Geo-Visual Fusion: An Enhanced Strategy for Drone Object Detection Based on High-Definition Map Context

Shenman Zhang [1], Qingshan Peng [1], Mengmeng Duan [1], Sheng He [1], Wei He [1], Yongsheng Yu [1], Xiaojun Fu [1], Wei Jiang [1], Ziliang Zhang [1], Jie Song [1], Lin Wen [1], Xiao Liu [1]
[1] Wuhan Geomatics Institute, Wansong Street Wuhan, China - smzhang@whu.edu.cn, 1284811905@qq.com, 2012102130001@whu.edu.cn, 2014301610342@whu.edu.cn, 19755870@qq.com, 18971203328@qq.com, 740317917@qq.com, whgi2017@qq.com, 2268442202@qq.com, songjie0613@126.com, 13464071@qq.com, 972842481@qq.com

Keywords: UAV Object Detection, Geo-Visual Fusion, High-Definition Maps, Contextual Reasoning, Semantic Compatibility.

Abstract: With the rapid advancement of Urban Air Mobility (UAM), vision-only UAV object detectors like YOLO often suffer from "context blindness" in complex urban canyons, leading to logical fallacies or missed occluded targets. To address these limitations, this paper proposes an innovative Geo-Visual Fusion (GVF) enhancement strategy. By leveraging high-definition (HD) city maps as deterministic geo-spatial priors, we introduce a Geo-spatial Contextual Reasoning (GCR) module to post-process raw visual outputs. This framework incorporates a Semantic Compatibility Matrix (SCM) to eliminate geographically implausible false positives and a Bayesian enhancement rule to boost the confidence of occluded targets. Experimental validation in the Baibuting Community, Wuhan, demonstrates that the GVF framework significantly outperforms the baseline YOLOv11, achieving perfect recall and precision in the test sequence. Furthermore, the 2D vector-based indexing ensures high computational efficiency for edge computing deployment on platforms like the DJI Dock 3. Finally, a closed-loop "reverse empowerment" mechanism for HD map updates is discussed. This work effectively bridges probabilistic computer vision and deterministic geospatial constraints for reliable UAV perception.

1. Introduction

With the rapid development of Urban Air Mobility (UAM) and the increasing demand for intelligent city management, Unmanned Aerial Vehicles (UAVs) have emerged as indispensable aerial intelligent perception platforms(Cohen et al., 2021; Sapkota et al., 2025). The integration of UAVs into complex operations—such as infrastructure inspection, disaster response, and autonomous logistics delivery—relies heavily on the accuracy and reliability of their onboard visual perception systems, particularly object detection algorithms(Zhao and Zheng, 2019).

The evolution of object detection in aerial platforms has transitioned from traditional hand-crafted feature extractors, such as HOG and SIFT (Dalal and Triggs, 2005), to modern deep learning paradigms. While convolutional neural networks (CNNs) have achieved unprecedented success on general datasets like COCO and ImageNet, the unique zenith and oblique perspectives of UAVs introduce significant intra-class variation and scale diversity (Xia et al., 2018). Recent scholarship has sought to mitigate these issues through multi-scale feature pyramids and attention-guided mechanisms (Liu et al., 2020). However, a persistent challenge remains: these models primarily focus on 'what' an object looks like (spectral and textural features) rather than 'where' it is logically situated within the urban fabric. This lack of topological awareness often results in high-confidence false positives in complex environments. For instance, in dense residential areas, the geometric similarity between roof-mounted technical equipment and ground vehicles leads to frequent misclassifications (Sun et al., 2022). Consequently, there is an emerging consensus that integrating non-visual, deterministic constraints—such as those provided by High-Definition (HD) maps — is essential for moving beyond probabilistic pixel matching toward cognitive-level scene understanding (Nex et al., 2018).

