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

Datao Tang[1,2] Xiangyong Cao[1,2*] Xuan Wu[1,2] Jialin Li[1,2] Jing Yao[5]
Xueru Bai[6] Dongsheng Jiang[7] Yin Li[7] Deyu Meng[2,3,4]
[1] School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China
[2] Ministry of Education Key Laboratory of Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an, 710049, China
[3] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China
[4] Pengcheng Laboratory [5] Chinese Academy of Sciences
[6] Xidian University [7] Huawei Technologies Ltd

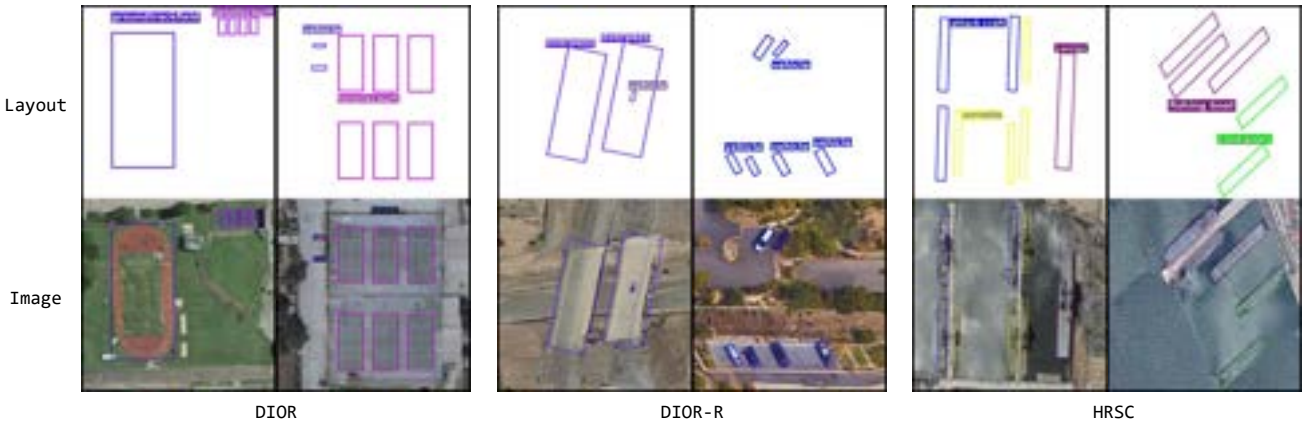

Figure 1. Generated images with our proposed AeroGen. AeroGen enables the input of both horizontal and rotated bounding box layout conditions, facilitating accurate remote sensing image layout generation.

Abstract

Remote sensing image object detection (RSIOD) aims to identify and locate specific objects within satellite or aerial imagery. However, there is a scarcity of labeled data in current RSIOD datasets, which significantly limits the performance of current detection algorithms. Although existing techniques, e.g., data augmentation and semi-supervised learning, can mitigate this scarcity issue to some extent, they are heavily dependent on high-quality labeled data and perform worse in rare object classes. To address this issue, this paper proposes a layout-controllable diffusion generative model (i.e. AeroGen) tailored for RSIOD. To our knowledge, AeroGen is the first model to simultaneously support horizontal and rotated bounding box condition generation, thus enabling the generation of high-quality synthetic images that meet specific layout and object category requirements. Additionally, we propose an end-to-end data augmentation framework that integrates a diversity-conditioned generator and a filtering mechanism to enhance both the diversity and quality of generated data. Experimental results demonstrate that the synthetic data produced by our method are of high quality and diversity. Furthermore, the synthetic RSIOD data can significantly improve the detection performance of existing RSIOD models, i.e., the mAP metrics on DIOR, DIOR-R, and HRSC datasets are improved by 3.7%, 4.3%, and 2.43%, respectively. The code is available at here.

*Corresponding author: caoxiangyong@mail.xjtu.edu.cn

1. Introduction

Object detection is a key technology for understanding and analyzing remote sensing images. It enables efficient pro-

cessing of large-scale satellite data to extract and identify critical information, such as land cover changes [42], urban development status [16], and the impacts of natural disasters [45]. Through object detection, researchers can automatically extract terrestrial targets from complex remote sensing images, including buildings, vehicles, roads, bridges, farmlands, and forests. This information can be further applied in environmental monitoring, urban planning, land use analysis, and disaster emergency management.

With the rapid development of deep learning, supervised learning-based object detection algorithms have made significant progress in remote sensing image analysis [49]. Although these algorithms can accurately locate and classify multiple objects in remote sensing images, they are heavily dependent on a large number of labelled training data. However, obtaining sufficient annotated data for remote sensing images is particularly challenging. Due to the presence of numerous and complex targets in remote sensing images, the manual annotation process is not only time-consuming and labour-intensive but also requires annotators to possess specialized knowledge, thus leading to high costs.

