remote sensing

MDPI

Article

Text2AIRS: Fine-Grained Airplane Image Generation in Remote Sensing from Nature Language

Yunuo Yang [1], Youwei Cheng [1], Jinlong Hu [1], Yan Xia [2] and Yu Zang [1,*]

[1] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China; yangyunuo555@stu.xmu.edu.cn (Y.Y.); youwei@stu.xmu.edu.cn (Y.C.)
[2] Computer Vision Group, TUM School of Computation, Information and Technology, Technical University of Munich (TUM), 80333 Munich, Germany
* Correspondence: zangyu7@126.com

Highlights

What are the main findings?

- This study proposes Text2AIRS, the first text-to-image generation method to explicitly incorporate ground sample distance (GSD) as a dual-level (prompt and feature) constraint, significantly enhancing the scale consistency and realism of synthesized fine-grained aircraft in remote sensing images.
- Extensive experiments demonstrate that Text2AIRS outperforms state-of-the-art methods on the Fair1M benchmark in terms of both perceptual quality and downstream utility for object detection, leading to notable improvements in detection accuracy.

What are the implications of the main findings?

- The work offers a promising and cost-effective solution to address data scarcity and class imbalance in fine-grained remote sensing analysis by generating high-fidelity, semantically-aligned synthetic data that adheres to physical imaging constraints.
- By transforming GSD from passive metadata into an active generative prior, the method establishes a new paradigm for physically-grounded remote sensing image synthesis, with potential to benefit various downstream tasks such as data augmentation and model robustness enhancement.

Abstract

Airplanes are the most popular investigation objects as a dynamic and critical component in remote sensing images. Accurately identifying and monitoring airplane behaviors is crucial for effective air traffic management. However, existing methods for interpreting fine-grained airplanes in remote sensing data depend heavily on large annotated datasets, which are both time-consuming and prone to errors due to the detailed nature of labeling individual points. In this paper, we introduce Text2AIRS, a novel method that generates fine-grained and realistic **A**irplane **I**mages in **R**emote **S**ensing from textual descriptions. Text2AIRS significantly simplifies the process of generating diverse aircraft types, requiring limited texts and allowing for extensive variability in the generated images. Specifically, Text2AIRS is the first to incorporate ground sample distance into the text-to-image stable diffusion model, both at the data and feature levels. Extensive experiments demonstrate our Text2AIRS surpasses the state-of-the-art by a large margin on the Fair1M benchmark dataset. Furthermore, utilizing the fine-grained airplane images generated by Text2AIRS, the existing SOTA object detector achieves 6.12% performance improvement, showing the practical impact of our approach.

Check for updates

Academic Editor: Xinghua Li

Received: 20 October 2025
Revised: 22 December 2025
Accepted: 27 December 2025
Published: 5 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: image generation; fine-grained; object detection

1. Introduction

Airplanes, as critical objects in remote sensing images (RSIs), play an essential role in many realistic applications, such as aircraft identification [1] and object tracking [2]. In recent years, many learning-based methods for interpreting airplanes in RSIs have shown impressive results, thanks to extensive training on large, annotated datasets. However, annotating airplanes, whether by labeling individual points or drawing 2D bounding boxes, is time-consuming and prone to errors. Thus, generating detailed airplane images in RSIs offers a promising solution, especially when real data is scarce. Moreover, creating airplane images in RSIs with diverse types and backgrounds helps models adapt better to various environmental conditions and geographical features [3].

Unlike natural-image settings, remote sensing imagery (RSI) is acquired from elevated, often orbital, viewpoints, typically hundreds to about one thousand kilometers above the Earth. This leads to large-format images containing many small, multi-oriented instances embedded in complex environments [4]. This setting poses several challenges: (i) limited expert-labeled data constrains detection accuracy; (ii) targets occupy few pixels relative to expansive scenes; and (iii) substantial scale variation exists across categories. High-altitude, nadir or off-nadir acquisition further induces pronounced appearance changes due to viewing geometry and illumination. A defining property of RSI is its approximately uniform spatial resolution within an image, commonly specified by the ground sample distance (GSD) [5,6]. GSD quantifies the on-ground spacing between pixel centers and reflects a sampling-limited component of image resolution. These factors collectively complicate high-fidelity generation of remote sensing images, where respecting GSD-driven scale consistency is essential.

To access the semantic content of remote sensing imagery (RSI), prior work has focused primarily on text annotation and image captioning. Early efforts introduced paired text descriptions to summarize scene semantics, as exemplified by the UCM-captions and Sydney-captions datasets, each providing multiple sentences per image to describe salient content [7]. Lu et al. established RSICD as a benchmark for RSI captioning and provided a comprehensive review of encoder–decoder approaches [8]. More recently, Cheng et al. proposed NWPU-Captions, a more challenging dataset designed to expand category diversity and description richness across 45 classes [9]. These advances have substantially advanced RSI captioning, yet comparatively less attention has been devoted to the inverse problem of text-conditioned RSI generation. In contrast to captioning, text-to-image generation aims to synthesize realistic imagery from natural-language prompts. In the RSI domain, Txt2Img-MHN employs Modern Hopfield Networks to produce photorealistic and semantically aligned images using RSICD captions [10,11]. ResBaGAN explores a GAN-based pipeline tailored for domain-specific synthesis, such as forest mapping [12]. RSDiff adopts diffusion modeling to generate RSI from text under diverse scenarios using RSICD descriptions [13]. Despite these promising steps, existing methods generally do not incorporate remote-sensing-specific constraints—most notably the ground sample distance (GSD)—nor do they explicitly target fine-grained aircraft synthesis, which limits their utility for downstream recognition tasks.

Despite recent progress in text-conditioned RSI synthesis, fine-grained aircraft generation remains underexplored. Fine-grained datasets provide higher-resolution evidence of subtle inter-class and intra-class differences, enabling nuanced discrimination among closely related subcategories and thus improving classification and detection precision. In

remote sensing, however, the ecosystem is dominated by coarse-grained detection datasets such as DOTA and DIOR [14,15]. Truly fine-grained resources are scarce, with notable exceptions including the 2020 Gaofen Challenge, Fair1M, and FGSD [16–18]. Even so, their detection performance often lags behind coarse-grained settings due to severe class imbalance, uneven instance counts, and substantial annotation costs stemming from very large images with sparse targets. For example, Fair1M exhibits pronounced long-tailed distributions across aircraft subcategories, which degrades detector performance. The need for expert interpretation further increases labeling expense, limiting accessibility and the broader applicability of fine-grained RSI analysis.

Motivated by the challenges of fine-grained analysis in remote sensing imagery (RSI), we propose Text2AIRS, a diffusion-based framework for generating realistic, fine-grained aircraft images. The core idea is to explicitly incorporate ground sample distance (GSD), a critical yet underutilized constraint in RSI synthesis. By leveraging GSD, our approach addresses two limitations of existing methods: inconsistent object scaling within a single image and the scarcity of fine-grained data. In RSI, spatial resolution is governed by GSD, implying that objects of the same category within one image should exhibit consistent pixel-scale properties. Conventional generative models rarely enforce this constraint, often producing size distortions and unrealistic content (see Figure 1). Therefore, we introduce a dual-level GSD integration strategy. At the prompt level, we append GSD information to captions to provide resolution-aware textual conditioning. At the feature level, we encode GSD as a deep feature and fuse it into the denoising U-Net via attention-guided modulation. This design calibrates the generative dynamics to the target spatial resolution and improves intra-image scale consistency. Text2AIRS transforms GSD from passive metadata into an active structural prior in the generation pipeline, thereby enhancing the precision and realism of synthesized, fine-grained aircraft. We conduct extensive evaluations against real images to assess visual fidelity and practical utility. Beyond perceptual quality, we demonstrate downstream benefits on object detection. Using Fair1M, which contains ten aircraft subcategories with severe class imbalance, we perform targeted augmentation by inserting GSD-consistent synthetic instances into the training set while keeping validation and test sets unchanged. For clarity and reproducibility of the augmentation protocol, the detection experiments in this work use horizontal bounding boxes. HBBs provide unambiguous, axis-aligned coordinates when inserting instances, whereas oriented bounding boxes would additionally require resolving insertion positions together with rotation angles and IOU-consistent placement, making the pipeline substantially more complex. Extending Text2AIRS to OBB-based augmentation and evaluation—so that orientation-sensitive cues are explicitly modeled—is an important direction we leave for future work. This HBB-based design leads to measurable improvements in detection accuracy, indicating that GSD-aware synthesis can effectively mitigate data scarcity and strengthen model robustness.

