Object Fidelity Diffusion for Remote Sensing Image Generation

Anonymous authors
Paper under double-blind review

Abstract

High-fidelity, controllable remote sensing layout-to-image generation is highly valuable for providing high-quality data for downstream object detection tasks. However, existing methods either rely on additional textual guidance, leading to geometric distortions, or require extra real-image references, limiting practical applicability. To address these challenges, we propose Object Fidelity Diffusion (OF-Diff), which leverages object layouts to extract structural shape priors and employs an online-distillation strategy to integrate complex image features. This allows the model to perform highly controllable, high-fidelity image generation at inference without relying on real-image references. Furthermore, we introduce DDPO to fine-tune the diffusion process, making the generated remote sensing images more diverse and semantically consistent. Comprehensive experiments demonstrate that OF-Diff outperforms state-of-the-art methods in the remote sensing across key quality metrics. Notably, the performance of several polymorphic and small object classes shows significant improvement. For instance, the mAP increases by 8.3%, 7.7%, and 4.0% for airplanes, ships, and vehicles, respectively.

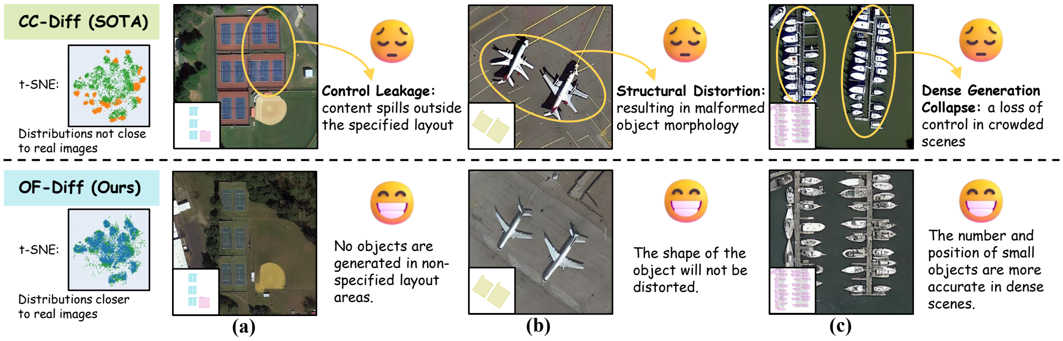

CC-Diff (SOTA)
t-SNE:
Distributions not close to real images
Control Leakage: content spills outside the specified layout
Structural Distortion: resulting in malformed object morphology
Dense Generation Collapse: a loss of control in crowded scenes
OF-Diff (Ours)
t-SNE:
Distributions closer to real images
No objects are generated in non-specified layout areas.
The shape of the object will not be distorted.
The number and position of small objects are more accurate in dense scenes.
(a)
(b)
(c)

Figure 1: Four critical failure modes in the State-of-the-Art (SOTA) method (CC-Diff): a distributional drift from real data, visualized by t-SNE; and (a) Control Leakage; (b) Structural Distortion; (c) Dense Generation Collapse. Our OF-Diff (2nd row) effectively resolves these issues.

1 Introduction

Synthesizing high-fidelity, spatially-controllable remote sensing (RS) images is a critical frontier for overcoming the data limitations that hinder downstream perception tasks like object detection (Yang et al., 2021; Zhang et al., 2020; Yang et al., 2019). Current RS generation methods, however, typically rely on either ambiguous text prompts (Khanna et al., 2023; Sebaq & ElHelw, 2024) or auxiliary conditions like semantic maps (Sebaq & ElHelw, 2024; Tang et al., 2024; Gong et al., 2024; Hu et al., 2025; Jia et al., 2025). While visually plausible, such guidance is fundamentally disconnected from the instance-level ground truth, failing to provide the precise control necessary for effective data augmentation.

In contrast, Layout-to-Image (L2I) generation conditioned on object bounding boxes offers a more robust solution for precise spatial control. This paradigm has been extensively studied in the natural

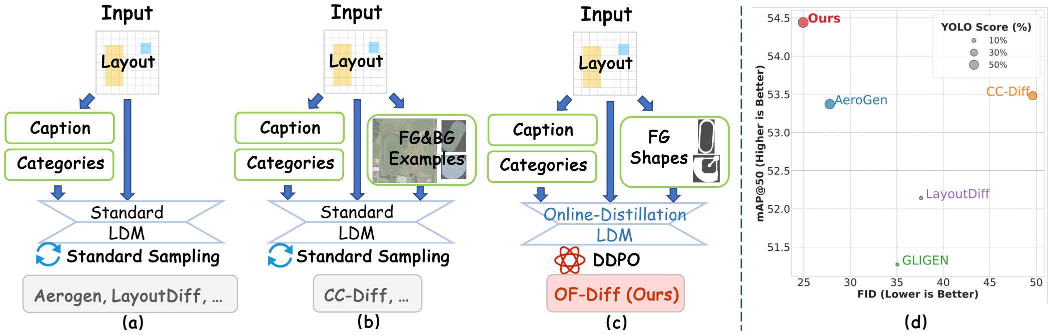


Figure 2: Comparison of OF-Diff with mainstream L2I methods. FG/BG stands for foreground/background. (a) Layout-conditioned baseline. (b) Added instance-based module, limited by quality/quantity of patches from ground truth. (c) OF-Diff enhances fidelity via shape extraction and DDPO, without patch reliance. (d) Results demonstrate superiority.

image domain—LayoutDiffusion (Zheng et al., 2023) treats it as a multi-modal fusion problem, GLIGEN (Li et al., 2023) enables open-world generation through additional control signals, and ODGen (Zhu et al., 2024) improves controllability by decoupling objects—yet its direct application to remote sensing (RS) imagery remains non-trivial due to expansive backgrounds, arbitrary object orientations, and densely packed scenes.

In RS layout-to-image generation, existing methods like AeroGen (Tang et al., 2025) and CC-Diff (Zhang et al., 2024) take different approaches. AeroGen, a coarse layout-conditioned model, suffers from limited spatial and shape control. In contrast, instance-level methods like CC-Diff achieve higher controllability and fidelity by referencing real instances, but this creates heavy dependence on the quality and quantity of real data, limiting generalization and flexibility.The images generated via CC-Diff diverge more markedly from the real remote sensing data distribution, aligning instead with the style characteristic of the model's pre-training corpus. We summarize common failure modes (see Figure 1), including control leakage, structural distortion, dense generation collapse and feature-level mismatch.

These deficiencies significantly degrade the performance on object detection tasks, limiting their practical application in intelligent RS interpretation. In this paper, we introduce **O**bject **F**idelity **Diff**usion Model (OF-Diff). It is designed to improve the shape fidelity and layout consistency of object generation in RS images. As shown in Figure 2, the existing L2I methods are mainly divided into two categories. The first is layout-conditioned baseline, as shown in Figure 2(a), like Aerogen and LayoutDiffusion. The second is the method with instance-based module, as shown in Figure 2(b), like CC-Diff. However, such methods require real instances and images as references during the sampling stage in order to generate high-quality synthetic images. In contrast, OF-Diff generates high-fidelity remote-sensing objects using only the foreground shape, and subsequently employs online-distillation to further align the outputs with real images, as shown in Figure 2(c). In addition, it fine-tunes the diffusion with DDPO, effectively enhancing the performance of downstream tasks for the generated images. The results in Figure 2(d) demonstrates the superiority of OF-Diff over other methods. Our contributions are summarized as follows:

- We introduce OF-Diff, an online-distillation controllable diffusion model with prior shape extraction, which improves generation fidelity while reducing reliance on real images, enhancing practical applicability.

- We propose a controllable generation pipeline that fine-tunes diffusion models with DDPO for remote sensing images, further boosting fidelity and diversity.

- Extensive experiments demonstrate that OF-Diff generates high-fidelity, layout- and shape-consistent images with dense objects, and serves as an effective enhancement for object detection tasks.

2 Related Work

2.1 Advances in Image Generation

Diffusion models (Dhariwal & Nichol, 2021; Ho et al., 2020; Kingma et al., 2021) have increasingly replaced Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Karras et al., 2021) and Variational Autoencoders (VAEs) (Kingma et al., 2013; Rezende et al., 2014) in image synthesis tasks due to their training stability and superior output quality. Recent advances in efficient samplers, such as DDIM (Song et al., 2021), Euler (Karras et al., 2022), and DPM-Solver (Lu et al., 2022), have further improved the practicality. Latent Diffusion Models (LDMs) (Rombach et al., 2022b), which operate in low-dimensional latent spaces, significantly reduce computational costs while preserving visual fidelity. The success of models like DALL·E2 (Ramesh et al., 2022) and Imagen (Saharia et al., 2022) demonstrates how this paradigm supports training on vast internet-scale datasets. As a result, diffusion-based approaches now provide a strong foundation for high-quality image generation.

2.2 Layout-to-Image Generation

Controllable image synthesis primarily includes text-to-image (T2I) and layout-to-image (L2I) generation. While T2I models (Nichol et al., 2022; Ramesh et al., 2022) achieve semantic alignment via textual prompts, L2I methods offer better spatial control. Recent works enhance layout conditioning through layout-as-modality designs (Zheng et al., 2023), gated attention (Li et al., 2023), and instance-wise generation (Wang et al., 2024; Zhou et al., 2024). However, these methods rely solely on coarse layout inputs (e.g., bounding boxes), which lack fine-grained shape information critical for synthesizing morphologically complex objects.

