AeroGen: Enhancing Remote Sensing Object Detection with Diffusion-Driven Data Generation

Supplementary Material

6. Limitations

Despite AeroGen's notable advancements in generating tiny objects and enhancing datasets, some limitations persist. Firstly, the quality of AeroGen-generated data decreases markedly as the number of instances grows, especially beyond 20 objects, highlighting the need for enhanced performance in complex scenarios. Additionally, AeroGen supports only fixed-category generation in closed environments, restricting flexibility and limiting its applicability to diverse open-domain requirements, particularly for fine-grained control, which necessitates further refinement. While CLIP has been integrated to reduce instance confusion, it shows limited capability in understanding tiny targets in remote sensing images, potentially undermining generation accuracy. Moreover, the final augmented dataset is directly merged with real data for training. However, optimizing this process and developing more effective training strategies remain critical challenges.

7. AeroGen Implementation Details

We utilized three datasets, including DIOR/DIOR-R, with a 1:1:2 division for the training, validation, and test sets. Each category is accompanied by tailored text descriptions and a uniform prompt template applied to the entire image, as shown in Tab. 10. In the HRSC dataset, the most detailed categories were employed at each stage of generation, with all 19 categories encoded within the training set.

Training, testing, and data augmentation were performed on all three datasets, with images consistently resized to 512×512 dimensions using cropping, resizing, and other preprocessing techniques. Experiments employed the original SD 1.5 weights, training the DIOR/DIOR-R dataset for 100 epochs and HRSC for 20 epochs using the AdamW optimizer with a learning rate of 1e-5. All experiments were performed on an NVIDIA GeForce RTX 4090 GPU with a batch size of 4. In experiments with the Layout Mask Attention module, layout information was incorporated during the initial diffusion phase (timesteps 0–500) and transitioned to the text-to-image generation stage (timesteps 500–1000) to ensure the quality of synthetic data, as shown in Fig. 6.

ID	Abbreviation	Category	Description
0	APL	airplane	airplane parked on the ground
1	APO	airport	busy airport
2	BF	baseballfield	green baseball field
3	BC	basketballcourt	outdoor basketball court
4	BR	bridge	long bridge
5	CH	chimney	tall chimney
6	DAM	dam	large dam
7	ESA	Expressway-Service-area	crowded expressway service area
8	ETS	Expressway-toll-station	busy expressway toll station
9	GF	golffield	well-maintained golf field
10	GTF	groundtrackfield	athletic ground track field
11	HA	harbor	bustling harbor
12	OP	overpass	elevated overpass
13	SH	ship	ship on the water
14	STA	stadium	large stadium
15	STO	storagetank	industrial storage tank
16	TC	tenniscourt	clay tennis court
17	TS	trainstation	crowded train station
18	VE	vehicle	vehicle on the road
19	WM	windmill	rotating windmill

Table 10. Category, abbreviation, and description Correspondence of the DIOR/DIOR-R dataset.

No.	Category ID	Category Name
0	100000001	aircraft_carrier
1	100000003	amphibious_assault_ship
2	100000004	submarine
3	100000005	destroyer
4	100000006	frigate
5	100000007	corvette
6	100000008	patrol_vessel
7	100000009	landing_ship
8	100000010	mine_sweeper
9	100000011	fast_attack_craft
10	100000012	supply_ship
11	100000013	medical_ship
12	100000015	research_vessel
13	100000016	fishing_boat
14	100000017	cruise_ship
15	100000018	cargo_ship
16	100000019	container_ship
17	100000020	tugboat
18	100000022	oil_tanker

Table 11. Category ID to name mapping of the HRSC dataset.

8. AeroGen Additional Implementation Details

8.1. Qualitative results

In this section, we complement the results with more visualisations on the three datasets, as shown in Fig. 8. In addition, we provided additional analyses of AeroGen's diversity to further demonstrate its generative capabilities, as illustrated in Fig.