In summary, our main contributionsare:

- We introduce Text2AIRS, a novel method that integrates ground sample distance (GSD) as a guiding constraint for generating remote sensing images. To the best of our knowledge, this is the first approach to incorporate GSD into the remote sensing image generation process, significantly enhancing the realism and consistency of the generated images.
- We extract fine-grained objects from remote sensing images and annotate them with detailed captions containing fine-grained information. This marks the first time that fine-grained instances are explicitly considered and utilized in remote sensing image generation, addressing a critical gap in the field.

- Extensive experimental results on both image generation and object detection tasks demonstrate the authenticity and practicality of the generated images. Our approach not only produces visually realistic outputs but also improves the performance of downstream object detection models, showcasing its potential for real-world applications.

The rest of this paper is organized as follows: Section 2 presents the related studies in object detection for remote sensing images. In Section 3, we introduce the proposed method in detail. Moreover, extensive experiments and discussions are presented in Section 4. Finally, Section 5 provides the conclusions.

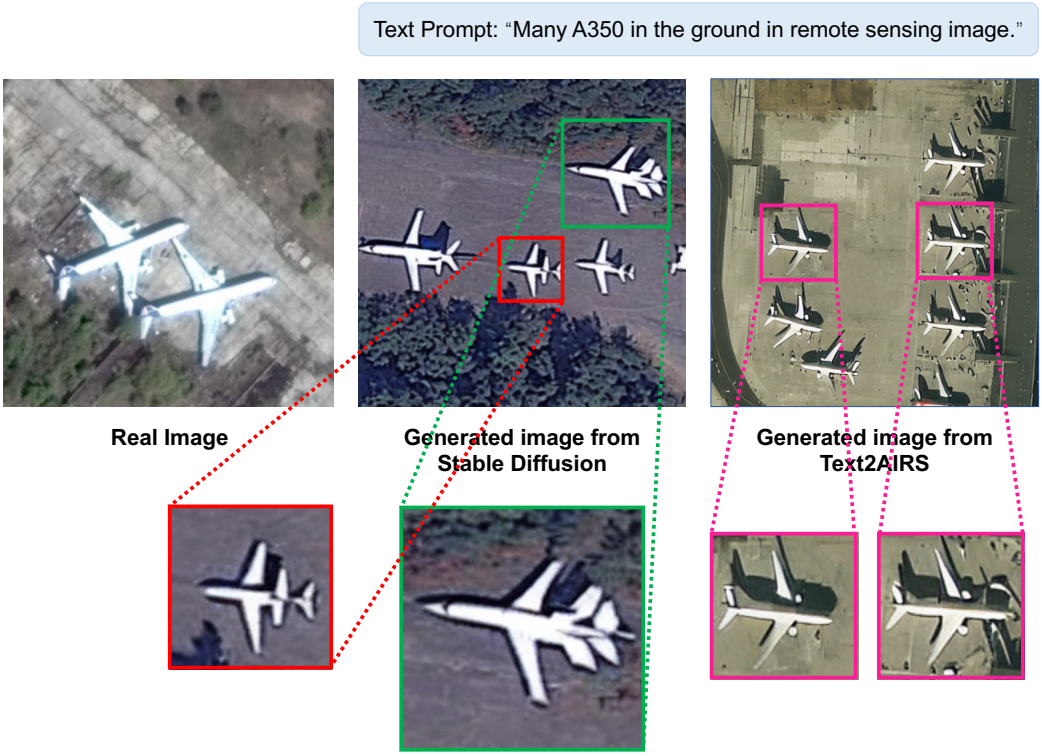


Figure 1. Remote sensing images. A genuine remote sensing image ideally portrays objects of the same class with consistent sizes, adhering to the limitations of the ground sample distance. However, inconsistencies in the size of the same object may arise in generated images. To address this issue, we introduced the ground sample distance as a constraint in the generation process, aiming to enhance the realism of generated remote sensing images.

2. Related Work

2.1. Text-to-Image Generation in Natural Images

Generating realistic images from natural language descriptions is an important and challenging task. Earlier research on text-to-image generation mainly focus on GAN-based methods [19–21]. Text-conditional GAN [19] is the first innovative work, where the generator is designed to generate realistic images from extracted text features and cheat the discriminator, while the discriminator is designed to identify real from fake images. It is only able to generate images with a size of 64 pixels by 64 pixels using text-conditional GAN. For larger realistic images, ref. [20] inputs key point locations as supporting information along with the generator. On the basis of this idea, researchers propose a generative adversarial what-where network capable of synthesizing images up to 128 pixels by 128 pixels. StackGAN [21] is another novel work in which the authors propose to synthesize images using stacked generators. In the first stage, despite a low

quality 64 pixels by 64 pixels image is generated, the second stage transforms it into a larger 256 pixels by 256 pixels image, adding refinement to the synthesis.

Aside from the GAN-based methods, models based on the transformer architecture have attracted a lot of attention in the field of image generation. These models utilise a self-attentive mechanism to capture correlations between different locations in an input sequence and can be easily extended to handle inputs of different lengths and scales. The CLIP model [22], proposed by OpenAI, integrates both images and text using contrastive learning techniques, enabling it to establish semantic understanding between textual descriptions and images. DALL·E [23], another OpenAI creation, creates images based on textual descriptions, understanding abstract concepts and scenes, and generating images that closely align with those descriptions. OpenAI has also developed the GPT series of language models, including models like GPT-3 [24] and GPT-4 [25], which are Transformer-based language models used primarily for the generation of text. Besides, diffusion models [26], like Imagen [27], Stable Diffusion [28], and Dreambooth [29] have shown promising results in various image generation tasks.

2.2. Text-to-Image Generation in Remote Sensing Images

Until now, most research has focused on natural image generation. However, remote sensing research has been immature due to the complexity of spatial distributions of different features and the large modal gap between textual descriptions and RSIs. It is still a challenge to generate realistic remote sensing images from text descriptions due to the high requirements for accuracy and credibility in remote sensing tasks.

The goal of Txt2Img-MHN [10] is to learn the most representative prototypes from text-image embeddings, implementing a coarse-to-fine learning strategy. These learned prototypes can then be utilized to represent more complex semantics in the task of generating images from text, thereby enhancing the capability of remote sensing image generation. ResBaGAN [12] is a method based on Generative Adversarial Networks designed to address the challenges of limited labeled data and class imbalances in remote sensing datasets. It constructs an advanced enhancement framework with properties such as autoencoder initialization and class balancing to better generate data. RSDiff [13] comprises two interconnected diffusion models: a Low-Resolution Generation Diffusion Model (LR-GDM) for generating low-resolution images from text, and a conditionally generated Super-Resolution Diffusion Model (SRDM). LR-GDM effectively synthesizes low-resolution images by computing the correlation of text embeddings and image embeddings in a shared latent space, capturing the essential content and layout of the desired scenes. Subsequently, SRDM takes the generated low-resolution image and its corresponding text prompts as input, efficiently producing high-resolution images.

