Geospatially Grounded Satellite Image Captioning with Improved Factual Consistency using Directional Scene Graphs and Large Language Models: An Automated Pipeline for Semantic Summarization of Satellite Images

Patil, S.,[1*] More, N.[2] and Mohite, N. [3*]
Department of Information Technology K. J. Somaiya School of Engineering (formerly known as K. J. Somaiya College of Engineering) Mumbai, India
E-mail: suchitrapatil@somaiya.edu,[1*] https://orcid.org/0000-0002-3418-806X
neelkamalsurve@somaiya.edu,[2] https://orcid.org/0000-0003-0964-9490
nikhil.mohitel@somaiya.edu,[3] https://orcid.org/0009-0002-1815-270X
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v22i10.5236

Abstract
Remote sensing image captioning has recently attracted significant attention for automated interpretation of satellite imagery; however, existing approaches often rely primarily on visual features and insufficiently model explicit geospatial relationships, resulting in factually inconsistent descriptions and limited spatial reasoning. To address this limitation, this paper proposes an automated end-to-end framework for geospatially grounded semantic summarization of remote sensing images by integrating object detection, geographic coordinate transformation, directional spatial reasoning, and large language models (LLMs). The proposed framework uses the YOLOv11 object detector trained and evaluated on the DIOR dataset to identify objects and then the affine transformation is applied to map pixel-level bounding boxes into an assumed WGS-84 coordinate system. A novel Directional Spatial Relationship Graph (DSRG) is then constructed to model adjacency, density-based clustering, transportation proximity, and object importance, thereby capturing both geometric and functional relationships among detected objects. The extracted spatial knowledge is organized using a Grounded Prompt Framework (GPF), which constrains LLMs to generate geographically grounded summaries based exclusively on detected objects and their spatial context. The framework is evaluated using object detection metrics (Precision, Recall, mAP), Geographic Intersection-over-Union (Geo-IoU), graph statistics, grounding score, hallucination rate, ROUGE, and BERTScore. Experimental results demonstrate that the proposed YOLOv11 model achieves a Precision of 0.893, Recall of 0.784, mAP@0.5 of 0.854, and mean Geo-IoU of 0.8685. The comparative analysis of different evaluated language models showed that the structured prompting strategy substantially improves factual consistency, with Gemini 2.5 Flash achieving a hallucination rate of 1.6% at the same time maintaining competitive grounding performance. A DIOR dataset does not provide actual georeferenced object annotations; geospatial consistency is evaluated using synthetic reference polygons. This represents a limitation of the present study. Overall, the proposed DSRG–GPF framework provides an effective approach for generating geospatially grounded and factually consistent semantic summaries of remote sensing imagery, indicating potential applications in geospatial intelligence and GIS-based decision-making applications.

Keywords: Directional Scene Graph, Large Language Models, Geospatial Projection, Object Detection, GeoIoU, Satellite Imagery

1. Introduction

Satellite imagery has become an indispensable source of information for numerous geospatial applications, including urban planning, environmental monitoring, disaster management, precision agriculture, transportation analysis, and infrastructure development.

The rapid growth in high-resolution Earth observation imagery has created an increasing demand for automated techniques capable of extracting meaningful semantic information from large volumes of remote sensing data.

Although advances in computer vision have significantly improved object detection and scene understanding, converting complex satellite imagery into accurate and human-readable textual descriptions remains a challenging problem because it requires not only object recognition but also contextual interpretation of spatial relationships. Early remote sensing image captioning methods primarily relied on CNN–RNN encoder decoder architectures that generated natural language descriptions directly from image features. More recently, vision transformers and vision-language foundation models have substantially improved caption fluency and semantic representation. Despite these advances, many existing approaches still focus mainly on visual appearance and frequently overlook explicit geographic context, object interactions, and functional spatial relationships that are essential for interpreting satellite imagery. Consequently, generated captions may incorrectly describe object counts, ignore important spatial arrangements, or introduce information that is not supported by the image. Scene graph representations have emerged as an effective mechanism for organizing objects and their relationships into structured knowledge graphs. Likewise, large language models (LLMs) have demonstrated remarkable capabilities in generating coherent textual descriptions from structured inputs.

However, conventional scene graphs used in computer vision mainly capture geometric relationships and are rarely designed for remote sensing applications where geographic location, transportation connectivity, object clustering, and functional land-use relationships are equally important. Furthermore, when LLMs receive insufficiently grounded visual information, they are prone to hallucinations, resulting in descriptions that contain unsupported objects or incorrect spatial reasoning. Based on the existing literature, three important research challenges remain. First, current remote sensing captioning methods rarely integrate explicit geospatial coordinate transformation with structured spatial reasoning. Second, existing scene graph approaches generally do not incorporate directional, functional, clustering, and role-based spatial relationships within a unified representation suitable for satellite imagery. Third, although structured prompting has been explored in vision-language research, integrating explicit directional and functional spatial relationships into remote sensing caption generation remains an area requiring further investigation. These limitations motivate the need for a framework that combines accurate object detection, geospatial reasoning, structured scene representation, and grounded prompt generation. To address these challenges, this paper proposes a fully automated end-to-end framework for geospatially grounded semantic summarization of remote sensing images. Initially, objects are detected using the YOLOv11 model trained on the DIOR dataset. Since the DIOR dataset does not provide real-world geographic coordinates, detected pixel-level bounding boxes are transformed into WGS-84 geographic coordinates using affine geospatial projection. A novel Directional Spatial Relationship Graph (DSRG) is then constructed to model adjacency, density-based clustering, transportation proximity, and object importance. The graph-derived spatial knowledge is subsequently organized into a Grounded Prompt Framework (GPF) that guides multiple large language models, including Gemini 2.5 Flash, LLaMA-3.1-8B, and LLaMA-3.3-70B, to generate geospatially grounded semantic summaries while reducing unsupported or hallucinated information. The major contributions of this work are summarized as follows:

- An end-to-end framework that integrates YOLOv11 object detection, affine geospatial projection, Directional Spatial Relationship Graph (DSRG) construction, and grounded prompt generation for LLM based semantic summarization of remote sensing imagery.
- A DSRG representation that combines directional adjacency, density-based clustering, transportation connectivity, and object importance to provide richer spatial context than conventional scene graphs.
- A Grounded Prompt Framework (GPF) that converts structured geospatial knowledge into LLM-compatible prompts, improving factual consistency and reducing hallucinations during summary generation.
- The comprehensive experimental evaluation on the DIOR dataset using object detection metrics, Geographic Intersection-over-Union (Geo-IoU), grounding score, hallucination rate, ROUGE, and BERTScore across multiple state-of-the-art large language models.

2. Literature Survey

2.1 Remote Sensing Image Captioning

Remote sensing image captioning has evolved considerably over the past decade. Early encoder decoder models based on CNN–RNN architectures demonstrated that natural image captioning techniques could be adapted to satellite imagery through datasets such as RSICD [1]. Although these methods generated fluent descriptions, they struggled to accurately represent complex object

distributions, spatial relationships, and object counts in high-resolution remote sensing scenes. These limitations motivated the development of more specialized vision-language approaches for Earth observation [2] and [3].

More recent vision-language foundation models, including CLIP-based frameworks and diffusion-based captioning approaches, have substantially improved semantic representation and caption fluency [2] [3] [4] and [5]. These models achieve higher BLEU, METEOR, and CIDEr scores by learning stronger multimodal representations [2] and [3]. However, because they primarily rely on visual feature alignment, they often fail to explicitly model geographic relationships and functional interactions among objects, limiting their applicability for geospatial reasoning tasks [2] and [6]. In parallel with caption generation research, significant progress has been achieved in remote sensing object detection [7] and [8]. Modern detectors such as YOLOv11, SSD, Faster R-CNN, and Cascade R-CNN have demonstrated increasingly accurate localization of objects in very high-resolution satellite imagery [7] and [8]. In particular, YOLOv11 combines an anchor-free detection strategy with improved feature extraction, attention mechanisms, and multi-scale feature aggregation, making it particularly effective for complex remote sensing scenes. Nevertheless, accurate object detection alone cannot provide semantic interpretation unless explicit spatial relationships among detected objects are modeled [9].

2.2 Spatial Reasoning and Scene Graphs

Spatial reasoning has become an important component of vision-language systems because it enables machine learning models to understand not only the presence of objects but also the relationships among them. Conventional scene graphs used in computer vision represent a scene as a graph consisting of object nodes connected through pairwise relationships such as *left of*, *right of*, *above*, *below*, *inside*, or *overlapping*. These representations have been widely adopted for visual reasoning, image captioning, and visual question answering because they provide structured contextual information beyond individual object detections [9]. In the remote sensing domain, several studies have extended scene graph representations by incorporating directional and topological relationships, such as *north of*, *adjacent to*, and distance-based spatial relationships between detected objects [9]. These approaches demonstrated improved semantic expressiveness compared with using only bounding box coordinates. However, most existing methods rely primarily on geometric relationships and template-based text generation, which limits their ability to produce rich and context-aware scene descriptions suitable for Large Language Models (LLMs) [2] [3] and [9].

The Scene-LLM framework further demonstrated that scene graph representations can significantly improve reasoning capabilities by converting 3D point-cloud data into graph structures before providing them to an LLM. Although this approach achieved strong performance for indoor scene understanding and visual question answering, it depends on detailed 3D annotations and computationally intensive encoders, making it unsuitable for large-scale satellite imagery where only 2D aerial images are typically available [10]. Unlike conventional scene graphs, the proposed Directional Spatial Relationship Graph (DSRG) extends the graph representation by integrating four complementary categories of spatial knowledge: (i) directional and topological neighbor relationships, (ii) density-based clustering relationships obtained using DBSCAN [11], (iii) functional relationships that associate nearby objects with transportation infrastructure, and (iv) object importance through Spatial Role labels (dominant, major, and minor). These additional relationships enable the graph to capture not only geometric arrangements but also functional context and object significance within the scene. These additional relationships enable the graph to capture not only geometric arrangements but also functional context and object significance within a scene, thereby providing a richer semantic representation for satellite image interpretation.

2.3 LLM Integration for Geospatial Tasks

Studies have investigated using large language models (LLMs) together with remote sensing data [2] [3] and [6]. For example, He showed that fluent generation of scene descriptions is possible from LLMs when structured input is given; however, their study demonstrated that structured scene representations improve the factual consistency of generated descriptions compared with unstructured feature inputs. Urban MLLM explored the potential of joint pre-training of satellite imagery and street-view imagery for urban scene understanding [12]. Their method yielded gains in scene classification and retrieval. However, the large amount of computation required hinders its practical use [12]. EagleVision introduced attribute-level understanding of objects using an attribute disentanglement module and achieved high recall in benchmark datasets [13]. While it performed well, it only focused on attribute level of objects without

providing any information about relationship understanding [13].

2.4 Research Gaps Addressed by Our Work

Based on the reviewed literature, three major research gaps remain in existing remote sensing image understanding methods.

First, existing scene graph approaches primarily model geometric or topological relationships between objects and rarely integrates multiple complementary forms of spatial knowledge within a unified representation [2] [3]and [9]. In particular, current methods generally do not combine directional, density-based, functional, and role-based relationships into a single graph structure suitable for satellite imagery [9] and [11]. To address this limitation, the proposed Directional Spatial Relationship Graph (DSRG) enriches conventional scene graphs by incorporating neighbour connectivity, object clustering, transport infrastructure relationships, and object importance information, thereby providing a more comprehensive semantic representation of complex aerial scenes.