2.3 Remote Sensing Image Synthesis

Synthesizing high-fidelity training data is crucial for advancing remote sensing (RS) object detection, a field critical to numerous applications, but often hampered by the scarcity of extensively annotated datasets. Despite its necessity, most generative models for RS imagery, such as DiffusionSat (Khanna et al., 2023) and RSDiff (Sebaq & ElHelw, 2024), still rely on coarse semantic guidance. While other approaches leverage diverse control signals (Tang et al., 2024) like OpenStreetMaps (Espinosa & Crowley, 2023), they are generally not optimized for the bounding box format central to object detection. This naturally motivates L2I approaches including AeroGen (Tang et al., 2025) and CC-Diff (Zhang et al., 2024), which have improved spatial accuracy and contextual consistency through layout-mask attention and FG/BG dual re-samplers. However, they suffer from limited controllability and heavy reliance on real data.

3 Method

3.1 Preliminary

Diffusion models (Song et al., 2021) aim to capture the underlying data distribution $p(x)$ by iteratively reconstructing data from a noisy representation that is initially sampled from a standard normal distribution. Denoising Diffusion Probabilistic Models (Ho et al., 2020) parameterize the model as the function $\epsilon_\theta(x_t, t)$ to predict the noise component of the sample x_t at any time step t. The training objective is to minimize the mean squared error (MSE) loss between the actual noise ϵ and the predicted noise $\epsilon_\theta(x_t, t)$:

$$\mathcal{L} = \mathbb{E}_{x_t, t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \epsilon_\theta \left(x_t, t \right) - \epsilon \right\|^2 \right]. \tag{1}$$

Stable Diffusion (SD) (Rombach et al., 2022b; Qiu et al., 2025) utilizes a pre-trained VQ-VAE Van Den Oord et al. (2017) to encode images into a lower-dimensional latent space, performing training on the latent representation z_0. In the context of conditional generation, given a text prompt c_t and task-specific conditions c_f, the diffusion training loss at time step t can be expressed as:

$$\mathcal{L} = \mathbb{E}_{z_t, t, c_t, c_f, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \epsilon - \epsilon_\theta \left(z_t, t, c_t, c_f \right) \right\|^2 \right]. \tag{2}$$

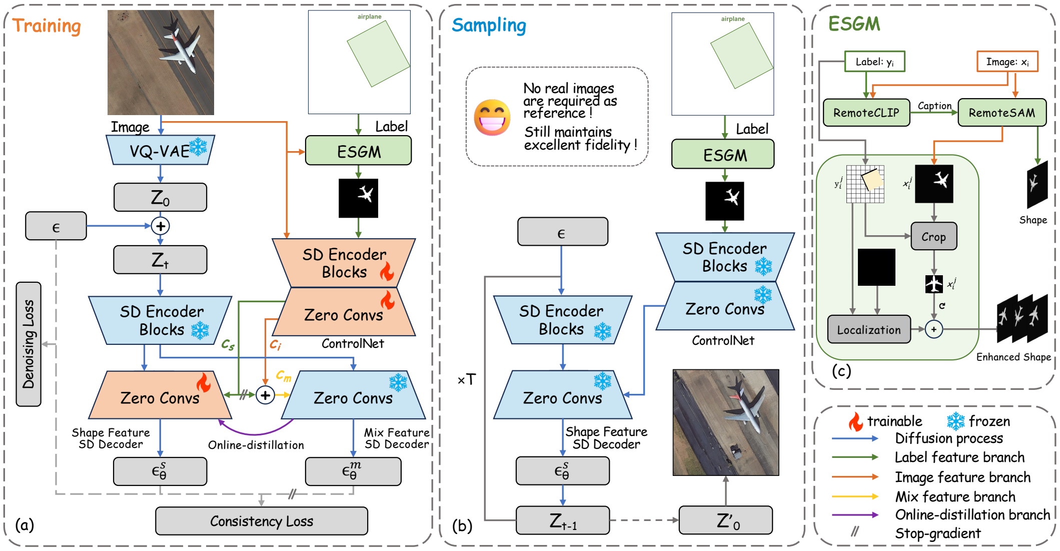


Figure 3: OF-Diff's overall architecture. (a) During training, object shape features extracted by ESGM and image features are processed by ControlNet, and the resulting information is used to update stable diffusion decoders via online-distillation. (b) During sampling, only the label and the shape feature stable diffusion decoder are used to generate synthetic images. (c) Architecture of the Enhanced Shape Generation Module (ESGM).

where $\mathcal{L}$ represents the overall learning objective of the complete diffusion model. This objective function is explicitly applied during the fine-tuning of diffusion models in conjunction with ControlNet (Zhang et al., 2023).

3.2 Architecture of OF-Diff

As illustrated in Figure 3(a), the training of OF-Diff requires both real images and their corresponding labels. **First**, for ControlNet, the real image and its label are processed by the Enhanced Shape Generation Module (ESGM) to extract the object mask. The image and mask are then fed into ControlNet to obtain the image feature c_i and the shape feature c_s. To enrich the structural-only shape prior with richer appearance and contextual cues from the image, we combine them into a mix-feature c_m, which will later serve as a teacher input in online-distillation. Concretely:

$$c_{\mathrm{m}} = \frac{n}{N} \cdot c_i + \mathrm{sg}\left[c_s\right], \tag{3}$$

where n denotes the current iteration number, and N is the total number of iterations. In order to enable the prediction conditioned on mix-feature to serve as a stable anchor point, to improve the morphological fidelity of the generation, we adopt a stop-gradient strategy (Chen & He, 2021) for c_s when calculating c_m.

Second, for Stable Diffusion, the input image is first compressed into latent space features z_0 by a pre-trained VQ-VAE. Then, it is concatenated with Gaussian noise ϵ to form z_t. After passing through the SD encoder blocks, the feature Z_t is fed into a dual-decoder architecture. One branch, the shape-feature SD decoder, conditions on c_s; the other, the mix-feature SD decoder, conditions on c_m. Their reconstruction losses are defined as L_s and L_M, respectively:

$$\mathcal{L}_s = \mathbb{E}\left[\left\|\epsilon_\theta^s - \epsilon\right\|^2\right], \epsilon_\theta^s = \epsilon_\theta\left(z_t, t, c_t, c_s\right), \tag{4}$$

$$\mathcal{L}_m = \mathbb{E}\left[\left\|\epsilon_\theta^m - \epsilon\right\|^2\right], \epsilon_\theta^m = \epsilon_\theta\left(z_t, t, c_t, c_m\right), \tag{5}$$

Third, for online distillation, the mix-feature SD produces more accurate predictions thanks to its stronger image prior, but needs real images, limiting diversity. In contrast, the shape-feature SD supports arbitrary label control but risks converging to low-fidelity local minima. To reconcile these trade-offs, we propose an online-distillation framework with a consistency loss L_c:

$$\mathcal{L}_c = \mathbb{E}\left[\|\epsilon_\theta^s - \mathrm{sg}\left[\epsilon_{\theta'}^{\mathrm{m}}\right]\|^2\right]. \tag{6}$$

Here, the prediction $\epsilon_{\theta'}^m$ from mix-feature SD decoder acts as a stop-gradient teacher signal, serving as an anchor to guide the prediction ϵ_θ^s from shape-feature SD decoder towards high-fidelity optima in parameter space.

The overall training objective is therefore:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_m + \lambda\mathcal{L}_c, \tag{7}$$

During the sampling phase, as illustrated in Figure 3(b), only the frozen ControlNet and the shape feature stable diffusion are utilized with arbitrary label prior control to synthesize RS images.

3.3 Enhanced Shape Generation Module

In natural imagery, perspective and scale changes prevent a unique geometric model for most objects. Conversely, remote-sensing objects display quasi-invariant shapes. For instance, courts are rectangular, chimneys and oil tanks circular, and airplanes bilaterally symmetric with a distinct nose and tail. This shape consistency enables the use of masks to impose strong controllability on image synthesis for remote sensing. To better leverage category labels for object shape extraction, we introduce the Enhanced Shape Generation Module (ESGM, see Figure 3(c)). During the training phase, ESGM uses paired images and labels to generate precise object masks. And at sampling time, it employs learned shape priors to synthesize diverse masks of object shape.

For the given image x_i and its bounding box y_i^j corresponding to category j ($j \in [1, N]$), we first utilize the RemoteCLIP (Liu et al., 2024) to generate a textual description of the object enclosed within the bounding box. With this description and the original image x_i, the RemoteSAM (Yao et al., 2025) then generates the corresponding shape masks $\{x_i^j\}$.

In the shape augmentation phase, each object mask x_i^j is cropped by its bounding box y_i^j, randomly rotated, and placed back onto a blank canvas to produce a shape-enhanced mask. During training, ESGM uses real image shapes; at sampling, it selects enhanced shapes from a lightweight mask pool collected during or after training. In our experiments, we use masks generated during training.