2.3. Ground Sample Distance

Ground sample distance (GSD) refers to the size of the smallest unit that can be distinguished in detail in remote sensing images, which is used to represent the detail of the ground instance resolution in the image. The ground sample distance of an image is generally higher when the area of the image represented by pixels is smaller. Currently, the GSD of remote sensing image is a very critical factor to consider when doing analyses of vegetation [30,31]. A vegetation analysis can be conducted by analyzing different ground sample distance images, which are of great significance to regional ecological environments and cannot be replaced by ordinary image information. Additionally, in areas such as traffic planning and vehicle navigation, remote sensing imagery can be used to extract road information [32,33]. It is appropriate to use images of different spatial resolutions for different scenarios, for example, an image with a resolution of 2 m can be used to monitor

land-cover and land-use change; an image with a resolution of 0.8 m can be used to monitor road networks in urban areas; and an image with a resolution of 0.5 m and 0.3 m can be used to investigate details such as road boundaries, turning arrows, and zebra crossings. Previous studies [5,6] show the significance of GSD in remote sensing images.

2.4. Data Augmentation

The principle of data augmentation is to increase variation as much as possible without changing the original semantic information of an image [3]. All data augmentation methods increase variation without changing the original semantic information, despite varying characteristics. Several scholars have conducted research on deep learning-based data augmentation in recent years. A data augmentation technique based on an image style transfer technique was proposed by Mikolajczyk et al. [34]. They reviewed and discussed various data augmentation methods, provided illustrative examples, and presented their own data augmentation method. In the data-based data augmentation method [35], the original data is used to increase the amount of data in different ways as it is necessary for the current deep learning model to require a large amount of data. Through geometric transformations [36], pixel values are mapped to new locations in an image. In geometric transformations, images are rotated, scaled, flipped, shifted, and cropped to generate new data. Despite the basic shape of the class remaining the same in general imagery, its position and orientation have changed. The sharpness transformation changes the sharpness of the image by sharpening and blurring it [37]. The sharpened RSI highlights more remote sensing information and is widely used in RSIs for small object detection. Overall, data augmentation enhances the generalizability, robustness, and performance of the models [38]. In scenarios with limited data, it diversifies datasets by providing a wider range of samples, greatly improving model performance, particularly for tasks with fewer training examples.

2.5. Object Detection

Object detection for remote sensing images plays an essential role in computer vision [6] and has significant importance in RSI analysis [14,17,39,40]. With the emergence of convolutional neural networks, object detection has been transformed from traditional to deep learning-based. Based on the deep learning method, there are two types of object detection: one-stage detection [41] and two-stage [42] detection. The main difference between the two methods is that the one-stage method uses convolutional neural networks directly to extract features to predict classification and localization, while the two-stage method divides the detection task into two phases: localization and recognition. The performance of object detection methods continues to improve and is increasingly applied to remote sensing and aerial images. Ding et al. [43] introduced a RoI Transformer to extracting rotationally invariant features. Yang et al. [44] developed the SCRDet to addresses boundary ambiguity in oriented bounding box regression by incorporating feature refinement and an IoU-consistent loss to improve localization of densely packed, arbitrarily oriented objects. Gliding Vertex [45] and RSDet [46] are both limit the offset on the corresponding side of horizontal bounding boxes and order the sequence of quadrilateral corners, RSDet [46] constrains offsets on the sides of horizontal proposals and imposes a consistent ordering of quadrilateral corners, thereby stabilizing rotation-aware regression and improving robustness to boundary discontinuities. ReDet [47] propose a novel module to extract rotation-invariant features from equivariant ones. Through assisted super resolution learning, SuperYOLO [48] integrates assisted super-resolution learning with a YOLO-style one-stage detector, fusing multi-modal information to enhance fine-detail recovery and perform high-resolution detection across multi-scale remote sensing targets. InsDist [49]

propose an instance-aware distillation method to derive efficient remote sensing object detectors. While the above mentioned methods perform very well in the remote sensing image target detection task, but not very ideal for the limited data samples.

3. Method

3.1. Overview

In this work, we present Text2AIRS, a diffusion-based framework tailored for fine-grained remote sensing image generation. As depicted in Figure 2, the architecture adopts a dual-branch conditioning scheme that integrates a GSD-aware visual pathway with a text semantic pathway, and couples them to a latent diffusion core in the style of Stable Diffusion. The goal is to synthesize high-fidelity aircraft imagery that is both semantically faithful to the prompt and consistent with the spatial-resolution constraints implied by the ground sample distance (GSD). A high-resolution RSI is first processed by a pretrained convolutional backbone to obtain multi-level feature maps that capture mid-level texture, edge density, and object-extent statistics correlated with GSD. A channel-attention module then reweights the feature channels to emphasize resolution-sensitive cues while suppressing confounders from cluttered backgrounds. The output of this branch is a compact GSD-conditioned representation that serves as an explicit structural prior for subsequent generation. The input description is tokenized and encoded by a text encoder to derive contextualized token embeddings. These embeddings capture object category, subtype, pose, and environmental attributes, and explicitly include GSD phrases to provide resolution-aware semantic conditioning. The resulting text features are used to guide the generative process via cross-attention. Following the Stable Diffusion paradigm, images are mapped to a latent space where iterative denoising is performed. At each denoising step, a U-Net operates on the latent representation and is conditioned jointly by the two branches: Cross-attention layers attend to the text embeddings to enforce semantic alignment, ensuring that fine-grained attributes described in the prompt are expressed in the synthesized content. Feature-level modulation injects the GSD-aware representation into the U-Net blocks via attention-guided gating and feature fusion. This transforms GSD from passive metadata into an active constraint that regulates intra-image scale, object extent in pixels, and frequency content across the denoising trajectory. The interaction between these conditioning signals is designed to be complementary: text guidance anchors the target semantics and fine-grained aircraft attributes, whereas the GSD prior calibrates spatial scale and suppresses implausible size distortions that commonly arise in generic text-to-image models. The denoiser, thus, converges toward latent codes that are both semantically coherent and resolution-consistent. After the diffusion process reaches a clean latent, an image decoder maps the latent back to the pixel domain. The resulting image aims to preserve high spatial detail, adhere to the GSD-implied object scale, and reflect the attributes specified by the textual description.

Overall, the dual-branch design provides disentangled yet synergistic conditioning: the text pathway contributes semantic specificity, and the visual/GSD pathway contributes resolution-aware structural priors. Coupled with the latent diffusion core, this design yields images that achieve improved intra-image scale consistency, enhanced fine-grained fidelity, and better suitability for downstream tasks such as data augmentation and aircraft detection.

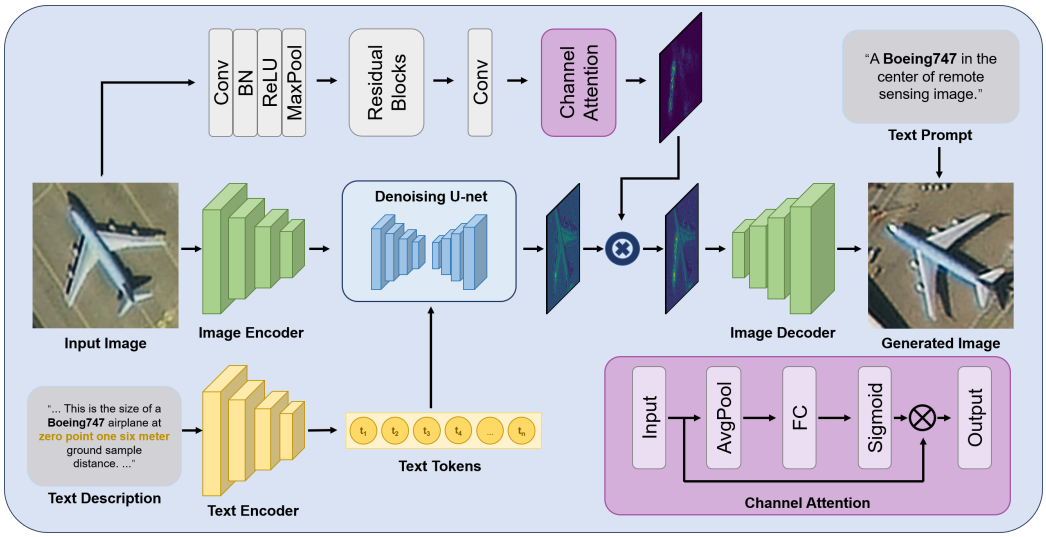


Figure 2. The framework of our proposed method. The first branch is the extraction of ground sample distance features, and the second branch is the image generation branch. The input of the first branch is the fine-grained remote sensing image, and the ground sample distance features are obtained through an attention layer. The input of the second branch is the fine-grained remote sensing image and its corresponding text description, which is a normal image generation process. The two features are fused to obtain new features and re-input into the network to generate remote sensing images.