The current state-of-the-art in object detection, largely driven by deep learning architectures like YOLO series and Faster R-CNN (Redmon et al., 2016; Ren and He, 2015), has demonstrated remarkable success on controlled benchmarks. These models, which leverage extensive training data for feature extraction, have pushed the boundaries of speed and accuracy.

However, deploying these vision-only models in real, complex urban environments exposes their fundamental limitations.While deep learning has revolutionized computer vision, its application in UAV-based urban sensing faces a 'semantic gap.' Unlike ground-based platforms that benefit from stable horizontal perspectives, UAVs operating in 'urban canyons' at oblique angles are frequently misled by building facade elements. This study addresses these 'context-blind' limitations by introducing a lightweight Geo-Visual Fusion (GVF) strategy that utilizes deterministic geo-spatial constraints to refine probabilistic visual outputs. Despite advances, these deep learning detectors remain inherently "context-blind" (Siris et al., 2021). They operate primarily based on local pixel and texture information, often failing to incorporate global scene semantics. To mitigate these limitations, recent studies have explored context-aware architectures. For instance, CAD-Net (Zhang et al., 2019) integrates global and local scene contexts through internal attention mechanisms to improve feature representation. However, such methods still rely heavily on visual patterns learned from training data and may remain susceptible to visual ambiguities in extreme conditions, such as severe occlusion or poor illumination. The absence of explicit geo-spatial constraints, even in context-aware neural networks, leads to two persistent challenges in urban aerial imagery:

1. Logical Fallacies (False Positives): Models are prone to making errors by violating real-world spatial constraints, such as mistaking geometrically similar features (e.g., specific windows or air-conditioner outdoor units) on building facades for mobile objects like vehicles (Ajmera et al., 2021). As illustrated in **Figure 1,** a vision-only baseline detector (YOLOv11) incorrectly identifies an air-conditioning outdoor unit on a building facade as a car due to its square geometric appearance.

Figure 1. Example of a vehicle detection(YOLO v11) logic error in UAV imagery.The air-conditioning outdoor unit enclosed by the red dashed box in the image was incorrectly detected as a car due to its square shape.

2. Insufficient Confidence (False Negatives): Occlusion, dynamic lighting, and small object sizes—common in aerial imagery—cause models to assign overly low confidence scores to objects that are genuinely present and correctly localized (e.g., partially occluded vehicles on a road). **Figure 2** provides a typical example of such a false negative error, where a vehicle under partial tree occlusion is completely missed by the baseline detector because its detection confidence falls below the operational threshold.

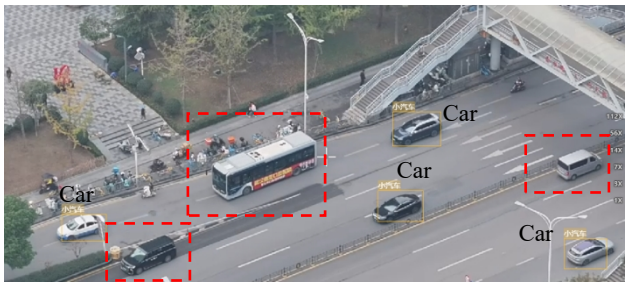

Figure 2. Example of low-confidence vehicle detection (YOLO v11) in UAV imagery. The vehicle enclosed by the red dashed box in the image was missed by the detector because its detection confidence was too low.

The core challenge, therefore, lies in transitioning from purely data-driven feature matching to context-aware semantic understanding. Unlike Unmanned Ground Vehicles (UGVs) that operate with a horizontal field of view, or satellite imagery that is strictly nadir, UAVs frequently operate at oblique angles in 'urban canyons.' This unique perspective makes them particularly susceptible to misidentifying building facade elements as ground objects—a specific 'logical fallacy' that our GVF strategy is uniquely designed to resolve. To address the limitations imposed by this context-blindness, this paper proposes an innovative Geo-Visual Fusion (GVF) enhancement strategy. Unlike prior works that primarily focus on architectural modifications to the neural networks (Bell et al., 2016) or rely on coarse contextual cues, our core idea is to systematically leverage the rich, immutable geo-spatial prior knowledge embedded within High-Definition (HD) city maps (Ma and Gao, 2023). This geographical information is utilized as a strong, non-visual, and deterministic contextual constraint to post-process and enhance the raw outputs of the UAV's visual detection system.