Second, most existing LLM-based remote sensing captioning methods rely on unstructured prompts, which limits the factual grounding of the generated descriptions [2][3] and [6]. The proposed Grounded Prompt Format (GPF) addresses this limitation by supplying the LLM with structured graph-derived spatial knowledge.

Third, previous studies have rarely performed a systematic comparison of geospatially grounded semantic summarization across multiple Large Language Models using a common evaluation framework [2] and [3]. This work addresses this gap by evaluating the proposed DSRG–GPF framework on several state-of-the-art LLMs using identical datasets, prompting strategies, and evaluation metrics [14] and [15].

3. Methodology

Figure 1 shows the whole pipeline. Major steps include: (i) Data Preprocessing and Tiling, (ii) Object Detection, (iii) Geo-Projection and Validation, (iv) Directional Scene Graph Construction, and (v) LLM-based Summarization.

3.1 Dataset Acquisition

A publicly available aerial object detection dataset, DIOR from the Hugging Face Hub was utilized as the main dataset benchmark in the research study [16]. The sample images are shown in Figure 2. The statement “load_dataset("HichTala/dior")” was used to load this dataset. This approach ensures reproducibility and eliminates manual data handling. Each image comes with annotation, which consists of structured annotations including labels and bounding boxes in format $[xmin, ymin, xmax, ymax]$.

3.2 Dataset Description

The DIOR (Dataset for Object Detection in Optical Remote Sensing Images) [16] is made up of a total of 23,463 images that are available for public use. These images contain 20 different classes, which fall into the following major class groupings: Large scale objects (for instance, airports and stadiums) and small-scale objects (for instance, vehicles and ships). The size of the objects, the scene presentation, and the conditions of imaging affect the quality of DIOR, thus making it a great choice for testing different multi-object detection systems in connection with remote sensing. The accuracy in the images is taken to be around 800 × 800 pixels with a spatial resolution ranging from 0.5 m to about 30 m [15].

However, in the particular case of the DIOR dataset derived from Hugging Face, no such partitions were provided. Hence, we perform our own custom split taking 80% of data for training while using 10% for validation and the last 10% for testing purposes. More than 18,770 images were used for training one model, 2,346 images were used for validation purposes, and 2,347 images were used for the purpose of testing. In the end, each separate subset was taken in order to optimize the parameters of the YOLOv11 model. Table 1 presents the 20 object categories included in the DIOR dataset. The significant variation in object scale and scene complexity enables comprehensive evaluation of object detection performance in diverse remote sensing scenarios. Furthermore, the public availability of the dataset enhances the reproducibility and comparability of the proposed framework. The DIOR dataset provides diverse object categories (Table 1) with significant scale variation, which is essential for training models for robust detection in remote sensing scenarios. Its public accessibility also helps make the research findings reproducible.

3.3 Dataset Preprocessing

Prior to training, the raw satellite images are preprocessed to ensure compatibility with the YOLO model. The preprocessing pipeline also improves object detection performance. The workflow consists of format conversion, image scaling, data augmentation, and tiled inference [10].

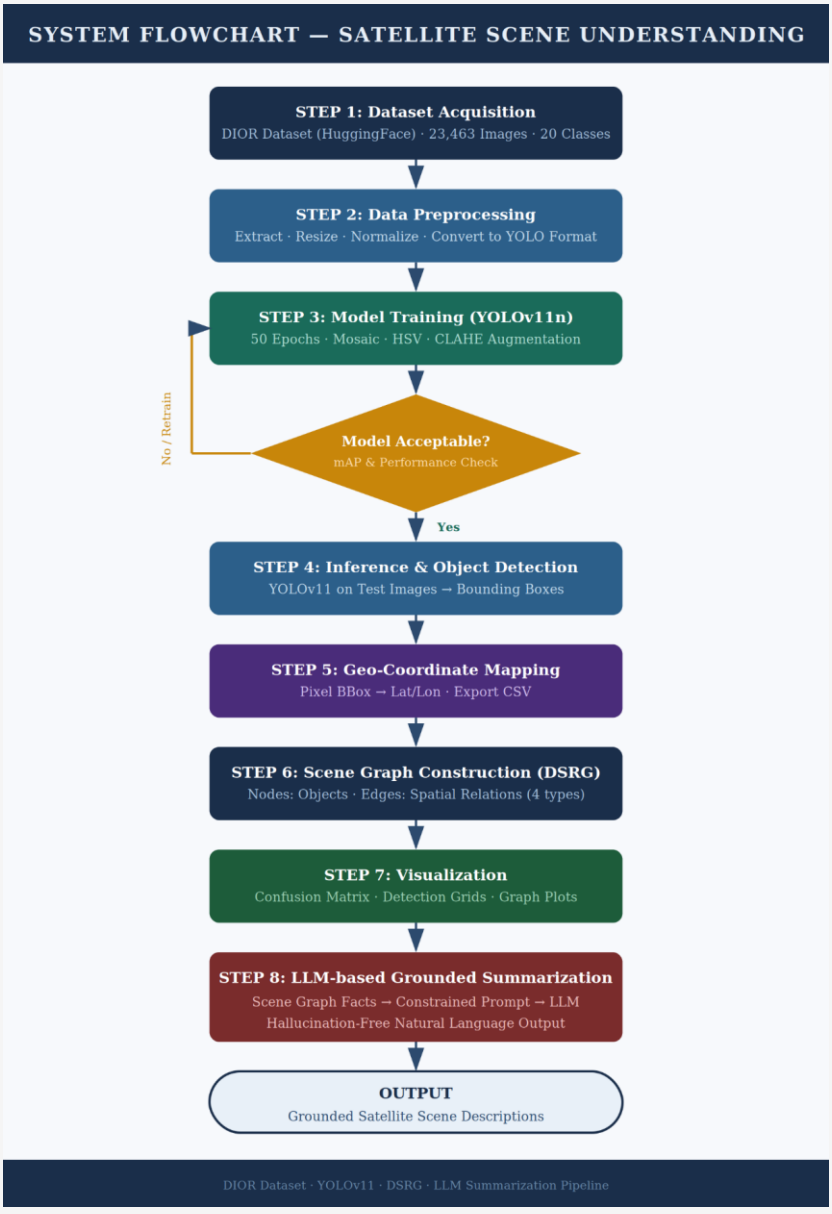


Figure 1: Overall workflow of the proposed DSRG–GPF framework for geospatially grounded satellite image summarization**.** The pipeline consists of five sequential stages: **(1)** satellite image preprocessing and object detection using YOLOv11, **(2)** geospatial coordinate transformation through affine projection, **(3)** construction of the Directional Spatial Relationship Graph (DSRG), **(4)** generation of a Grounded Prompt Format (GPF) containing spatial and semantic information, and **(5)** Large Language Model (LLM)-based generation of factually grounded scene summaries

Table 1: DIOR dataset object categories

ID	Category	ID	Category	ID	Category	ID	Category
0	Airplane	5	Chimney	10	Ground Track Field	15	Storage Tank
1	Airport	6	Dam	11	Harbor	16	Tennis Court
2	Baseball Field	7	Expressway Service Area	12	Overpass	17	Train Station
3	Basketball Court	8	Expressway Toll Station	13	Ship	18	Vehicle
4	Bridge	9	Golf Field	14	Stadium	19	Windmill

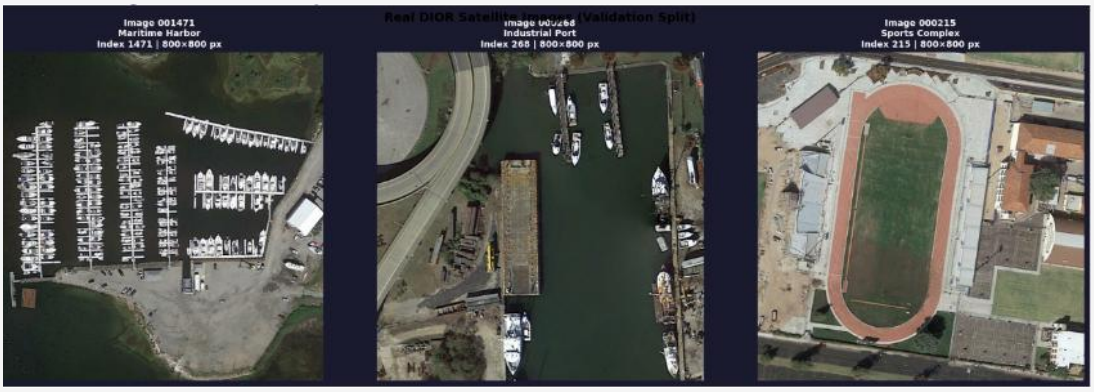


Figure 2: Representative satellite images from the DIOR dataset illustrating the diversity of object categories and scene types used in this study. The dataset includes various remote sensing environments such as airports, harbours, bridges, sports facilities, vehicles, and industrial areas, providing a comprehensive benchmark for object detection and semantic scene understanding

3.3.1 Format conversion

The annotations from the DIOR dataset are provided in absolute terms (x_min,y_min,x_max, y_max). These values are normalized to fit the YOLO format as shown in Equation 1:

$$x_c = \frac{x_{\min} + x_{\max}}{2W}, \quad y_c = \frac{y_{\min} + y_{\max}}{2H}$$

$$W = \frac{x_{\min} - x_{\min}}{W}, \quad h = \frac{y_{\max} + y_{\min}}{H}$$

Equation 1

Where:

x_c = normalized coordinates of the box center along the horizontal directions.

y_c = normalized coordinates of the box center along the Vertical directions

w = normalized width of the box relative to the image size

h = normalized height of the box relative to the image size

W = correspond to the overall width of the image

H = correspond to the overall height of the image

(x_{min}, y_{min}) is the top-left corner of the box, and (x_{max}, y_{max}) is the bottom-right corner

Pixel coordinates are normalized to relative values in the range 0–1, making the annotations independent of image resolution. This normalization enables the model to process images of different sizes without requiring changes to the annotation format. Consequently, the model achieves improved scale invariance and maintains consistent detection performance across satellite images with varying resolutions.

3.3.2 Image scaling

All images are resized to 640×640 pixels with bicubic interpolation. This standardization is done in accordance with YOLO's input specifications while ensuring that minute details essential for recognizing smaller objects are retained [8].

3.3.3 Data augmentation

Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to improve local image contrast. The technique performs histogram equalization within small image regions while preventing excessive amplification of image noise. As a result, faint objects such as vehicles and ships become more distinguishable in satellite imagery, thereby improving detection performance [5].The intensity obtained after the transformation is defined in Equation 2:

$$T(k) = \frac{L-1}{MN} \sum_{i=0}^{k} h_{clip}(i)$$

Equation 2

Where:

$T(k)$ = intensity value

L = number of gray levels in the image

M, N = dimensions of the local area (tile)

$h_{clip(i)}$ = clipped histogram value at intensity i

This transformation improves detection recall by enhancing faint object boundaries [17].