3.2. Encoder and Decoder

The proposed model architecture incorporates an image encoder, an image decoder, and a text encoder derived from CLIP [22], as these components play a pivotal role in determining the quality of the synthesized images. The image encoder is designed to transform the input image into a high-dimensional semantic representation vector that encapsulates the content and semantic information of the image. Typically, it is implemented using a pre-trained visual model, such as the Vision Transformer, equipped with feature extraction layers capable of capturing rich and meaningful features from the input image. The image decoder, on the other hand, is tasked with reconstructing an image from the learned semantic representation vector, ensuring the generated image is visually similar to the original input. This reconstruction process not only validates the quality of the decoder but also encourages the image encoder to learn more expressive, meaningful, and discriminative representations by aligning the input and reconstructed images.

The text encoder is responsible for transforming natural language input into a fixed-dimensional vector representation, where semantically similar texts are closer in the embedding space, and dissimilar texts are positioned farther apart. This process begins with tokenizing the input text, where each token is mapped to an embedding vector. These token embeddings are then processed through the encoder layers of the model, which progressively capture and encode the semantic information of the text. The image and text encoders operate synergistically, learning a shared embedding space by contrasting the representations of images and text. This shared space facilitates the model's ability to understand and align the semantic relationships between the two modalities.

To achieve this alignment, a contrastive loss function ($\mathcal{L}_c$) is employed, which is essential for learning the relationship between image and text representations. The objective of this loss is to maximize the similarity score of positive image-text pairs while minimizing the similarity score of negative pairs. This ensures that the model learns robust and discriminative embeddings, enabling it to effectively capture and leverage the semantic alignment between images and text for downstream tasks.

$$\mathcal{L}_c = -\log \frac{\exp(\sin(f(\text{image}), g(\text{text})))}{\sum_{j=1}^{N} \exp(\sin(f(\text{image}), g(\text{other_texts}[j])))} \tag{1}$$

where *sim* represents the similarity measurement function, $f(\text{image})$ and $g(\text{text})$ are the image encoder and text encoder, respectively.

3.3. Text-to-Image

The quality of synthesized images is fundamentally dependent on the functionality and effectiveness of both the image encoder and decoder. Building upon the advancements in diffusion models presented in [50], this study adopts stable diffusion as the foundational framework to model the latent space of images. The core principle of diffusion models lies in an iterative process where noise signals are progressively refined to generate high-quality data samples. This generative process is grounded in probabilistic modeling, wherein the pixel values of the synthesized data are incrementally updated to closely approximate those of real-world data distributions. The diffusion process typically comprises two key phases: the forward process and the reverse process [26]. In the forward process, an initial image or feature vector undergoes a gradual transformation by the incremental addition of Gaussian noise at each timestep. This step-wise introduction of randomness progressively alters the image, driving it toward a noisy latent representation. Conversely, the reverse process begins from this noisy state and iteratively denoises the representation by applying learned transformations, reconstructing the data to resemble realistic images. By modeling this bidirectional process, diffusion models effectively capture the latent structure of the data, enabling the generation of samples that are both semantically meaningful and visually realistic. This process can be represented as $X_t = X_{t-1} + \epsilon_t$, here, X_t is the state at time t, and ϵ_t is Gaussian noise added at each step. After the completion of the forward process, a backward process is carried out to gradually denoise and restore the clarity of the image or feature vector. This process can be represented as $X_{t-1} = X_t - \epsilon_t$, where X_{t-1} is the state at time $t-1$ and ϵ_t is the Gaussian noise added in the same step. This step progressively removes the noise introduced in the forward process, gradually bringing the image back to a clearer state. The forward and backward processes can be iterated until specific termination conditions are met. Finally, based on the results of the diffusion process, the ultimate image or feature vector can be generated.

In the foreword process, the original image satisfies the distribution of $x_0 \sim Q(x_0)$, the Gaussian noise is sequentially added to the samples over a time period of T, and get a series of $x_1, ..., x_T$.

$$Q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

$$Q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} Q(x_t \mid x_{t-1}) \tag{3}$$

where $\{\beta_t \in (0,1)\}_{t=1}^{t}$ is the variance at each step of sampling, as time t increases, x_0 becomes more and more ambiguous until t approaches T and x_T approaches an isotropic Gaussian distribution [26]. Based on this, x_t at any t-step can be calculated by the original x_0:

$$Q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)\mathbf{I}) \tag{4}$$

where $\alpha_t = \prod_{i=1}^{t}(1-\beta_t)$, $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$, and ϵ is the standard Gaussian noise. A neural network ϵ_θ is used to predict noise in the optimization process based on the above theory. The denoising model $\epsilon_\theta(x_t, t)$ starts with a input noisy x_t at time t. The model predicts the noise ϵ by mean square error loss:

$$\mathcal{L}_t = \left[\left\| \epsilon - \epsilon_\theta(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon, t) \right\|_2^2 \right] \quad (5)$$

The input of the generation network is an image **I** and a text description **T** as shown in Figure 2 and the output is a generated image by giving a text promote.

3.4. Channel Attention Layer

Channel attention mechanisms are widely employed in convolutional neural networks (CNNs) to enhance the network's ability to focus on salient features by adaptively reweighting the importance of different channels in the feature maps [51]. The central objective of channel attention is to enable the model to dynamically prioritize specific channels, thereby amplifying critical information while suppressing irrelevant or less significant features [52]. By leveraging channel-wise dependencies, this mechanism effectively guides the network to extract and refine information at multiple semantic levels, improving its capacity to capture complex patterns and subtle variations within the input data. Through the dynamic learning of channel weights, channel attention enhances the representational power of the network, allowing it to better adapt to intricate and high-dimensional image tasks. Additionally, this adaptive focus contributes to improved perceptual understanding, making the network more robust and efficient in handling challenging visual scenarios. The ability to selectively emphasize critical features not only optimizes the overall performance of the model but also aligns the network's computations with the underlying structure of the data, solidifying its role as a key component in modern deep learning architectures.