3.4 DDPO fine-tuning

To enhance the diversity of the distribution of data generated by the fine-tuned model and maintain better consistency with the distribution of real images Schulman et al. (2015; 2017), denoising diffusion policy optimization (DDPO) Black et al. (2023) is applied in the post-training of OF-Diff. DDPO regards the denoising process of the diffusion model as a multi-step Markov decision process (MDP) (for a detailed derivation, please refer to the Appendix A.2). To optimize the policy $\pi(a_t \mid s_t)$ so as to maximize the cumulative reward $\mathbb{E}_{\tau\sim p(,\cdot|\pi)}!\left[\sum_{t=0}^{T} R(\mathbf{s}_t, \mathbf{a}_t)\right]$, the gradient $\hat{g}$ is computed as follows:

$$\hat{g} = \mathbb{E}\left[\sum_{t=0}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \mid c,\ t,\ \mathbf{x}_t)}{p_{\theta'}(\mathbf{x}_{t-1} \mid c,\ t,\ \mathbf{x}_t)} \cdot r(\mathbf{x}_0, c) \cdot \nabla_\theta \log p_\theta(\mathbf{x}_{t-1} \mid c,\ t,\ \mathbf{x}_t)\right] \tag{8}$$

$$r(\mathbf{x}_0, c) = \big(KNN(\mathbf{x}_0, \mathbf{x}_0) - \omega KL(\mathbf{x}_0, \mathbf{x}_0')\big) \tag{9}$$

The reward functions based on K-Nearest Neighbor (KNN) and KL divergence are introduced, respectively, to optimize the diversity of generated data and the distribution consistency between generated data and real data. ω is the weight parameter, and $\mathbf{x}_0'$ is the real image in the dataset. Following standard practice, we compute the KNN in the low-dimensional embedding space of CLIP's image encoder. The implementation details are shown in Appendix A.2.

4 Experiments

4.1 Experimental Settings

Datasets. DIOR-R (Cheng et al., 2022), the rotated variant of DIOR (Li et al., 2020), contains 20 categories annotated with oriented bounding boxes; we follow the official 1:1:2 split for training/validation/testing. **DOTA-v1.0** (Xia et al., 2018) includes 15 categories featuring dense scenes and small objects. We crop the images from DOTA to 512 × 512 following MMRotate (Zhou et al., 2022), discarding those without valid objects. **HRSC2016** (Liu et al., 2017) is a high-resolution ship detection dataset with a multi-level hierarchical taxonomy. We use the finest-grained level, consisting of 26 detailed ship categories. The experiments of this dataset are reported in Appendix A.4 Unless otherwise specified, we train the diffusion model on the trainset. For downstream detection, we use the trainset annotations as layout and mix generated samples with the real trainset, and report evaluation results on the testset.

Implementation Details. We train OF-Diff separately on each dataset (DIOR/DOTA), based on the Stable Diffusion 1.5 (Rombach et al., 2022a) pretrained model. Only the ControlNet and shape feature SD decoder are fine-tuned, while all other modules remain frozen. The weighting coefficient λ of the consistency loss is set to 1, the k value in KNN is set to 50, and the weight ω of the KL divergence is set to 2. Training is performed using the AdamW optimizer with a learning rate of 1e-5. The global batch size is set to 64, and training runs for 100 epochs.

Benchmark Methods. We compare our method with state-of-the-art L2I generation models for both remote sensing (AeroGen (Tang et al., 2025), CC-Diff (Zhang et al., 2024)), and natural images (LayoutDiffusion (Zheng et al., 2023), GLIGEN (Li et al., 2023)). For a fair comparison, all models are re-trained using our dataset settings, following their official training details respectively.

Evaluation Metrics. To more comprehensively evaluate the effectiveness of OF-Diff, we adopt a total of 13 metrics spanning 4 different evaluation aspects.

- **Generation Fidelity.** We use **FID** (Heusel et al., 2017) and **KID** (Bińkowski et al., 2018) to assess perceptual quality, along with **CMMD** (Jayasumana et al., 2024), which measures CLIP feature distances between generated and real images to evaluate layout alignment.
- **Layout Consistency.** We report **CAS** (Ravuri & Vinyals, 2019) using a pretrained classifier to assess object recognizability, and **YOLOScore** by applying a pretrained Oriented R-CNN (Xie et al., 2021) (w/. Swin Transformer backbone (Liu et al., 2021), MMRotate) to generated images for instance-level consistency.
- **Shape Fidelity.** To assess the geometric quality of generated instances, we perform pairwise comparisons with ground-truth shapes. Each instance pair is cropped, resized to 64×64, and converted to edge maps. We compute five metrics: **IoU**, **Dice**, Chamfer Distance (**CD**), Hausdorff Distance (**HD**), and **SSIM** (Wang et al., 2004).
- **Downstream Utility.** We train a detector on mixed real and generated images and report **mAP**$_{50}$, **mAP**$_{75}$, and overall **mAP** on real test data using Oriented R-CNN (Swin backbone) with a batch size of 24 on 8×NVIDIA 4090 GPUs.

4.2 Qualitative Results

Comparative Results. Figure 4 compares the generation results of OF-Diff with other methods. OF-Diff not only generates more realistic images but also has the best controllability. For instance, in the first two cases, OF-Diff successfully controlled the number and layout information of the generated objects. The third and fourth cases demonstrate the accuracy of OF-Diff in generating small targets, which other algorithms fail to do accurately. The last case shows the superiority of OF-Diff over other algorithms when generating objects with complex shapes such as airplane.

Diversity Results. The images generated by OF-Diff consistently present plausible textures and realistic object shapes, as shown in Figure 6 in Appendix. For instance, airplanes rendered at different orientations maintain coherent semantic relationships with their surrounding environments. Even in small-object scenes (some of which are grayscale remote-sensing images from the DOTA dataset), OF-Diff can still generate visually faithful and geometrically accurate results.

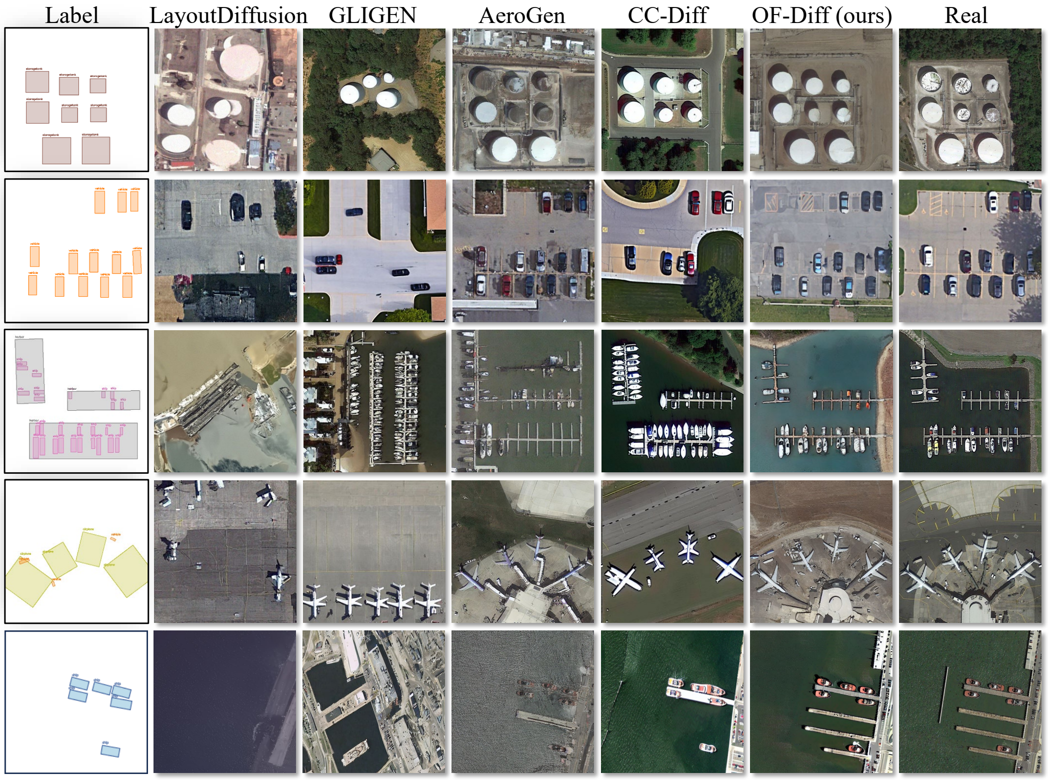


Figure 4: Qualitative results on DIOR, DOTA and HRSC2016. OF-Diff is more realistic and fidelity compared to other methods.