By integrating the UAV's "real-time visual perception" with the city's "known geo-spatial structure," we enable a crucial transition from simply "seeing" to contextually "understanding" the urban scene.

As a municipal institute affiliated with the Urban Planning Bureau of a major city, we possess comprehensive, high-precision mapping data of the entire urban area. This unique data foundation makes it feasible to develop and deploy the Geo-Spatial Contextual Reasoning (GCR) Module, allowing our model to reason effectively based on the real geo-spatial context.

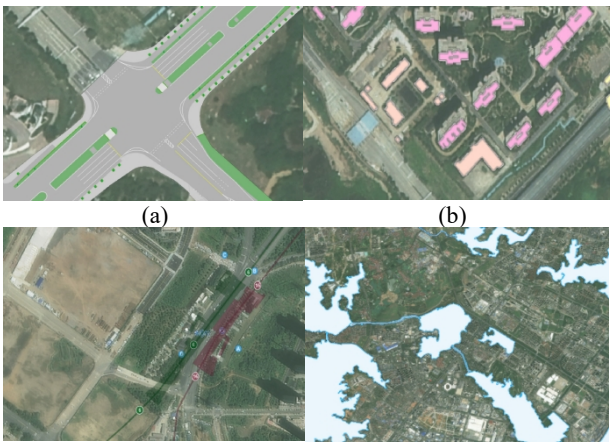

Figure 3. Shows an example of a high-precision map. (a) High-precision road vectors, (b) high-precision building vectors, (c) high-precision rail transit facility vectors, and (d) high-precision water-body vectors.

2. Methodology

The framework of this strategy mainly comprises three core modules: the overall workflow and modules of the proposed Geo-Visual Fusion (GVF) framework are schematically depicted in **Figure 4**.

2.1 Real-time geo-spatial Registration Module

This is the foundation for achieving graph-visual fusion. This module first acquires the UAV's high-precision GPS/IMU data and, in conjunction with camera intrinsic parameters, back-projects every pixel (or detected object bounding box) in the real-time UAV imagery onto the unified geographic coordinate system of the city(Liu et al., 2022). The projection formula is as follows:

$$X = (Z_a - Z_s) \cdot \frac{a_1(x - x_0) + a_2(y - y_0) - a_3 f}{c_1(x - x_0) + c_2(y - y_0) - c_3 f} + X_s$$
$$Y = (Z_a - Z_s) \cdot \frac{b_1(x - x_0) + b_2(y - y_0) - b_3 f}{c_1(x - x_0) + c_2(y - y_0) - c_3 f} + Y_s \quad (Equation\ 1)$$

where x_0 and y_0 are the pixel coordinates of the principal point, f is the focal length, $a_1 \sim c_3$ are the direction cosines, X_s, Y_s, Z_s are the geographic coordinates of the camera, x and y are the pixel coordinates of detected object, Z_a is the elevation of the detected object. By utilizing the vector data from the HD map (such as building outlines and road boundaries), we can accurately "localize" every object in the imagery, determining its geo-spatial semantic environment (e.g., whether it is on the "road layer," "building facade layer," or "park/green space layer"). Examples of these high-precision municipal vector

layers, including road networks, building footprints, rail transit facilities, and water bodies, are explicitly detailed in **Figure 3**.

The precision of the back-projection is highly sensitive to the temporal synchronization between the global navigation satellite system (GNSS), the inertial measurement unit (IMU), and the camera shutter. To mitigate projection drift at a flight altitude of 120m, our system implements a timestamp-based interpolation method for UAV pose estimation. This ensures that the transformed geographic coordinates *(X, Y)* accurately align with the high-precision vector layers of the HD map.