3.3.4 Mosaic augmentation

In mosaic augmentation, four images are used to create one training sample. This enhances the variety of objects and allows for multi-scale object detection. The resulting mosaic image is generated using Equation 3:

$$M(x,y)=\begin{cases} I_1(x,y) & ;x<x_c, y<y_c \\ I_2(x-x_c,y) & ;x\geq x_c, y<y_c \\ I_3(x,y-y_c) & ;x<x_c, y\geq y_c \\ I_4(x-x_c,y-y_c) & ;x\geq x_c, y\geq y_c \end{cases}$$

Equation 3

Where:

$M(x,y)$ = resulting mosaic image
I_1, I_2, I_3, I_4 = input images
x_c, y_c = center coordinates where images are merged
x, y = location of pixels in the output image

Mosaic augmentation exposes the model to multiple objects within a single training sample. This improves robustness to varying object scales, dense object distributions, and diverse spatial contexts commonly found in remote sensing imagery [2].

3.3.5 Tiled inference

High-resolution satellite images are divided into 640 × 640-pixel tiles before inference. Object detection is performed independently on each tile. The detections from all tiles are then merged using Non-Maximum Suppression (NMS) to obtain the final prediction results [3].

3.4 Object Detection

Object detection is the core component of the proposed pipeline, responsible for identifying and localizing objects in satellite images. This work evaluated YOLOv8 and YOLOV11 both and then adopted YOLOv11 due to its improved performance.

3.4.1 YOLOv8: Initial model

YOLOv8 is a single-stage, anchor-free object detector consisting of a CSPDarknet backbone with C2f modules, a PANet neck, and a decoupled detection head. Distribution Focal Loss (DFL) is employed to improve bounding box regression accuracy. The coordinate of the predicted bounding box is defined in Equation 4:

$$\hat{x}=\sum_{i=0}^{R-1} softmax(z_i)\cdot i$$

Equation 4

Where:

$\hat{y}$ = predicted bounding box coordinate
z_i = predicted logits for bin i
R = number of discrete bins used for regression

Total loss is given by Equation 5:

$$L=\lambda_{box}L_{CIoU}+\lambda_{cls}L_{BCE}+\lambda_{dfl}L_{DFL}$$

Equation 5

Where:

L_{CIoU} = accuracy of bounding box regression
L_{BCE} = evaluate how well the model distinguishes between object classes.
L_{DFL} = distribution focal loss
$\lambda_{box}, \lambda_{cls}, \lambda_{dfl}$ = weighting coefficients of loss terms

3.4.2 YOLOv11: proposed model

YOLOv11 performs object detection using an enhanced C3k2 backbone, the Convolutional Block Attention Module (CBAM), and optimized feature aggregation. These architectural improvements enhance feature representation and improve object detection precision. CBAM Attention mechanism is given by following Equation 6:

$$M_c(F)=\sigma\left(MLP(AvgPool(F))+MLP(MaxPool(F))\right)$$
$$F'=M_c(F)\otimes F$$
$$M_s(F')=\sigma(Conv_{7\times7}([AvgPool(F'),MaxPool(F')]))$$
$$F''=M_s(F')\otimes F'$$

Equation 6

Where:

F = represent the input feature map
$M_c(F)$ = emphasize important feature channels,
$M_c(F')$ = spatial attention mechanism that focuses on relevant regions within the feature map
σ = normalize the attention values
$\otimes$ = combine the attention maps with the input features.

The overall loss functions are designed to jointly optimize both localization and classification performance. The loss functions used in this study are defined in Equation 7.

$$L=\lambda_{box}L_{CIoU}+\lambda_{cls}L_{BCE}+\lambda_{dfl}L_{DFL}$$
$$L_{CIoU}=1-IoU+\frac{\rho^2(b,b_{gt})}{c^2}+\alpha v$$

Equation 7

Where:

IOU = intersection over union
$\rho(b, b_{gt})$ = distance between center points of prediction and ground truth box
c = diagonal length of smallest bounding box
αv = aspect ratio consistency

3.5 Evaluation Metrics used in YOLO

The YOLO algorithm's performance is assessed based on object detection metrics that are common for the PASCAL VOC and COCO datasets.

3.5.1 Precision and recall

Precision and Recall measure detection accuracy and completeness. The precision and recall are determined from Equations 8 and 9, respectively.

$$Precision = \frac{TP}{TP+FP}$$

Equation 8

$$Recall = \frac{TP}{TP+FN}$$

Equation 9

Where:

TP = represents correctly detected objects
FP = represent incorrect detections
FN = correspond to missed objects

High precision indicates that the model produces fewer false-positive detections. High recall indicates that fewer objects are missing during detection [7].

3.5.2 Mean Average Precision (AP)

Average Precision (*AP)* is used to summarize the model's performance by measuring the area under the precision–recall curve. The average precision is expressed in Equation 10:

$$AP = \sum \left(R_n - R_{n-1}\right) P_n$$

Equation 10

Where:

P_n = precision at step n
R_n = recall at step n

The *AP* is first computed for each object class. The mean Average Precision (*mAP*) is then obtained by averaging the *AP* values across all object categories as presented in Equation 11:

$$mAP = \frac{1}{C} \sum_{c=1}^{c} AP_c$$

Equation 11

Where:

C = total number of classes
AP_c = average precision for class c

mAP@0.5 metric considers detections correct when the Intersection over Union (IoU) exceeds 0.5. While *mAP@0.5:0.95* in Equation 12 provides a more comprehensive evaluation by averaging results across a range of IoU thresholds.

$$mAP@0.5:0.95 = \frac{1}{10} \sum_{t=0.50}^{0.95} mAP_t$$

Equation 12

3.5.3 Automatic Mixed Precision (AMP)

Automatic Mixed Precision (*AMP*) is employed to improve training efficiency by combining different numerical precisions. It uses FP16 and FP32 precisions [12]:

- FP16 for faster computation
- FP32 for stable loss and gradient updates

Automatic Mixed Precision reduces GPU memory consumption and accelerates model training. At the same time, it preserves numerical stability and detection accuracy, making it suitable for training YOLOv11 on large-scale remote sensing datasets [8].

3.5.4 Inference and object detection

After training, both the YOLO models were used for inference on satellite images to generate bounding box detections. The input image is processed at a fixed resolution of 640×640, where the model predicts object locations, confidence scores, and class probabilities. Predictions with confidence above a defined threshold are retained for further processing. To remove duplicate detections, Non-Maximum Suppression (NMS) [2] is applied, which keeps the most confident bounding box while eliminating overlapping ones based on an IoU threshold. For large satellite images, tiled inference is used by dividing the image into overlapping patches. Detection is performed on each patch separately, and the results are mapped back to the original image. Finally, NMS is applied across all tiles to produce accurate and consistent object detections [18].

3.5.5 Output Format for detected objects

The detected objects are stored in a database containing the image ID, object category, bounding box coordinates, confidence score, and tile offsets. This structured output serves as the input for the subsequent geospatial coordinate conversion stage [14].

3.6 Geo-Projection and Spatial Consistency Validation

In the second stage, the coordinates of all the bounding boxes are converted into geographical coordinates through an affine transformation which is computed from the raster metadata. For each DIOR image, the following geographic extent is assumed, and the affine transformation in Equation 13 is applied:

$$T = \begin{bmatrix} a & 0 & c \\ 0 & e & f \\ 0 & 0 & 1 \end{bmatrix}$$

Equation 13

Where:

a, e = pixel resolution in longitudinal and latitudinal directions

c, f = geographic coordinates of the image origin

In this constructed a matrix, parameters a and e represent pixel resolution in longitudinal and latitudinal directions respectively. Each detected bounding box (x_{min}, y_{min}, x_{max}, y_{max}) is projected from pixel space into geographic space using this transformation, yielding a geospatial polygon defined in the EPSG: 4326 (WGS-84) coordinate system [13]. To evaluate spatial consistency of the projection process, the Geographic Intersection-over-Union (*Geo-IoU*) in Equation 14 is computed.

$$Geo - IoU = \frac{Area\left(P_{pred} \cap P_{ref}\right)}{Area\left(P_{pred} \cup P_{ref}\right)}$$

Equation 14

Where:

P_{pred} = projected detection polygon

P_{ref} = reference (synthetic ground truth) polygon

Where, P_{pred} is the projected detection polygon and P_{ref} is a synthetically generated geo-referenced counterpart. Polygon intersection and union areas are computed using Shapely-based geometric operations directly in geographic space [19].

The DIOR dataset does not provide georeferenced object annotations or geographic coordinate metadata for the annotated bounding boxes. Consequently, direct validation of projected object locations against real-world geographic coordinates is not possible. To enable quantitative evaluation of the proposed geospatial projection framework, synthetic reference polygons were generated by applying controlled geometric transformations to the projected object boundaries. These synthetic reference polygons simulate realistic geospatial uncertainties commonly seen during remote sensing data processing, including orthorectification errors, sensor positioning inaccuracies, map projection distortions, and minor spatial shifts introduced during image tiling. Rather than representing true field-measured object locations, they provide a controlled reference for evaluating the consistency and robustness of the proposed projection algorithm under different levels of spatial perturbation [11].

Accordingly, the Geo-IoU metric reported in this study measures the ability of the proposed framework to preserve geometric consistency under simulated geospatial uncertainty. It should therefore be interpreted as a measure of projection robustness rather than an assessment of absolute geographic positioning accuracy. Validation against real georeferenced datasets remains an important direction for future work.

3.7 Directional Scene Graph Construction

The DSRG constitutes the primary innovation proposed in this work. In this work we have focused on four major spatial relationships: AdjacentTo, ClusteredWith, NearTransportInfra, SpatialRole, as these are more appropriate for the mages in DIOR dataset and are designed to capture spatial relationships at multiple scales and multiple types, such as topological nearness, cluster formation, functional nearness, and relative significance [1]. The DSRG represents objects and spatial relationships between them represented as a graph $G=(V, E)$, where each vertex $v_i \in V$ represents an object and contains attributes such as its class label (c_i), centroid coordinates (lon_i, lat_i), and geographic bounding polygon (P_i) with area (A_i) expressed in square meters [12]. Each directed edge (v_i, v_j, r) $\in E$ represents a spatial relationship between a pair of objects v_i and v_j, where r is selected from the set of relation types.

3.7.1 AdjacentTo

Two objects v_i and v_j are deemed as neighbors when their geo-projected bounding box polygons either overlap or intersect, based on the adjustable buffer distance (ϵ) shown in Equation 15:

$$\varepsilon = 0.15 \min\left(w_A, w_B\right)$$

Equation 15

Where w_A and w_b denote the widths of the two objects (in degrees).

The buffer distance is defined as 15% of the smaller object width to accommodate minor localization inaccuracies introduced during object detection and

geospatial projection while preserving meaningful spatial relationships. Using a relative buffer rather than a fixed distance enables the criterion to adapt automatically to objects of different sizes. Smaller objects therefore require stricter spatial precision, whereas larger objects can tolerate proportionally larger localization uncertainty. This value was selected based on empirical observations during preliminary experiments to balance missing valid neighbour relationships and introducing false spatial adjacencies. Polygon intersection is evaluated using standard geospatial operations (e.g., the Shapely intersects () function). Adjacency relationships are represented as undirected edges, where a single edge is added for each object pair (v_i, v_j) such that $i < j$.

3.7.2 ClusteredWith

For modeling clustered objects which belong to the same class type and are spatially located near to each other like ships, storage tanks, and sports courts, we use DBSCAN (Density-Based Spatial Clustering of Applications with Noise), on a per-class basis for each scene [2]. The algorithm used to compute the *ClusteredWith* Relation is given below:

Algorithm:

1.) For each class (c) present in scene s, construct the set: $C_c = \{(lon_i, lat_i) \mid class_i = c\}$

2.) Apply DBSCAN clustering on C_c with parameters: ε = 0.003°, where the value 0.003° corresponds to approximately 330 m at the reference latitude, and *minsample* =2.