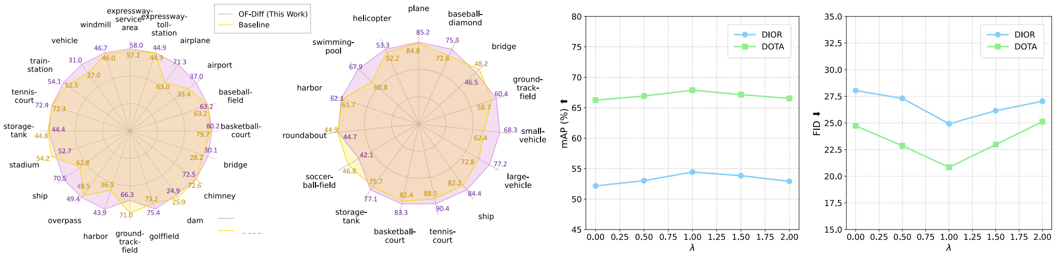
Table 1: Quantitative comparison with SOTA methods on DIOR and DOTA. We evaluate performance on **generation fidelity** (FID, KID, CMMD), **layout consistency** (CAS, YOLOScore) and **trainability** (mAP). OF-Diff demonstrates superior overall performance.

Method	**DIOR Dataset**						**DOTA Dataset**					
	FID↓	KID↓	CMMD↓	CAS↑	YOLOScore↑	mAP_{50}	FID↓	KID↓	CMMD↓	CAS↑	YOLOScore↑	mAP_{50}
LayoutDiff	37.60	0.015	0.447	70.32	7.01	52.14	21.73	0.015	0.288	77.56	21.43	66.75
GLIGEN	35.06	**0.010**	0.622	76.41	6.51	51.27	39.79	0.026	0.357	76.19	15.58	66.10
AeroGen	27.78	0.013	0.563	81.69	55.38	53.37	26.65	0.017	0.298	81.91	44.85	67.09
CC-Diff	49.62	0.024	0.685	**82.61**	42.17	53.48	32.40	0.019	0.279	81.63	49.62	66.52
Ours	**24.92**	0.011	**0.312**	82.55	**58.99**	**54.44**	**20.84**	**0.014**	**0.271**	**83.79**	**55.68**	**67.89**

4.3 Quantitative Results

Generation Fidelity and Consistency. We compared OF-Diff with state-of-the-art generation methods in remote sensing, including layoutDiffusion (Zheng et al., 2023), GLIGEN (Li et al., 2023), AeroGen (Tang et al., 2025), and CC-Diff (Zhang et al., 2024). The performance of these methods is reported in Table 1. OF-Diff achieved nearly the best performance in both generation fidelity metrics (FID, KID, CMMD) and layout consistency metrics, especially on the DOTA dataset. Additional results are available in the appendix A.4 for the HRSC2016 dataset.

Trainability of Object Detection. Following the data enhancement protocol in (Chen et al., 2023), we double the training samples using OF-Diff and assess detection results with the expanded dataset. As shown in Table 9 in Appendix A.7, OF-Diff performs the best on both DIOR and DOTA with mAP improved by 2.2% and 1.94% compared to baseline,respectively. Notably, the performance of several polymorphic and small object classes shows significant improvement. According to Figure 5 (a) and (b), the AP_{50} increases by 8.3%, 7.7%, and 4.0% for airplane, ship, and vehicle on DIOR, and 7.1%, 5.9% and 4.4% for swimming pool, small vehicle, and large vehicle on DOTA.



(a) Per-class AP_{50} on DIOR Dataset

(b) Per-class AP_{50} on DOTA Dataset

(c) Effect of λ on mAP

(c) Effect of λ on FID

Figure 5: AP_{50} on DIOR and DOTA.

Table 2: Object-Shape Fidelity on Canny Edge Maps. We measure the morphological similarity between generated and ground-truth instances by computing IoU, DICE, Chamfer Distance (CD), Hausdorff Distance (HD), and SSIM.

Method	DIOR Dataset					DOTA Dataset				
	IoU↑	Dice↑	CD↓	HD↓	SSIM↑	IoU↑	Dice↑	CD↓	HD↓	SSIM↑
LayoutDiff	0.0497	0.0908	12.037	25.962	0.1667	0.0402	0.0748	15.229	30.202	0.2194
GLIGEN	0.055	0.1002	12.257	25.850	0.1652	0.0645	0.1182	10.432	23.196	0.1967
AeroGen	0.0855	0.153	8.209	20.314	0.2142	0.0863	0.1536	8.1386	20.687	0.2261
CC-Diff	0.0891	0.1582	8.0909	20.066	0.1963	0.0692	0.1255	9.6226	21.247	0.2171
Ours	**0.1009**	**0.1763**	**7.6579**	**19.459**	**0.2691**	**0.1205**	**0.2045**	**6.6317**	**17.311**	**0.2938**

Object-Shape Fidelity. We measure the morphological similarity between the generated instances and ground truth by calculating the Intersection over Union (IoU), DICE coefficient, Chamfer distance (CD), Hausdorff distance (HD), and Structural Similarity Index (SSIM), based on the Canny Edge Map. As shown in Table 2, the results demonstrate that OF-Diff attains state-of-the-art performance in all evaluation metrics for object-shape fidelity. Specifically, we first convert the rotated bounding box (R-Box) to a horizontal bounding box (H-Box) and crop the instance with a 20% padding to ensure the full object is captured. The cropped patches are then resized to 64×64 pixels, and their shapes are extracted using cv2.Canny. For a detailed qualitative comparison, Figure 11 in Appendix A.8 visualizes the instance patches and their corresponding edge maps from different methods. Each image set is ordered as follows: Ground Truth, OF-Diff, AeroGen, CC-Diff, GLIGEN, and LayoutDiff, demonstrating our method's superior ability to adhere to object shapes.

Adaptability of Unknown Layout. To evaluate robustness of these methods, we also generate images based on the unknown layouts during the training phase. According to Table 3, for unknown layout, OF-Diff performs well in terms of generation fidelity, layout consistency, and trainability. In downstream tasks, OF-Diff still delivers a 1.54% mAP gain over the second-best method.

The Detailed Results of Downstream. Table 10 and 11 in Appendix A.8 report the average precision (AP) obtained by the competing generative methods over multiple categories in the downstream tasks. From Tables 10, it can be observed that OF-Diff (ours) achieves a clear advantage in several categories. For instance, OF-Diff achieves superior performance on Airplane (71.3%), Golf Field (75.4%) and Ship (70.5%), with improvements of approximately 5% to 10% over the second-best method. For a few other categories, OF-Diff does not deliver the top AP, yet the gap to the best result remains marginal. Table 11 shows that, on the DOTA dataset, OF-Diff obtains the highest AP in roughly half of all categories and still delivers notable gains in categories such as Small-vehicle (68.3%), Ship (84.4%) and Swimming-pool (67.9%).

4.4 Ablation Study

We assessed the impact of different modules on image generation semantic consistency and downstream trainability by OF-Diff in Table 4. We found that the images generated with captions are more in line with semantic consistency and human aesthetics, but the fidelity of these images decreases. This is equivalent to the data distribution deviating from the real dataset and being more inclined towards the data distribution during pre-training. We conduct human/GPT assessments and

Table 3: Quantitative comparison on the unknown layout dataset during training (DIOR Val).

Method	**Unknown Layout during Training**							
	FID↓	KID↓	CMMD↓	CAS↑	YOLO Score↑	mAP↑	mAP$_{50}$↑	mAP$_{75}$↑
LayoutDiff	44.58	0.018	0.539	29.34	10.37	30.41	53.07	32.07
GLIGEN	39.56	0.013	0.444	66.36	2.13	30.06	52.68	31.29
AeroGen	28.62	0.013	0.276	80.78	46.36	32.98	55.11	34.26
CC-Diff	49.92	0.024	0.513	78.01	**51.74**	32.49	53.72	35.39
Ours	**24.18**	**0.012**	**0.271**	**83.34**	49.59	**33.02**	**56.65**	**36.17**

Table 4: Ablation study: impact of ESGM, Online-distillation L_c, and DDPO on semantic consistency (CAS) and downstream trainability (YOLOScore and mAP$_{50}$).

ESGM	L_c	DDPO	FID ↓	KID ↓	CMMD ↓	CAS ↑	YOLOScore ↑	mAP$_{50}$ ↑
✗	✗	✗	42.59	0.029	0.965	80.27	41.20	52.13
✓	✗	✗	24.87	0.012	0.428	82.16	55.08	52.76
✗	✓	✗	36.25	0.021	0.596	81.57	46.27	53.14
✗	✗	✓	41.26	0.027	0.815	81.06	42.53	53.41
✓	✓	✗	24.98	**0.010**	0.313	82.30	57.83	54.31
✓	✗	✓	25.78	0.013	0.368	82.37	58.26	54.17
✗	✓	✓	37.98	0.025	0.613	81.91	47.74	53.21
✓	✓	✓	**24.92**	0.011	**0.312**	**82.55**	**58.99**	**54.44**

a fine-grained feature analysis in Appendix A.7, which collectively reveal the nature of this trade-off. Therefore, the ablation experiments for each module were conducted based on the absence of caption input. The contribution of each module to the enhancement of image generation fidelity is evaluated by incorporating additional components into the diffusion model with online-distillation. DDPO indicates whether to fine-tune the trained diffusion model through reinforcement learning. Results show that Enhanced Shape Generation Module (ESGM), Online-Distillation (L_c) and the DDPO based on KNN and KL Divergence effectively improve the performance metrics. Notably, ESGM can substantially improve the YOLOScore by over 10%. In addition, we vary the weighting coefficient λ in the consistency loss (Eq. 7) to assess its impact on mAP and FID. As shown in Figure 5 (c) and (d), both metrics are optimal at $\lambda = 1$.