Although traditional data augmentation methods [4] (e.g., rotation and scaling) and enhancement techniques suitable for object detection (e.g., image mirror [15], object-centric cropping [21], and copy-paste [7]) can increase data diversity to some extent, they do not address the fundamental issue of insufficient data. The emergence of generative models [11, 25] provides a new solution to this problem. Currently, in the field of natural images, numerous high-performance generative models [27, 31] have been developed, capable of generating high-quality images from text conditions and also achieving significant progress in layout control. For remote sensing images, the application of generative models is usually combined with specific tasks, such as change detection [46], semantic segmentation [34] and road extraction [33]. These studies have been highly successful in utilizing data obtained from generative models to augment real-world datasets, thereby enhancing the performance of target models in downstream tasks. Therefore, utilizing generative diffusion models to fit the distribution of existing datasets and generate new samples to enhance the diversity and richness of remote sensing datasets is a feasible solution.

In this paper, we focus on the remote sensing image object detection (RSIOD) task and construct a layout generation model (i.e., AeroGen) specifically designed for this task. The proposed AeroGen model allows for the specification of layout prior conditions with horizontal and rotated bounding boxes, enabling the generation of high-quality remote sensing images that meet specified conditions, thus filling a gap in the research field of RSIOD. Based on the AeroGen model, we further propose a conditional generation-based end-to-end data augmentation framework. Unlike pipeline-style data augmentation schemes in the natural image domain [43], our proposed pipeline is implemented by directly synthesizing RSIOD data through conditional generative models, thus eliminating the need for additional instance-pasting procedures. By introducing a diversity-conditioned generator and generation quality evaluation, we further enhance the diversity and quality of the generated images, thereby achieving end-to-end data augmentation for downstream object detection tasks. Moreover, we also design a novel filtering mechanism in this data augmentation pipeline to select high-quality synthetic training images, thus further boosting the performance.

In summary, the contributions of our work are threefold:

- We propose a layout-controllable diffusion model (i.e., AeroGen) specifically designed for remote sensing images. This model can generate high-quality RSIOD training datasets that conform to specified categories and spatial positions. To our knowledge, AeroGen is the first generative model to support layout conditional control for both horizontal and rotated bounding boxes.
- We design a novel end-to-end data augmentation framework that integrates the proposed AeroGen generative model with a layout condition generator as well as an image filter. This framework can produce synthetic RSIOD training datasets with high diversity and quality.
- Experimental results show that the synthetic data can improve the performance of current RSIOD models, with improvements in mAP metrics by 3.7%, 4.3%, and 2.43% on the DIOR, DIOR-R, and HRSC datasets, respectively. Notably, the performance in some rare object classes also significantly improves, e.g., achieving improvements of 17.8%, 14.7%, and 12.6% in the GF, DAM, and APO categories, respectively.

2. Related Work

2.1. Diffusion Models

Diffusion models [11, 25, 31], renowned for their stable training and superior generative quality, are increasingly supplanting Generative Adversarial Networks (GANs) [30, 40] in generative tasks. While text-guided diffusion models excel at producing realistic images, their reliance on brief text prompts often limits precise control over spatial composition and personalized generation. To overcome this, researchers have integrated diverse control mechanisms, expanding applications to layout-guided generation [17, 36, 39, 44, 47], style transfer [6, 37], image restoration [5], and video synthesis [12]. For layout-to-image tasks, methods like ReCo [39] enforce object relational constraints via cross-attention modulation, LayoutDiffusion [44] injects bounding box coordinates through adaptive layer normalization, and GLIGEN [17] employs gated self-attention to dynamically align layouts with text semantics. MIGC

[47] further enhances compositional accuracy by decoupling object- and scene-level denoising. Built on the efficient latent diffusion framework (LDM) [31], these approaches demonstrate how diffusion models balance fine-grained control with computational efficiency, establishing them as versatile tools for complex generation tasks.

2.2. Task-Oriented Data Generation

Generative models are increasingly used to synthesize training data for tasks such as object detection [1, 18, 41, 48], semantic segmentation [34], and instance segmentation [43]. Beyond natural image synthesis, these models adapt to specialized domains like remote sensing and medical imaging via fine-tuning. For example, GeoDiffusion [1] employs text-guided geometric control to generate spatially aligned detection datasets, while DiffusionEngine [41] leverages diffusion models as scalable data engines for object detection. Domain-specific advancements include SSL-driven embedding generation for histopathology [9], joint image-segmentation synthesis in SatSynth [34], and hyperspectral image super-resolution [26]. Remote sensing applications further benefit from models like CRS-Diff [33] and DiffusionSat [14], which utilize multi-modal inputs for tasks like road extraction. However, remote sensing image object detection (RSIOD) lacks dedicated generative solutions. To address this, we propose a layout-controllable model supporting both rotated and horizontal bounding boxes for high-precision synthetic remote sensing imagery.