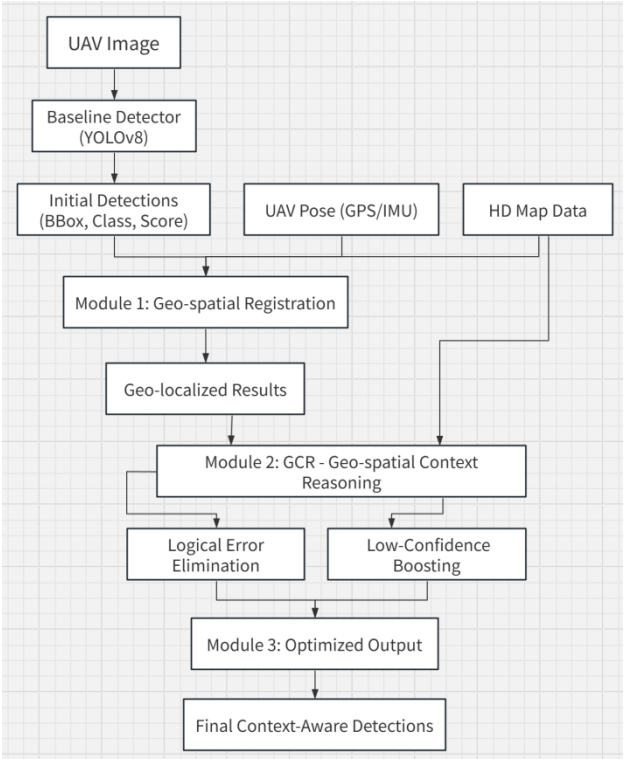

Figure 4. System framework diagram.

2.2 Geo-spatial Contextual Reasoning (GCR) Module

This is the core of the strategy. After the initial object detection model (e.g., YOLOv11) outputs preliminary results (class, confidence score, bounding box), the GCR module intervenes. This module contains two key sub-strategies:

2.2.1 Logical Error Elimination: We construct a Semantic Compatibility Matrix (Table 1). This matrix defines the logical relationships between different object classes (e.g., vehicle, pedestrian, container, tree) and different geo-spatial layers (e.g., road, sidewalk, building roof, water body). For instance, if the system detects a "van" with a confidence of 0.9, but the geo-spatial registration module shows it is located on the "building facade" layer, the GCR module will determine this as a logical fallacy and either force its confidence to zero or mark it as an "invalid detection." Similarly, if a park bench is misidentified as a "container," this detection will be suppressed because the compatibility between "container" and the "park/green space" layer is extremely low, as shown in **Table 1**.

Geo-spatial Layer	Vehicle	Pedestrian	Air-conditioner outdoor unit
Road	High	Low	0
Sidewalk	Medium	High	0
Building Facade	0	0	High
Water Body	0	Low	0

Table 1. The Semantic Compatibility Matrix (SCM)

2.2.2 Low-Confidence Boosting:
Contrary to error elimination, this strategy is used to enhance "reasonable" low-confidence detections. We formalize this enhancement using a Bayesian approach (Sheikh et al., 2005), treating the spatial information derived from the HD map as a strong prior probability.

Our objective is to compute the posterior probability, $P(C \mid V, G)$, which represents the probability that an object belongs to a class C (e.g., "vehicle"), given both the visual evidence V (from the baseline detector) and the geo-spatial context G (from the HD map).
According to Bayes' theorem, this posterior probability is defined as:

$$P(C|V,G) = \frac{P(V|C,G) \cdot P(C|G)}{P(V|G)} \quad (Equation\ 2)$$

Where:

- $P(C \mid G)$ is the geo-spatial prior probability. This is the core contextual knowledge provided by the HD map. For instance, the prior probability of an object being a "vehicle" given its location on a "road" layer ($P(C = "Vehicle"|$G = "On Road") is very high, whereas $P(C = "Vehicle"|$G = "On Building Facade" is effectively zero.