3.) For each pair of objects (v_i, v_j) assigns the same cluster label *cluster_id* ≠ -1, add an edge: (v_i, v_j, *ClusteredWith*) ∈ E.

4.) Objects with *cluster_id* ≠ -1 are treated as noise and are not assigned any *ClusteredWith* relations.

The DBSCAN parameters were selected to capture spatially adjacent objects that typically form compact groups in remote sensing imagery, such as ships, storage tanks, and sports courts. The neighborhood radius of approximately 330 m was found to effectively preserve local spatial groupings while avoiding the merging of nearby but independent object clusters. These parameter values were determined through preliminary empirical evaluation and were retained consistently throughout all experiments [9].

There are several advantages of using DBSCAN over other clustering approaches, such as K-means and hierarchical clustering. First, DBSCAN does not necessarily need prior knowledge of the number of clusters, which is necessary when processing images with randomly placed satellites; second, DBSCAN can cluster datasets regardless of their shapes, therefore helping recognize clusters of an irregular shape, for instance, ships located in ports; and third, outliers will be recognized, thus avoiding clustering single objects [3]. All these advantages make it perfect choice for our work.

3.7.3 NearTransportInfrastructure

NearTransportInfrastructure represents domain-specific knowledge, which states that proximity between a non-transport object and transport infrastructure denotes a functional connection between the land uses. For instance, a storage tank found in proximity to the harbor indicates its connection with an industrial port area, whereas a sports ground found in proximity to the vehicle denotes a sports complex with parking facilities [20]. Distance calculation between the two objects employs the Haversine great-circle distance presented in Equation 16:

$$d_{ij} = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi_i \cos\phi_j \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

Equation 16

Where $\emptyset$ and λ denote the latitude and longitude (in radians), and R is the mean radius of the earth which is approximately 6,371,000 m.

A directed edge (v_i, v_j, *NearTransportInfra*) is added to the graph if the distance (d_{ij}) meet the criteria in Equation 17:

$$d_{ij} \leq 0.25 d_{diag}$$

Equation 17

Where, d_{diag} is the Euclidean diagonal length of the spatial tile which is approximately 1,957 m.

The proximity threshold was defined as 25% of the tile diagonal to establish meaningful functional relationships between transportation infrastructure and nearby objects while avoiding long-range associations that may not represent actual spatial interactions. A relative threshold was preferred over a fixed geographic distance because it scales consistently with the spatial extent of the processed image tile and remains applicable across different scenes. In DIOR dataset airplane, ship, vehicle, bridge, overpass, expressway_service_area, expressway_toll_station, train_station, harbor are considered as the transport infrastructure class objects and rest of the classes in the DIOR dataset are taken as non-transportation class objects.

3.7.4 SptialRole

The SpatialRole capability involves labeling each detected object according to its importance level based on the geographical boundary of the object compared to the biggest object in the same scene. The landmark objects can be differentiated using Equation 18 in LLM [4].

$$role(v_i) = \begin{cases} dominant, if\ A_i \geq 0.80 A_{\max} \\ major\quad , if\ 0.30 A_{\max} \leq A_i < 0.80 A_{\max} \\ minor\quad , if\ A_i < 0.30 A_{\max} \end{cases}$$

Equation 18

Where, A_{max} is the maximum object area within the scene.

The thresholds of 0.80 and 0.30 partition objects into dominant, major, and minor categories according to their relative spatial footprint within each scene. These thresholds were selected to distinguish large landmark objects from medium-sized contextual objects and smaller supporting objects, thereby enabling the Large Language Model to generate more informative and context-aware scene descriptions. The thresholds were determined through preliminary empirical observations to provide consistent role assignments across diverse DIOR scenes. Unlike other relational predicates that depend on pairwise spatial context, the *SpatialRole* attribute can be assigned independently to every detected object in the scene [5]. Summary of all four DSRG relation types with alogirthm used and key parameters is shown in Table 2.

Parameter Selection: The heuristic parameters adopted in the DSRG construction were selected through preliminary empirical experiments and domain knowledge of remote sensing imagery. Adaptive thresholds were preferred over fixed values whenever possible to improve generalization across different object scales and scene layouts. Although the selected parameters provided stable graph construction throughout the experimental dataset, a comprehensive sensitivity analysis was beyond the scope of the present study. Investigating the influence of these parameters on graph topology and semantic summarization performance will be considered in future work.

3.8 Geospatial Prompt Generation Framework

Geospatial Prompt Generation Framework (*GPF*) is an organized prompt generation template comprising five blocks capturing all knowledge of spatial information like geo-projection, scene graph, and DSRG relationship etc. produced by the pipeline. The organized prompt is optimized for being concise and token-efficient and serves as an input to several LLMs to create a five-block geospatial intelligence report [14]. The formal representation of the framework is defined in Equation 19:

$$P = (C, O, R, K, T)$$

Equation 19

Where, C denotes the context, O represents object-level information with geographic coordinates, R contains DSRG relations, K summarizes cluster information, and T defines the task instructions along with anti-hallucination constraints [11]. It is detailed with content and role in Table 3.

Table 2: Summary of DSRG relation types, algorithms, and dataset statistics

Relation Type	Algorithm	Key Parameter
AdjacentTo	Shapely polygon intersection	Buffer $\varepsilon = 0.15 \times \min(w_a, w_\beta$ in geographic degrees)
ClusteredWith	DBSCAN density clustering	$\varepsilon = 0.003°$ (≈ 330 m), *min_samples* = 2
NearTransportInfra	Haversine distance	Threshold = 25% of tile diagonal (≈ 1,957 m)
SpatialRole	Area-ratio thresholding	Major threshold = 80% of max area

Table 3: Five blocks of the GPF schema and their roles

Block	Symbol	Contents	Role
[CONTEXT]	*C*	Scene ID, CRS (WGS-84), tile bounds, object count, source	Anchors geographic context
[OBJECTS]	*O*	Class(lat, long), area, role, confidence	Provides object-level details
[DSRG RELATIONS]	*R*	Directional, adjacency, proximity relations	Encodes spatial relationships
[CLUSTERS]	*K*	DBSCAN cluster summary (ID, count, centroid)	Captures grouping patterns
[TASK]	*T*	Output format + anti-hallucination rules	Controls LLM output

Table 4: LLMs evaluated in the GPF framework

Model	Provider	Parameters	API
Gemini 2.5 Flash	Google AI Studio	Large multimodal	Gemini API (max tokens = 1,800)
LLaMA-3.1-8B	Meta / Groq	8B	Groq API (max tokens = 1,800)
LLaMA-3.3-70B	Meta / Groq	70B	Groq API (max tokens = 1,800)

3.8.1 Five-Section output structure

The *TASK* block instructs the LLM to generate a structured five-section report for each scene where information about each section is detailed below:

Section 1 Detected Objects: Lists all detected objects along with their class labels, counts, and confidence scores.

Section 2 Geo-Reference Data: Provides geographic coordinates, object area, and assigned spatial roles for each detected object.

Section 3 Scene Graph Relations: Describes directional and topological relationships derived from DSRG.

Section 4 Spatial Analysis: Interprets spatial patterns such as directional alignment, functional zones, and clustering behavior.

Section 5 Natural Language Summary: Generates a clear and human-readable description of the scene.

3.8.2 LLM Selection and API integration

The selection of LLMs in this work is carefully designed to evaluate how differences in model architecture, parameter size, and reasoning ability affect geospatial summarization performance. Instead of relying on a single model, three models are chosen Gemini 2.5 Flash [6], LLaMA-3.1-8B [19], and LLaMA-3.3-70B [15] to create a balanced comparison across efficiency, scalability, and accuracy. Gemini 2.5 Flash is selected for its strong reasoning capabilities, multimodal support, and ability to generate concise and well-structured outputs. It is particularly effective in tasks that require strict factual correctness and minimal hallucination, which is essential in geospatial applications such as infrastructure analysis and disaster monitoring. Its efficient design also allows faster inference while maintaining high-quality outputs [2].

To complement this, two variants from the LLaMA family are included to analyze the effect of model scale within the same architecture. LLaMA-3.1-8B represents a lightweight and resource-efficient model that can be deployed in constrained environments. Its inclusion helps evaluate whether smaller models can still produce meaningful summaries when guided by structured prompts. In contrast, LLaMA-3.3-70B is a large-scale model with significantly higher capacity, enabling deeper reasoning and better contextual understanding. Comparing these two models allows the study to isolate the impact of parameter size on summarization quality and factual consistency [6].

All models are accessed through their respective APIs. Gemini via Google AI Studio and LLaMA models via Groq ensuring reliable and efficient inference. A consistent token limit of 1800 tokens is applied across all models to ensure that each generates complete outputs without truncation (Table 4). Additionally, all models receive the same structured input generated through the Geospatial Prompt Framework (GPF), ensuring a fair and controlled comparison. All the models receive the same input in terms of the prompts structured in a similar way. The token limit is set to 1800 tokens to guarantee that all five sections are produced by all three LLMs without exceeding the restrictions. The parsing of the outputs is carried out through section header matching by means of regular expressions. Three LLMs are assessed to determine the impact of their sizes and architectures on geospatial summarization performance.

3.8.3 LLM-Based Grounded Summarization

One of the major contributions of this work is the use of well structured, constraint-driven prompting to guide the behaviour of large language models. The pipeline converts outputs from object detection, geo-projection, and scene graph construction into well-defined prompts that include object counts, geographic coordinates, spatial relationships, and clustering information. These prompts explicitly restrict the models to use only the provided data. This approach significantly reduces hallucination and ensures that generated summaries remain fully grounded in the input data. As a result, the models produce clear and human-readable descriptions that accurately reflect object distribution, spatial patterns, and functional characteristics of the scene [17]. This improves both the reliability and interpretability of the system, making it suitable for real-world applications such as urban planning, environmental monitoring, and decision support systems. To maintain consistency in evaluation, the generated outputs are further processed using section-based parsing techniques. Predefined headers are identified using regular expressions, allowing systematic extraction and comparison of different output components [14].

Overall, the combination of diverse LLM selection and structured prompting provides a comprehensive framework for evaluating grounded summarization. It highlights an important insight: while larger models generally offer better reasoning capabilities, carefully designed prompts can significantly improve the performance of smaller models, enabling a practical balance between accuracy, efficiency, and scalability.

3.8.4 Metrics for evaluation of entire system

The performance of the designed model is analyzed through four stages that involve evaluation of individual components of the entire system. Geospatial Assessment: Spatial precision is analyzed based on the Geo-IoU (Eq. 13) criterion calculated as a mean and median metric, and as Geo-IoU metrics at threshold values of 0.50 and 0.75. Also, the proportion of bounding boxes mapped to the CRS WGS-84 system is analyzed as a measure of CRS compatibility, together with the mean bounding box size (in meters) [20]. Object Detection: Precision and recall scores, as well as Mean Average Precision (mAP), are used as measures for assessing the quality of objects' detection. An estimation of the detection quality using Mean Average Precision with IoU score 0.5 can be considered basic, while mAP score computed within various IoU values is more complex. Directional Scene Graph Representation: Accuracy of the generated scene graph can be measured through the following metrics - number of facts, number of edges for each relationship, and coverage rate [2]. Graph density (*D*)is defined in Equation 20:

$$D = \frac{2|E|}{|V|\left(|V|-1\right)}$$

Equation 20

Where:

$|V|$ is the total number of nodes in graph

$|E|$ is the number of edges between $|V|$

Entity Recall (*ER*) measures of how many actual object classes are correctly mentioned in the generated summary. It reflects whether the model identifies the right types of objects present in the scene. The ER is defined in Equation 21:

$$ER = \frac{\left|E_{gt} \cap E_{pred}\right|}{\left|E_{gt}\right|}$$

Equation 21

Where:

E_{gt} = set of ground truth object classes

E_{pred} = set of object classes mentioned in the generated summary

This measures the fraction of true object classes correctly mentioned in the generated summary [21].