2.3. Generative Data Augmentation

Existing methods typically combine synthetic and real data directly for downstream training, though some approaches enhance data utility through quality filtering. For instance, Auto Cherry-Picker [2] optimizes synthetic data selection, while X-Paste [43] employs a copy-paste strategy with CLIP-based filtering to boost instance segmentation. DriverGen [8] further analyzes synthetic data distribution and proposes a multi-stage pipeline to improve diversity. Closest to our work, ODGEN [48] generates multi-object scenes via object-wise synthesis to mitigate domain gaps. However, our method focuses on remote sensing object detection, synthesizing images end-to-end using a conditional generator without manual instance pasting. We additionally introduce a diversity-conditioned generator coupled with a dual-criteria (diversity-quality) filtering mechanism, significantly enhancing data utility for downstream tasks.

3. AeroGen

In this section, we introduce AeroGen, a layout-conditional diffusion model for enhancing remote sensing image data. The model consists of two key components: (a) a remote sensing image layout generation model in Sec. 3.1 that allows users to generate high-quality RS images based on predefined layout conditions, such as horizontal and rotated boxes; (b) a generation pipeline in Sec. 3.2 that combines a diffusion-model-based diversity-conditional generator, which produces diverse layouts aligned with physical conditions, with a data-filtering mechanism to balance the diversity and quality of synthetic data, improving the utility of the generated dataset.

3.1. Layout-conditional Diffusion Model

The model weights, obtained through comprehensive fine-tuning on a remote sensing dataset based on LDM [31, 33], are adopted for RS study. In the original text-to-image diffusion model, the conditional position information is combined with the text control condition, and layout-based remote sensing image generation is achieved by establishing a unified position information encoding along with a corresponding dual cross-attention network, as shown in Fig. 2. Building on the latest research advances [17, 47], combined with the regional layout mask-attention strategy, control accuracy is improved, particularly for small target regions.

Layout Embedding. As shown in Fig. 2(a), each object's bounding box or rotational bounding box is uniformly represented as a list of eight coordinates, i.e., $\mathbf{x} = [x_1, y_1, \ldots, x_4, y_4]$, ensuring a consistent representation between horizontal and rotated bounding boxes. Building on this, Fourier [24] encoding is employed to convert these positional coordinates into a frequency domain vector representation, similar to GLIGEN [17]. We use a frozen CLIP text encoder [28] to obtain fixed codes $\mathbf{c}$ for different categories, which serve as layout condition inputs. The Fourier-encoded coordinates are then fused with the category encodings using an additional linear layer to produce the layout control input:

$$\mathbf{h} = \text{Linear}([\gamma(\mathbf{x}); \mathbf{c}]), \tag{1}$$

where $[\gamma(\mathbf{x}); \mathbf{c}]$ denotes the concatenation of Fourier-coded coordinates and category codes, and $\text{Linear}(\cdot)$ represents the linear transformation layer. In this manner, spatial location and category information are effectively combined as layout control tokens.

Layout Mask Attention. In addition to traditional token-based control [39, 44], recent studies [13, 47] indicate that direct semantic embedding based on feature maps is also an effective method for layout guidance. In the denoising process of a diffusion model, the injection of conditional information is gradual, enabling local attribute editing at the noise level. To this end, conditionally encoded noise region steering is employed and combined with a cropping step for improved layout precision. As shown in Fig. 2(b), each bounding box is first transformed into a 0/1 mask M, and category attributes are obtained through CLIP encoding. During each denoising step, the mask attention network provides additional layout guidance. The process is

Figure 2. AeroGen's overall architecture. (a) The layout embedding module combines bounding box coordinates with vectorized semantic information using Fourier and MLP layers. This encodes layout information to facilitate control, with the prompt description processed by a CLIP text encoder for global conditional guidance. (b) The injection of layout information at the noise level is demonstrated, where a local mask governs the injection position of the layout information, allowing for finer layout control. (c) The overall architecture and training process of AeroGen. At each timestep, the image being denoised first passes through a layout information injection module, which enhances layout conditional guidance.

expressed as follows: for each denoised image Q and category encoding K, V, the mask M is used for attention computation according to the following equation:

$$\mathcal{Q} = \sum_{i=1}^{n} M_i \cdot \text{softmax}\left(\frac{QK_i^\top}{\sqrt{d_k}} + \mathcal{M}_i\right) V_i, \tag{2}$$

where M represents the corresponding bounding box mask, and $\mathcal{M}_i$ derived from M_i as the attention mask. This method enables precise manipulation of local noise characteristics during the diffusion generation process, offering finer control over the image layout.

