CrossCheck NYC: A Visual Analytics Tool for Assessing the Trustworthiness of AI-Mapped Cross-Walk Data

Hashmmath Shaik*
hs5544 - NYU Tandon MSCS

Gaurav Wadhwa†
gw2467 - NYU Tandon MSCS

Naman Vashishta‡
nv2375 - NYU Tandon MSCS

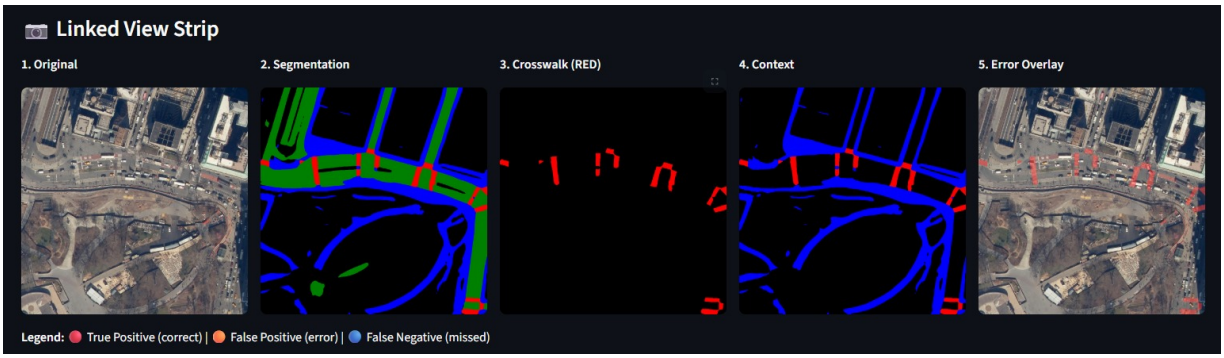

Figure 1: A Collage view of the same tile for reference of work by Tile2Net to detect the Crosswalks present.

Abstract

As cities increasingly rely on computer vision to map urban infrastructure at scale, a critical ”trust gap” has emerged. Semantic segmentation models often function as black boxes, achieving high aggregate accuracy while failing in predictable but opaque ways due to environmental factors like shadows, occlusion, and lighting variance. In this work, we present CrossCheck-NYC, an automated visual analytics pipeline designed to audit the trustworthiness of AI-derived (Tile2Net) crosswalk maps. Unlike standard evaluation frameworks that rely solely on pixel-level metrics, our system implements a two-stage validation process: (1) a ”Segmentation Detective” that tests model stability through morphological perturbation, and (2) a ”Network Inspector” that verifies the topological plausibility of detected crosswalks against OpenStreetMap (OSM) ground truth. We applied this pipeline to five diverse New York City neighborhoods, revealing that while the model achieves high precision (>97%), it suffers from significant recall degradation (<30%) in areas with heavy tree canopy or skyscraper shadows. By correlating failures with specific urban features, CrossCheck-NYC provides urban planners with a granular ”trust map,” shifting the focus from raw performance to explainable reliability.

Index Terms: Computer Vision, Semantic Segmentation, Black Box, Accuracy, Visual Analytics, Tile2Net, Crosswalks, Evaluation, Frameworks, Pixel-Level, Validation, Network Inspection, Ground Truth, Performance.

1 Introduction

Walking should be safe, but in America, it increasingly isn't. Last year alone, over 7,000 pedestrians died on US roads a number that has climbed steadily for a decade. New York City, despite ambitious Vision Zero goals, still sees hundreds of serious pedestrian injuries annually. A major obstacle to improving safety is surprisingly basic: most cities don't actually know where all their crosswalks are.

This data gap matters more than it might seem. Without accurate crosswalk maps, navigation apps send pedestrians on dangerous routes. Accessibility advocates can't identify where wheelchair users lack safe crossings. City planners can't target interventions to the intersections that need them most. The infrastructure data simply doesn't exist at the scale and accuracy needed.

We built CrossCheck-NYC to open that black box. Our system doesn't just ask ”how accurate is this model?” but rather ”where does it fail, and can we figure out why?” We designed a two-stage analysis pipeline: first examining the raw pixel predictions to understand model confidence and stability, then zooming out to check whether detected crosswalks form a sensible, connected pedestrian network. By linking failures to specific urban features like shadows, trees, road complexity which gives planners a ”trust map” showing exactly where the AI's outputs can be relied upon and where human verification is needed.