3.5. Ground Sample Distance Feature Extraction

Ground sample distance as an important attribute of remote sensing is calculated by:

$$g = \frac{d_{geo}}{n} \quad (6)$$

where the g is the GSD, d_{geo} is the actual distance on earth and n is the number of pixels of the instance in the image. Geospatial Data Abstraction Library (GDAL) [53] is used to extract the GSD information for each image. Considering the special shooting angle of remote sensing images, the same category should have the same scale in one image. When the generation process is not restricted, diffusion models can develop the ability to generate objects that appear realistic to the human eye. However, the same category may have different sizes in one image which is seriously anamorphic as a remote sensing image. By incorporating ground sample distance (GSD) into the image generation process, we enable the creation of more realistic and contextually accurate objects for remote sensing applications. In our approach, GSD is integrated into the generative model through two complementary mechanisms. First, as a text-to-image framework, textual descriptions play a critical role in guiding the generation process. To incorporate GSD effectively, we classify the ground sample distances of the original dataset into three discrete categories: 0.3 m, 0.6 m, and 0.8 m. For this study, we focus exclusively on airplane categories within the dataset, as aircraft images captured via remote sensing typically correspond to a GSD of 0.8 m. The textual description of GSD is dynamically adjusted to reflect changes in image size, ensuring consistency between the input description and the generated imagery. Second, GSD is encoded as a deep feature and seamlessly integrated into the generation pipeline. To achieve this, we design a GSD feature extraction network based on ResNet50, where the average pooling layer and fully connected layer are replaced by a convolutional layer for enhanced feature granularity. As depicted in Figure 2, the extraction network takes an input image **I** and outputs a feature representation. To further refine the extracted GSD feature, we incorporate a channel attention module, which dynamically adjusts

attention weights across channels. This ensures that the GSD feature is neither diluted nor overshadowed during the feature fusion process, preserving its integrity and contribution to the overall generation. GSD manifests in the image plane as characteristic statistics of spatial frequencies, edge densities, and object extents (in pixels). A moderately deep convolutional backbone with residual connections is well suited to capture such mid-level structure. ResNet50 provides sufficient receptive field and depth to encode scale- and texture-sensitive patterns that correlate with spatial resolution, without incurring the optimization difficulties of very deep networks. By employing these dual mechanisms—GSD textual guidance and deep feature encoding—we ensure that the generative model not only adheres to realistic spatial constraints but also captures the nuanced dependencies introduced by GSD. This dual integration ultimately enhances the accuracy, realism, and fidelity of the generated remote sensing imagery.

$$\mathbf{F}'_{Gen} = Conv(concat(\mathbf{F}_{Gen}, \mathbf{weight} \cdot \mathbf{F}_{GSD})), \tag{7}$$

where $\mathbf{weight} = Sigmoid(Att(\mathbf{F}_{GSD}))$, F_{Gen} is original feature from encoder of diffusion model, F_{GSD} is GSD feature. The incorporation of GSD into image generation through these two ways facilitates the generation of a more realistic and reliable image for remote sensing applications.

4. Experiments and Results

4.1. Datasets

FAIR1M2.0 is a large-scale dataset that was designed to refine object detection and recognition in remote sensing images. The images in the FAIR1M were collected from Gaofen satellites and Google Earth with a ground sample distance range of 0.3 m to 0.8 m. FAIR1M contains more than a million instances and more than 15,000 images. Five categories were annotated: airplanes, ships, vehicles, courts, and roads, with 37 subcategories within each category. There are 10 types for airplanes, 8 specific categories for ships, 9 specific categories of vehicles, 4 categories for courts, 3 categories for roads and 'other-airplane', 'other-ship', and 'other-vehicle' for objects that do not belong to the previously defined specific object types. Our experiments concentrate on airplanes, including Boeing 737, Boeing 777, Boeing 747, Boeing 787, Airbus A220, Airbus A321, Airbus A330, Airbus A350, COMAC C919, and COMAC ARJ21. The categories 'other-airplane', 'other-ship', and 'other-vehicle' are used for objects that do not belong to the previously defined specific object types. Images range in size from 1000 × 1000 to 10,000 × 10,000 and contain objects of various sizes, orientations, and shapes. Zeng et al. [54]'s dataset was screened for plane categories on Fair1m and then trained and tested. On this basis, we removed the category 'other-airplane', resulting in 1196 images in the training set and 364 images in the test set. This dataset is referred to as Small Fair1m in this paper. Both small fair1m and fair1m2.0 were used for the experiment of object detection.

4.2. Evaluation Metrics

To evaluate the quality of samples generated, we compute Fréchet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM), respectively. FID measures the distance between feature representations of real and generated images. LPIPS quantifies perceptual similarity between images based on learned representations, using deep neural networks to evaluate local image patches' similarity in a perceptually consistent manner. PSNR measures the noise level between images, evaluating image quality by comparing the maximum possible signal value to the mean squared error of the images. SSIM gauges

the similarity between images based on luminance, contrast, and structure, providing a comprehensive assessment of perceptual image quality.

4.3. Implementation Details

A classification network is used to extract GSD features. A ResNet50 architecture (excluding the fully connected layer) is selected for the classification network, with a channel attention layer following the final convolution layer to enhance classification performance. The classification head for feature extraction is a fully connected layer. We use the pretrained models from Pytorch 1.12.1 to initialize the GSDN weights and regulate the stride of the last layer according to the image size. We use an initial learning rate of 0.01 and a step learning rate decay schedule to modify $lr \leftarrow 0.1 \cdot lr$ every 50 epochs. This neural network architecture was trained on the FAIR1M dataset over the course of 100 epochs, using a batch size of 32. The optimization algorithm used is stochastic gradient descent, with a momentum value of 0.9 to improve convergence.

For the generation process, we selected Stable Diffusion as the pretrained foundational model due to its robust performance in image synthesis tasks. Using the FAIR1M dataset, we first filtered the dataset to focus on ten specific aircraft models. For each selected instance, we cropped images to a size of 256 × 256 pixels, centering on the aircraft. Instances were excluded if the aircraft's center was located within 100 pixels of the image boundary to ensure sufficient contextual information around the target object.

From our analysis of the dataset, we observed that aircraft categories such as Boeing 787, Boeing 777, Boeing 747, A350, and A330 typically span approximately 100 pixels at a ground sample distance (GSD) of 0.8 m, whereas the remaining five classes occupy approximately 64 pixels under the same GSD. To standardize input sizes for further processing, we cropped these objects individually to dimensions of 100 × 100 pixels or 64 × 64 pixels, depending on their original resolution. However, given that such small image sizes are insufficient for effective training, we resampled all cropped images to a uniform resolution of 512 × 512 pixels.

Additionally, the GSD was recalibrated for the resampled images to reflect the new scaling. Specifically, objects with an original resolution of 100 pixels were assigned a rescaled GSD of 0.16 m, while those with an original resolution of 64 pixels were adjusted to a GSD of 0.13 m. Each object was annotated with its corresponding rescaled GSD to provide accurate semantic guidance during the generation process. This dual adjustment of image resolution and GSD ensures consistency between the input annotations and the synthesized outputs, thereby enhancing the realism and precision of the generated imagery.

To improve realism and reduce the domain gap between isolated cutouts and operational scenes, we composite the resampled aircraft onto background tiles drawn from FAIR1M images that do not contain the same target instance. Backgrounds are chosen to be semantically compatible with aviation contexts, including runways, taxiways, and apron areas, and are sampled to provide diverse textures and illumination. For each composite, we align the aircraft orientation to the dominant background heading (for example, along runway or taxiway directions) and apply soft alpha matting to blend the fuselage and wings with minimal boundary artifacts. We also harmonize color and contrast between the foreground and background to suppress visible seams and, when appropriate, ensure the aircraft is placed in plausible locations within the scene layout. A small portion of compositions intentionally use non-runway paved areas to encourage contextual robustness in downstream tasks, while always maintaining the recalibrated GSD and the physical scale relationship established in the previous steps.

For the training of the text-to-image generation model, we utilized the resampled images, each paired with a caption containing the specific aircraft category and a description

of the corresponding ground sample distance (GSD). The training process was configured with 100 epochs and a batch size of 4, ensuring sufficient iterations for the model to learn fine-grained details. The maximum length of the input sentence in the CLIP model was set to 50 tokens, and the gradient accumulation steps were set to four to balance memory efficiency and computational stability.

The learning rate was initialized with a warm-up phase, gradually increasing to 0.0001 over the first 10,000 steps, after which it remained constant throughout the remainder of the training. Given the varying availability of samples across categories—some exceeding 10,000 instances—we opted to generate 1000 images per category to ensure a balanced output distribution and to mitigate class imbalance in downstream tasks.