AeroGen Architecture. In AeroGen, the text prompt serves as a global condition and is integrated with layout control tokens via a dual cross-attention mechanism. The output is computed as:

$$\text{Out} = \Psi(Q, K^g, V^g) + \lambda \cdot \Psi(Q, K^l, V^l), \tag{3}$$

where Ψ represents the cross-attention mechanism. K^g and V^g are the keys and values of the global text condition, while K^l and V^l are the layout control tokens. λ balances the influence of global and layout conditions.

The overall loss function for AeroGen combines both the global text condition and layout control, defined as:

$$\mathcal{L} = \mathbb{E}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}^g, \mathbf{c}^l)\right\|^2\right], \tag{4}$$

where $\mathbf{x}_t$ represents the noisy image at time step t, $\mathbf{c}^g$ is the global text condition, and $\mathbf{c}^l$ is the layout control.

3.2. Generative Pipeline

The layout generative pipeline, as illustrated in Fig. 3, is divided into five stages: label generation, label filter, image generation, image filter, and data augmentation. Each generation step is followed by a corresponding screening step to ensure synthesis quality.

Label Generation. Inspired by recent cutting-edge research [34], we adopt a denoising diffusion probabilistic model (DDPM [11]) to learn the conditional distribution and directly sample from it to obtain layout labels, thereby avoiding conflicts in layout conditions that may arise from random synthesis approaches. The specific method is illustrated in Fig. 3, where a labelling matrix $\mathbf{M}_L$ is first constructed. This matrix contains all categories of conditions with dimensions $H \times W \times N$, where H and W represent

Figure 3. Overview of the pipeline based on AeroGen. By fitting the conditional distribution using a diffusion model, we expand a diverse set of layout conditions and combine them with AeroGen to generate synthetic data. Additionally, we introduce two filters to eliminate low-quality synthetic conditions and images, further ensuring the semantic consistency and layout consistency of the synthetic images. Finally, we incorporate synthetic images alongside real images in the training set to improve the performance of downstream tasks.

the height and width of the images, respectively, and N denotes the number of condition categories. For each condition corresponding to the target frame of the image, the value within the target frame region is set to 1, while the values in the remaining regions are set to -1. This process is formally represented as:

$$\mathbf{M}_L(i,j,k) = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{R}_k, \\ -1, & \text{if } (i,j) \notin \mathcal{R}_k, \end{cases} \tag{5}$$

where $i \in \{1,\ldots,H\}$, $j \in \{1,\ldots,W\}$, and $k \in \{1,\ldots,N\}$, with $\mathcal{R}_k$ denoting the target area for the k-th category. Next, this conditional distribution is fitted using a DDPM-based generator G_θ. The loss function is based on the mean square error (MSE):

$$\mathcal{L} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\mathbf{M}_L^{(t)}, t)\|^2\right], \tag{6}$$

where $\mathbf{M}_L^{(0)}$ represents the original layout matrix, $\mathbf{M}_L^{(t)}$ represents the noise matrix at the t-th time step, and $\epsilon_\theta(\mathbf{M}_L^{(t)}, t)$ denotes the model's predicted noise at step t.

Label Filter and Enhancement. The label data sampled from the generator may not always align with real-world intuition or effectively guide image generation. Therefore, we propose a normal distribution-based filtering mechanism to screen the generated bounding box information, ensuring that the data conform to the distribution characteristics of real labels. The label filter assumes that the attributes of the bounding boxes (e.g., area) follow a normal distribution ($\mathcal{N}(\mu_X, \sigma_X^2)$) and introduces the following constraint: $\frac{(X-\mu_X)}{\sigma_X} \leq \epsilon$, where ϵ determines the filter's strictness, thereby ensuring that generated bounding boxes fall within a realistic and feasible range. Synthetic pseudo-labels and genuine a priori labels are filtered to form a comprehensive layout condition pool through additional enhancement strategies, including scaling, panning, rotating, and flipping.

Image Generation. The synthetic bounding box labels are obtained based on the pool of layout conditions. The corresponding synthetic images are generated using the layout-guided diffusion model through the image generation process described in Sec. 3.1. The model uses these bounding box labels to guide the generation, ensuring that the image content matches the generated layout conditions.

Image Filter. Since the images generated by the diffusion model do not consistently meet high-quality or predefined layout requirements, a screening mechanism is implemented to evaluate both the quality of the generation and the consistency of the layout. The consistency of the semantic and layout is evaluated using the CLIP model [20] and a ResNet101-based classifier [10]. Synthetic images are then filtered by calculating their CLIP scores and minimum classification accuracies, which are compared against predefined thresholds to select the final filtered images.