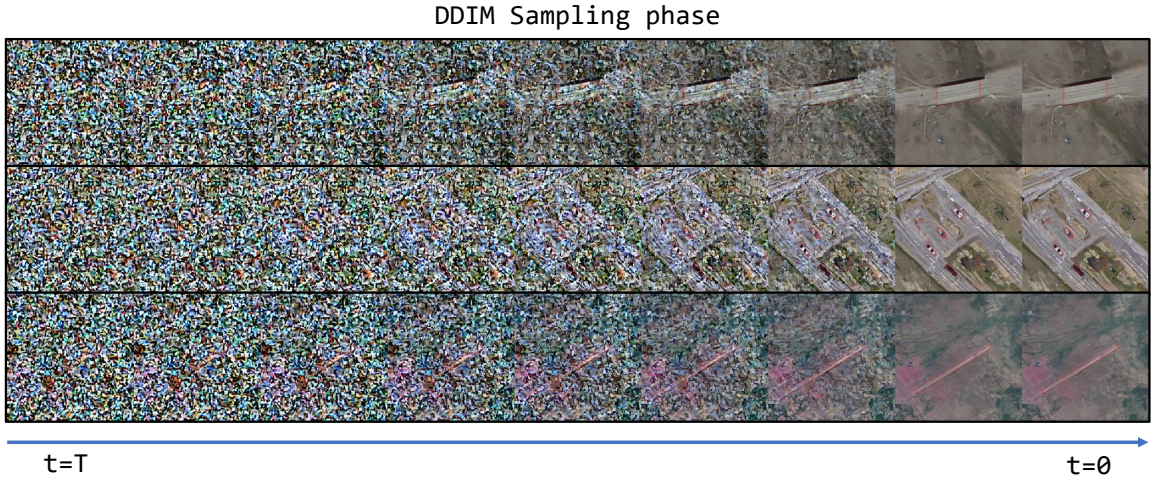

Figure 6. Visualisation of the AeroGen generation process.

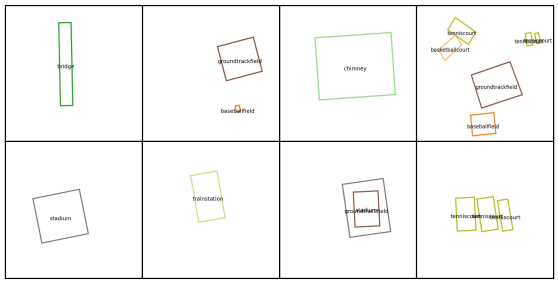

Figure 7. Synthesis of layout information based on DDPM.

9. Pipeline Enhancements Implementation Details

9.1. Label Generator

The visualisation of the conditions obtained based on the layout condition generator is shown in Fig. 7.

We use the following diversity enhancement strategy to get more diverse layout tags.

9.2. Additional experiments

We conducted comparative analysis with main diffusion models, including ReCo, LayoutDiffusion, GLIGEN, MIGC, and ODGEN. The evaluation was performed on COCO benchmark data to ensure consistent comparison conditions. As evidenced in Tab. 12, our AeroGen framework demonstrates competitive performance despite being designed without specific optimization for handling significant object occlusion in natural images.

9.3. Image Filter

This section describes the image filtering process. The CLIP model computes the cosine similarity sim_{clip} between the global text description and the synthetic image. The pretrained ResNet-101 model, trained on bounding box regions from the original dataset, evaluates the classification confidence for each region. The steps are as follows:

The visual experimental results are shown in the Fig. 10

Algorithm 1 Bounding Box Data Augmentation with Transformations

Require: Original bounding box coordinates $P = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\}$, image dimensions $(width, height)$, number of variants $num\_variants$
Ensure: Augmented coordinates $\{P'\}$
  for each variant **do**
    Sample random transformation parameters: scale s, translation (t_x, t_y), rotation θ, horizontal flip f_h, vertical flip f_v, and stretch factor $stretch$
    Set image center $C = (width/2, height/2)$
    Initialize augmented coordinates $P' = P$
    for each point (x_j, y_j) in P **do**
      Apply transformations:
        Scaling: $(x_j, y_j) \leftarrow (x_j - C_x, y_j - C_y) \cdot s + (C_x, C_y)$
        Translation: $(x_j, y_j) \leftarrow (x_j + t_x, y_j + t_y)$
        Rotation: $(x_j, y_j) \leftarrow R(\theta) \cdot (x_j - C_x, y_j - C_y) + (C_x, C_y)$
        Horizontal flip: if f_h, then $x_j \leftarrow width - x_j$
        Vertical flip: if f_v, then $y_j \leftarrow height - y_j$
        Stretch: if horizontal, $x_j \leftarrow (x_j - C_x) \cdot stretch + C_x$; if vertical, $y_j \leftarrow (y_j - C_y) \cdot stretch + C_y$
    end for
    Store augmented coordinates P' for this variant
  end for
  return Augmented coordinates $\{P'\}$