4.5 Discussion

As shown in Figure 10 in Appendix A.3, the inclusion of additional captions as input has a significant impact on the outcomes of image generation. Specifically, incorporating captions enhances the aesthetic appeal of the generated images, resulting in richer and more visually pleasing color compositions. However, this improvement comes at a cost: similar to CC-Diff, it leads to a deviation of the generated data distribution from that of the original real data. In contrast, when no additional captions are provided as input, although the generated images may appear less aesthetically refined, their data distribution remains closer to that of real images. A user study from both human and GPT-5 in Table 8 have confirmed this. Further analysis of the generated-image distribution and the impact of aesthetics on performance is provided in the Appendix A.7.

5 Conclusion

Existing image generation methods struggle to precisely generate dense small objects and those with complex shapes, such as numerous small vehicles and airplanes in remote sensing images. To address this, we introduce OF-Diff, an online-distillation controllable diffusion model with prior shapes extraction and DDPO. During the training phase, we extract the prior shape of objects to enhance controllability and use a online-distillation diffusion with parameter sharing to improve the model's learning ability for real images. Therefore, in the sampling phase, OF-Diff can generate images with high fidelity without real images as references. Finally, we fine-tune the diffusion by DDPO that combines KNN and KL divergence to make the synthesized images more realistic and consistent. Extensive experiments demonstrate the effectiveness and superiority of OF-Diff in generating small and difficult objects with complex structures and dense scenes in remote sensing.

6 Ethics statement

This work adheres to the ICLR Code of Ethics. Our study focuses on generating synthetic remote sensing images using diffusion models, and does not involve human subjects or personally identifiable information. All datasets used are publicly available and appropriately licensed for research purposes. We have taken care to avoid introducing harmful biases or misrepresentations in the generated images. While OF-Diff aims to improve object fidelity and controllability in synthetic data for downstream tasks such as object detection, we acknowledge that misuse of generated images could have unintended consequences. Researchers using this technology should ensure ethical and lawful application, and consider potential societal impacts. No conflicts of interest or undisclosed sponsorship influenced this work.

7 Reproducibility Statement

We have made efforts to ensure the reproducibility of our results. The main paper provides detailed descriptions of the OF-Diff architecture, online-distillation training strategy, and DDPO fine-tuning process. All datasets used are publicly available, and the data preprocessing and layout-to-image preparation steps are described in the main text and Appendix. Hyperparameters, training schedules, and evaluation metrics are fully documented in the supplementary materials. Additionally, we provide anonymized source code and configuration files to facilitate replication of our experiments.

References

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

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. Oct 2023.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. doi: 10.1109/cvpr46437.2021.01549. URL `http://dx.doi.org/10.1109/cvpr46437.2021.01549`.

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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Miguel Espinosa and Elliot J Crowley. Generate your own scotland: Satellite image generation conditioned on maps. *arXiv preprint arXiv:2308.16648*, 2023.

Ziyang Gong, Zhixiang Wei, Di Wang, Xianzheng Ma, Hongruixuan Chen, Yuru Jia, Yupeng Deng, Zhenming Ji, Xiangwei Zhu, Naoto Yokoya, et al. Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *arXiv preprint arXiv:2410.22629*, 2024.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Xiaoxing Hu, Ziyang Gong, Yupei Wang, Yuru Jia, Gen Luo, and Xue Yang. Earth-adapter: Bridge the geospatial domain gaps with mixture of frequency adaptation. *arXiv preprint arXiv:2504.06220*, 2025.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

Yuru Jia, Valerio Marsocci, Ziyang Gong, Xue Yang, Maarten Vergauwen, and Andrea Nascetti. Can generative geospatial diffusion models excel as discriminative geospatial foundation models? *arXiv preprint arXiv:2503.07890*, 2025.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606*, 2023.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

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.

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*, pp. 22511–22521, 2023.

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*, 62:1–16, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods*, volume 2, pp. 324–331. SciTePress, 2017.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.

Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, and Yongxin Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15672–15681, 2025.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

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 (CVPR)*, pp. 10684–10695, June 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022b. doi: 10.1109/cvpr52688.2022.01042. URL `http://dx.doi.org/10.1109/cvpr52688.2022.01042`.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ahmad Sebaq and Mohamed ElHelw. Rsdiff: Remote sensing image generation from text using diffusion model. *Neural Computing and Applications*, 36(36):23103–23111, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=St1giarCHLP`.

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.

Datao Tang, Xiangyong Cao, Xuan Wu, Jialin Li, Jing Yao, Xueru Bai, Dongsheng Jiang, Yin Li, and Deyu Meng. Aerogen: Enhancing remote sensing object detection with diffusion-driven data generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3614–3624, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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*, pp. 6232–6242, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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 (ICCV)*, pp. 3520–3529, October 2021.

Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3163–3171, 2021.

Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. Remotesam: Towards segment anything for earth observation. *arXiv preprint arXiv:2505.18022*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Mu Zhang, Yunfan Liu, Yue Liu, Yuzhong Zhao, and Qixiang Ye. Cc-diff: enhancing contextual coherence in remote sensing image synthesis. *arXiv preprint arXiv:2412.08464*, 2024.

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.

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*, pp. 22490–22499, 2023.

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 (CVPR)*, pp. 6818–6828, June 2024.

Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi Lyu, et al. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 7331–7334, 2022.

Jingyuan Zhu, Shiyu Li, Yuxuan Andy Liu, Jian Yuan, Ping Huang, Jiulong Shan, and Huimin Ma. Odgen: Domain-specific object detection data generation with diffusion models. *Advances in Neural Information Processing Systems*, 37:63599–63633, 2024.

A Appendix

A.1 The Use of Large Language Models (LLMs)

Large Language Models (LLMs) were used in this work exclusively for language polishing and improving the clarity of writing. No LLM was used for generating scientific ideas, experimental design, or data analysis. All technical content, results, and interpretations presented in this paper are solely the work of the authors. The authors take full responsibility for the content of this paper, including any text refined with the assistance of LLMs.

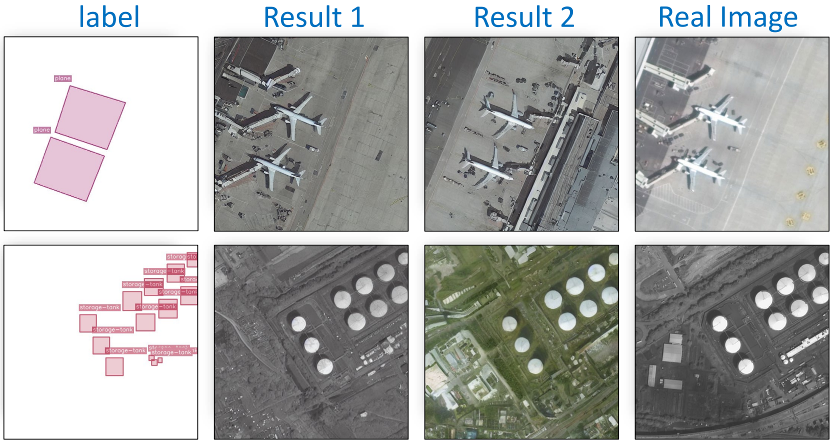


Figure 6: The diversity of different results from the same OF-Diff model.

A.2 Reinforcement Learning Strategy

The mapping relationship is defined as follows:

$$\pi(a_t \mid s_t) \triangleq p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, c) \tag{10}$$

$$P(s_{t+1} \mid s_t, a_t) \triangleq (\delta_c,\ \delta_{t-1},\ \delta_{\mathbf{x}_{t-1}}) \tag{11}$$

$$\rho_0(s_0) \triangleq \big(p(c),\ \delta_T,\ \mathcal{N}(0, I)\big) \tag{12}$$

$$R(s_t, a_t) \triangleq \begin{cases} r(\mathbf{x}_0, c), & \text{if } t = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

where δ_z denotes the Dirac delta distribution whose probability density is zero everywhere except at z. The symbols s_t and a_t represent the state and action at time t, respectively. Specifically, s_t is defined as the tuple composed of the condition c, the time step t, and the noisy image $\mathbf{x}_t$ at that time, whereas a_t is defined as the noisy image $\mathbf{x}_{t-1}$ from the preceding time step. The policy is denoted by $\pi(a_t \mid s_t)$, the transition kernel by $P(s_{t+1} \mid s_t, a_t)$, the initial state distribution by $\rho_0(s_0)$, and the reward function by $R(s_t, a_t)$.

For detailed DDPO policy, we employ a ResNet101 pre-trained on ImageNet-1K as our feature extraction model, and utilize KNN and KL divergence to compute both the diversity among generated images and their similarity to real images. Let X denote the set of generated images, Y represent the real images, where $x_i \in X$, $y_j \in Y$, and M is our feature extraction model.

The KNN reward is calculated as follows: 1) First, we extract features from X using model M: $F_x = M(X)$. 2) For each feature vector $f_x^i \in F_x$, we compute its K-nearest neighbors among all feature vectors $f_x^j \in F_x$. The KNN reward for x_i is the average of these K nearest neighbor distances, denoted as $KNN(f_x^i, F_x)$. In our implementation, we set K to 50.