Data Augmentation. The synthetic data serves as a complementary dataset alongside the real dataset, and both are utilized as training data for downstream target detection model training.

4. Experiments

In this section, we conducted extensive experiments to verify the generative capabilities of AeroGen and its auxiliary data augmentation ability to support downstream RSIOD

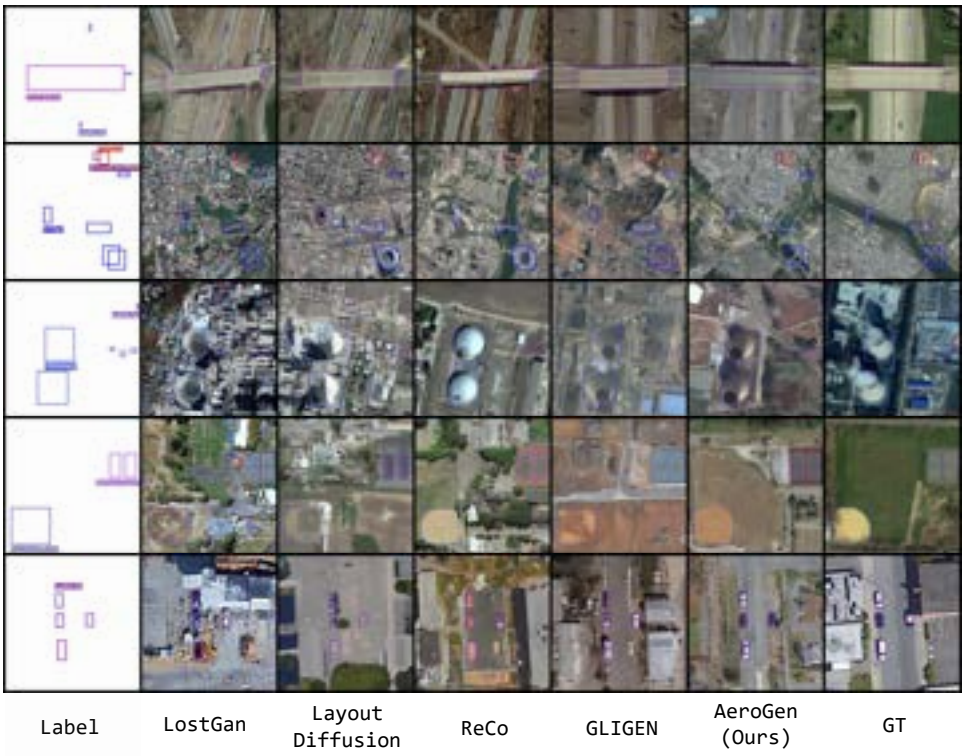

Figure 4. Visualization comparison of the generated image by different methods on the DIOR dataset.

Dataset	Modality	Images	Objects	Categories
DIOR [16]	HBB	23,463	192,518	20
DIOR-R [3]	OBB	23,463	192,518	20
HRSC [22]	OBB	1,061	2,976	19

Table 1. Statistical information of the benchmark RSIOD datasets.

tasks. Specifically, we assessed the performance of our layout generation model AeroGen from both quantitative and qualitative perspectives. Subsequently, we performed data augmentation experiments on three datasets (i.e., DIOR, DIOR-R, and HRSC) to verify the effectiveness of synthetic data generated by our AeroGen model in improving the performance of downstream object detection tasks.

4.1. Implementation Details

Data Preparation. An overview of the three datasets is provided in Tab. 1. Notably, the DIOR and DIOR-R datasets [3] share the same image data but differ in annotation format, with DIOR using bounding boxes and DIOR-R using rotated bounding boxes. HRSC [22] is a Remote Sensing dataset for ship detection, with image sizes ranging from 300×300 to 1500×900 pixels.

Training Details. We trained our AeroGen separately on each dataset for 100 epochs. During training, we used the AdamW optimizer [23] with a learning rate of 1e-5. Only the attention layers of UNet and the Layout Mask Attention (LMA) are updated, while the remaining weights are inherited from the fine-tuned LDM in RS data [33].

Evaluation Metrics. For the quantitative analysis of generated images, we used the FID score to evaluate the visual quality of the generated images and employed Classification Score (CAS) [29] and YOLO Score [19] to measure the layout consistency of the generated images. In the data augmentation experiments, we assessed object detection model performance based on mAP50 and mAP50-95 (mAP) metrics to evaluate their overall quality.

