions about the most effective integration point for multimodal context.

Finally, we evaluate two label structures. The flat label structure treats the task as a single-stage, three-class classification problem. Each sample is assigned one of the mutually exclusive categories: no bike lane, designated bike lane, or protected bike lane. In contrast, the hierarchical label structure decomposes the classification task into two sequential binary decisions. The first stage determines whether a bike lane is present (i.e., no bike lane vs. any bike lane). If present, the second stage classifies it as either designated or protected. This structure is operationalized through two stacked heads in the model: the first head outputs logits for the presence detection, and the second head is conditionally activated to distinguish between the two bike lane types.

3.1.1. Data

To construct a publicly accessible and geographically diverse dataset for bike lane classification, we identified candidate road segments across 28 major U.S. cities (listed in Appendix B) using OpenStreetMap (OSM). Within the OSM schema, bicycle infrastructure is annotated using the cycleway key, where segments labeled as 'cycleway = lane' represent bike lanes that are part of the roadway and typically separated from vehicular traffic by painted markings. While segments labeled as 'cycleway = track' denote physically separated bike lanes with barriers such as curbs or bollards. From these tagged segments, we randomly sampled road locations across both categories. For each sampled segment, we extracted the geographic midpoint and collected three complementary images: one satellite image and two opposing street-view images.

Satellite images were collected using the Google *Maps Static API* at a zoom level of 21, with each 640 × 640-pixel image centered on the target location. Using the Google Street View Static API, we captured two street view images per location from opposite directions along the roadway. Street view images were captured using a 120-degree field of view with a pitch of −30° and retrieved at a 640 ×

640-pixel resolution to maintain consistency with the corresponding satellite imagery.

Following image collection, we conducted manual annotation to verify and refine labels through visual inspection of each image set. Our annotation process focused on identifying visual evidence of bike lane infrastructure and classifying locations into three categories: no bike lanes, designated bike lanes, and protected bike lanes. For designated bike lanes, we identified six recurring visual patterns that indicate their presence (Fig. 4). These lanes typically run along road edges and are characterized by specific pavement markings following the distinct configurations. The most common involves dual parallel white lines delineating bike lane boundaries, with some variations incorporating diagonal hatching or chevron patterns between solid lines. A third pattern positions the bike lane between the travel lane and parked cars, using similar dual solid line markings. The remaining patterns employ single solid outer boundary lines combined with various visual reinforcements, including colored pavement, painted bicycle symbol, and closely spaced dashed buffer lines that are shorter and more densely packed than standard vehicle lane separators.

Protected bike lanes are distinguished from designated lanes by continuous physical barriers. Our dataset revealed two predominant configurations: raised infrastructure including concrete curbs or raised medians creating vertical separation, and on-street parking serving as a buffer between cyclists and moving traffic (Fig. 5). We excluded marginal physical elements such as occasional poles or small intermittent curbs from the protected category, classifying such infrastructure as designated bike lanes due to insufficient continuous physical separation.

To ensure balanced classification, we included randomly selected locations with no visible bike lane infrastructure in the same 28 cities. These locations were sampled from OSM road segments lacking cycleway annotation and manually verified for absence of relevant markings or barriers. The completed dataset encompasses 1800 unique street segments: 1036 locations with no bike lanes, 526 with designated bike lanes, and 238 with protected bike lanes. With three images per location, the dataset contains 5400 total images. For model training, we split each label's dataset into training and validation sets using a 7:3 ratio and upsampled the training datasets for protected and designated bike lane classes due to class imbalance.

4. Results

4.1. Comparison of unimodal and multimodal approaches

Before presenting the evaluation of the six multimodal model configurations, we first compare the multimodal approach against unimodal baselines that use either satellite imagery or street view imagery alone. For a fair comparison, we follow the same training configurations and backbone architecture as in the multimodal experiments. The unimodal satellite model requires only a single image and is therefore implemented with feature-level fusion under both flat and hierarchical label structures. The unimodal street view model requires two images with different headings and is evaluated across the two architectural dimensions described in Section 2.2, namely fusion level and label structure.