The KL reward is calculated as follows: 1) We extract features from both X and Y using model M: $F_x = M(X)$ and $F_y = M(Y)$. 2) For each feature vector $f_x^i \in F_x$ and $f_y^j \in F_y$, we compute $KL(f_x^i, f_y^i)$ for each i, and use $-KL(f_x^i, f_y^i)$ as the KL reward for x_i.

In summary, the reward for a generated image x_i is computed as:

$$r_x^i \triangleq KNN(f_x^i, F_x) - wKL(f_x^i, f_y^i) \tag{14}$$

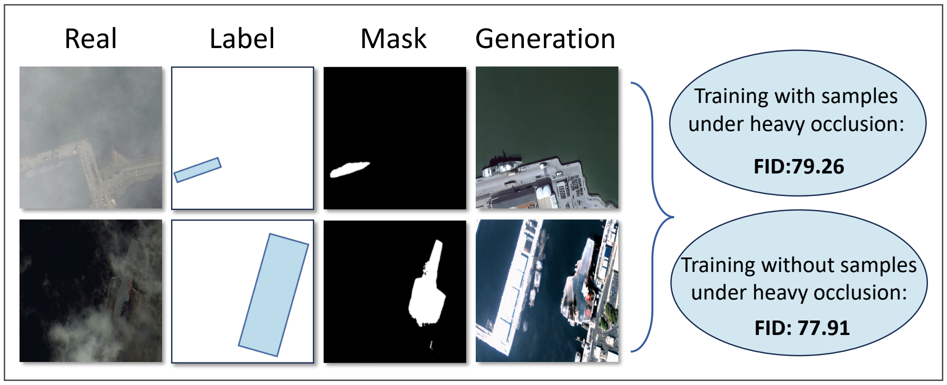

Real
Label
Mask
Generation
Training with samples under heavy occlusion:
FID:79.26
Training without samples under heavy occlusion:
FID: 77.91

Figure 7: The results of OF-Diff in handling objects under heavy occlusion. The results indicate that while severe occlusion does indeed cause a certain degree of degradation in the quality of the target mask extracted by ESGM, it has little impact on the generated quality FID.

A.3 Analysis and Discussion

According to the current experimental results, adding the DDPO strategy does not simultaneously outperform previous results on all metrics. Using reinforcement learning strategies can indeed improve the performance of downstream tasks, but it does not necessarily improve the quality of image generation simultaneously. In other words, reinforcement learning strategies can also purposefully improve the quality of image generation, but this may come at the cost of not improving the performance of downstream tasks.

The proposed OF-Diff injects object shape masks extracted from the image layout as controllable conditions into the diffusion model, which effectively enhances object fidelity and improves the generation of small objects. However, this also makes the model dependent on the quality of the extracted shape masks. We analyze the impact of a distorted mask on the model's generated results. Specifically, we selected cases such as objects under heavy occlusion to examine the model's generation performance. Based on the analysis results in Figure7, we found that even under severe occlusion conditions, ESGM still demonstrates strong object mask extraction and generation capabilities. However, when the generated mask shape exhibits certain anomalies, it does produce objects matching that distorted shape. Nevertheless, this does not affect the overall FID and trainability of the generated images. Although the shapes we currently extract may exhibit edge anomalies in the object mask due to occlusion and other issues, complete errors are extremely rare.

A.4 Quantitative Results on HRSC2016 Dataset

Table 5 reports the comparative results on HRSC2016, where our method consistently achieves strong performance. Although it ranks second on CMMD, CAS, and YOLOScore—which mainly reflect aesthetic quality or local recognizability—it attains the best results on FID and KID, which measure distribution fidelity, as well as on the most crucial downstream metric, mAP50, outperforming the second-best method by +1.5%. This indicates that our generated data preserves the real remote-sensing distribution more faithfully and thus provides more effective support for downstream tasks. A more detailed analysis is provided in Appendix A.7.

Table 5: Fidelity and Downstream Performance on HRSC2016

Method	**HRSC216 Dataset**					
	FID↓	KID↓	CMMD↓	CAS↑	YOLOScore↑	mAP$_{50}$ ↑
LayoutDiff	120.68	0.152	1.763	24.51	2.51	56.97
GLIGEN	92.92	0.037	0.634	35.41	5.03	39.72
AeroGen	97.44	0.055	**0.51**	39.62	16.4	47.68
CC-Diff	84.55	0.035	0.681	**45.27**	**32.42**	62.57
Ours	**77.91**	**0.026**	0.573	42.19	30.97	**64.1**

Table 6: The impact of real and generated images at different ratios on mAP for downstream tasks.

Data Composition	mAP (%)
100% Generated	45.67 (-7.17)
50% Real + 50% Generated	50.74 (-2.10)
100% Real	52.84
100% Real + 50% Generated	53.92 (+1.08)
100% Real + 100% Generated	54.38 (+1.54)
100% Real + 200% Generated	54.74 (+1.90)
100% Real + 300% Generated	54.82 (+1.98)

Table 7: The data on the computational cost of training OF-Diff compared to the key baselines.

Models	Train GPU Mean Memory (MB)	Train GPU Hours	Inference Mean Time/Sample (s)
LayoutDiff	29232	41.33	<1s
GLIGEN	14186	57.76	5.18
AeroGen	27634	49.52	1.85
CC-Diff	13668	38.01	3.96
OF-Diff	27340	44.27	3.42

A.5 THE mAP EVOLUTION GIVEN DIFFERENT AMOUNTS OF SYNTHETIC AND REAL DATA

We conduct multiple experiments on trainability using different quantities of real and generated data. The results are shown in the Table 6. Experimental results indicate that using only 100% synthetic data struggles to achieve downstream task performance comparable to real data. However, this also demonstrates that even without a single real image, relying solely on synthetic images can enable object detection algorithms to achieve a mAP of 45.67%. Furthermore, training with a larger volume of generated images can effectively enhance the model's object detection capabilities. However, when the amount of generated data reaches three times that of real data (based on the generation setting described in the paper), downstream performance shows little further improvement.

A.6 THE COMPUTATIONAL COST

We provide the data on the computational cost of training OF-Diff compared to the key baselines in the Table .7. Experimental results indicate that although OF-Diff is not the most optimal in terms of training costs (GPU memory and GPU hours) and inference time, it remains nearly the second-best among these methods and does not incur high computational costs.

A.7 AESTHETIC–DOWNSTREAM PERFORMANCE CONFLICT

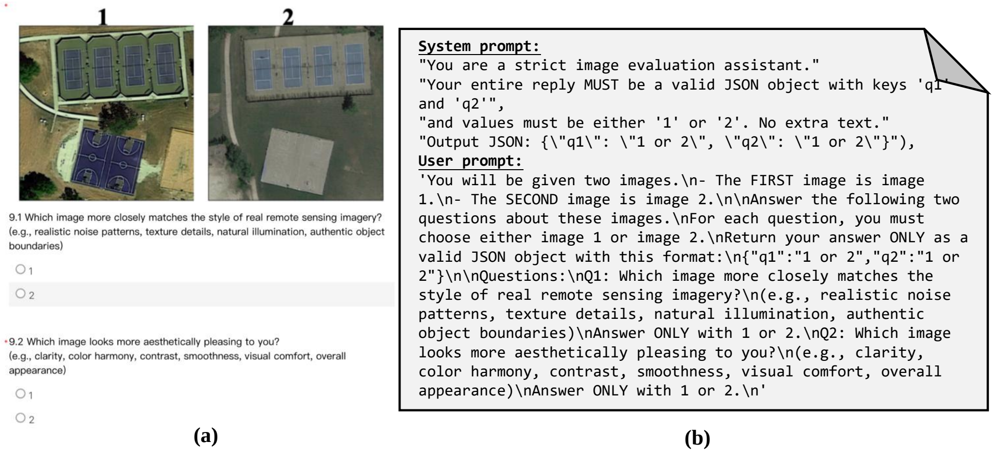
To further reveal the potential conflict between aesthetic quality and downstream task performance, we conduct a three-part analysis consisting of questionnaire evaluation, downstream performance comparison, and feature-level visualization.