- $P(V \mid C, G)$ is the likelihood of observing the visual evidence V given the class C and context G. We make a simplifying assumption that the visual evidence is conditionally independent of the geo-spatial context given the class, i.e., $P(V \mid C, G) \approx P(V \mid C)$. This term $P(V \mid C)$ is represented by the original confidence score output by the baseline detector (e.g., YOLOv11).

- $P(V \mid G)$ is the marginal likelihood of the evidence, which acts as a normalization constant.

Given this formulation, our confidence-boosting update rule for a detected object i is simplified to a proportional relationship, where the final score is the product of the initial visual confidence and the geo-spatial prior:

$$Score_{posterior}(i) \propto Score_{detector}(i) \cdot P(C_i|G_i) \quad (Equation\ 3)$$

This Bayesian enhancement directly addresses the problem of low-confidence false negatives. For example, if an object on the "road" layer is identified as a "vehicle" with a low original confidence of 0.2 (perhaps due to occlusion or poor lighting), the GCR module will multiply this by a high prior probability (e.g., $P(C = "Vehicle" \mid G = "On\ Road" = 0.9$). This significantly raises its posterior confidence, effectively solving the problem of "it should be a vehicle, but the confidence is low".

2.3 Optimized Result Output Module

After processing by the GCR module, the system outputs a list of objects that has been purified by the "geo-spatial context." Compared to the raw detection results, this list exhibits a lower

False Positive Rate (FPs) and a higher True Positive Rate (TP, especially for low-confidence targets).

3. Results and Discussion

3.1 Experimental Setup

To evaluate the efficacy of the proposed GVF framework, we conducted a field validation in the Baibuting Community, Jiang'an District, Wuhan, covering an area of approximately 1.5 km^2. This site provides a representative urban canyon environment with high-rise residential structures and complex road networks, as schematically represented in the experimental area diagram in **Figure 5**. The experimental data was acquired using a DJI Dock 3 (UAV airport). The UAV operated at a flight altitude of 120 meters above ground level (AGL), with the gimbal camera maintained at a 45-degree downward pitch angle to capture high-definition video streams.The baseline detector utilized was YOLOv11 pre-trained on the VisDrone dataset, with a detection confidence threshold set at 0.35. Temporal consistency was maintained using the ByteTrack (Zhang et al., 2022) algorithm. The ancillary geo-spatial data comprised high-precision vector layers (road networks and building footprints) produced from municipal surveying projects.

Figure 5. Schematic diagram of the experimental area. The gray area in the diagram represents high-precision road data collected through surveying and mapping, with lines indicating road centerlines, lane markings, road signs, etc. The white area represents building outline data, while the green area represents urban vegetation and green spaces.

3.2 Quantitative and Qualitative Analysis

The validation was performed on a 2.2-minute flight sequence. The integration of the Geo-Spatial Contextual Reasoning (GCR) module demonstrated significant improvements in both precision and recall by addressing two primary failure modes of vision-only detectors.

A. Elimination of Logical Fallacies (False Positives)
During the flight, the baseline YOLOv11 model incorrectly identified two air-conditioning (AC) outdoor units—located on the upper floors of high-rise buildings—as "vehicles" for a duration of 5 seconds(The video has 30 frames per second). This error stemmed from the geometric and textural similarity between AC units and vehicles when viewed from an aerial perspective at a similar altitude. By applying the Semantic Compatibility Matrix (SCM), the GVF framework identified that these detections were geo-registered onto building facades rather than road layers. Consequently, these geographically implausible false positives were successfully suppressed, as shown in Table 1.