Table 2 reports the average performance of the three approaches—multimodal imagery, unimodal satellite, and unimodal street view—based on mean accuracy, macro-averaged precision, recall, and F1-score across five runs with different random seeds. Detailed

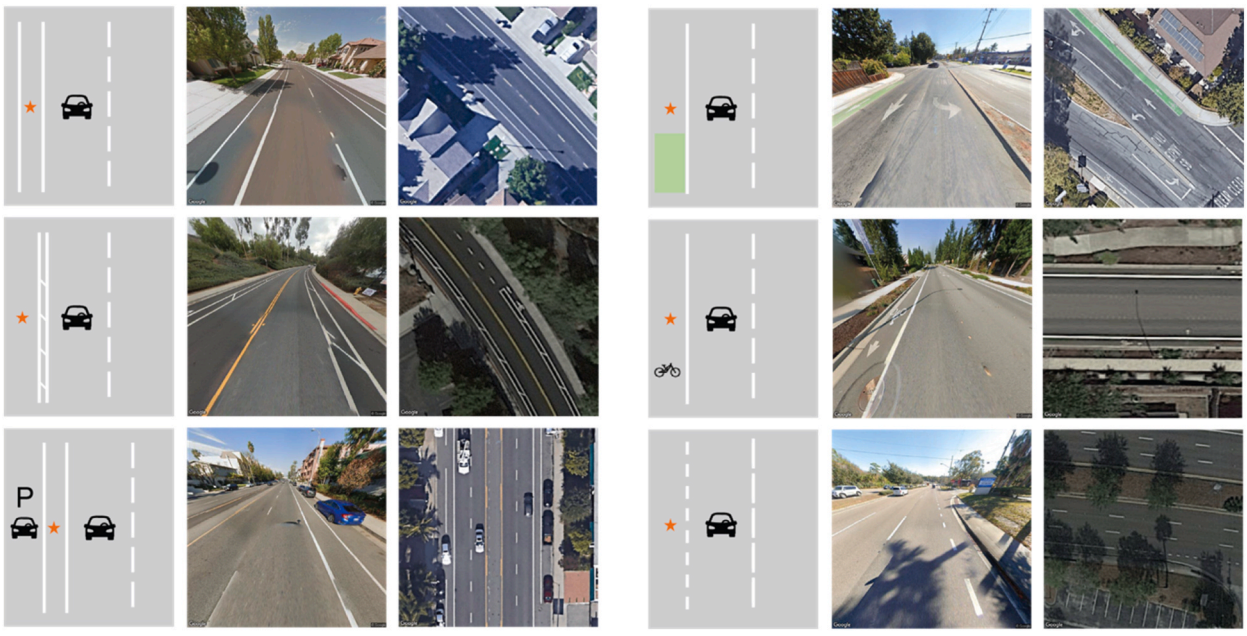

Fig. 4. Visual patterns of designated bike lanes. Each case shows: schematic diagram, street view, and satellite view of different bike lane configurations.

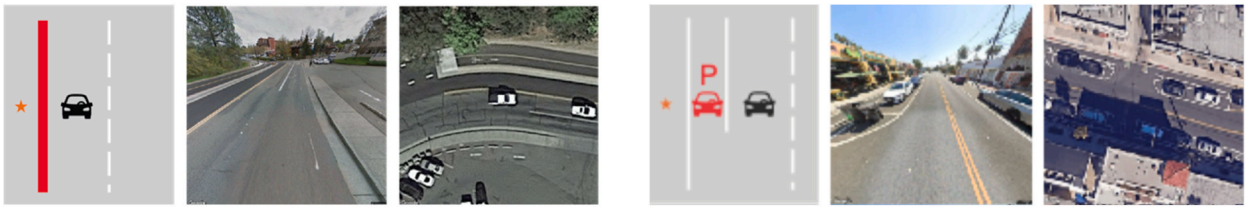

Fig. 5. Visual patterns of protected bike lanes.

results for each unimodal configuration are provided in Appendix C, while performance for the multimodal variants is reported in Table 3.

The comparison confirms that multimodal integration yields a performance gain relative to using either imagery source alone. While the street view-only model performs substantially better than the satellite-only model, which reflects the importance of ground-level detail for identifying lane markings and protective features, the multimodal model achieves the highest overall accuracy and other metrics. These findings demonstrate that combining complementary perspectives enhances classification robustness, which aligns with prior studies showing that dual perspective models outperform single-modality approaches in urban feature detection (Luttrell et al., 2024; Ning et al., 2022; Zhang et al., 2025).