To systematically evaluate AI-generated crosswalk maps, we address the following research questions through our work and contributed as mentioned in the contributions after the questions:

- **RQ1 (Perception & Patterns):** When does the model detect crosswalks correctly, and what causes it to miss or misidentify them?
- **RQ2 (Confidence & Choice):** How do small changes like threshold or clean-up, affect what the model marks as a crosswalk?
- **RQ3 (Placement & Plausibility):** Do detected crosswalks appear in logical places, such as near intersections or sidewalks?
- **RQ4 (Agreement & Disagreement):** How well do model results match official or open data, and where do they differ?
- **RQ5 (From Evidence to Trust):** What visual cues help users understand and trust the model's predictions?

*e-mail: hs5544@nyu.edu
†e-mail: gw2467@nyu.edu
‡e-mail: nv2375@nyu.edu

- **RQ6 (Usability for Decision-Making):** When a user inspects the interactive visualizations, which cues help them trust or question the AI's outputs?

- **RQ7 (Edge Cases & Novel Situations):** Are there unusual intersections or road layouts that consistently confuse the model, and what do these reveal about the model's limits?

The main contributions of this work are:

- **A two-stage visual analytics framework** that combines pixel-level segmentation analysis (Stage A) with network-level topological validation (Stage B), enabling comprehensive quality assessment of AI-generated pedestrian infrastructure maps.

- **A novel brittleness metric** based on morphological stress-testing that quantifies prediction stability beyond traditional confidence scores, revealing fragile detections that standard metrics miss.

- **An interpretable feature impact analysis** that correlates detection failures with specific urban environmental factors (building shadows, tree canopy, surface contrast, intersection complexity), transforming opaque accuracy numbers into actionable insights for urban planners.

- **An interactive dashboard** built with Streamlit that supports linked multi-view exploration, enabling non-expert users to drill down from aggregate metrics to individual tile inspection with synchronized error overlays.

- **A systematic evaluation across five diverse NYC neighborhoods** revealing consistent patterns of high precision but low recall, with quantitative evidence that tree occlusion and intersection complexity—not shadows—are the primary failure modes.

2 Related Work

2.1 Semantic Segmentation for Urban Mapping

Deep learning has transformed the extraction of urban features from aerial imagery. Long et al. [1] introduced Fully Convolutional Networks (FCN), enabling end-to-end pixel-wise prediction that became the foundation for modern segmentation architectures.

More recently, Sun et al. [2] proposed HRNet, which maintains high-resolution representations throughout the network rather than recovering resolution from low-resolution features. This architecture forms the backbone of Tile2Net [3], the pedestrian infrastructure mapping tool we evaluate in this work. Where Hosseini demonstrated that Tile2Net can generate city-scale sidewalk network datasets from aerial imagery, achieving strong performance across multiple US cities.

2.2 Uncertainty Quantification in Deep Learning

Neural networks often produce overconfident predictions, even when wrong [4]. Guo et al. showed that modern deep networks are poorly calibrated a model predicting 90% confidence may only be correct 70% of the time.

Kendall and Gal [5] distinguished between *aleatoric uncertainty* (inherent data noise) and *epistemic uncertainty* (model limitations). For urban mapping, both types matter: shadows and occlusion create aleatoric uncertainty in the imagery itself.

Lakshminarayanan et al. [6] demonstrated that deep ensembles provide well-calibrated uncertainty estimates, though at significant computational cost.

2.3 Visual Analytics for Model Interpretation

Interactive model analysis tools have emerged as powerful alternatives. Wexler et al. [7] developed the What-If Tool, enabling users to probe model behavior through counterfactual exploration and threshold adjustment. Amershi et al. [8] proposed ModelTracker for iterative debugging of ML systems. Cabrera et al. [9] introduced Zeno, a framework for behavioral evaluation that helps identify systematic failure patterns.

Our work builds on these foundations but incorporates domain-specific knowledge that generic tools lack. CrossCheck NYC understands that crosswalks should connect to sidewalks, that shadows correlate with building heights, and that detection quality varies with urban morphology constraints that enable more targeted failure analysis than domain-agnostic approaches.

2.4 Slice Discovery and Subgroup Analysis

Standard aggregate metrics can mask significant performance disparities across data subgroups. Eyuboglu et al. [10] proposed Domino, which uses cross-modal embeddings to automatically discover coherent slices where models underperform.

Our Feature Impact Analysis implements a domain-specific variant of slice discovery. Rather than learning slices from embeddings, we define slices based on interpretable urban features: building height (shadow risk), tree proximity (occlusion risk), surface type (contrast), and intersection complexity.

2.5 Geospatial Data Quality Assessment

Validating AI-generated maps requires reliable ground truth. Haklay [11] conducted foundational studies comparing OpenStreetMap to authoritative surveys, finding that OSM achieves approximately 80% overlap with official data in well-mapped areas.