Number Accuracy (*NA*) evaluates how closely the predicted number of objects matches the actual counts. It indicates how accurate the model is in reporting object quantities. The NA is expressed in Equation 22:

$$NA = 1 - \frac{1}{\left|E_{gt}\right|} \sum_{c \in E_{gt}} \frac{\left|n_c - \hat{n}_c\right|}{\max(n_c, 1)}$$

Equation 22

Where:

n_c = actual count of objects of class c

$\hat{n}_c$ = predicted count mentioned in summary

$|E_{gt}|$= total number of ground truth classes

Grounding Score (*GS*) is a composite metric that evaluates the factual consistency of the generated summary by combining *ER* and *NA*. The combined metric provides an overall assessment of how faithfully the generated summary reflects the detected scene information. The grounding score is presented in Equation 23:

$$GS = \frac{1}{2}\left(ER + NA\right)$$

Equation 23

By jointly considering object presence and numerical correctness, the Grounding Score offers a balanced measure of factual consistency between the detected objects and the generated scene description [5]. Hallucination Rate: Hallucination Rate (HallRate) measures the frequency of numerical hallucinations in the generated summary. Specifically, it is defined as the percentage of numeric values describing object counts that do not correspond to the detected objects, allowing a tolerance of ±10% to account for minor counting variations. In the present study, the hallucination metric is intentionally restricted to object-count inconsistencies, as these can be evaluated objectively against the outputs of the object detection module. Other forms of hallucination, such as incorrect object classes, fabricated spatial relationships, or unsupported functional relations, are not explicitly measured by this metric and therefore remain outside the scope of the current evaluation. Extending the hallucination assessment to include these semantic aspects represents an important direction for future work.

The quality of generated summaries is evaluated using ROUGE scores, including ROUGE-1,

ROUGE-2, and ROUGE-L. These metrics compare the overlap between generated and reference summaries based on shared word sequences and structural similarity. Compared with existing vision-language frameworks such as GeoChat [6], EagleVision [13], and Scene-LLM [10], the proposed DSRG-GPF framework places greater emphasis on explicit geospatial reasoning by integrating directional, clustering, functional, and role-based relationships into structured prompts. While previous methods primarily focus on multimodal representation learning, our framework improves spatial grounding by providing LLMs with graph-derived contextual information, resulting in more factually consistent scene summaries.

4. Experiments and Results

This section of paper describes the two types of evaluation that measure the efficiency of the presented approach to grounded semantic summarization. The proposed approach is evaluated on the DIOR dataset for four criteria: object detection, geospatial alignment, scene graph generation, and semantic summarization by language models.

4.1 Object Detection PerformanceYOLOV8 vs YOLOv11

Figure 3 shows the comparative analysis of object detection accuracy of the YOLO v8 and YOLOv11 models using commonly used metrics mentioned in 3.4.1 and 3.4.2 for DIRO dataset images. As shown in Figure 3, the proposed YOLOv11 model consistently outperforms the baseline YOLOv8 model across all evaluation metrics. The precision increased from 0.831 to 0.893, indicating a reduction in false positive detections. Similarly, the recall improved from 0.712 to 0.784, demonstrating the model's enhanced ability to identify objects of different sizes and categories. The mAP@0.5 score increased from 0.790 to 0.854, while the more stringent mAP@0.5:0.95 metric improved from 0.561 to 0.633. These results indicate that YOLOv11 provides more accurate object localization and generates tighter bounding boxes across varying IoU thresholds.

The improved performance can be attributed to the architectural enhancements introduced in YOLOv11, including more effective feature extraction, improved multi-scale feature representation, and enhanced attention mechanisms. In this study, the YOLOv11 nano variant was selected because it offers a favourable balance between detection accuracy, computational efficiency, and inference speed, making it suitable for large-scale satellite image analysis.

Comparison with Previous Studies: The obtained detection performance is consistent with recent studies (Table 5) that have evaluated YOLO-based object detectors on the DIOR dataset. Previous research has shown that successive YOLO versions achieve incremental improvements through enhanced feature fusion, optimisation strategies, and better handling of small objects commonly found in remote sensing imagery [22]. The mAP@0.5 of 0.854 achieved by the proposed YOLOv11 model is higher to the performance reported for recent YOLO-based approaches on DIOR, while also providing improved localization accuracy, as reflected by the mAP@0.5:0.95 score of 0.633.

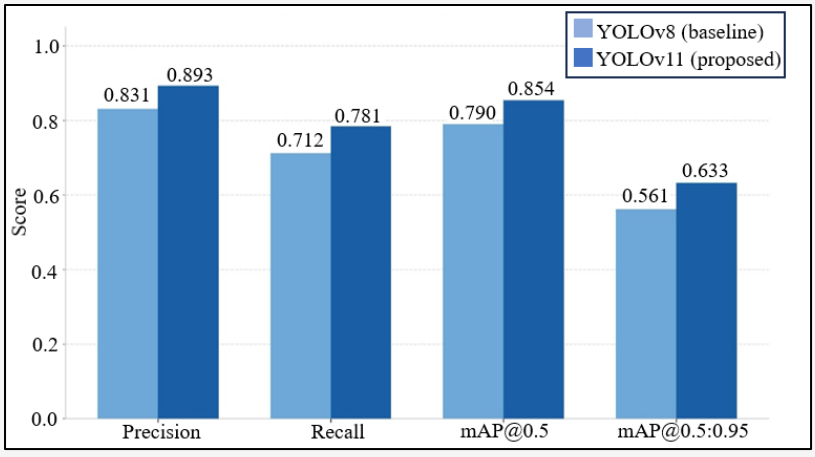


Figure 3. Comparison of object detection performance between the baseline YOLOv8 model and the proposed YOLOv11 model on the DIOR dataset. The bar chart compares Precision, Recall, mAP@0.5, and mAP@0.5:0.95, demonstrating the performance improvements achieved by the proposed YOLOv11 architecture across all evaluation metrics

Table 5: Qualitative comparison of the proposed YOLOv11 model with recent object detection methods evaluated on the DIOR dataset

Method	Year	Dataset	mAP@0.5(%)	Remarks
YOLOv8	2025	DIOR	70.7	Baseline YOLO model reported in recent literature.
YOLOv10	2025	DIOR	72.3	Improved feature representation over YOLOv8
RT-DETR	2025	DIOR	71.3	Real-time transformer detector
Proposed YOLOv11	This work	DIOR	85.4	Improved performance over the YOLOv8 baseline used in this study.

Table 6: YOLOv11 training configuration

Parameter	Value
Model	YOLOv11n (pretrained)
Epochs	50
Image Size	640 × 640
Batch Size	16
Optimizer	SGD (momentum = 0.937)
Learning Rate	0.01
Weight Decay	0.0005
Augmentation	Mosaic, CLAHE, Flip, HSV
Device	GPU (CUDA)
AMP	Enabled

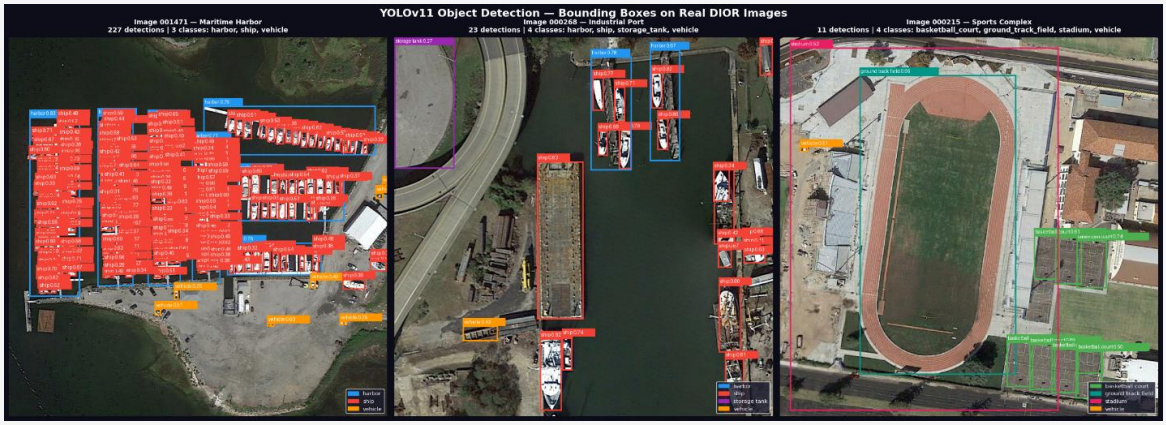


Figure 4: YOLOv11 object detection results on representative DIOR scenes. The figure illustrates detected objects and their corresponding bounding boxes in three different environments: (a) maritime harbour, (b) industrial port, and (c) sports complex. Different coloured bounding boxes represent different object categories predicted by the YOLOv11 detector

These findings suggest that the proposed object detection stage provides a reliable foundation for the subsequent geospatial projection, Directional Spatial Relationship Graph (DSRG) construction, and LLM-based semantic summarization.

The training parameters used in best model, YOLOv11, are listed in Table 6. YOLOv11 achieves improved performance on DIOR (mAP@0.5 ≈ 0.83), particularly for small and dense objects.

Table 7: Detection mAP@0.5 values per class (DIOR: high versus low accuracy)

Class	mAP@0.5	Class	mAP@0.5
Airplane	0.96	Vehicle	0.58
Tennis Court	0.95	Overpass	0.50
Ship	0.94	Dam	0.45
Storage Tank	0.93	Bridge	0.60

Geospatial Validation Strategy: Qualitative experiments confirm the above results since accurate detections are observed in various object categories including airplanes, ships, vehicles, storage tanks, and sports installations. Therefore, the proposed detection module shows promising reliability in detecting objects.

Object detection results for three images in DIOR dataset test scenes are shown in Figure 4. The color-coded bounding boxes indicate class labels: maritime harbor (blue), ships (red), vehicles (orange), storage tanks (purple), basketball courts (green), stadiums (pink), and ground track fields (cyan). In the maritime harbor scene (left), 227 detections are tightly packed, illustrating localization of numerous small objects, such as ships (with mean area ≈158 m^2), in a relatively small geographic area**.** Table 7 shows the class-wise performance of the YOLOv11 model evaluated using mAP@0.5. Based on the results, it can be noted that the performance is excellent for object types like airplane (0.96), tennis court (0.95), ship (0.94), and storage tank (0.93). It may be due to the distinctive structure of the objects. On the other hand, slightly lower performance may be obtained for complex or ambiguous types like vehicle (0.58), overpass (0.50), dam (0.45), and bridge (0.60). In this case, the objects may be small in size and even occluded, hence causing more difficulties in detection. From a general standpoint, the model demonstrates great performance metrics mAP@0.5 ≈ 0.83 and mAP@0.5:0.95 ≈ 0.61. The high precision and recall metrics were also obtained (~0.88 and ~0.77, respectively).