(1) **Human/GPT questionnaire study.** As shown in Fig. 8, we design two targeted questions:

Q1. Which image more closely matches the style of real remote-sensing imagery? (e.g., realistic noise patterns, texture details, natural illumination, authentic object boundaries)

Q2. Which image looks more aesthetically pleasing? (e.g., clarity, color harmony, contrast, smoothness, visual comfort, overall appearance)

We invite 8 PhD researchers and 8 remote-sensing experts to participate, and additionally perform 3 rounds of GPT-5 evaluation. For each class in DIOR, we randomly sample one pair of images generated with vs. without captions (from the same ground truth), resulting in 20 image pairs. Each pair is randomly shuffled to avoid positional bias. The results are shown in Table 8, each



9.1 Which image more closely matches the style of real remote sensing imagery?
(e.g., realistic noise patterns, texture details, natural illumination, authentic object boundaries)

○ 1

○ 2

*9.2 Which image looks more aesthetically pleasing to you?
(e.g., clarity, color harmony, contrast, smoothness, visual comfort, overall appearance)

○ 1

○ 2

(a)

```
System prompt:
"You are a strict image evaluation assistant."
"Your entire reply MUST be a valid JSON object with keys 'q1'
and 'q2'",
"and values must be either '1' or '2'. No extra text."
"Output JSON: {\"q1\": \"1 or 2\", \"q2\": \"1 or 2\"}"),
User prompt:
'You will be given two images.\n- The FIRST image is image
1.\n- The SECOND image is image 2.\n\nAnswer the following two
questions about these images.\nFor each question, you must
choose either image 1 or image 2.\nReturn your answer ONLY as a
valid JSON object with this format:\n{"q1":"1 or 2","q2":"1 or
2"}\n\nQuestions:\nQ1: Which image more closely matches the
style of real remote sensing imagery?\n(e.g., realistic noise
patterns, texture details, natural illumination, authentic
object boundaries)\nAnswer ONLY with 1 or 2.\nQ2: Which image
looks more aesthetically pleasing to you?\n(e.g., clarity,
color harmony, contrast, smoothness, visual comfort, overall
appearance)\nAnswer ONLY with 1 or 2.\n'
```

(b)

Figure 8: Aesthetic evaluation questionnaire design for generated images: (a) human experts, (b) GPT-5.

value represents the average frequency with which the corresponding option was selected across all questionnaires.

Table 8: Single-choice results from human experts and GPT-5 (averaged over multiple annotators or repeated evaluations).

Option	Human experts		GPT-5	
	Q1	Q2	Q1	Q2
w./ caption	6.57	**11.21**	2.33	**15.33**
w./o. caption	**13.43**	8.79	**17.67**	4.67

Both human experts and GPT consistently prefer the caption-conditioned images in terms of aesthetics, but find them less similar to real remote-sensing imagery. In contrast, images generated without captions appear less visually appealing but better preserve the real-world texture and structural characteristics needed for downstream tasks.

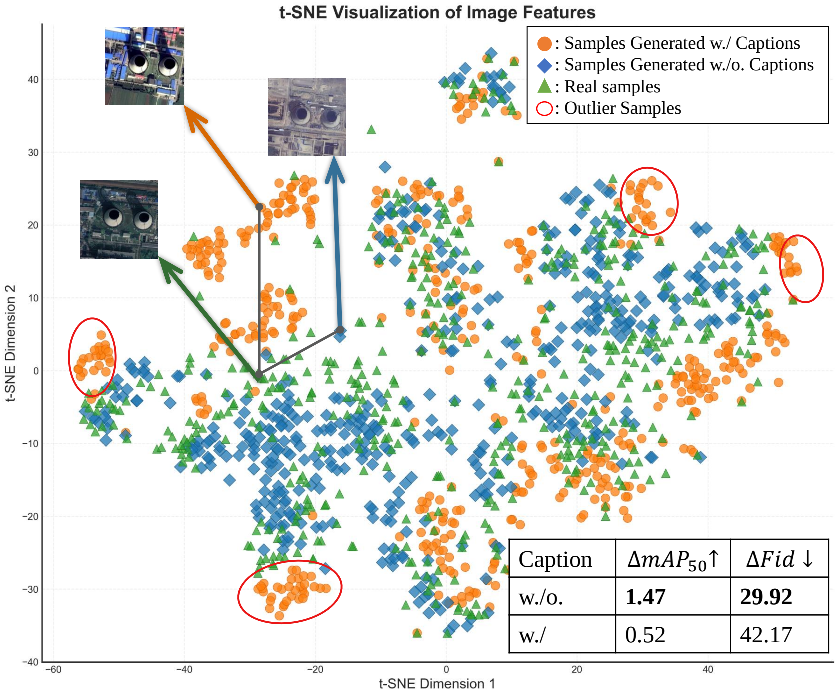
(2) Downstream performance comparison. On the DIOR dataset, as shown in the table at the bottom-right of Fig. 9, adding captions reduces the downstream improvement $\Delta\mathrm{mAP}_{50}$ by 1.15 and also leads to a significantly higher FID. Combined with finding (1), this reveals that caption-guided generation tends to over-beautify images—masking the natural imperfections of remote-sensing imagery—and consequently harms downstream performance.

(3) Feature-level visualization. We also visualize features using t-SNE in Fig. 9. We observe that adding captions produces more outliers, whereas samples generated without captions align more closely with the GT distribution, indicating higher fidelity.

Taken together, these findings suggest that models should remain faithful to the inherent quirks and imperfections of the original remote-sensing data, rather than generating overly "idealized" or aesthetically enhanced imagery. Incorporating captions risks amplifying the latter behavior.

Additional examples are provided in Fig. 10.

A.8 More Qualitative and Quantitative Results



Caption	$\Delta mAP_{50}\uparrow$	$\Delta Fid\downarrow$
w./o.	**1.47**	**29.92**
w./	0.52	42.17

Figure 9: t-SNE feature analysis of generated samples w./ and w./o. captions. Incorporating captions produces a large number of outlier samples, lowers fidelity (higher FID), and further degrades downstream performance (lower ΔmAP_{50}).

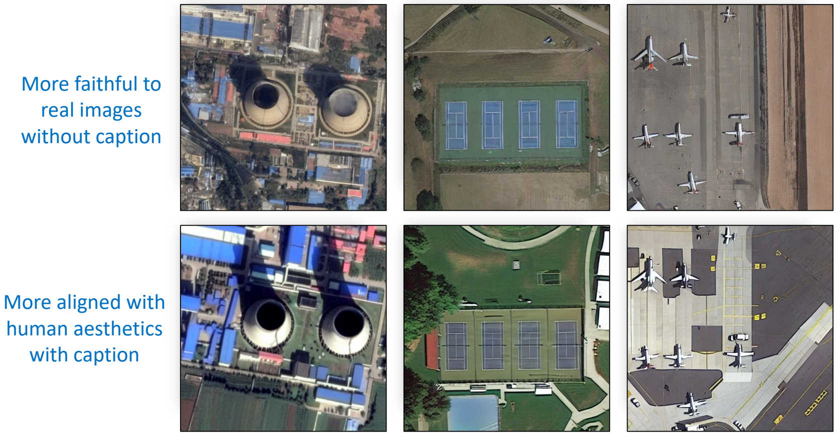

More faithful to real images without caption
More aligned with human aesthetics with caption

Figure 10: The influence of caption on the generation of images in terms of being more realistic and more aesthetically pleasing.

Table 9: Trainability (↑) comparison on DIOR and DOTA. 'Baseline' denotes accuracy with the unaugmented dataset.

Method	**DIOR Dataset**			**DOTA Dataset**		
	mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}
Baseline	30.51	52.84	32.10	38.09	66.31	38.44
LayoutDiff	29.81	52.14	30.36	38.91	66.75	40.37
GLIGEN	28.48	51.27	29.21	38.84	66.10	40.24
AeroGen	31.53	53.37	33.60	38.45	67.09	39.07
CC-Diff	31.82	53.48	33.97	38.51	66.52	39.02
Ours	**32.71**	**54.44**	**34.05**	**40.03**	**67.89**	**42.20**