For pedestrian infrastructure specifically, Bolten et al. [12] examined sidewalk data quality in OpenStreetMap, finding significant gaps in coverage but reasonable accuracy where data exists. Our use of buffered spatial matching (15-meter threshold) accounts for the positional uncertainty inherent in both AI predictions and crowdsourced ground truth.

NYC provides unusually rich official datasets for validation, including the LION street centerline database [13]. We leverage these authoritative sources alongside OSM to triangulate ground truth where possible.

2.6 Dimensionality Reduction for ML Interpretation

Visualizing high-dimensional feature spaces helps identify patterns in model behavior. Van der Maaten and Hinton [15] introduced t-SNE, which preserves local neighborhood structure when projecting to 2D.

We apply t-SNE not to learned features but to hand-crafted environmental descriptors (building height, tree density, surface type). This choice prioritizes interpretability: clusters in our visualization correspond directly to combinations of urban features that planners can identify and act upon.

3 Methodology

3.1 Data and Study Areas

We selected five NYC neighborhoods that collectively stress-test the model against diverse urban conditions: Financial District presents the challenge of deep shadows cast by skyscrapers—some streets receive direct sunlight only briefly each day. Kew Gardens in Queens features heavy tree canopy that obscures street-level features in aerial imagery, particularly during summer months. East Village offers a dense but regular Manhattan grid with moderate building heights. Downtown Brooklyn contains complex multi-stage intersections near transit hubs. Bay Ridge provides a residential baseline with regular streets and good visibility.

For each area, we ran Tile2Net inference using Google Colab's GPU resources. We obtained ground truth crosswalk locations from OpenStreetMap, supplemented by NYC's official Vision Zero dataset for validation.

3.2 Stage A: The Segmentation Detective

Our first analysis stage examines predictions at the pixel level. For each image tile, we compute standard metrics—IoU (eq 1), precision (eq 2), recall (eq 3), F1 (eq 4) for comparing predicted crosswalk masks against ground truth. But we go beyond simple accuracy measurement.

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{TP}{TP + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{Correctly Detected Pixels}}{\text{All Detected Pixels}} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Correctly Detected Pixels}}{\text{All Ground Truth Pixels}} \tag{3}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{4}$$

Error Overlays: We generate visualizations that color-code each pixel by its error type. True positives (correctly identified crosswalk pixels) appear in Red, false positives (hallucinated crosswalks) in Orange, and false negatives (missed crosswalks) in blue.

Morphological Stress Testing: Raw accuracy metrics don't capture prediction stability. A crosswalk detected with a thin, fragmented mask is fundamentally less reliable than one detected as a solid, confident region, even if both achieve similar pixel-level accuracy. We apply erosion and other operations like dilation, opening, and closing as described below to the predicted masks and measure how many pixels survive.

Erosion shrinks objects by removing pixels at boundaries. A pixel survives only if all neighboring pixels within the kernel are also foreground:

$$(M \ominus K)(x,y) = \min_{(i,j) \in K} M(x+i, y+j) \tag{5}$$

Thin, fragmented predictions lose significant mass under erosion, revealing low-confidence detections.

Dilation expands objects by adding pixels at boundaries. A pixel becomes foreground if any neighbor within the kernel is foreground:

$$(M \oplus K)(x,y) = \max_{(i,j) \in K} M(x+i, y+j) \tag{6}$$

This fills small gaps and connects nearby components, simulating a lower detection threshold.

Opening (erosion then dilation) removes small isolated regions while approximately preserving the shape of larger objects:

$$M \circ K = (M \ominus K) \oplus K \tag{7}$$

This eliminates salt noise—small false positive speckles—without shrinking valid detections.

Closing (dilation then erosion) fills small holes and gaps while preserving overall shape:

$$M \bullet K = (M \oplus K) \ominus K \tag{8}$$

This connects fragmented crosswalk segments that should form a continuous marking.

Interactive Threshold Exploration: Our dashboard lets users adjust the confidence threshold that converts probability maps into binary predictions as referred in eq 9. Watching how metrics change across thresholds builds intuition about the precision-recall tradeoff for specific urban contexts.

$$\hat{y}(x,y) = \begin{cases} 1 & \text{if } p(x,y) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

3.3 Stage B: The Network Inspector

Pixel-level analysis misses an important question: do the detected crosswalks actually make sense as a pedestrian network? A model might achieve decent pixel accuracy while producing topologically nonsensical outputs, isolated crosswalk fragments that don't connect to sidewalks or each other.