4.2. Model evaluation

Table 3 presents the performance of the six model configurations, reported as the mean and standard deviation of accuracy, macro-averaged precision, recall, and F1-score across five training runs with different random seeds. Per-class performance metrics for each model are described in Appendix D. Overall, Model 6, which combines late-stage concatenation, decision-level fusion, and hierarchical label structure, achieved optimal performance with the highest accuracy (0.886), F1-score (0.855), and precision (0.870). These results indicate that both the fusion strategy and the hierarchical label structure meaningfully contribute to improved classification performance.

Model 5, which shares the same fusion and concatenation design but uses a flat label structure, demonstrated superior performance with an F1-score of 0.841 and accuracy of 0.876 compared to all configurations except Model 6. This result suggests that decision-level fusion offers consistent benefits regardless of label structure, though integration with a hierarchical design may slightly enhance accuracy. The hierarchical model's decomposition into two binary tasks (bike lane detection followed by type classification) may enable finer-grained learning, particularly when classes are visually similar.

Late-stage concatenation consistently outperformed mid-stage concatenation across all metrics. For example, Model 1 (late-stage, feature-level, flat) outperformed Model 2 (mid-stage, feature-level, flat) in both accuracy and F1-score. A similar trend is observed between Model 3 and Model 4. This finding supports the hypothesis that allowing each modality to be processed independently before integration yields more informative and specialized feature representations.

Fusion level also plays an important role. Models using decision-level fusion (Models 5 and 6) marginally outperform those using feature-level fusion (Models 1 and 3) in terms of both accuracy and precision. This suggests that aggregating predictions from independently trained classifiers allows the model to balance modality contributions more adaptively and reduce sensitivity to noisy or misaligned input. In particular, Model 6 incorporates a learnable weighted fusion mechanism, where modality-specific weights are optimized during training to reflect the relative contribution of each modality to classification. The final weights learned were 0.3216 for Street View 1, 0.3220 for Street View 2, and 0.3564 for satellite imagery. These values indicate a modest preference for satellite imagery, likely due to its superior ability to capture spatial layout and broader contextual cues, while still maintaining balanced contributions from both ground-level views.

4.2.1. Application to Atlanta Roads

To assess the practical utility of our best-performing model (Model 6: late-stage, decision-level fusion with hierarchical structure), we applied it to a real-world dataset of road segments in Atlanta, Georgia. Rather than including all city roads, we focused on segments prioritized for Complete Street design under the Complete Streets Policy Guideline proposed by the regional planning agency for the Atlanta region (Atlanta Regional Commission, 2019), which promotes multimodal infrastructure and the development of bike lanes. Following the guideline, we calculated a priority score for each census block group based on estimated walking and bicycling demand and propensity, identified arterial road segments intersecting high-priority areas, and selected those exceeding 100 m in length within

Table 2
Average performance of unimodal versus multimodal models.

Model	Accuracy	Precision	Recall	F1
Unimodal – satellite	0.792	0.764	0.746	0.748
Unimodal – street view	0.851	0.835	0.833	0.826
Multimodal	0.863	0.836	0.829	0.829

Table 3
Performance comparison of the six model configurations. Metrics are reported as mean values with standard deviations in parentheses, averaged over five independent training runs with different random seeds.

Model	Accuracy	Precision	Recall	F1
1	0.872 (0.021)	0.851 (0.034)	0.846 (0.017)	0.846 (0.022)
2	0.829 (0.020)	0.794 (0.026)	0.790 (0.024)	0.785 (0.026)
3	0.872 (0.025)	0.845 (0.042)	0.844 (0.029)	0.842 (0.036)
4	0.845 (0.013)	0.804 (0.024)	0.817 (0.021)	0.808 (0.019)
5	0.876 (0.021)	0.853 (0.033)	0.834 (0.030)	0.841 (0.030)
6	0.886 (0.017)	0.870 (0.023)	0.845 (0.026)	0.855 (0.019)

the top 10 % of the overall demand distribution. This process resulted in 1002 prioritized segments.