During the generation phase, the inference process was configured with 100 sampling steps to ensure high-quality image synthesis. Additionally, a guidance scale of 7.5 was employed to enhance the fidelity of the generated images, striking a balance between adherence to the textual prompt and diversity in the outputs. This carefully designed training and inference pipeline ensures that the model effectively leverages the annotated GSD and category information to produce remote sensing imagery that is realistic and semantically consistent.

For object detection, initially, we generated several random background images and then inserted 25 randomly generated images containing aircraft into each background image, ensuring that all aircraft have the same ground sample distance. The training epoch was set to 120 and the batch size was set to four. The optimization algorithm used was stochastic gradient descent, with a momentum value of 0.9. The learning rate warmup was set to 0.0025 for 800 steps and then kept constant, and the modify $lr \leftarrow 0.1 \cdot lr$ epoch was set to 80 and 100. The experiments in this paper were conducted in PyTorch using NVIDIA RTX 3090 GPUs.

4.4. Comparison and Ablation Studies in Generation Result

We evaluated the quality of the generated images using four metrics: FID, LPIPS, PSNR, and SSIM, where lower FID and LPIPS values indicate greater fidelity and perceptual similarity, and higher PSNR and SSIM values reflect improved noise reduction and structural consistency. As shown in Table 1, 100 real and 100 generated images per category were evaluated, with their similarity calculated and averaged. Despite minor inconsistencies in object orientation impacting FID values, our method achieved significant improvements across all metrics, as summarized in Table 2. Specifically, FID decreased by 113.44, demonstrating better alignment between real and generated image distributions; LPIPS reduced by 0.25, indicating finer perceptual detail retention; PSNR increased by 2.98, reflecting enhanced image fidelity; and SSIM improved by 0.24, confirming greater structural accuracy. These results validate the effectiveness of our method in producing high-quality, realistic images that closely align with real remote sensing imagery. Although slight variations in object orientation impacted certain metrics, the overall performance improvements across all categories underscore the robustness of our approach. Importantly, the metrics confirm that our integrated use of ground sample distance (GSD) and the novel architecture significantly enhances the generation process, offering both perceptual and structural improvements. In conclusion, the substantial reductions in FID and LPIPS, combined with the increases in PSNR and SSIM, highlight the strength of our model in generating authentic and high-fidelity remote sensing images. These improvements not only validate the generative capabilities of our framework but also reinforce its applicability to downstream tasks, such as data augmentation for object detection.

Table 1. Evaluation metrics of image generation. The generated images are from Txt2Img-MHN (VQGAN) [10], Stable Diffusion [28], and our method.

Category	FID ↓			LPIPS ↓			PSNR ↑			SSIM ↑		
	VQGAN	SD	Text2AIRS	VQGAN	SD	Text2AIRS	VQGAN	SD	Text2AIRS	VQGAN	SD	Text2AIRS
Boeing737	222.76	194.83	**148.92**	0.40	0.46	**0.37**	12.66	13.19	**14.98**	0.63	0.52	**0.64**
Boeing747	301.49	207.06	**148.93**	0.53	0.66	**0.43**	13.16	10.02	**14.67**	**0.52**	0.17	0.51
Boeing777	267.88	260.27	**145.14**	0.54	0.65	**0.31**	12.43	10.00	**13.19**	**0.51**	0.17	0.48
Boeing787	288.04	241.60	**152.82**	0.50	0.69	**0.37**	13.49	10.52	**14.73**	**0.57**	0.17	0.52
A350	286.71	241.42	**153.83**	0.56	0.64	**0.26**	12.30	10.24	**13.23**	**0.50**	0.19	0.43
A330	259.85	234.97	**115.46**	**0.39**	0.53	**0.39**	12.70	10.42	**14.09**	**0.56**	0.20	0.53
A321	258.32	164.65	**118.34**	0.40	0.47	**0.31**	12.87	12.25	**14.34**	0.66	0.55	**0.70**
A220	192.99	201.47	**148.88**	0.32	0.46	**0.19**	15.18	13.93	**16.83**	0.68	0.55	**0.71**
ARJ21	340.90	302.50	**142.16**	0.40	0.53	**0.35**	13.93	12.69	**14.03**	**0.64**	0.50	0.59
C919	340.56	303.26	**123.09**	0.41	0.57	**0.21**	13.07	10.91	**13.92**	0.63	0.45	**0.76**

↑ indicates that higher values are better, and ↓ indicates that lower values are better. Bold values represent the best performance for each metric within each specific category.

Table 2. Supplement to Ablation studies. SD is short for Stable Diffusion, Text is short for textual description with GSD, Fea. is short for encoded GSD feature. FID, LPIPS, PSNR and SSIM are evaluation metrics for generation, while mAP, mAP_{50} and mAP_{75} are for detection.

SD	+ Text	+ Fea.	FID	LPIPS	PSNR	SSIM	mAP	mAP_{50}	mAP_{75}
			-	-	-	-	63.21	69.43	68.59
✓			253.20	0.57	11.42	0.35	64.49	70.41	70.27
✓	✓		170.49	0.44	13.56	0.54	64.78	71.09	70.24
✓		✓	231.25	0.57	11.76	0.37	66.71	73.06	72.35
✓	✓	✓	**139.76**	**0.32**	**14.40**	**0.59**	**67.35**	**74.01**	**73.49**

Note: The checkmark (✓) indicates that the corresponding component is included in the model. Bold values represent the best result for each evaluation metric.

Table 1 illustrates that, while the generated images from all three methods may not appear ideal, Text2AIRS exhibits distinct advantages across various categories, performing well in FID, LPIPS, and PSNR. Notably, nearly every category demonstrates favorable results in these metrics. But in terms of SSIM, images generated by Txt2Img-MHN (VQGAN) outperform the others. As shown in Figure 3, although the data do not represent the actual generation well in SSIM, the resulting object appears to be complete to the human eye and is very similar to the original image. It is crucial to note that the four aforementioned indicators should not be employed as the sole criteria for assessing the quality of generation. From a different perspective, if these generated images contribute to applications in the field of remote sensing, the authenticity and reliability of the generated images can also be emphasized.

To assess the effectiveness of each component in Text2AIRS, we conducted a series of ablation experiments with different baselines. Table 2 summarizes the results. In the table, 'SD' denotes the text-only Stable Diffusion baseline without any GSD conditioning, 'Text' denotes Stable Diffusion conditioned on captions that include GSD descriptions, and 'Fea.' denotes Stable Diffusion further conditioned on the encoded GSD feature. The evaluation metrics for generation include FID, LPIPS, PSNR, and SSIM, while those for object detection encompass mAP, mAP_{50}, and mAP_{75}. Using the generated image for data enhancement yields a significant improvement in detection accuracy. Specifically, compared to the baseline, the mean Average Precision (mAP) increases by 1.57 when GSD text information is incorporated into the image and by 0.29 when GSD text information is omitted. The introduction of the encoded depth feature of GSD results in a mAP increase of 2.57 compared to using only text descriptions and a 4.14 increase compared to the baseline. Additionally, mAP_{50} and mAP_{75} show varying degrees of growth. In comparison to using only text descriptions, mAP_{50} and mAP_{75} increased by 2.92 and 3.25, respectively.



Figure 3. Partial categories of real and generated images in Fair1M. The generated images are from Txt2Img-MHN (VQGAN) [10], Stable Diffusion [28], and our Text2AIRS.

4.5. Application

We selected the object detection task as an application to evaluate the effectiveness of the generated images, with visualization results presented in Figure 4. Image interpretation in remote sensing is inherently challenging due to the small scale of target objects, the complexity of surrounding environments, and the limited availability of annotated datasets. Manual annotation, particularly for fine-grained datasets, is resource-intensive and costly. To address this, we incorporated the generated remote sensing images containing airplanes into the training set as a data augmentation strategy to mitigate the issue of limited labeled samples.