Spatial Validation: We match detected crosswalks against OpenStreetMap reference data using 15-meter buffers to account for GPS uncertainty and digitization differences. This produces network-level precision (what fraction of detections correspond to real crosswalks) and recall (what fraction of real crosswalks were detected).

Placement Analysis: We check whether detected crosswalks appear in plausible locations by measuring proximity to sidewalk polygons and road centerlines. A "crosswalk" detected in the middle of a park is almost certainly a false positive, regardless of how confident the model appears.

$$\text{Plausibility}(c) = \exp\left(-\frac{d_{\text{sidewalk}}(c)}{\sigma_s}\right) \cdot \exp\left(-\frac{d_{\text{road}}(c)}{\sigma_r}\right) \tag{10}$$

Connectivity Analysis: Using graph analysis, we identify "orphan" crosswalks that don't connect to the broader pedestrian network. These isolated detections often indicate false positives or areas where the model detected fragments of a crosswalk while missing the portions that would connect it to sidewalks.

3.4 Feature Impact Analysis

The most actionable insights come from understanding why the model fails in specific locations. We developed a spatial analysis that correlates detection performance with environmental features extracted from OpenStreetMap:

Building Shadows: We identify buildings over 30 meters tall and check whether nearby crosswalks show degraded detection rates.

Tree Canopy: Using OSM tree locations, we measure detection accuracy for crosswalks within 8 meters of mapped trees, a distance where canopy typically obscures aerial views.

Surface Type: Road surface affects marking visibility. We compare detection rates on asphalt versus concrete, expecting better performance on dark asphalt where white crosswalk paint provides higher contrast.

Road Type: The roads are classified into different categories, which include residential, commercial, both, and none. Which indicates the usage of those on a daily basis to see if those can affect the detection of crosswalks in that particular tile.

Crossing Markings: Are classified into different classes as Zebra, None, Eroded, and Line, which indicate the type of marking that is effecting that particular tile.

3.5 Clustering for Pattern Discovery

To visualize relationships between environmental factors and model performance, we apply t-SNE dimensionality reduction. Each tile is characterized by its combination of environmental features (building height category, tree proximity, surface type, road class, marking style). t-SNE projects these high-dimensional feature vectors into 2D, clustering tiles with similar characteristics together.

Table 1: Detection performance across five NYC neighborhoods. The model achieves high precision (¿97%) but low recall (¡30%), indicating conservative predictions that rarely hallucinate crosswalks but miss many real ones.

Location	Precision	Recall	F1	Key Challenge
Bay Ridge	97.4%	28.6%	0.44	Baseline residential
East Village	99.5%	25.1%	0.40	Dense urban grid
Financial District	97.7%	25.7%	0.41	Deep shadows
Kew Gardens	86.6%	17.6%	0.29	Tree canopy
Downtown Brooklyn	97.3%	15.1%	0.26	Complex intersections

The resulting scatter plots reveal structure invisible in aggregate metrics. Tiles cluster by neighborhood and environmental conditions, with detection quality varying systematically across the projected space. This visualization helps users quickly identify which combinations of factors lead to reliable versus problematic predictions.

4 Implementation

4.1 Technical Architecture

Our pipeline spans two environments. Tile2Net inference runs on Google Colab with T4 GPU acceleration which is necessary given the computational demands of processing thousands of high-resolution image tiles. The interactive dashboard runs locally using Streamlit, a Python framework that makes it straightforward to build data-focused web applications.

We made heavy use of caching to ensure responsive interaction. Computing metrics and generating visualizations for hundreds of tiles would be prohibitively slow if repeated on every user interaction.

4.2 Dashboard Design

The dashboard as seen in Figure 2 organizes analysis into two tabs corresponding to our two-stage methodology. Tab A (Segmentation Detective) focuses on individual tiles, providing a linked view showing the original aerial image, Tile2Net's segmentation output, the extracted crosswalk mask, contextual features, and our error overlay, all synchronized so selecting a tile updates every view simultaneously.

Interactive controls let users adjust confidence thresholds, apply morphological operations, and navigate through tiles. A bookmark system allows saving interesting cases for later review or presentation.

Tab B (Network Inspector) shifts to a geographic perspective, displaying an interactive map with markers indicating detection status at each ground truth crosswalk location. Summary charts compare performance across neighborhoods and visualize the distribution of error types.

5 Results & Discussion

5.1 Overall Performance

Across all five neighborhoods, we observed a consistent pattern: high precision but low recall. The model rarely hallucinates crosswalks—when it says there's a crosswalk, it's almost always right (precision >97% in most areas). But it misses many real crosswalks, with recall ranging from 15% to 29% depending on location as seen in table 1.