B. Recovery of Low-Confidence Detections (False Negatives)
Statistical analysis revealed 85 ground-truth vehicles within the monitored road segments. The baseline detector correctly identified 73 vehicles but missed 12 targets. Detailed analysis reveals that 8 of these missed vehicles were localized by the baseline model but assigned initial confidence scores ranging from 0.12 to 0.31, falling below the detection threshold. The remaining 4 vehicles were entirely missed (zero confidence) due to extreme environmental factors. Through Bayesian Enhancement, the GVF module successfully boosted the posterior confidence of the 8 low-score candidates above the 0.35 threshold and recovered the others by leveraging temporal consistency and spatial priors. Through Bayesian Enhancement, the GVF module leveraged the high spatial prior of the road layer to boost the posterior confidence of these candidates above the 0.35 threshold. This resulted in the recovery of all 12 missed targets, effectively increasing the recall rate. The detailed quantitative performance comparison between the baseline YOLOv11 and our proposed GVF framework is summarized in **Table 2**.

Method	Correct Detections	Missed (FN)	False Positives (FP)	Recall	Precision
Baseline (YOLOv11)	73	12[Note 1]	2	85.9 %	97.3%
Proposed GVF	85	0	0	100%	100%

Table 2. Performance comparison between the baseline YOLOv11 and the proposed GVF framework in the Baibuting Community experiment. Note 1: Among the 12 missed targets, 8 candidates were localized with initial confidence scores between 0.12 and 0.31, while the remaining 4 were entirely bypassed by the visual detector due to extreme environmental factors.

3.3 Discussion on System Robustness and Limitations

The efficacy of the Geo-Visual Fusion (GVF) strategy is inherently coupled with the quality of the multi-source ancillary data and the precision of the spatial transformation pipeline. During practical deployment, three primary factors may introduce geometric uncertainty:

Multi-Source Positioning Errors: Beyond the horizontal drift of GPS/IMU systems, the accuracy of the Digital Elevation Model (DEM) is critical for precise image-to-map projection. Errors in the underlying DEM can lead to significant vertical-to-horizontal displacement during the coordinate transformation, causing the detected objects to be incorrectly mapped onto the HD map layers.

Map Latency and Dynamics: The 'static' nature of HD maps presents a challenge in rapidly evolving urban landscapes. If the geo-spatial database is not updated to reflect recent structural changes (e.g., new construction or demolished buildings), the Semantic Compatibility Matrix (SCM) might provide outdated priors, potentially leading to the suppression of valid detections.

Sensor Alignment: Minor boresight misalignments between the camera and the navigation sensors can amplify projection errors at higher flight altitudes.

Future work will explore the integration of real-time Simultaneous Localization and Mapping (SLAM) and online DEM refinement techniques to improve the system's resilience against sensor noise and geo-spatial data obsolescence.

3.4 Computational Overhead Analysis

The GVF framework is designed for high efficiency, functioning as a lightweight post-processing engine rather than a compute-intensive deep learning refinement. By utilizing 2D vector-based spatial queries and optimized indexing structures (such as R-trees), the system converts complex scene understanding into low-complexity spatial joins. In experimental evaluations, the additional inference latency was under 2 ms per frame, a negligible overhead for real-time applications.

This minimal computational footprint makes the framework ideal for edge computing platforms with constrained resources, such as the DJI Dock 3 or NVIDIA Jetson series. Because it relies on CPU-efficient spatial logic rather than consuming additional GPU VRAM, the GVF strategy can be implemented as a "plug-and-play" enhancement. It significantly improves detection reliability without necessitating hardware upgrades or compromising the UAVs operational flight time due to increased power consumption.

4. Conclusion and Discussion

In this research, we have developed and validated a lightweight, plug-and-play enhancement framework that integrates static high-definition (HD) map priors with real-time UAV visual perception. By bridging the inherent "semantic gap" in vision-only models, our Geo-Visual Fusion (GVF) strategy effectively addresses the long-standing "context-blindness" problem that plagues Unmanned Aerial Vehicles operating in complex urban canyons. Leveraging the comprehensive municipal geomatics resources of our institute, we have demonstrated that deterministic geospatial constraints can serve as a powerful supervisor for probabilistic deep learning outputs, significantly reducing logical fallacies such as the mis-identification of building facade elements as vehicles.