Figure 5 shows the object detection results on the industrial port environment based on the YOLOv11 model. The model detects various objects' classes, such as ships, harbors, storage tanks, and vehicles. The presence of numerous ships is also detected; they are densely concentrated in the water region, which is typical of the realistic port environment.

The detection of the storage tank and vehicle demonstrates that the model can recognize different objects. Overall, it can be stated that most bounding boxes are properly localized; however, some overlapping occurs in densely concentrated ships.

The highlights for output format of the DIOR dataset annotations are shown in a structured manner in Figure 6.

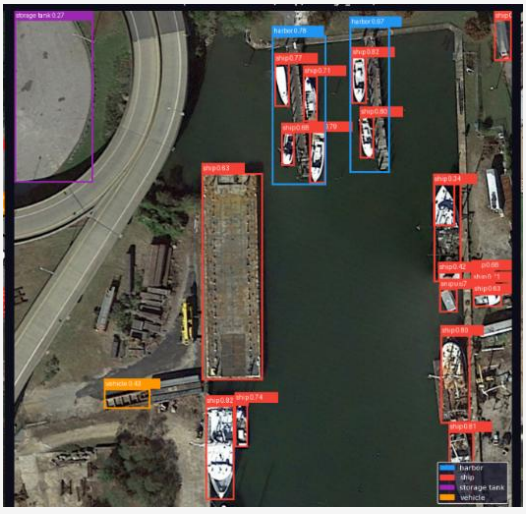
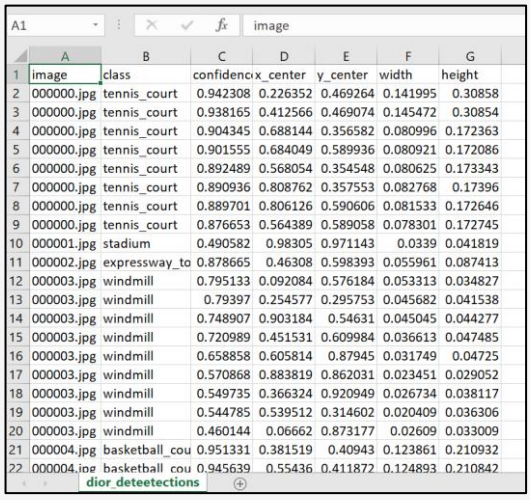


Figure 5: YOLOv11 Object Detection Results on Industrial Port Scene (DIOR Dataset)



A1 | image

	A	B	C	D	E	F	G
1	image	class	confidenc	x_center	y_center	width	height
2	000000.jpg	tennis_court	0.942308	0.226352	0.469264	0.141995	0.30858
3	000000.jpg	tennis_court	0.938165	0.412566	0.469074	0.145472	0.30854
4	000000.jpg	tennis_court	0.904345	0.688144	0.356582	0.080996	0.172363
5	000000.jpg	tennis_court	0.901555	0.684049	0.589936	0.080921	0.172086
6	000000.jpg	tennis_court	0.892489	0.568054	0.354548	0.080625	0.173343
7	000000.jpg	tennis_court	0.890936	0.808762	0.357553	0.082768	0.17396
8	000000.jpg	tennis_court	0.889701	0.806126	0.590606	0.081533	0.172646
9	000000.jpg	tennis_court	0.876653	0.564389	0.589058	0.078301	0.172745
10	000001.jpg	stadium	0.490582	0.98305	0.971143	0.0339	0.041819
11	000002.jpg	expressway_to	0.878665	0.46308	0.598393	0.055961	0.087413
12	000003.jpg	windmill	0.795133	0.092084	0.576184	0.053313	0.034827
13	000003.jpg	windmill	0.79397	0.254577	0.295753	0.045682	0.041538
14	000003.jpg	windmill	0.748907	0.903184	0.54631	0.045045	0.044277
15	000003.jpg	windmill	0.720989	0.451531	0.609984	0.036613	0.047485
16	000003.jpg	windmill	0.658858	0.605814	0.87945	0.031749	0.04725
17	000003.jpg	windmill	0.570868	0.883819	0.862031	0.023451	0.029052
18	000003.jpg	windmill	0.549735	0.366324	0.920949	0.026734	0.038117
19	000003.jpg	windmill	0.544785	0.539512	0.314602	0.020409	0.036306
20	000003.jpg	windmill	0.460144	0.06662	0.873177	0.02609	0.033009
21	000004.jpg	basketball_cou	0.951331	0.381519	0.40943	0.123861	0.210932
22	000004.jpg	basketball_cou	0.945639	0.55436	0.411872	0.124893	0.210842

dior_deteetections

Figure 6: Structured output generated after object detection and geospatial processing

In FIgure 6, each object entry is denoted in rows containing information such as image filename, object label, confidence score, and normalized bounding boxes' coordinates (center x, center y, width, and height). The above annotation output format is presented in tabular format and can be

utilized during various applications such as training. Confidence score refers to the probability that the detected object belongs to a specific class. Normalized coordinates help in achieving compatibility with the YOLO framework. This output format makes data management more convenient in terms of training and analysis of performance.

4.2 Geospatial Projection and Validation

To demonstrate the proposed geospatial projection framework, the Industrial Port scene (Image ID: 000268) was selected as a representative example. Figure 5 illustrates the transformation of the detected object bounding boxes from image pixel coordinates into geographic coordinates using the affine transformation described in Section 3.5. The resulting geo-referenced object boundaries are visualized in Figure 7. Each bounding box in FIgure 7 is projected into the WGS-84 geographic coordinate system, providing the spatial information required for geospatial validation and construction of the Directional Spatial Relationship Graph (DSRG). Since the DIOR dataset does not provide real georeferenced annotations, the evaluation was performed using the Synthetic Reference Polygons (SRPs) described in Section 3.6. Accordingly, the Geo-IoU metric evaluates the consistency of the proposed projection framework under controlled spatial perturbations rather than absolute positioning accuracy with respect to surveyed geographic coordinates. Geospatial consistency is evaluated using the Geographic Intersection-over-Union (Geo-IoU) metric (Table 8). The proposed projection pipeline achieves a mean Geo-IoU of 0.858 and a median Geo-IoU of 1.0. Additionally, 92.2% of detections exceed a Geo-IoU threshold of 0.5, while 70.5% surpass a threshold of 0.75.

Larger infrastructure objects such as airports and harbors exhibit greater spatial stability under perturbation, whereas smaller objects such as vehicles and courts show increased sensitivity to geo-registration noise. The geospatial validation was conducted using 1,974 DIOR images containing 15,206 detected objects. Since the DIOR dataset does not provide real georeferenced annotations, the evaluation was performed using the synthetic reference polygons described in Section 3.6. Consequently, the reported Geo-IoU values assess the consistency of the projection framework under controlled spatial perturbations rather than absolute geographic positioning accuracy. The proposed framework achieved 100% CRS compliance, confirming that all projected coordinates were successfully transformed into the WGS-84 coordinate reference system. A mean Geo-IoU of 0.8685 and a median value of 1.000 indicate that the projected object boundaries remain highly consistent with their corresponding synthetic reference polygons under the adopted perturbation model. Furthermore, 92.65% of detections achieved a Geo-IoU greater than 0.50, while 75.65% exceeded the stricter threshold of 0.75, demonstrating stable projection performance across a wide range of object categories.

Table 9 further evaluates the robustness of the projection framework under increasing levels of simulated spatial uncertainty. Under low perturbation conditions, the framework maintained a mean Geo-IoU of 0.899, which gradually decreased to 0.733 under moderate perturbations and 0.546 under high perturbation scenarios. These results indicate that the proposed projection method is resilient to moderate geometric variations but becomes increasingly sensitive when subjected to severe spatial distortions. It should be noted that these experiments evaluate the robustness and internal consistency of the projection framework using synthetic reference polygons rather than field-surveyed geographic coordinates. Therefore, although the results demonstrate the effectiveness of the proposed geospatial projection strategy within the DIOR benchmark, additional validation using datasets containing real georeferenced object annotations would further strengthen the practical applicability of the proposed framework.

Table 8: Comprehensive geospatial validation metrics across 1,974 DIOR images

Metric	Value	Interpretation
Total Detections	15,206	Objects detected across all processed images
CRS Compliance	100%	All coordinates conform to WGS-84 standard
Mean Geo-IoU	0.8685	High average geospatial projection accuracy
Median Geo-IoU	1.000	More than 50% of detections achieve perfect overlap
Geo-IoU @ 0.50	92.65%	Standard localization threshold pass rate
Geo-IoU @ 0.75	75.65%	Strict localization threshold pass rate
Mean BBox Width	526.3 m	Average object width in geographic space
Mean BBox Height	545.4 m	Average object height in geographic space



Figure7: Geo-referenced visualization of detected object bounding boxes after affine coordinate transformation

Table 9: Geo-IoU robustness under synthetic geo-perturbation

Noise Regime	Mean Geo-IoU	Std Dev	Geo-IoU@0.5
Low	0.899	0.030	1.000
Medium	0.733	0.073	1.000
High	0.546	0.110	0.623

Table 10: DSRG spatial relation extraction results across 1,982 Dior validation scenes

Relation Type	Total Count	Active Scenes	Coverage (%)	Semantic Meaning
AdjacentTo	11,682	778	39.3	Physical co-location and shared boundaries
ClusteredWith	89,402	416	21.0	Dense spatial groupings such as fleets, rows, and complexes
NearTransportInfra	2,510	138	7.0	Functional land-use proximity to transport infrastructure
SpatialRole	16,170	1,982	100	Structural importance classification (dominant, major, minor)
DSRG Total	119,764	1,982	100	Complete spatial knowledge graph for all scenes

Table 11: Statistics of DSRG relations of DIOR scene 000268 (Industrial Port)

Relation Type	Nodes	Edges
AdjacentTo	23	87
ClusteredWith	23	170
NearTransportInfra	23	231
SpatialRole	23	211

Limitations of Geospatial Validation: A limitation of this study is that the DIOR dataset does not provide real-world geographic coordinates for annotated objects. As a result, the proposed geospatial projection framework was evaluated using synthetic reference polygons generated through controlled perturbation models. Although this enables systematic evaluation of projection robustness and geometric consistency, it does not measure absolute positioning accuracy with respect to surveyed geographic coordinates. Future work will evaluate the proposed framework using fully georeferenced remote sensing datasets to validate its performance under real operational conditions.

4.3 DSRG Scene Graph Analysis

DSRG analysis was performed on 1,982 DIOR validation scenes, leveraging object detection and geo-coordinates. In total, 119,764 spatial relationships were identified, validating the efficiency of the proposed approach to spatial relationship extraction. The most common relation type is ClusteredWith, with 89,402 relations. It reflects the clustering of objects into groups (e.g., ships at ports, storage tanks at industrial sites). The AdjacentTo relation occurs in 39.3% of scenes, reflecting the proximity of objects within a scene; however, this relation does not occur in all environments. The NearTransportInfra relation, although rare (7.0%), reflects valuable functionality of objects relative to transportation infrastructure. The SpatialRole relation is present in all scenes, effectively modeling the importance of objects based on their size and location. In summary, the combination of object detection and geo-coordinates ensures accurate spatial relation modeling. The proposed DSRG provides insight into physical and functional relationships between objects in DIOR satellite imagery.