4.2. Image Quality Results

Quantitative Evaluation. We used a bounding box condition defined by four extreme coordinates and conducted both training and testing on the DIOR dataset. We compared AeroGen with state-of-the-art layout-to-image generation methods, including LostGAN [32], ReCo [39], LayoutDiffusion[44], and GLIGEN [17]. The performance of these methods on three metrics is reported in Tab. 2. To ensure fairness, we initialized all methods with identical SD weights (Stable Difusion 1.5) and trained them on the DIOR dataset for the same number of epochs. Our method outperformed other methods across all the metrics.

Furthermore, we evaluated AeroGen and GLIGEN on

Method	Dataset	Modality	FID ↓	CAS ↑	YOLO Score ↑
LostGAN [32]	DIOR [16]	HBB	57.10	46.02	14.3/27.3/15.2
Layout Diffusion [44]			45.31	56.98	20.0/37.4/19.3
ReCo [39]			42.56	55.42	21.1/40.7/23.1
GLIGEN [17]			41.31	63.50	25.8/44.4/27.8
AeroGen (Ours)			**38.57**	**76.84**	**29.8/54.2/31.6**
GLIGEN [17]†	DIOR-R [3]	OBB	48.43	58.89	24.6/41.6/25.1
AeroGen (Ours)			**35.07**	**74.13**	**29.6/57.6/32.0**
GLIGEN [17]†	HRSC [22]	OBB	66.69	43.35	23.4/44.7/26.3
AeroGen (Ours)			**45.86**	**51.19**	**27.1/51.0/27.6**

Table 2. Quantitative results of the generated images by different methods. For the Oriented Bounding Box (OBB) modality, we replicated the GLIGEN† method to account for the rotated bounding box. The best results are in **bold**.

the DIOR-R and HRSC datasets with rotated bounding boxes, where AeroGen consistently excelled. Notably, the original GliGen method does not support rotated bounding box conditions; therefore, we modified the layout encoding (as shown in Fig. 2 (a)) and retrained the model.

Qualitative Evaluation. Fig. 4 compares the results of AeroGen with those of other methods. AeroGen shows superior layout consistency and an enhanced capability for generating small objects. Besides, we present experimental results on natural images in the supplemental material.

4.3. Data Augmentation Experiments

We synthesized augmented data on three RSIOD datasets to enhance model training. For the DIOR/DIOR-R datasets, we generated 10k, 20k, and 50k synthetic samples respectively for the RSIOD task. Following the same proportion, 2k, 4k, and 10k synthetic samples were created for the HRSC dataset. The experiments were conducted using the OBB branch configurations of two established RSIOD frameworks: the unified YOLOv8 [35] and Oriented R-CNN [38]. Table 3 demonstrations that incorporate synthetic data consistently enhance downstream task performance across all experimental settings.

We visualize the mAP scores for different categories in detail, as shown in Fig. 5. In most categories, results incorporating enhancements significantly outperform those without them, particularly in rarer categories, achieving improvements of 17.8%, 14.7%, and 12.6% in the GF, DAM, and APO categories, respectively. AeroGen, enhanced on YOLOv8 by leveraging original training data and 10k synthetic images, achieves superior performance over competitors on the DOTA dataset in Tab. 4.

4.4. Ablation Study

Ablation of Enhanced Methods. We compared synthetic data enhancement methods with traditional approaches, including the basic enhancement techniques of Flip and Copy-Paste [7] for target detection tasks, as shown in Tab. 5. The target detection model trained on synthetic data performs significantly better than when trained with traditional

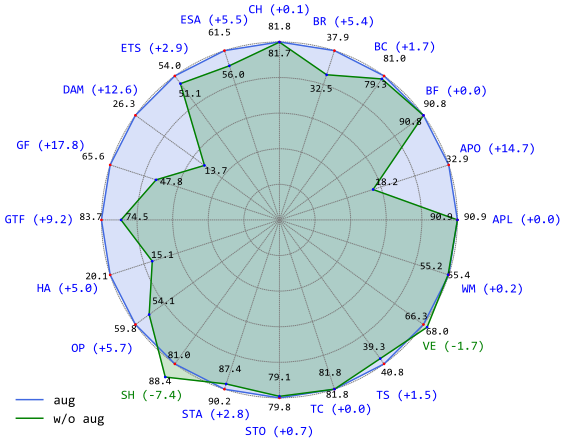

Figure 5. Comparison of mAP50 across each category on the DIOR-R dataset under the setting of data augmentation with 50k generation images (aug) and without augmentation (w/o aug).

methods, demonstrating the effectiveness of the generative model for data enhancement.

Ablation of Different Modules. We assessed the impact of different modules on image quality generated by AeroGen in Tab. 6. The contribution of each module to the enhancement of image quality is evaluated by incorporating additional components into the original SD model. Results show that Layout Mask Attention (LMA) effectively captures global semantic information and preserves layout consistency, while adding Dual Cross Attention (DCA) further enhances performance, particularly in YOLO Score, indicating improved regional target generation. Overall, the model performs best when both LMA and DCA are used.