For each road segment, we collected three images using the same protocol as the training set. Ground-truth labels were established through manual inspection supplemented by Google Maps and street view imagery. While OSM data were initially considered, bicycle infrastructure coverage in Atlanta proved extremely limited, with tags available for only a small fraction of road segments. Moreover, as a crowdsourced database, OSM's untagged segments do not necessarily indicate the absence of bicycle facilities but rather reflect gaps in community documentation. We therefore relied on systematic manual verification to ensure completeness and accuracy of the validation dataset. This approach enabled rigorous assessment of model performance while demonstrating that our framework can provide more comprehensive infrastructure coverage than existing crowdsourced data sources. Of the total segments, 856 had no bike lane, 128 featured designated bike lanes, and 18 included protected bike lanes (Fig. 6). To prevent potential data leakage and ensure independence between training and application, road segments present in the training dataset were excluded from this application dataset.

Using a prediction confidence threshold of 0.9 applied to the final fused class probabilities, the model classified 519 segments as having no bike lane, 164 as designated, and 8 as protected. Among these predictions, the number of true positives was 515 for the class of no bike lane, 99 for designated lanes, and 7 for protected lanes. These values correspond to class-wise precision scores of 0.992 (no bike lane), 0.603 (designated), and 0.875 (protected). Although the model demonstrated excellent performance in identifying segments without bike lanes, its precision was significantly lower for segments that contained designated or protected infrastructure.

A closer examination of the misclassifications revealed several limitations. First, many errors were due to differences in visual characteristics between the Atlanta's road environment and the training data. For instance, several protected bike lanes in Atlanta utilized barrier types, such as tall barricades or concrete traffic barrier, not commonly represented in the training set. Designated lanes also exhibited low visual clarity due to faded markings, pavement cracks, or partially visible bike symbols, all of which contributed to

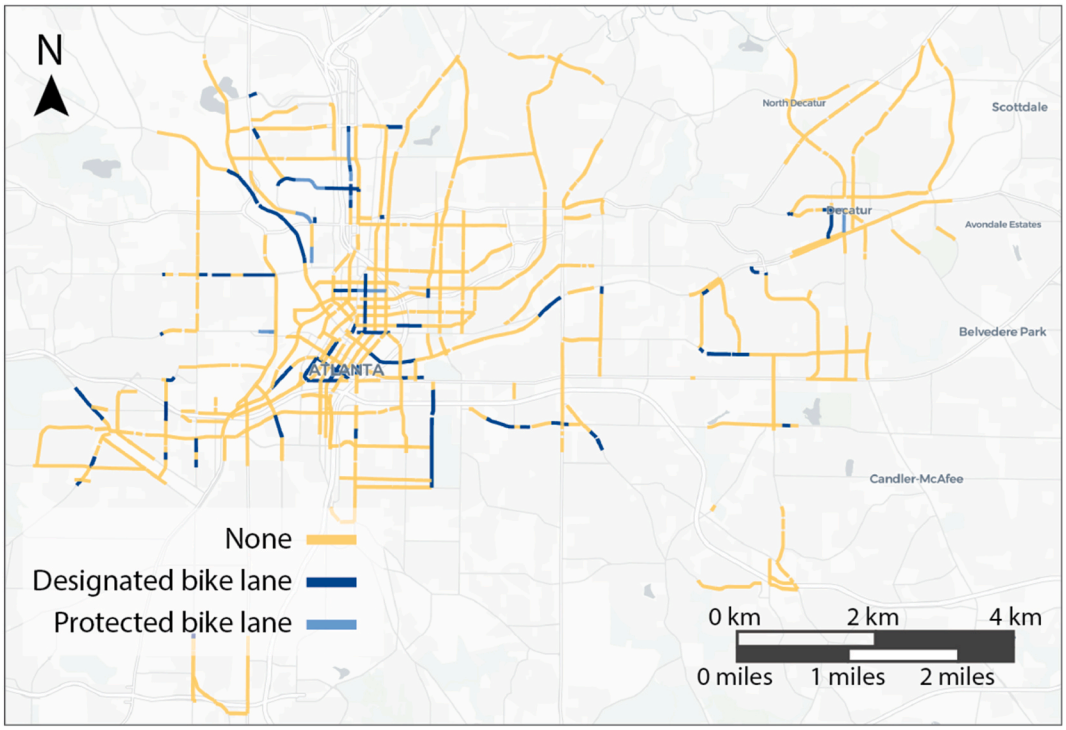

Fig. 6. Map of bike lane infrastructures in Atlanta, GA.

misclassification. These issues were particularly evident in residential and suburban areas where maintenance appeared lower and lane indicators were often incomplete or only partially captured in the imagery.