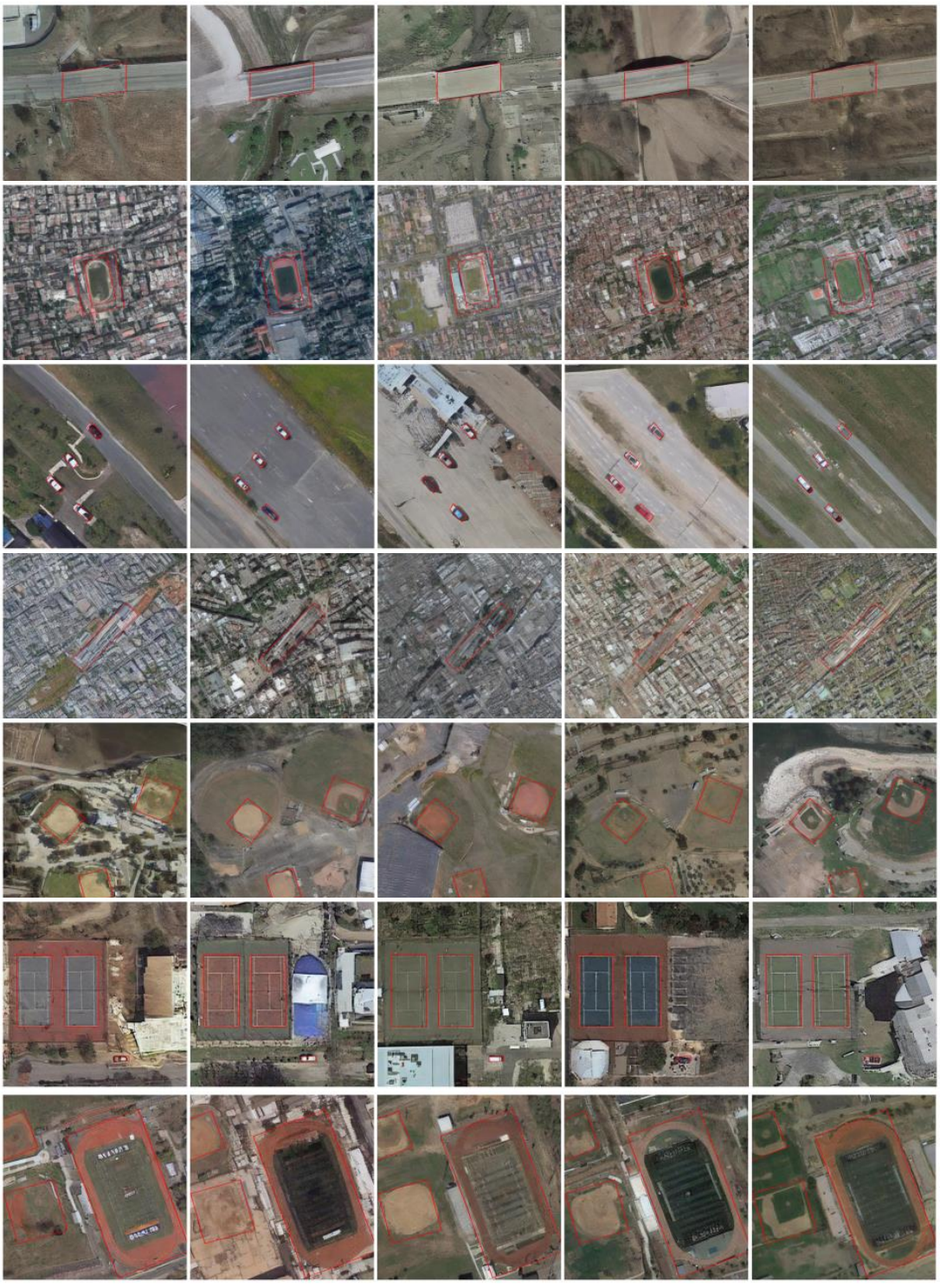

Figure 8. Visualization of AeroGen's diversity results. Each row of images originates from the same layout and exhibits variations in the corresponding background areas.

10. Analysis of Data Distribution

The real DIOR data and corresponding synthetic data are visualized using UMAP downscaling. In this visualization,