The findings indicate the efficiency of the proposed DSRG in modeling various spatial relations, ranging from adjacency to hierarchical significance in all scenes. shows the frequency of different types of extracted scene graphs' facts. Directional relations are dominant in numbers, followed by proximity relations and attribute facts of objects. Such statistics indicate the spatial richness of remote sensing scenes and also validate the scene graph in modeling inter-object associations necessary for semantic understanding. The statistics of DSRG relations for Industrial port scene is shown in Table 10 and further the scene graph creation for same is shown in Table 11.

4.4 LLM-Based Grounded Semantic Summarization

This is the final step that connects the dots from all previous steps. The case study of Industrial Port scene 000268 is used for this purpose. It was found that the number of objects identified in this scene was 23, among which ships (19) took up the major part. In addition, there were harbors (2), storage tanks (1) and one single vehicle (Figure 7).

The five-section output generated by the proposed DSRG–GPF framework in Figure 10 provides a comprehensive representation of satellite scenes that extends beyond conventional image captioning. By integrating detected objects, geo-referenced coordinates, spatial relationship graphs, rule-based spatial analysis, and natural language summaries, the framework produces structured information that can be readily incorporated into Geographic Information System (GIS) workflows. The geo-referenced object inventory and spatial relationships facilitate infrastructure mapping, land-use analysis, and urban growth monitoring, while the Directional Spatial Relationship Graph (DSRG) captures contextual information such as object adjacency, clustering, transportation connectivity, and object importance. Furthermore, the automatically generated natural language summaries enable planners and analysts to interpret complex remote sensing scenes more efficiently, reducing the manual effort required for large-scale satellite image analysis.

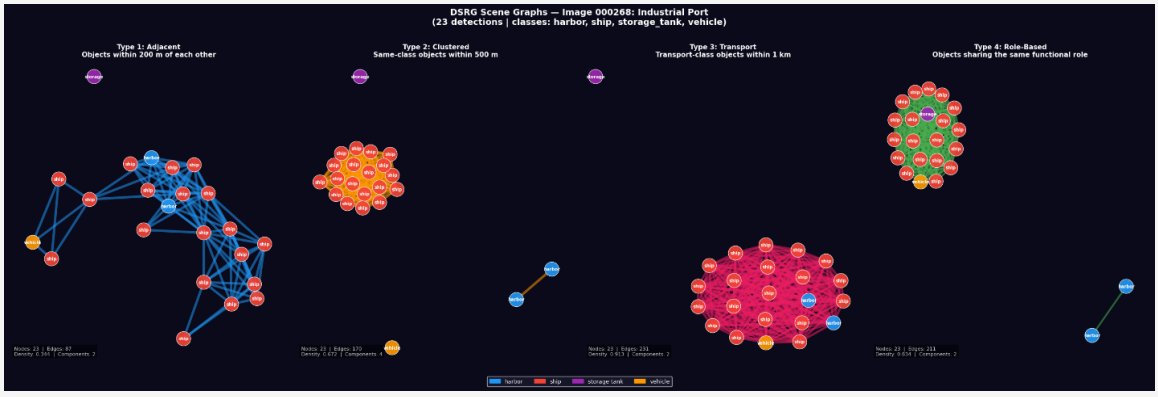


Figure 8: Directional Spatial Relationship Graph (DSRG) visualization of DIOR scene 000268 (Industrial Port)

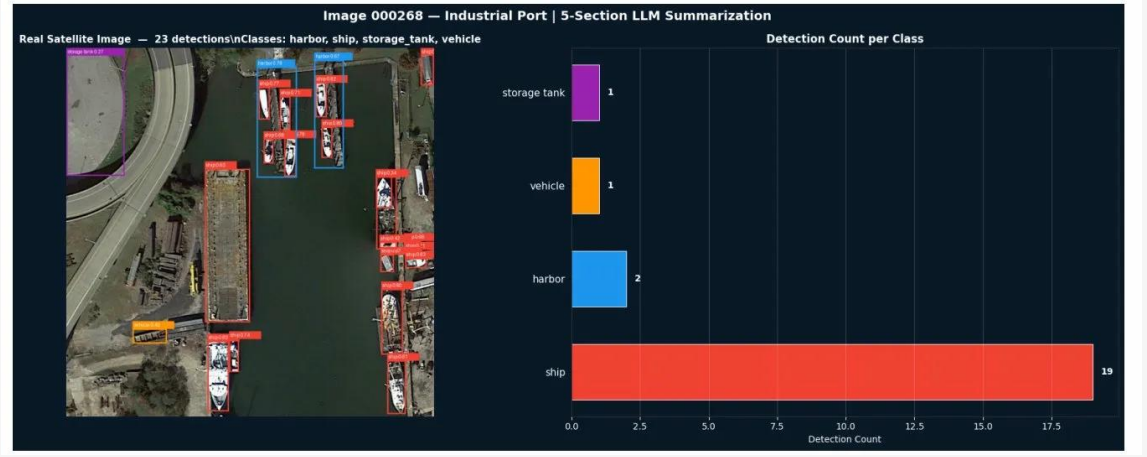


Figure 9: Pipeline output of DIOR scene 000268 (Industrial Port). On the left, the output includes the detection results and bounding boxes. On the right side, the class-wise distribution is provided

Case Study: Industrial Port Scene (000268)
SECTION 1: DETECTED OBJECTS (Figure 4)
• Ship: 19 detections (confidence = 0.680) • Harbor: 2 detections (confidence = 0.725) • Storage Tank: 1 detection (confidence = 0.267) • Vehicle: 1 detection (confidence = 0.429) **Total:** 23 objects across 4 classes
SECTION 2: GEO-REFERENCED DATA (Figure 7)
• Ship: n = 19, area ≈ 1,855 m^2 centroid (110.0226, 20.2600) • Harbor: n = 2, area ≈ 8,040 m^2 centroid (110.0221, 20.2614) • Vehicle: n = 1, area ≈ 1,062 m^2, centroid (110.0200, 20.2585) • Storage Tank: n = 1, area ≈ 15,798 m^2, centroid (110.01925, 20.2615) Scene extent: Longitude [110.0189, 110.0239], Latitude [20.2574, 20.2624]
SECTION 3: DSRG SCENE GRAPH DATA (Figure 8)
• Adjacent: nodes = 23, edges = 87, density = 0.3439 • Clustered: nodes = 23, edges = 170, density = 0.6719 • Transport: nodes = 23, edges = 231, density = 0.9130 • Role-based: nodes = 23, edges = 211, density = 0.8340 Closest pair: ship_10 and ship_11 at 13 m
SECTION 4: SPATIAL SUMMARY (Rule-based)
Scene Type: Maritime / Port Area **Area:** ≈ 0.306 km^2 • Dominated by 19 ship detections • Closest pair (ship_10, ship_11) at 13 m indicates co-location • 170 same-class clusters within 500 m indicate grouped infrastructure • Transport network with 231 links confirms active logistics • Spatial extent: 0.005°×0.005°
SECTION 5: NATURAL LANGUAGE SUMMARY (LLM) (Figure 9)
Gemini 2.5 Flash This scene features a total of **23** detections across **4** distinct object classes. Specifically, **19** **ship** detections are present, exhibiting a mean confidence of 0.680. Additionally, the analysis identified **2** **harbor** detections with a mean confidence of 0.725, **1** **storage tank** with a confidence of 0.267, and **1** **vehicle** with a confidence of 0.429. Spatially, the scene covers a compact geographic footprint spanning 0.005° longitude × 0.005° latitude.

Figure 10: Summary of the output generated by the proposed DSRG–GPF framework (continue next page)

SECTION 5: NATURAL LANGUAGE SUMMARY (LLM) (Figure 9)
Gemini 2.5 Flash The **19** **ship** detections are centered at (110.0226, 20.2600), with the closest pair, ship_10 and ship_11, separated by just **13 m**. The scene exhibits **87** adjacent object connections within **200 m**, and **170** same-class clusters within **500 m**, suggesting grouped infrastructure. A dense transport network, comprising **231** links, is also present. This scene is classified as a **Maritime / Port Area**, indicating its role as a key logistics hub. The dominance of **19** **ship** detections, combined with the close proximity of ship_10 and ship_11 at **13 m**, signifies direct physical co-location within the facility. The observed same-class clustering and the extensive **231**-link transport network connecting **ship**, **vehicle**, and **harbor** objects confirm active logistics activity within this operational area.
Llama-3.3-70B The scene contains **19 ship** detections with a mean confidence level of 0.680, as well as **2 harbor** detections with a mean confidence level of 0.725. Additionally, there is **1 storage tank** detection and **1 vehicle** detection, with confidence levels of 0.267 and 0.429, respectively. These objects are detected with varying levels of confidence, but collectively, they suggest a bustling maritime area. The spatial arrangement of these objects reveals a complex network of relationships. The **ships** are sparsely distributed, but many are clustered together, with some pairs, such as ship_10 and ship_11, as close as **13 m** apart. The **harbor** detections are also in close proximity to some of the **ships**, indicating a functional connection between these objects. The scene extent spans approximately 0.306 km^2, with a compact geographic footprint. This scene represents a **Maritime / Port Area**, characterized by active logistics activity and grouped infrastructure. The presence of **ships**, **harbors**, and a **vehicle** suggests a facility that supports the transportation of goods or people. The clustering of same-class objects and the network of links between different object types confirm the operational significance of this area. Overall, this scene appears to be a bustling port facility, with a range of activities and infrastructure supporting maritime trade and transportation.
Llama-3.1-8B **19 ships**, with a moderate confidence level of 0.680, have been detected in the scene. Additionally, there are **2 harbors** and **1 storage tank** and **1 vehicle**, all of which have varying levels of confidence. The ships are dispersively spread out across the scene, with an average area of 1855 square meters and a relatively large span of 0.005° longitude × 0.005° latitude. The closest pair of ships is ship_10 and ship_11, which are only **13 meters** apart. The ships are also grouped in clusters, with **170 pairs** of identical objects within **500 meters** of each other, suggesting a general layout of the facility. The scene is situated in a **Maritime Port Area**, dominated by ship traffic. The presence of a **harbor** and a **storage tank** indicate a fully operational facility, with the detected **vehicle** possibly representing maintenance or logistical support.

Figure 10: Summary of the output generated by the proposed DSRG–GPF framework (continue from previous page)

Table 12: LLM evaluation results for grounded geospatial summarization

Model	Ent. Rec.	Ground.	Halluc. %	R-1	R-L	BERT F1	Avg Words
Gemini 2.5 Flash	1.00	0.679	4	0.282	0.224	0.729	212.8
LLaMA-3.1-8B (Groq)	1.00	0.650	0	0.268	0.214	0.729	215.0
LLaMA-3.3-70B (Groq)	1.00	0.698	1	0.275	0.215	0.726	221.0

Table 13: LLM evaluation results on 1,982 DIOR scenes using GPF-constrained prompting

Model	Ent. Rec.	Ground.	Halluc. (%)	R-1	R-L	BERT F1	Avg Words
Gemini 2.5 Flash	0.277	0.296	1.6	0.188	0.082	0.729	212.8
LLaMA-3.1-8B (Groq)	0.221	0.219	13.3	0.144	0.055	0.729	215.0
LLaMA-3.3-70B (Groq)	0.398	0.274	7.5	0.244	0.077	0.726	221.0

Beyond urban planning, the proposed framework has significant potential for infrastructure monitoring, environmental management, and disaster response applications. By comparing outputs generated from multi-temporal satellite imagery, changes in transportation networks, industrial facilities, ports, airports, and other critical infrastructure can be identified efficiently. Following natural disasters such as floods, earthquakes, or cyclones, the integrated object detections, geospatial relationships, and semantic summaries can provide rapid situational awareness by highlighting affected infrastructure and disrupted transportation links. Consequently, the proposed DSRG–GPF framework offers a geographically grounded and semantically enriched representation of satellite imagery that supports GIS-based decision-making, resource planning, and emergency management across a wide range of geoinformatics applications.