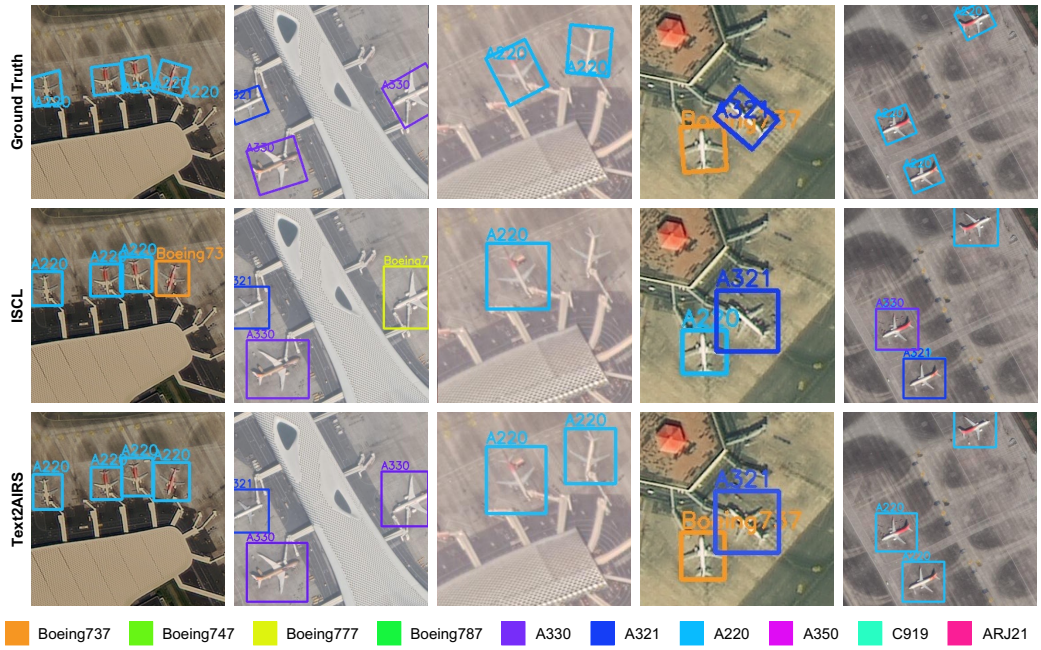


Figure 4. Visualisation results on Fair1M of object detection.

The generated images were randomly placed onto a background map of the same size as those in the training set, with corresponding annotations generated simultaneously. Given the directional variability of objects in remote sensing images, the orientations of objects in the generated samples were also arbitrary. To simplify annotation and reduce labor costs, we adopted horizontal bounding boxes (HBBs) as the labeling format. Importantly, the ground sample distance (GSD) of objects within each generated image was kept constant to maintain consistency. We further leveraged the ISCL framework, which selectively refines instance switching to ensure a smaller number of instances can effectively participate in training. Additionally, ISCL employs data augmentation to improve detection performance. For fairness, the baseline for our experiments was set to Faster R-CNN, as used in ISCL, alongside two additional baselines for comparison. To evaluate our method comprehensively, 100 instances per category were generated for the Small-Fair1M dataset, and 1000 instances per category were generated for Fair1M2.0. The initial distribution of instances is illustrated in Figure 5, providing a clear view of the dataset composition. This approach demonstrates the potential of utilizing generated images for data enhancement in remote sensing object detection, reducing dependence on manual annotation and improving the overall effectiveness in detecting fine-grained objects.



Figure 5. The distribution of the number of instances per category. The first line is the initial data distribution and the second line is the added distribution. The first column is the small Fair1M and the second column is the entire Fair1M.

Table 3 shows comparisons of the airplane subset of Fair1M following Zeng et al., with "other-airplane" removed, yielding 1196 training images and 364 test images. Object detection is evaluated with HBBs, reporting per-class AP and mAP. Abbreviations: B = Boeing, A = Airbus, C = COMAC; columns are B737, B747, B777, B787, A220, A321, A330, A350, ARJ21, and C919. In Table 3, it is demonstrated that our Text2AIRS, based on Faster R-CNN, can achieve a result of 67.35 on the Small Fair1M dataset. A comparison of different data augmentation methods was also conducted. The term Txt2Img-MHN refers to the use of VQGAN as a tool for generating remote sensing images. We use some of the instances in the validation set and rotated the samples to achieve a Faster R-CNN + Aug as it is the common data augmentation method. Text2AIRS is designed by adding a ground sample distance constraint to assist with image generation. It should be noted that all three algorithms used the same 100 instances. Compared to the other two augmentation methods, our method improves the effect more effectively. In comparison to the baseline, 7 of the 10 categories have shown an improvement from the baseline. Based on the ISCL, 4 of the 10 categories have shown positive outcomes. As there is only one image of C919 in the Small Fair1M test set, and the right-wing of C919 in this image can barely be seen, the detection result of most of the methods for this category is 0. We also choose different baselines for the experiments to avoid bias in the results according to Table 4. Table 4 results on the airplane categories using the official splits and evaluation protocol; all other training and evaluation settings match Table 3. We reported per-class AP and mAP. Text2AIRS also leads on this larger dataset, demonstrating the method's transferability and robustness. Based on the results of these two networks, it can be concluded that the results have been greatly improved after the addition of the generated images by Text2AIRS.

Table 3. Comparisons on Small Fair1m. B is short for Boeing, A is short for Airbus, C is short for COMAC.

Method	B737	B747	B777	B787	A220	A321	A330	A350	ARJ21	C919	mAP
Faster R-CNN [42]	70.92	55.98	31.32	86.61	73.68	78.57	78.29	66.34	90.20	0.00	63.21
Mask R-CNN [55]	69.83	36.57	35.23	81.55	74.74	80.40	81.18	63.07	**98.51**	0.00	62.11
Cascade R-CNN [56]	70.68	56.14	46.10	83.89	72.62	78.90	74.15	63.07	93.92	0.00	63.95
DetectoRS [57]	71.56	55.74	29.80	84.80	73.42	78.09	83.53	59.70	94.32	0.00	63.10
MADet [58]	**74.64**	55.64	43.62	83.23	**78.94**	81.58	83.58	64.22	94.79	0.00	66.02
DDQ DETR [59]	70.32	56.26	**58.21**	**93.04**	78.68	81.51	81.64	48.55	95.17	**0.78**	66.42
ISCL [54]	72.04	**67.99**	44.28	84.96	78.68	81.58	82.19	62.63	95.74	0.00	67.01
Txt2Img-MHN [10]	71.23	40.56	38.48	90.21	73.21	**83.76**	82.62	63.07	97.30	0.00	64.04
Faster R-CNN + Aug	68.17	52.79	52.68	91.00	74.36	80.27	74.83	63.07	93.83	0.00	65.10
Text2AIRS	70.52	61.05	53.44	84.03	76.11	82.08	**83.82**	**69.50**	93.01	0.00	**67.35**

Note: Bold values indicate the best results for each category, and underlined values represent the second-best results.

Table 4. Comparisons on Fair1m2.0. B is short for Boeing, A is short for Airbus, C is short for COMAC.

Method	B737	B747	B777	B787	A220	A321	A330	A350	ARJ21	C919	mAP
Faster R-CNN [42]	84.39	91.62	88.64	89.2	86.51	89.32	92.64	88.65	88.94	83.60	88.35
Cascade R-CNN [56]	84.00	91.97	88.60	89.67	86.19	89.90	90.36	91.14	90.96	87.49	89.03
Mask R-CNN [55]	**85.09**	92.99	90.46	90.35	**87.89**	90.04	**93.26**	91.56	**93.46**	80.61	89.57
ISCL [54]	83.20	**93.14**	89.31	90.05	86.14	89.01	92.81	**91.70**	92.68	**89.76**	89.78
Txt2Img-MHN [10]	82.99	91.61	88.34	87.57	86.61	89.23	91.95	88.81	90.48	89.23	88.68
Text2AIRS	84.79	91.97	**90.70**	**90.79**	87.60	**90.59**	92.83	91.61	91.72	87.54	**90.01**

Note: Bold values indicate the best results for each category, and underlined values represent the second-best results.