Second, misclassification occasionally occurred for roads with painted shoulders that resembled designated lanes but were not formally marked as bicycle facilities. To minimize this issue, such ambiguous cases were excluded from the training dataset during manual annotation. Only segments showing clear evidence of a designated facility such as bicycle icons, colored pavement, or distinctive double-line markings, were retained. We also observed that these visual cues are more consistently visible near intersections where bike lanes begin or end, while midblock segments sometimes lack explicit markings. Future data collection efforts could benefit from incorporating imagery closer to intersections to reduce this uncertainty.

The model's performance on bike lane classes was further constrained by the limited visual diversity within those categories in the training dataset. Although class imbalance was addressed through upsampling, this strategy merely increased the frequency of existing examples without introducing new design variations. Expanding the dataset to capture a broader range of design patterns, geographic contexts, and visual conditions would improve robustness and applicability across diverse environments. Furthermore, future work could include additional facility types such as shared-lane markings (sharrows). While this study focused on protected and designated lanes because of their well-documented safety benefits, other facility types remain highly relevant for planners and researchers. Additionally, this framework has the potential to extract more detailed attributes of bicycle infrastructure. For instance, by comparing opposite-direction street view images, it may be possible to infer whether a facility supports bidirectional travel. Incorporating such attributes would enhance the utility of automated inventories for network planning.

5. Conclusion

This study developed and evaluated a multimodal deep learning framework for classifying bike lane infrastructure through the integration of street view and satellite imagery. Through systematic evaluation of six model configurations across three architectural dimensions (modality concatenation stage, fusion level, and label structure), we determined that a late-stage, decision-level fusion model with hierarchical labeling achieved optimal performance. Application to 1002 road segments in Atlanta demonstrated the framework's practical utility for detecting protected, designated, and non-existent bike lanes. While the model exhibited reliable performance in identifying segments without bike lane infrastructure, classification accuracy diminished for segments containing bike lanes. Misclassifications of protected bike lanes frequently stemmed from unfamiliar design elements, particularly barrier configurations absent from the training data, whereas designated lane detection challenges primarily arose from diminished visual clarity attributable to faded markings and pavement deterioration.

Several limitations warrant consideration. The model's generalizability remains constrained by the geographic and typological scope of the training dataset. As evidenced in the Atlanta case study, performance degraded when confronted with barrier configurations or degraded lane markings underrepresented in the training corpus, underscoring the imperative for dataset expansion to encompass broader design variability, surface conditions, and regional contexts. Additionally, dependence on proprietary Google data presents accessibility challenges for cities in developing regions where coverage remains incomplete, thereby limiting the framework's immediate applicability in contexts where automated, cost-effective infrastructure mapping would yield the greatest benefit.

Notwithstanding these limitations, this research makes several contributions to advancing remote sensing and transportation planning. The framework demonstrates the potential of fusing multimodal imagery to extract fine-grained urban features that are difficult to capture using aerial or street-level imagery alone. The findings illustrate how multimodal integration can enhance the detection of bike lane by leveraging the complementary strengths of different data sources. For transportation planning applications, the framework provides a scalable method for documenting bicycle infrastructure, offering value for municipalities with limited resources. The ability to distinguish between infrastructure types supports data-driven evaluations of safety, equity, and network connectivity, helping planners assess the quality and distribution of existing facilities. Methodologically, the systematic architectural comparison shows that late-stage, decision-level fusion paired with hierarchical labeling yields the most effective configuration for multimodal bike lane classification. These insights provide planners and policymakers with a clearer understanding of how automated systems can be designed to capture the nuances between designated and protected lanes, which is an important distinction for evaluating safety, equity, and long-term network completeness. Ultimately, this work demonstrates the potential of multimodal deep learning to support sustainable, evidence-based transportation planning through automated infrastructure documentation.