Ablation of Augment experiments. We further analyze the filtering strategies and data augmentation techniques in the generation pipeline, including diverse generation strategies, filtering strategies for layout conditions, and filtering strategies for layout and semantic consistency of images. We use the synthetic data generated in various ways as enhancement data and conduct enhancement experiments on

Gen Data	mAP ↑	mAP50 ↑
0	54.22	72.69
10k	55.62	74.79
20k	56.78	76.31
50k	**57.92**	**77.10**

(a) DIOR

Gen Data	mAP ↑	mAP50 ↑
0	37.39	60.21
10k	39.81	62.39
20k	41.12	63.33
50k	**41.69**	**64.12**

(b) DIOR-R

Gen Data	mAP ↑	mAP50 ↑
0	63.49	90.28
2k	64.12	91.79
4k	64.78	92.31
10k	**65.92**	**93.10**

(c) HRSC

Table 3. The enhancement effects of various scales of synthetic data on downstream tasks across DIOR, DIOR-R, and HRSC datasets.

Dataset	Baseline	LostGAN	ReCo	LayoutD	GLIGEN	AeroGen
DIOR	72.69	69.49	72.89	73.07	73.39	**74.79**
DOTA1.0	70.31	60.52	68.56	68.91	70.22	**71.38**

Table 4. Enhancement experiment on original train set and 10k synthetic images on YOLOv8 Model.

Strategy	mAP ↑	mAP50 ↑
Flip	37.39	60.21
CopyPaste [7]	38.25	61.79
AeroGen	41.32	63.98
CopyPaste + Flip [7]	38.75	62.11
AeroGen + Flip	**41.69**	**64.12**

Table 5. Comparison of different augmentation strategies on the DIOR-R dataset with 50k synthetic images.

LMA	DCA		FID ↓	CAS ↑	YOLO Score ↑
	Local	Global			
✗	✗	✗	82.11	18.48	1.3/3.9/1.1
✓	✗	✗	61.50	50.11	25.3/46.5/27.2
✗	✓	✓	66.29	40.71	16.5/29.2/17.7
✓	✗	✓	51.42	55.26	25.8/46.9/27.3
✓	✓	✗	40.76	73.89	28.5/52.7/29.3
✓	✓	✓	**38.57**	**76.84**	**29.8/54.2/31.6**

Table 6. Ablation study of different modules in the AeroGen model on the DIOR dataset.

Layout Diversity			Image Quality		Metrics	
Synthesis	Filter	Augment	Semantic	Layout	mAP ↑	mAP50 ↑
✓	✓	✓	✓	✓	**41.69**	**64.12**
✗	✗	✓	✓	✓	41.31	63.47
✓	✗	✓	✓	✓	40.92	62.41
✓	✓	✗	✓	✓	39.62	61.32
✓	✓	✓	✗	✓	40.27	62.13
✓	✓	✓	✓	✗	37.05	60.03

Table 7. Ablation study results analyzing the impact of diverse generation and filtering strategies in the synthesis pipeline on layout labeling and image filtering consistency.

the DIOR-R datasets and the experimental results are shown in Tab. 7. As can be seen, each component in the generation pipeline contributes positively.

Train Set	Redet	F-RCNN	O-RCNN
DIOR	59.21	57.32	60.21
DIOR+10k Synthetic Data	**62.03**	**60.11**	**62.39**

Table 8. Comparison of detection models with data augmentation.

Dataset	mAP	mAP50
HRSC-100	35.20 / 37.35	70.25 / 73.60
HRSC-100+2k	42.35 / 44.48	78.17 / 80.05
HRSC-100+5k	**46.23 / 48.02**	**81.98 / 84.23**

Table 9. Few-shot experiment with 100 images (HRSC-100).

To systematically evaluate the effectiveness of data augmentation strategies across diverse detection frameworks, we conducted comprehensive experiments on the DIOR dataset. As demonstrated in Tab. 8, the proposed AeroGen framework achieved significant and consistent performance improvements in all architectures evaluated. To further validate its utility in data-scarce scenarios, we constructed a simulated few-shot training subset (HRSC-100) augmented with AeroGen-generated synthetic samples. Experimental results in Tab. 9 reveal that AeroGen robustly enhances detection accuracy under limited training data conditions.