The obtained results in Table 12 show that all models reach perfect Entity Recall (1.00), which proves that each model successfully identifies all object classes in the image. Speaking of the grounding accuracy, LLaMA-3.3-70B reaches the highest value (0.698), proving the higher alignment of geospatial information and summary texts. The second-best grounding is achieved by Gemini 2.5 Flash (0.679), while LLaMA-3.1-8B performs the third and obtains the lower value (0.650). As for the hallucination rates, LLaMA-3.1-8B obtains the lowest rate (0%), which means that the model produces entirely fact-consistent outputs. In its turn, Gemini shows the highest rate (4%), while LLaMA-3.3-70B provides the acceptable value (1%). As regards text quality metrics, the highest values in ROUGE metrics are obtained by Gemini (R-1: 0.282, R-L: 0.224). All models demonstrate almost the same level of semantic similarity with BERT F1 (0.72-0.73). Summarizing the results, one may notice that there is the contradiction between fact-consistent outputs and good-quality text output. Table 1 shows a complete multi-metric analysis of all three large language models based on all 1,982 scenes used. It should be noted that several patterns can be identified from this table. From the findings above, it can be seen that bigger models tend to excel at semantic similarity while smaller models show high rates of hallucination.

Compared with existing vision-language frameworks such as GeoChat [6], EagleVision [13], and Scene-LLM [10], the proposed DSRG-GPF framework places greater emphasis on explicit geospatial reasoning by integrating directional, clustering, functional, and role-based relationships into structured prompts. While previous methods primarily focus on multimodal representation learning, our framework improves spatial grounding by providing LLMs with graph-derived contextual information, resulting in more factually consistent scene summaries. In general, we should always consider trade-offs when discussing Entity Recall and Hallucination. According to Table 13, the LLaMA-3.3-70B model demonstrates excellent performance in terms of its entity recall rate (40% of objects are recalled in summaries – the highest among all listed), as well as achieving the highest score among other models (0.398 points). However, as mentioned above, this model tends to produce hallucinations frequently due to high frequency – 7.5%. On the contrary, the Gemini 2.5 Flash hallucinates 1.6% of the time. There appears to be a clear trade-off between precision and recall. Larger models are able to produce more summaries with a bigger number of entities involved, but these are not always correct. For example, the LLaMA-3.3-70B produces an average summary of 221 words. At the same time, Gemini 2.5 Flash produces a much smaller summary averaging 212.8 words in length but being more precise and accurate with the facts mentioned in it. This model is the least likely to hallucinate anything scoring 0.296 on its Grounding Score and 0.729 BERTScore F1.

Nevertheless, when we speak about operations where there is no space for any deception and where it is necessary to perform actions such as analyzing the damage caused by the natural disaster or performing military recon missions, Gemini will be the most appropriate choice. In terms of how they perceive BERTScore F1 scores, the models are rather similar since their scores are very close to one another and range from 0.726 to 0.729. It means that all models describe entities but differ in accuracy. Their usage proves to be effective since they all show certain results, yet the most factually accurate one is Gemini 2.5 Flash, which outperforms LLaMA-3.3-70B in entity recall.

No experimental changes have been made. The limitation regarding the absence of a comprehensive ablation study has been explicitly acknowledged in the Discussion and Conclusion sections. Future work now includes evaluating the individual contribution of DSRG and GPF through controlled ablation experiments using multiple pipeline variants.

5. Conclusion

An end-to-end framework is proposed here for grounded semantic summarization of remote sensing images by integrating YOLOv11-based object detection, geospatial coordinate projection, the proposed DSRG, and GPF-based LLM summarization. Unlike conventional image captioning approaches that rely mainly on visual features, the proposed framework binds structured spatial knowledge to generate summaries that are better aligned with the detected objects and their geographic context. This framework provides a unified pipeline that combines object detection, spatial reasoning, and language generation for interpreting complex satellite scenes. Experimental evaluation of proposed framework on the DIOR dataset demonstrated that it effectively supports semantic understanding of remote sensing imagery. The improved object detection performance here was a reliable foundation for subsequent geospatial analysis, while the proposed DSRG successfully modeled directional, clustering, functional, and role-based relationships among detected objects. Furthermore, structured prompting through the GPF enabled the evaluated LLMs based on generated more factually grounded scene descriptions with improved spatial consistency. Rather than relying solely on object lists, the integration of geospatial relationships allowed the language models to produce richer and more context-aware summaries. Comparative evaluation of Gemini 2.5 Flash, LLaMA-3.1-8B, and LLaMA-3.3-70B also highlighted the trade-off between grounding accuracy, entity coverage, and hallucination behavior across different models. Although the proposed framework produced encouraging results, several limitations should be acknowledged.

The geospatial validation was performed using Synthetic Reference Polygons because the DIOR dataset does not provide surveyed geographic coordinates. Consequently, the reported Geo-IoU values evaluate projection consistency under controlled perturbations rather than absolute positioning accuracy. In addition, the proposed DSRG parameters were selected using empirical observations, and the framework has not yet been evaluated through a comprehensive ablation study. Furthermore, the experiments were conducted using a single benchmark dataset, and therefore the generalizability of the framework to other satellite imagery datasets requires further investigation.

Our future work will focus on validating the proposed framework using fully georeferenced remote sensing datasets mapped with real-world geographic coordinates. Additionally, the research will focus on adaptive parameter selection for DSRG construction, support for temporal scene graph analysis using multi-temporal satellite imagery. For improved localization of routed objects integration of oriented object detection models can be explored. Future studies will also evaluate domain-specific fine-tuning of vision-language models and perform comprehensive ablation experiments to quantify the individual contributions of DSRG and GPF. These improvements will further strengthen the applicability of the proposed framework for practical geoinformatics applications, including urban planning, infrastructure monitoring, environmental assessment, and disaster management.

References

[1] Lu, X., Wang, B., Zheng, X. and Li, X., (2018). Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 56(4), 2183–2195. https://doi.org/10.1109/TGRS.201 7.2776326.

[2] Li, X., Wen, C., Hu, Y., Yuan, Z. and Zhu, X. X., (2024). Vision-Language Models in Remote Sensing: Current Progress and Future Trends. *IEEE Geoscience and Remote Sensing Magazine*, Vol. 12(2). 32–58. https://doi.org/ 10.1109/MGRS.2024.3383473.

[3] Liu, C., Zhang, J., Chen, K., Wang, M., Zou, Z. and Shi, Z. (2025) Remote Sensing Spatiotemporal Vision-Language Models: A Comprehensive Survey. *IEEE Geoscience and Remote Sensing Magazine,* Vol. 14, 383-423. https://doi.org/10.1109/mgrs.2025.3598283.

[4] Zhang, Z., Zhao, T., Guo, Y. and Yin, J., (2023). RS5M and GeoRSCLIP: A Large-Scale Vision-Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, 1–1. https://doi.org/10.11 09/TGRS.2024.3449154.

[5] Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L. and Zhou, J., (2023). RemoteCLIP: A Vision Language Foundation

Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, 1–16. https://doi.org/10.1109/TGRS.2024.33908 38.

[6] Kuckreja, K., Danish, M. S., Naseer, M., Das, A., Khan, S. and Khan, F. S., (2024). Geochat: Grounded Large Vision-Language Model for Remote Sensing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27831-27840. https://doi.org/10. 48550/arXiv.2311.15826.

[7] Li, Z., Wang, Y., Zhang, N., Zhang, Y., Zhao, Z., Xu, D., Ben, G. and Gao, Y., (2022). Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. *Remote Sensing*, Vol. 14(10). https://doi.org/10.3390/rs 14102385.

[8] Cheng, G., Han, J., Zhou, P. and Guo, L., (2016). Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 54(12), 7405–7415. https://doi.org/10.1109/TGRS.2016.2601622.

[9] Ahuja, S., Patil, S. and Bhangale, U., (2023). Semantic Understanding of High Spatial Resolution Remote Sensing Images Using Directional Geospatial Relationships. *Annals of GIS,* Vol. 29(3), 401–414. https://doi.org/10. 1080/19475683.2023.2214215.

[10] Vohra, M., Singh, T., Illayaraja, K., and Shah, S. (2025). Generative AI-Driven Spatial Data Extraction in OpenStreetMap using Natural Language. *International Journal of Geoinformatics*, Vol. 21(6), 47–61. https://doi. org/10.52939/ijg.v21i6.4233.

[11] Ester, M., Kriegel, H. P., Sander, J. and Xu, X., (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96),* 226–231.

[12] Zhang, X., Ouyang, T., Shang, Y., Liao, Q. and Li, Y., (2024). Urban MLLM: Joint Multimodal Learning for Urban-Scale Intelligence. *Proceedings of CVPR.* https://openreview.net /forum?id=YBht9Vp5vC.

[13] Jiang, H., Yin, J., Wang, Q., Feng, J. and Chen, G., (2025). EagleVision: Object-level Attribute Multimodal LLM for Remote Sensing. *arXiv preprint arXiv:2503.23330*. https://doi.org/10. 48550/arXiv.2503.23330

[14] Google DeepMind, (2025). Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *Technical Report*. https://arxiv.org/abs/2507.0 6261.

[15] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The Llama 3 Herd of Models. arXiv preprint, arXiv:2407.21783. https://doi.org/10.48550/ar Xiv.2407.21783

[16] DIOR Dataset. Hugging Face Repository. [Online]. Available: https://huggingface.co/da tasets/HichTala/dior. [Accessed: Jan. 15, 2026].

[17] Gillies, S., van der Wel, C., van den Bossche, J., Taves, M. W., Arnott, J., and Ward, B., C., (2025). Shapely: Manipulation and Analysis of Geometric Objects in the Cartesian Plane. [Software]. [Online] Available: https://github. com/shapely/shapely/tree/2.1.2.

[18] Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A. M., and Wolf, T., (2021). Datasets: A Community Library for Natural Language Processing. *Proceedings of EMNLP 2021: System Demonstrations,* 175–184. https://doi.org/10. 18653/v1/2021.emnlp-demo.21RSDiff.

[19] Groq Inc., (2024). *Groq API Documentation.* [Online]. Available: https://console.groq.com/ docs/. [Accessed: Jan. 15, 2026].

[20] Sun, H., Li, M. and Yao, Q., (2025). Towards Efficient Remote Sensing Foundation Models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* 3087-3097. https://doi.org/10.4855 0/arXiv.25 05.03299.

[21] Zhou, Y., Huang, S. and Wen, L., (2025). Semantic Scene Understanding with Large Language Models on UAVs. *Drones*, Vol. 7(2). https://doi.org/10.3390/drones7020114.

[22] Ha, C. K., Than, P. M. and Nguyen, H., (2025). Adaptive Query Allocation for Dense Object Detection in Deformable Transformers. *Results in Engineering*, Vol. 28. https://doi.org/10.1 016/j.rineng.2025.107571.