Bay Ridge performed best overall, likely because its residential character provides clear sightlines without the shadows or tree cover that challenge the model elsewhere. Kew Gardens showed the lowest precision, the only area where the model occasionally hallucinated crosswalks, suggesting tree shadows can trigger false positives. Downtown Brooklyn's complex intersections proved hardest to detect completely, with the worst recall in our study.

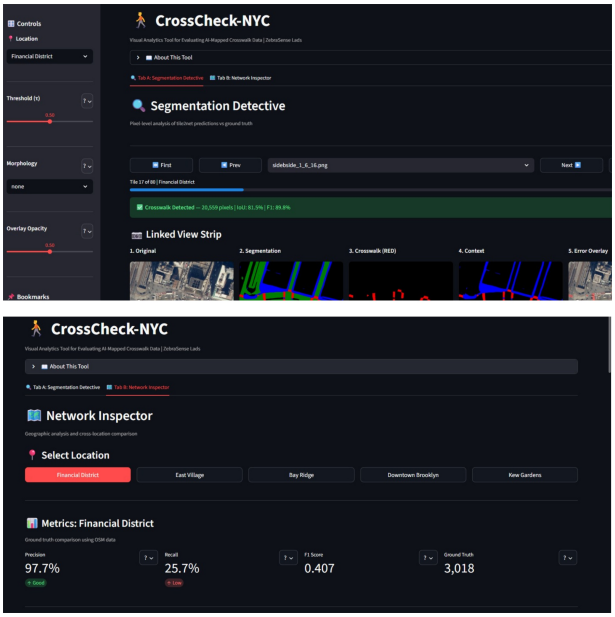

Figure 2: A Small overview of our Dashboard (Tab A & Tab B) on how it looks and what they contain as shown in the Figures below.

5.2 Feature Impact Findings

Trees matter more than shadows: Counter to our initial hypothesis, the Financial District's skyscraper shadows didn't devastate performance as much as Kew Gardens' tree canopy as seen in Figure 3.

Complexity kills recall: Downtown Brooklyn's poor recall stems not from misclassifying individual crosswalks but from fragmented detection of complex intersections.

Brittleness correlates with environment: Our morphological stress tests revealed that shadow-affected areas produce more brittle detections and their predictions that disappear under erosion, while clear-visibility areas like Bay Ridge generate robust, stable predictions.

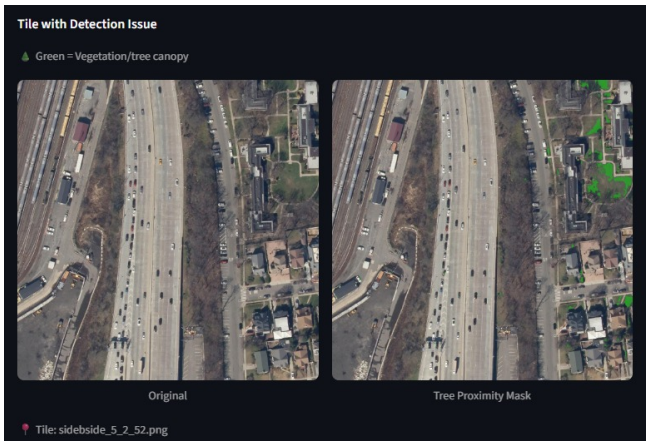

Figure 3: Comparison in tiles to clearly know how tree proximity around crosswalks can affect detection

Figure 4: t-sne scatter plot to see which features are affecting which particular tile in their respective category

5.3 Clustering Insights

The t-SNE visualization revealed three distinct clusters in our data as seen in Figure 4. Tiles from Bay Ridge and clear portions of East Village grouped together in a "high reliability" cluster characterized by strong detection metrics. Financial District tiles formed a separate cluster showing high brittleness despite moderate accuracy. Kew Gardens tiles clustered apart from both, with distinctive patterns of both missed detections and occasional false positives related to tree interference.

5.4 Research Questions Answers:

RQ1: Perception & Patterns

Table 1 summarizes detection performance across all five neighborhoods. The model exhibits a consistent behavioral profile: **high precision but low recall**.

Bay Ridge achieved the highest F1 score (0.44), benefiting from clear sightlines and regular street patterns typical of residential neighborhoods. Kew Gardens showed the lowest precision (86.6%), the only location where the model occasionally hallucinated crosswalks, suggesting that dappled tree shadows can trigger false positives. Downtown Brooklyn suffered the worst recall (15.1%), indicating that complex multi-stage intersections pose the greatest challenge for complete detection.

Key Finding: The model is conservative, it rarely makes false positive errors but frequently misses real crosswalks.