5. Conclusion

This paper introduces AeroGen, a layout-controllable diffusion model designed to enhance remote sensing image datasets for target detection. The model comprises two primary components: a layout generation model that creates high-quality remote sensing images based on predefined layout conditions, and a data generation pipeline that incorporates a diversity of condition generators for the diffusion model. The pipeline employs a double filtering mechanism to exclude low-quality generation conditions and images, thereby ensuring the semantic and layout consistency of the generated images. By combining synthetic and real images in the training set, AeroGen significantly improves model performance in downstream tasks. This work highlights the potential of generative modeling in enhancing the datasets of remote sensing image processing tasks.

Acknowledgment. This work was supported by the National Natural Science Foundation of China under Grant 62192781, Grant 62272375, Grant 62272374, Grant 62250009, Grant 62137002, Grant 62276208, and Grant 62425113, in part by Natural Science Foundation of Shaanxi Province under Grant 2024JC-JCQN62, in part by Natural Science Basic Research Program of Shaanxi Province under Grant 2024JC-JCQN-02, in part by Project of China Knowledge Center for Engineering Science and Technology, in part by Project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for "The Belt and Road" Training in MOOC China," in part by the Major Key Project of PCL under Grant PCL2024A06, and in part by Tianyuan Fund for Mathematics of the National Natural Science Foundation of China under Grant 12426105.

References

[1] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. In *ICLR*, 2024. 3

[2] Yicheng Chen, Xiangtai Li, Yining Li, Yanhong Zeng, Jianzong Wu, Xiangyu Zhao, and Kai Chen. Auto cherry-picker: Learning from high-quality generative data driven by language. *arXiv preprint arXiv:2406.20085*, 2024. 3

[3] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 6, 7

[4] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021. 2

[5] Hyungjin Chung, Eun Sun Lee, and Jong Chul Ye. Mr image denoising and super-resolution using regularized reverse diffusion. *IEEE Transactions on Medical Imaging*, 42(4): 922–934, 2022. 2

[6] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8795–8805, 2024. 2

[7] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017. 2, 7, 8

[8] Chengxiang Fan, Muzhi Zhu, Hao Chen, Yang Liu, Weijia Wu, Huaqi Zhang, and Chunhua Shen. Divergen: Improving instance segmentation by learning wider data distribution with more diverse generative data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3986–3995, 2024. 3

[9] Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8532–8542, 2024. 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[13] Chengyou Jia, Minnan Luo, Zhuohang Dang, Guang Dai, Xiaojun Chang, Mengmeng Wang, and Jingdong Wang. Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2480–2488, 2024. 3

[14] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[15] Mate Kisantal. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019. 2

[16] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 2, 6, 7

[17] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2, 3, 6, 7

[18] Yuhang Li, Xin Dong, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. A simple background augmentation method for object detection with diffusion model. *arXiv preprint arXiv:2408.00350*, 2024. 3

[19] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13819–13828, 2021. 6

[20] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 5

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amster-*

dam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37. Springer, 2016. 2

[22] Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping Yang. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE geoscience and remote sensing letters*, 13(8): 1074–1078, 2016. 6, 7

[23] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[25] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[26] Li Pang, Datao Tang, Shuang Xu, Deyu Meng, and Xiangyong Cao. Hsigene: A foundation model for hyperspectral image generation. *arXiv preprint arXiv:2409.12470*, 2024. 3

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

[29] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019. 6

[30] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[32] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019. 6, 7

[33] Datao Tang, Xiangyong Cao, Xingsong Hou, Zhongyuan Jiang, Junmin Liu, and Deyu Meng. Crs-diff: Controllable remote sensing image generation with diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 3, 6

[34] Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27695–27705, 2024. 2, 3, 4

[35] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6. IEEE, 2024. 7

[36] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024. 2

[37] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 2

[38] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3520–3529, 2021. 7

[39] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 2, 3, 6, 7

[40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[41] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J Ma. Diffusionengine: Diffusion model is scalable data engine for object detection. *arXiv preprint arXiv:2309.03893*, 2023. 3

[42] Xin Zhang, Liangxiu Han, Lianghao Han, and Liang Zhu. How well do deep learning-based methods for land cover classification and object detection perform on high resolution remote sensing imagery? *Remote Sensing*, 12(3):417, 2020. 2

[43] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In *International Conference on Machine Learning*, pages 42098–42109. PMLR, 2023. 2, 3

[44] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 2, 3, 6, 7

[45] Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made dis-

asters. *Remote Sensing of Environment*, 265:112636, 2021. 2

[46] Zhuo Zheng, Stefano Ermon, Dongjun Kim, Liangpei Zhang, and Yanfei Zhong. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[47] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024. 2, 3

[48] Jingyuan Zhu, Shiyu Li, Yuxuan Liu, Ping Huang, Jiulong Shan, Huimin Ma, and Jian Yuan. Odgen: Domain-specific object detection data generation with diffusion models. *arXiv preprint arXiv:2405.15199*, 2024. 3

[49] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 2