The core contribution of this work lies in the conceptual shift from isolated visual processing to integrated geo-spatial reasoning. Unlike traditional methods that require computationally expensive retraining or massive new datasets, our proposed GCR module provides a computationally efficient post-processing solution. Experimental results from the Baibuting Community in Wuhan confirm that this fusion strategy not only eliminates high-confidence false positives but also successfully recovers occluded targets by analyzing their geographic plausibility. The framework’s low latency and minimal hardware requirements make it exceptionally suitable for deployment on edge computing devices, such as the DJI Dock 3, supporting long-term, autonomous urban sensing missions.

Looking forward, the GVF framework demonstrates robust scalability across a wide array of intelligent urban governance applications. Beyond object detection, the principles of geo-visual fusion can be extended to infrastructure health monitoring, real-time traffic flow analysis, and automated hazard detection in sensitive urban zones. Furthermore, our discussion on the "reverse empowerment" mechanism opens a new pathway for closed-loop HD map maintenance, where UAV perception data can be used to identify and update dynamic changes in the urban fabric. Future research will focus on integrating multi-temporal map data and exploring more sophisticated Bayesian inference models to further refine the synergy between the digital twin and the physical world, ultimately contributing to the development of more reliable and intelligent Urban Air Mobility (UAM) ecosystems.

Acknowledgments

This work was supported by the Wuhan Science and Technology Achievement Transformation Project titled "Multi-scenario Application Demonstration of Satellite Data" (Grant No. 2024031103010837).

References

Ajmera, F., Meshram, S., Nemade, S., et al., 2021. Survey on object detection in aerial imagery. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). IEEE, pp. 1050-1055.

Bell, S., Zitnick, C. L., Bala, K., Girshick, R., 2016. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2874-2883.

Cohen, A. P., Shaheen, S. A., Farrar, E. M., 2021. Urban air mobility: History, ecosystem, market potential, and challenges. IEEE Transactions on Intelligent Transportation Systems, 22(9), pp. 6074-6087.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, Vol. 1, pp. 886-893.

Liu, Y., et al., 2022. A Real-Time Registration Algorithm of UAV Aerial Images Based on Feature Matching. Remote Sensing, 14(9), 2056.

Liu, Z., Gao, G. H., Lin, L., et al., 2020. Learning graph-based center-point representations for object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 58(12), pp. 8549-8562.

Ma, H., et al., 2023. High-definition maps for autonomous driving: A survey and future directions. IEEE Transactions on Intelligent Vehicles, 8(2), pp. 1184-1202.

Nex, F., et al., 2018. Optimization of OpenStreetMap Building Footprints Based on Semantic Information of Oblique UAV Images. Remote Sensing, 10(4), 624.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems (NIPS), 28, pp. 91-99.

Sapkota, R., Roumeliotis, K. I., Karkee, M., 2025. UAVs Meet Agentic AI: A Multidomain Survey of Autonomous Aerial Intelligence and Agentic UAVs. arXiv preprint arXiv:2506.08045.

Sheikh, Y., Shah, M., 2005. Bayesian object detection in dynamic scenes. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, Vol. 1, pp. 74-79.

Siris, A., Jiao, J., Tam, G. K. L., et al., 2021. Scene context-aware salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4156-4166.

Sun, X., Wang, P., Yan, Z., et al., 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 184, pp. 116-130.

Xia, G. S., Bai, X., Ding, J., et al., 2018. DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3974-3983.

Zhang, G., Lu, S., Zhang, W., 2019. CAD-Net: A Context-Aware Detection Network for Objects in Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 57(12), pp. 10012-10024.

Zhang, Y., Sun, P., Jiang, N., et al., 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In: European Conference on Computer Vision (ECCV).

Zhao, Z. Q., Zheng, P., Xu, S., et al., 2019. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11), pp. 3212-3232.