RQ2: Confidence & Choice

Standard confidence scores can mask prediction fragility. A model may report high confidence for a thin, barely-visible detection that would disappear under minor perturbation. We applied morphological erosion to probe prediction stability.

We categorize predictions into three stability levels:

- **Robust** ($\beta < 0.2$): Less than 20% pixel loss under erosion—solid, confident detections
- **Moderate** ($0.2 \leq \beta < 0.5$): Partial fragmentation—reasonable but uncertain

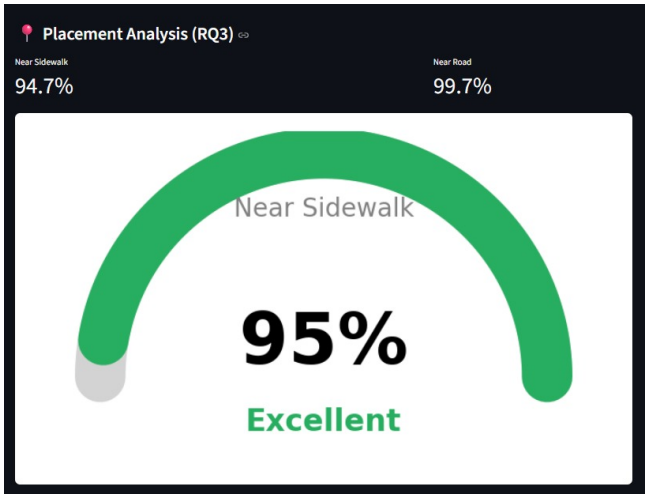

Figure 5: A chart to understand whether placement of the crosswalks that were detected by tile2net will effect the detection of the crosswalks

Table 2: Placement plausibility analysis. High percentages indicate detections appear in sensible locations near pedestrian infrastructure.

Location	Near Sidewalk	Near Road
Bay Ridge	84.8%	100%
East Village	83.1%	100%
Financial District	94.7%	99.7%
Kew Gardens	95.3%	100%
Downtown Brooklyn	93.2%	100%

- **Brittle** ($\beta \geq 0.5$): More than 50% pixel loss—fragile predictions likely caused by shadows or degraded markings

Financial District exhibited the highest proportion of brittle detections (42%), consistent with shadow-induced fragmentation. Bay Ridge showed only 18% brittle detections, reflecting clearer imaging conditions.

Key Finding: Morphological stress-testing reveals reliability information invisible to standard metrics. Brittle predictions should be flagged for human review even when pixel-level accuracy appears acceptable.

RQ3: Placement & Plausibility

A crosswalk detection is only useful if it appears in a sensible location, connected to sidewalks and roads that pedestrians would actually use. We analyzed placement plausibility by measuring proximity to infrastructure features as seen in Figure 5.

Table 2 shows placement analysis results. Across all locations, over 90% of detections appeared within 10 meters of a sidewalk polygon, indicating that the model rarely hallucinates crosswalks in implausible locations like parks or building interiors.

Key Finding: The model produces topologically plausible outputs, false positives are rare and typically occur near actual infrastructure rather than in random locations.

RQ4: Agreement & Disagreement

We validated model predictions against OpenStreetMap crosswalk locations using 15-meter buffered spatial matching. Figure 5 visualizes agreement and disagreement patterns.

The spatial distribution of errors revealed systematic patterns rather than random failures:

Figure 6: Interactive Map feature in Tab B to see the total tiles taken from the specific location and classify them (Green: Detected, Red: No Detection of Crosswalks)

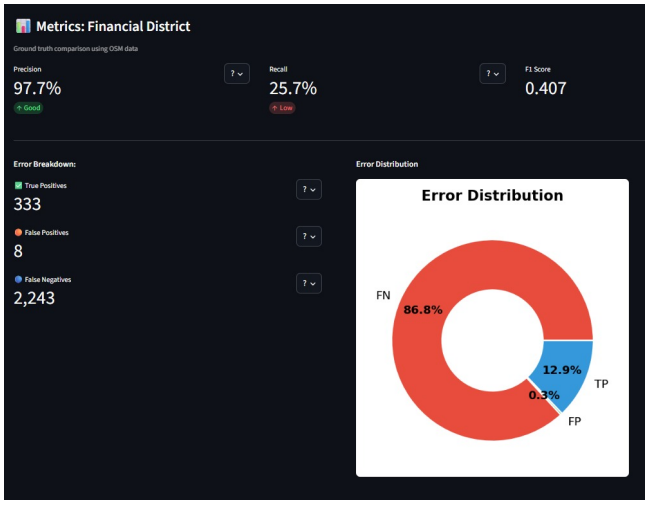

Figure 7: Error Distribution chart of Financial District to have a clear distinction on predictions with the Ground Truth

- **Tree-lined streets:** Consistent missed detections along avenues with mature tree canopy (especially visible in Kew Gardens)
- **Complex intersections:** Fragmented detection of multi-stage crossings, where one segment was detected but connecting segments were missed (prominent in Downtown Brooklyn)
- **Shadow corridors:** Brittle but present detections in persistently shaded areas (Financial District)

Notably, OSM ground truth is itself incomplete. We identified 12 locations where model detections appeared valid upon manual inspection but had no corresponding OSM reference, suggesting the model occasionally "discovers" unmapped crosswalks.