The trained model checkpoint with the highest classification accuracy is available on GitHub (https://github.com/GT-CURA/complete_streets/tree/main/step2_elements/bike_lane), along with instructions for running the model.

CRediT authorship contribution statement

Seung Jae Lieu: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bon Woo Koo:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Uijeong Hwang:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Subhrajit Guhathakurta:** Writing – review & editing, Supervision, Conceptualization.

Ethical statement

The authors hereby declare that the manuscript "Automated Detection and Classification of Bike Lanes Using Multimodal Imagery" represents their original work, which has not been previously published nor is under consideration elsewhere. All authors have made

substantial contributions to the study and take full responsibility for its content. The research was conducted in accordance with the highest ethical standards, and all data used were obtained from publicly available and properly cited sources. The authors confirm that there are no conflicts of interest and that this submission fully complies with the ethical guidelines and publication policies of *Remote Sensing Applications: Society and Environment*.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rsase.2025.101817.

Data availability

Data will be made available on request.

References

Agulto, J.M., Noroña, Y.L., Gutierrez, G.C., Benabaye, E., Franco, G., Laguna, A.F., Cu, J., Ilao, J., Cordel, M., 2023. Development of a computer vision-based road physical feature extraction. In: 2023 IEEE International Conference on Big Data (BigData), pp. 1470–1475. https://doi.org/10.1109/BigData59044.2023.10386810.

Aldred, R., Dales, J., 2017. Diversifying and normalising cycling in London, UK: an exploratory study on the influence of infrastructure. J. Transport Health 4, 348–362. https://doi.org/10.1016/j.jth.2016.11.002.

Antwi, R.B., Kimollo, M., Takyi, S.Y., Ozguven, E.E., Sando, T., Moses, R., Dulebenets, M.A., 2024. Turning features detection from aerial images: model development and application on florida's public roadways. Smart Cities 7 (3). https://doi.org/10.3390/smartcities7030059. Article 3.

Antwi, R.B., Lawson, P.L., Kimollo, M., Ozguven, E.E., Moses, R., Dulebenets, M.A., Sando, T., 2025. Automated detection of pedestrian and bicycle lanes from high-resolution aerial images by integrating image processing and artificial intelligence (AI) techniques. ISPRS Int. J. GeoInf.

Atlanta Regional Commission, 2019. Walk, bike, thrive!: regional workbook for complete streets supplement report. https://cdn.atlantaregional.org/wp-content/uploads/arc-complete-streets-workbook-webview.pdf.

Audebert, N., Le Saux, B., Lefèvre, S., 2017. Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (Eds.), Computer Vision – ACCV 2016. Springer International Publishing, pp. 180–196. https://doi.org/10.1007/978-3-319-54181-5_12.

Buehler, R., Dill, J., 2016. Bikeway networks: a review of effects on cycling. Transp. Rev. 36 (1), 9–27. https://doi.org/10.1080/01441647.2015.1069908.

Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., Zhang, Q., Qiu, G., 2018. Integrating aerial and street view images for urban land use classification. Remote Sens. 10 (10), 1553. https://doi.org/10.3390/rs10101553.

Caulfield, B., Brick, E., McCarthy, O.T., 2012. Determining bicycle infrastructure preferences – a case study of Dublin. Transport. Res. Transport Environ. 17 (5), 413–417. https://doi.org/10.1016/j.trd.2012.04.001.

Clark, C., Mokhtarian, P., Circella, G., Watkins, K., 2019. User preferences for bicycle infrastructure in communities with emerging cycling cultures. Transp. Res. Rec. 2673 (12), 89–102. https://doi.org/10.1177/0361198119854084.

Dill, J., 2009. Bicycling for transportation and health: the role of infrastructure. J. Publ. Health Pol. 30 (1), S95–S110. https://doi.org/10.1057/jphp.2008.56.

Ding, X., Fan, H., Gong, J., 2021. Towards generating network of bikeways from mapillary data. Comput. Environ. Urban Syst. 88, 101632. https://doi.org/10.1016/j.compenvurbsys.2021.101632.

Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q., 2020. A survey on ensemble learning. Front. Comput. Sci. 14 (2), 241–258. https://doi.org/10.1007/s11704-019-8208-z.

Fan, R., Li, J., Li, F., Han, W., Wang, L., 2022. Multilevel spatial-channel feature fusion network for urban village classification by fusing satellite and streetview images. IEEE Trans. Geosci. Rem. Sens. 60, 1–13. https://doi.org/10.1109/TGRS.2022.3208166.

Fang, F., Zeng, L., Li, S., Zheng, D., Zhang, J., Liu, Y., Wan, B., 2022. Spatial context-aware method for urban land use classification using street view images. ISPRS J. Photogrammetry Remote Sens. 192, 1–12. https://doi.org/10.1016/j.isprsjprs.2022.07.020.

Filho, A., Shimabukuro, M., Poz, A.D., 2023. Deep learning multimodal fusion for road network extraction: context and contour improvement. IEEE Geoscience and Remote Sensing Letters 20, 1–5. https://doi.org/10.1109/LGRS.2023.3291656.

Gibril, M.B.A., Kalantar, B., Al-Ruzouq, R., Ueda, N., Saeidi, V., Shanableh, A., Mansor, S., Shafri, H.Z.M., 2020. Mapping heterogeneous urban landscapes from the fusion of digital surface model and unmanned aerial vehicle-based images using adaptive multiscale image segmentation and classification. Remote Sens. 12 (7). https://doi.org/10.3390/rs12071081. Article 7.

Guo, W., Xu, C., Jin, S., 2024a. Fusion of satellite and street view data for urban traffic accident hotspot identification. Int. J. Appl. Earth Obs. Geoinf. 130, 103853. https://doi.org/10.1016/j.jag.2024.103853.

Guo, Z., Xu, R., Feng, C.-C., Zeng, Z., 2024b. PIF-Net: a deep point-image fusion network for multimodality semantic segmentation of very high-resolution imagery and aerial point cloud. IEEE Trans. Geosci. Rem. Sens. 62, 1–15. https://doi.org/10.1109/TGRS.2023.3342477.

Handy, S., van Wee, B., Kroesen, M., 2014. Promoting cycling for transport: research needs and challenges. Transp. Rev. 34 (1), 4–24. https://doi.org/10.1080/01441647.2013.860204.

Hoang, T.M., Nguyen, P.H., Truong, N.Q., Lee, Y.W., Park, K.R., 2019. Deep RetinaNet-Based detection and classification of road markings by visible light camera sensors. Sensors. 19 (2). https://doi.org/10.3390/s19020281. Article 2.

Hoffmann, E.J., Wang, Y., Werner, M., Kang, J., Zhu, X.X., 2019. Model fusion for building type classification from aerial and street view images. Remote Sens. 11 (11), 1259. https://doi.org/10.3390/rs11111259.

Hosseini, M., Miranda, F., Lin, J., Silva, C.T., 2022. CitySurfaces: city-Scale semantic segmentation of sidewalk materials. Sustain. Cities Soc. 79, 103630. https://doi.org/10.1016/j.scs.2021.103630.

Hull, A., O'Holleran, C., 2014. Bicycle infrastructure: can good design encourage cycling? Urban, Planning. Trans. Res. 2 (1), 369–406. https://doi.org/10.1080/21650020.2014.955210.

Hwang, U., Guhathakurta, S., 2023. Exploring the impact of bike lanes on transportation mode choice: a simulation-based, route-level impact analysis. Sustain. Cities Soc. 89, 104318. https://doi.org/10.1016/j.scs.2022.104318.

Ito, K., Biljecki, F., 2021. Assessing bikeability with street view imagery and computer vision. Transport. Res. C Emerg. Technol. 132, 103371. https://doi.org/10.1016/j.trc.2021.103371.

Jones, T., 2012. Getting the British back on bicycles—The effects of urban traffic-free paths on everyday cycling. Transp. Policy. 20, 138–149. https://doi.org/10.1016/j.tranpol.2012.01.014.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS J. Photogrammetry Remote Sens. 145, 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006.