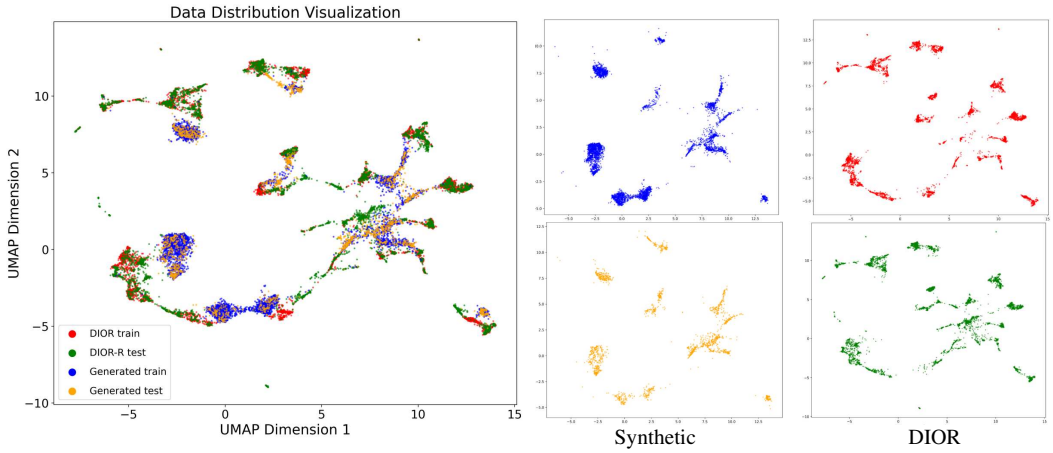

Figure 9. Visualization of data distributions on real-world data and synthetic data.

Metric	ReCo	LayoutDiffusion	GLIGEN	MIGC	ODGEN†	AeroGen
FID ↓	26.19	25.93	24.53	22.10	**16.16**	24.32
mAP ↑	5.32	4.33	5.92	6.58	**9.70**	6.42
mAP50 ↑	12.74	9.71	13.20	14.89	**18.90**	14.68

Table 12. Results on COCO dataset. †: taken from original paper.

Algorithm 2 Synthetic Image Processing with Confidence Filtering and CLIP Similarity

Require: Synthetic image I, layout annotations B, global text description T, classifier model M_c, CLIP model M_{clip}, confidence threshold τ
Ensure: Filtered data
  Compute $\text{sim}_{\text{clip}} = M_{clip}(I, T)$
  Initialize confidence list $C = []$
  Set high-confidence flag $f_{\text{high}} = \text{True}$
  for each bounding box $b \in B$ **do**
    Crop and resize region I_b from I based on b
    Transform I_b to tensor and pass through M_c
    Compute confidence $c = \max M_c(I_b)$
    Append c to C
    if $c < \tau$ **then**
      Set $f_{\text{high}} = \text{False}$ and break
    end if
  end for
  if $C \neq \emptyset$ **then**
    Calculate $\text{avg}(C)$, $\max(C)$, and $\min(C)$
  end if
  if $f_{\text{high}} = \text{True}$ **then**
    Mark I as filtered
  end if
  return Filtered data

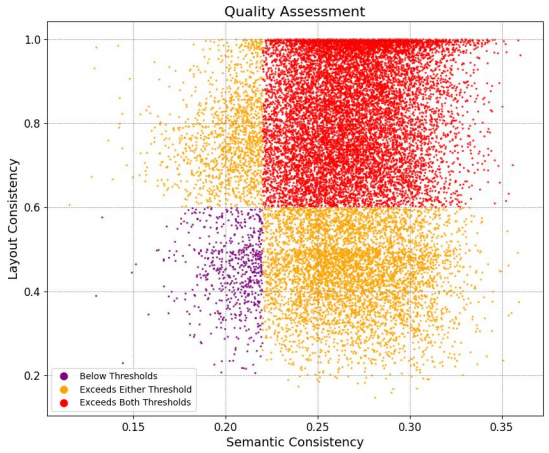

Figure 10. In the image filtering data visualization, the data in the top right corner represents the final filtered synthetic data.

the synthetic training set represents the original labels encountered by the generative model, while the layout labels in the test set are novel to AeroGen. The result shown as in Fig. 9. Given that DIOR is a dataset with clear inter-category relationships, the visualization area effectively reflects these relationships. However, the synthetic data does not align perfectly with the real data and appears more concentrated under the influence of specific categories.

11. Details of target detection model training

In the target detection experiments, we train and evaluate models for horizontal bounding box (HBB) and rotating bounding box (OBB) detection, respectively. For the HBB modality, we selected the one single-stage detection model YOLOv8, set an initial learning rate of 0.01, and conducted 50 training epochs to optimize performance on the validation set. Subsequently, the model's performance was evaluated on the test set. For the OBB modality, we utilized Oriented R-CNN for rotating bounding box target detection, employing the official implementation from mmrotate. The optimizer applied stochastic gradient descent (SGD) with a learning rate of 0.0025, momentum set to 0.9, and a weight decay of 0.0001. A segmented learning rate descent strategy (STEP) was employed, featuring 500 warm-up iterations and a warm-up start ratio of 0, followed by 30 training epochs.