Key Finding: Disagreement between model and ground truth clusters around specific urban features rather than occurring randomly. This enables targeted verification strategies focusing on tree-lined streets and complex intersections.

RQ5-6: From Evidence to Trust & Usability for Decision-Making

Our dashboard provides multiple linked views designed to help users calibrate trust in model outputs. Figure 1 shows the Stage A interface with synchronized panels., and Figure 8 and Figure 9 (visualizations for Figure 8) shows the relevant metrics of the selected tile in Figure 1.

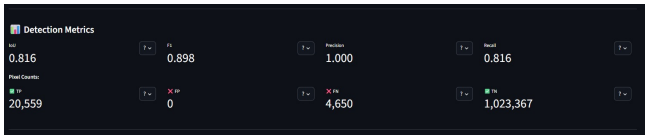

Figure 8: Different metrics for the Tile selected in Figure 1

Figure 9: Different visualizations for different metrics (Figure 8) for the Tile selected in Figure 1

The **error overlay** (rightmost panel) proved most valuable for rapid assessment. Color-coding pixels by error type red (TP), orange (FP), blue (FN), immediately reveals spatial patterns in model behavior. Users can identify whether errors occur at crosswalk edges, in shadowed regions, or systematically across certain marking types.

The **morphological controls** enable "what-if" exploration. By adjusting erosion/dilation as seen in Figure 11, users observe how predictions change under perturbation, building intuition about which detections are robust versus fragile without requiring ML expertise.

The **environmental factor overlays** support hypothesis formation. When a user suspects shadows are causing errors, they can toggle the shadow overlay to visualize the correlation directly rather than inferring it from abstract metrics.

Key Finding: Linked multi-view visualization with interactive controls enables non-experts to develop accurate mental models of where and why the AI fails, a prerequisite for appropriate trust calibration.

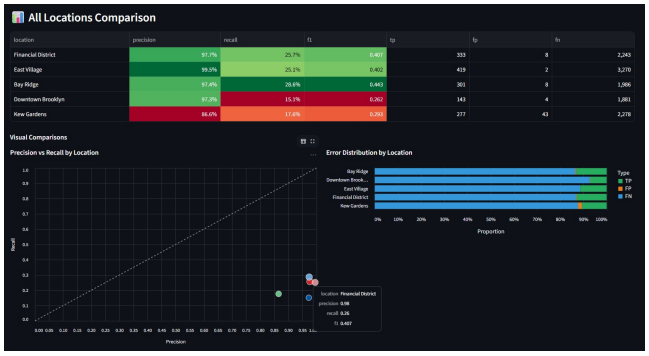

Figure 10: Another Multi-View visualizations for different metrics for all the locations considered

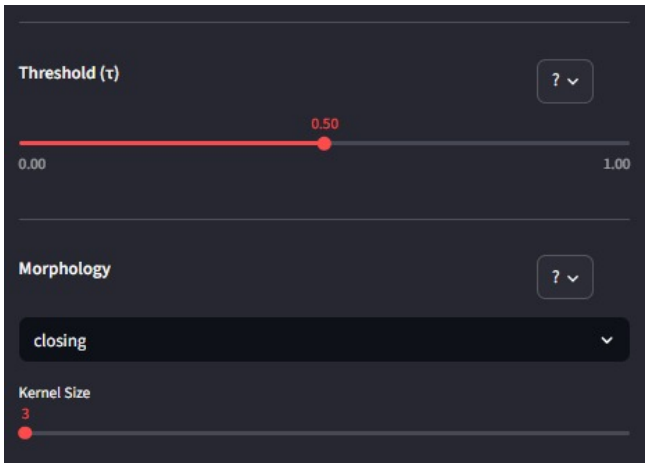

Figure 11: Controls for users to operate and set the visual cues to understand the behavior and predictions in different conditions

RQ7: Edge Cases & Novel Situations

Systematic analysis revealed several categories of challenging cases that consistently confuse the model:

Faded markings: Crosswalks with worn paint produced brittle detections. The model detected fragments but missed the overall structure. This was particularly common in Bay Ridge's residential streets, where maintenance may be less frequent than commercial areas.