Li, B., Gao, J., Chen, S., Lim, S., Jiang, H., 2024. DF-DRUNet: a decoder fusion model for automatic road extraction leveraging remote sensing images and GPS trajectory data. Int. J. Appl. Earth Obs. Geoinf. 127, 103632. https://doi.org/10.1016/j.jag.2023.103632.

Lieu, S.J., Guhathakurta, S., 2025. A novel approach for estimating sidewalk width from street view images and computer vision. Environ. Plan. B Urban Anal. City Sci., 23998083251369602 https://doi.org/10.1177/23998083251369602.

Liu, D., Jiang, Y., Wang, R., Lu, Y., 2023. Establishing a citywide street tree inventory with street view images and computer vision techniques. Comput. Environ. Urban Syst. 100, 101924. https://doi.org/10.1016/j.compenvurbsys.2022.101924.

Luo, H., Wang, Z., Du, B., Dong, Y., 2024. A deep cross-modal fusion network for road extraction with high-resolution imagery and LiDAR data. IEEE Trans. Geosci. Rem. Sens. 62, 1–15. https://doi.org/10.1109/TGRS.2024.3360963.

Luttrell, J., Zhang, Y., Zhang, C., 2024. Automatically detect crosswalks from satellite view images: a deep learning approach with ground truth verification. Inter. J. Trans. Sci. Technol. 16, 165–176. https://doi.org/10.1016/j.ijtst.2024.01.006.

Mattyus, G., Wang, S., Fidler, S., Urtasun, R., 2016. HD maps: fine-grained road segmentation by parsing ground and aerial images, 3611–3619. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Mattyus_HD_Maps_Fine-Grained_CVPR_2016_paper.html.

Ning, H., Ye, X., Chen, Z., Liu, T., Cao, T., 2022. Sidewalk extraction using aerial and street view images. Environ. Plan. B Urban Anal. City Sci. 49 (1), 7–22. https://doi.org/10.1177/2399808321995817.

Pucher, J., Dill, J., Handy, S., 2010. Infrastructure, programs, and policies to increase bicycling: an international review. Prev. Med. 50, S106–S125. https://doi.org/10.1016/j.ypmed.2009.07.028.

Senlet, T., Elgammal, A., 2012. Segmentation of occluded sidewalks in satellite images. Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 805–808. https://ieeexplore.ieee.org/abstract/document/6460256.

Shackelford, A.K., Davis, C.H., 2003. A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas. IEEE Trans. Geosci. Rem. Sens. 41 (9), 1920–1932. https://doi.org/10.1109/TGRS.2003.814627.

Shen, H., Lin, Y., Tian, Q., Xu, K., Jiao, J., 2018. A comparison of multiple classifier combinations using different voting-weights for remote sensing image classification. Int. J. Rem. Sens. 39 (11), 3705–3722. https://doi.org/10.1080/01431161.2018.1446566.

Singh, G., Guleria, K., Sharma, S., 2024. Deep learning-based fine-tuned convolutional neural network model for google map image analysis. In: 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), pp. 437–442. https://doi.org/10.1109/ICDICI62993.2024.10810791.

Workman, S., Zhai, M., Crandall, D.J., Jacobs, N., 2017. A Unified Model for near and Remote Sensing, pp. 2688–2697. https://openaccess.thecvf.com/content_iccv_2017/html/Workman_A_Unified_Model_ICCV_2017_paper.html.

Yang, Q., Cai, J., Feng, T., Liu, Z., Timmermans, H., 2021. Bikeway provision and bicycle commuting: city-level empirical findings from the US. Sustainability 13 (6), 3113. https://doi.org/10.3390/su13063113.

Zhang, Y., Luttrell, J., Zhang, C., 2025. How to detect occluded crosswalks in overview images? Comparing three methods in a heavily occluded area. Inter. J. Trans. Sci. Technol. 17, 148–160. https://doi.org/10.1016/j.ijtst.2024.04.001.

Zhao, K., Li, S., Zhou, L., Sun, J., Hao, S., 2023. When complementarity meets consistency: weighted collaboration fusion constrained by consistency between views for multi-view remote sensing scene classification. Int. J. Rem. Sens. 44 (23), 7492–7514. https://doi.org/10.1080/01431161.2023.2285741.