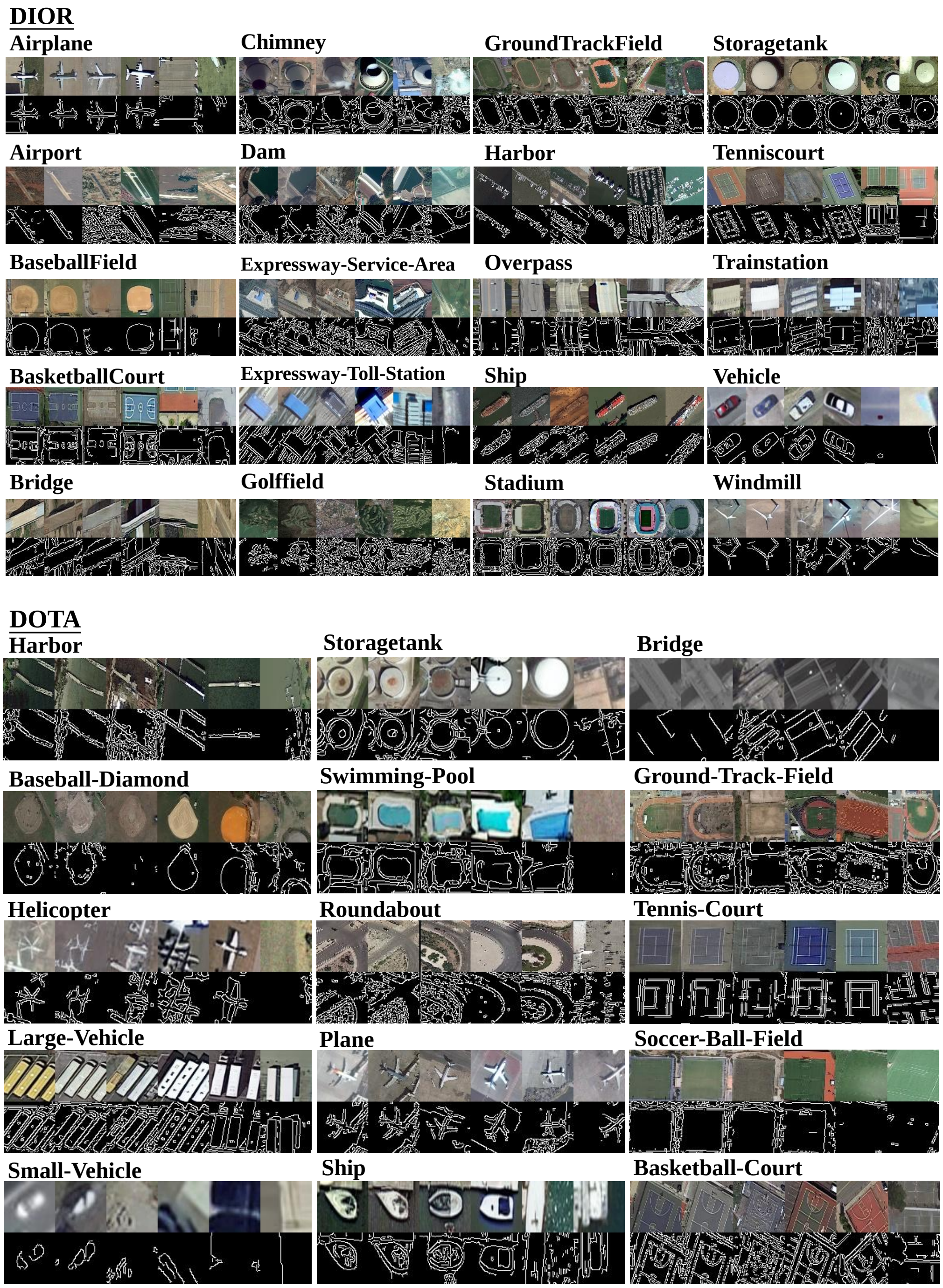


Figure 11: Comparison of generated instance patches and their Canny edge maps for the same bbox on the DIOR and DOTA dataset. Each image set is ordered as follows: Ground Truth, OF-Diff, AeroGen, CC-Diff, GLIGEN, and LayoutDiff.

Method	DIOR Dataset									
	Expressway Service-area	Expressway toll-station	Airplane	Airport	Baseball field	Basketball court	Bridge	Chimney	Dam	Golf Field
LayoutDiff	53.1	44.8	62.8	29.4	63.2	79.6	25.9	72.6	22.4	69.3
GLIGEN	52.7	44.8	62.6	26.7	63.0	79.6	25.2	72.6	19.5	67.4
AeroGen	**58.1**	**45.2**	63.1	32.7	**63.4**	**81.0**	29.5	72.6	21.1	69.1
CC-Diff	53.5	45.1	62.9	**38.4**	63.3	79.9	29.3	**72.7**	**27.6**	70.5
Ours	58.0	44.9	**71.3**	37.0	63.2	80.2	**30.1**	72.5	24.9	**75.4**

Method	DIOR Dataset									
	Ground Track-field	Harbor	Overpass	Ship	Stadium	Storage Tank	Tennis Court	Trainstation	Vehicle	Windmill
LayoutDiff	**71.2**	32.8	43.9	62.9	59.0	**52.5**	72.4	52.1	26.9	46.0
GLIGEN	70.1	30.3	45.8	62.8	56.8	52.0	**72.5**	49.0	26.9	45.3
AeroGen	71.0	42.7	**50.7**	62.9	56.6	44.5	**72.5**	52.6	**31.4**	**46.7**
CC-Diff	64.6	43.1	49.0	63.0	**61.7**	44.7	72.4	**54.4**	27.0	46.5
Ours	66.3	**43.9**	49.4	**70.5**	52.7	44.4	72.4	54.1	31.0	**46.7**

Table 10: Detailed downstream trainability results (measured by average precision) on the DIOR dataset.

Method	DOTA Dataset							
	Plane	Baseball-diamond	Bridge	Ground Track-field	Small-vehicle	Large-vehicle	Ship	Tennis-court
LayoutDiff	80.4	74.2	**48.8**	59.9	62.9	72.7	82.5	89.6
GLIGEN	87.0	72.9	47.3	56.4	63.7	73.1	82.7	90.1
AeroGen	86.1	**77.3**	48.6	58.6	64.5	**78.1**	82.5	83.3
CC-Diff	**87.2**	73.4	47.1	57.8	64.3	73.9	82.6	89.2
Ours	85.2	75.3	46.5	**60.4**	**68.3**	77.2	**84.4**	**90.4**

Method	DOTA Dataset						
	Basketball-court	Storage-tank	Soccer-ball-field	Roundabout	Harbor	Swimming-pool	Helicopter
LayoutDiff	78.9	76.3	**46.3**	**47.6**	60.9	62.1	**57.9**
GLIGEN	79.6	76.5	42.2	43.1	60.7	62.3	53.8
AeroGen	77.3	79.9	44.8	46.6	59.4	62.6	56.4
CC-Diff	79.0	**82.7**	42.7	43.1	58.9	62.7	52.9
Ours	**83.3**	77.1	42.1	44.7	**62.1**	**67.9**	53.3

Table 11: Detailed downstream trainability results (measured by average precision) on the DOTA dataset.

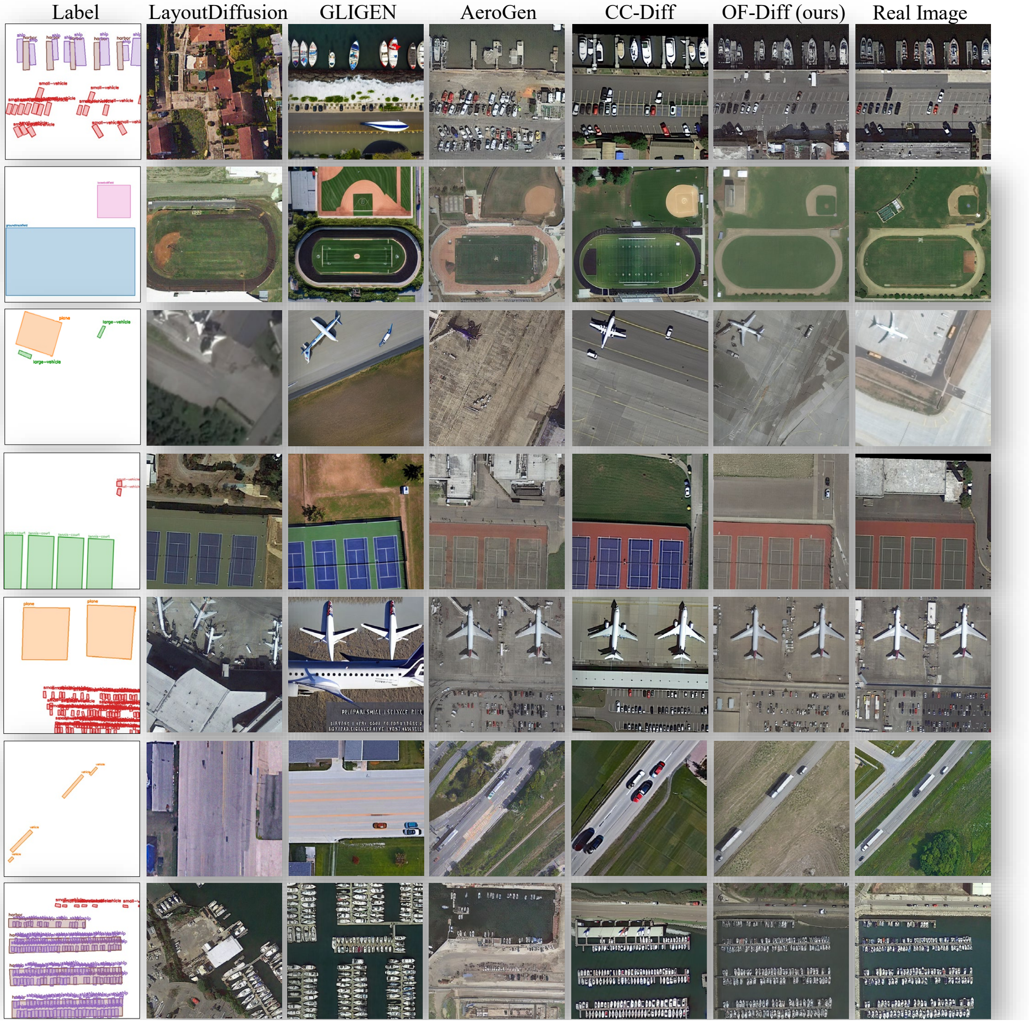


Figure 12: Additional qualitative results on DIOR and DOTA. The results demonstrate that OF-Diff has certain superiority and accuracy in generating small objects, and it also has an advantage in generating the shapes of objects. For instance, the aircraft target in the third row is generated more accurately by OF-Diff, with a more realistic structure. The small vehicles in the fourth and fifth rows and the large vehicle in the sixth row are also more accurately generated. Additionally, the small ship in the seventh row is generated with greater accuracy.