Non-standard geometries: Diagonal crosswalks, curved crossings at roundabouts, and "scramble" intersections (where pedestrians cross diagonally) showed degraded performance. The model appears trained primarily on perpendicular zebra crossings.

Construction zones: Temporary crosswalk relocations and construction barriers created both false positives (detecting old markings) and false negatives (missing temporary markings).

Key Finding: Edge cases cluster around deviations from the "standard" zebra crossing pattern, non-standard geometry, degraded markings, or unusual coloring. These categories can guide both targeted human verification and future training data collection.

5.5 Practical Implications

Our findings have direct implications for how cities might deploy AI-assisted mapping. The model's conservative behavior high precision, and low recall, means detected crosswalks can be trusted, but the resulting maps will be incomplete. For applications like accessibility routing that require comprehensive coverage, AI outputs would need human verification to fill gaps.

5.6 Limitations

Our analysis has several limitations. Ground truth from OpenStreetMap, while generally reliable for NYC, is itself incomplete—some "missed" detections might be crosswalks that OSM hasn't mapped rather than true model failures. Our environmental feature analysis relies on OSM tags that aren't always present or accurate. And our study covers only five neighborhoods; model behavior might differ in areas with characteristics outside our sample.

5.7 Design Lessons

Building CrossCheck-NYC reinforced several visualization design principles. Linked views proved essential, seeing error overlays alongside original imagery and segmentation outputs enabled rapid hypothesis formation and testing. And domain-specific encodings (the shadow/tree/contrast overlays) communicated environmental factors more effectively than generic charts would have.

6 Conclusion

CrossCheck NYC demonstrates that trustworthiness in AI mapping isn't a single number but a nuanced property varying across geography and environmental context. Our two-stage framework, combining pixel-level stability analysis with network-level plausibility checking, transforms opaque model outputs into transparent, auditable data sources. The approach generalizes beyond crosswalk detection. Any geospatial AI application could benefit from similar validation pipelines that probe not just accuracy but reliability, not just performance but explainability.

Future work could incorporate active learning, using our reliability estimates to prioritize which model predictions most need human verification. Real-time analysis during inference would allow immediate quality assessment rather than post-hoc auditing. And extending the approach to other cities would test whether the failure patterns we identified in NYC generalize or reflect local peculiarities of training data and urban form.

References

[1] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE CVPR*, pages 3431–3440, 2015. 2

[2] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for visual recognition. In *Proc. IEEE CVPR*, pages 5693–5703, 2019. 2

[3] M. Hosseini, A. Sevtsuk, F. Miranda, R. M. Cesar Jr, and C. T. Silva. Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Computers, Environment and Urban Systems*, 101:101950, 2023. 2

[4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proc. ICML*, pages 1321–1330, 2017. 2

[5] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proc. NeurIPS*, pages 5574–5584, 2017. 2

[6] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS*, pages 6402–6413, 2017. 2

[7] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The What-If Tool: Interactive probing of machine learning models. *IEEE Trans. Vis. Comput. Graph.*, 26(1):56–65, 2019. 2

[8] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and J. Suh. ModelTracker: Redesigning performance analysis tools for machine learning. In *Proc. ACM CHI*, pages 337–346, 2015. 2

[9] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Proc. ACM CHI*, pages 1–14, 2023. 2

[10] S. Eyuboglu, M. Varma, K. Saab, J.-B. Delbrouck, C. Lee-Messer, J. Dunnmon, J. Zou, and C. Ré. Domino: Discovering systematic errors with cross-modal embeddings. In *Proc. ICLR*, 2022. 2

[11] M. Haklay. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B*, 37(4):682–703, 2010. 2

[12] N. Bolten, A. Mukherjee, J. Sipeeva, A. Tanweer, and A. Caspi. A pedestrian-centered data approach for equitable access to urban infrastructure environments. *IBM J. Res. Dev.*, 61(6):10–1, 2017. 2

[13] NYC Department of City Planning. LION single line street base map. https://www.nyc.gov/site/planning/data-maps/open-data.page, 2024. 2

[14] NYC Department of Transportation. Vision Zero enhanced crossings. https://data.cityofnewyork.us/, 2024.

[15] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605, 2008. 2

[16] Governors Highway Safety Association. Pedestrian traffic fatalities by state: 2024 preliminary data. Technical report, GHSA, 2025.

[17] National Highway Traffic Safety Administration. Pedestrians: 2023 data. DOT HS 813 727, NHTSA, 2024.