In Table 3, we compare the performance of various methods on Fair1M2.0. Considering that the number of instances for each category in Fair1M2.0 is basically all above 1000, with the greatest number reaching 5000, we use 1000 generated instances in each of the categories. In comparison to previous results, our model has demonstrated exceptional performance, reaching a state-of-the-art mean average precision of 90.01. Notably, specific categories within the model displayed remarkably high-performance results such as ARJ21 and C919, which have the fewest instances compared to other categories. The AP of ARJ21 increased by 2.78 and the AP of C919 increased by 3.94. It is worth noting that although we did not exceed the accuracy of ISCL for five of the categories, the training time of ISCL is significantly greater than that of our method. The ISCL algorithm takes 137.3 min per epoch to train, while we need only 11.4 min. The ISCL model requires 1004 MB, whereas ours requires only 315 MB. We obtained the best or second-best results among the 6/10 categories.

Furthermore, to minimize methodological bias and ensure a fair assessment, we evaluate our approach against multiple detector backbones rather than relying on a single baseline. As summarized in Table 5, we conduct ablation studies on both Mask R-CNN and Cascade R-CNN under identical training and evaluation protocols. Across both architectures, the incorporation of our proposed method yields consistent and substantial performance gains, indicating that the observed improvements are robust to the choice of detector and not attributable to model-specific idiosyncrasies. In the case of Mask R-CNN, mAP increased by 5.60, mAP_{50} increased by 5.47, and mAP_{75} increased by 6.18, respectively. A significant improvement has also been made to the Cascade R-CNN, of which the mAP has been increased by 6.12, the mAP_{50} by 5.50, and the mAP_{75} by 5.89. The visualisation results are shown in Figure 6.

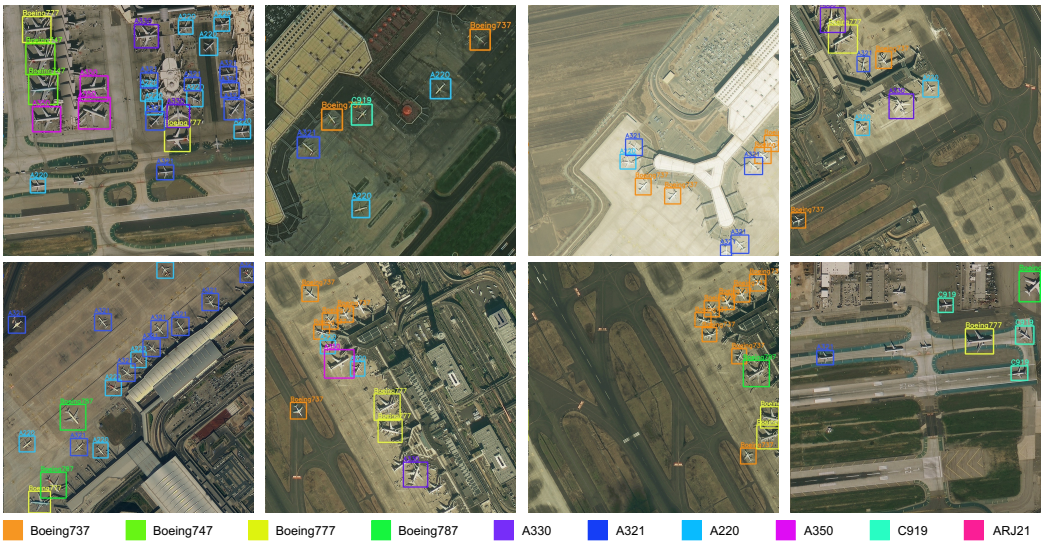


Figure 6. Visualisation results on Fair1M2.0 of object detection.

Table 5. Ablation studies on Mask R-CNN and Cascade R-CNN.

Baseline	+GSD	mAP	mAP_{50}	mAP_{75}
Mask-RCNN		62.11	67.85	67.08
	✓	67.71 (+5.60)	73.32 (+5.47)	73.26 (+6.18)
Cascade-RCNN		63.95	70.00	69.47
	✓	70.07 (+6.12)	75.5 (+5.50)	75.36 (+5.89)

Note: The checkmark (✓) indicates that the corresponding component is included in the model.

4.6. Discussion

This paper proposes a novel approach to text-to-image generation for remote sensing imagery by introducing ground sample distance (GSD) as a critical constraint. The GSD is explicitly annotated in captions and encoded as deep features, guiding the model to produce images that are consistent with real-world spatial resolutions. This design also mitigates a key bottleneck in remote sensing—limited availability of high-quality text–image pairs—by leveraging a low-cost, scalable annotation that is both physically meaningful and broadly available. Our experiments demonstrate that GSD-aware conditioning improves visual realism and semantic fidelity, validating the effectiveness of the proposed method.

We focused on aircraft as the primary object category because of their comparatively larger footprint in overhead imagery and easily distinguishable morphology, which facilitates controlled evaluation of scale, pose, and background context. This choice provides a robust starting point for assessing GSD-aware generation prior to tackling smaller, more cluttered categories such as vehicles and ships. We also acknowledge that uncertainty in object orientation complicates downstream oriented bounding box (OBB) detection; to reduce annotation cost and simplify evaluation, we adopt horizontal bounding boxes (HBBs) in this study, while OBB-based detection remains a priority for future work.

Despite these strengths, several limitations deserve explicit discussion and point to concrete extensions. First, the current setup lacks explicit orientation control during generation: while captions encode category and GSD, they do not reliably constrain aircraft heading or roll/yaw, which can hinder OBB training and orientation-sensitive applications. A promising direction is to incorporate geometric metadata (for example, heading or runway axis) as an auxiliary conditioning signal, or to add lightweight control modules that steer pose without sacrificing diversity. Second, the model does not simulate physical sensor properties such as point spread function, modulation transfer function, noise characteristics, or atmospheric effects; this can lead to minor mismatches in texture frequency

and edge sharpness compared with specific satellites. Future work could integrate simple sensor models at training or inference time, or finetune on sensor-specific subsets to improve transfer.

Third, our experiments rely exclusively on RGB imagery, which may bias color statistics and compress fine-grained textures that are salient in remote sensing. Extending the framework to multispectral or SAR modalities and training or adapting a remote-sensing-specific VAE could better preserve domain-typical structures and spectra. Fourth, although we introduced background selection and scene composition to reduce the gap between isolated objects and operational scenes, the present work still emphasizes object-centric generation rather than full-scene synthesis with geographic coherence. Scaling to larger tiles that maintain consistent spatial layout, land-cover distribution, and co-occurrence patterns (for example, runways, taxiways, stands, and service roads) is an important step toward downstream tasks such as detection, segmentation, and change monitoring.

Finally, data coverage and caption design remain practical constraints. While GSD annotations help standardize scale, category- and site-specific diversity (illumination, material, markings, and clutter) is not exhaustively represented. Iterative dataset expansion, more structured captions that encode background type and lighting, and modest domain-adaptation techniques can further improve robustness.

In summary, injecting GSD as a physically grounded constraint demonstrably enhances remote-sensing text-to-image generation, particularly for aircraft. Addressing the identified limitations—explicit pose/orientation control, lightweight sensor modeling, multispectral extensions, remote-sensing–aware VAEs, and geographically coherent scene synthesis—constitutes a clear path for future research. These directions will broaden applicability to smaller, more complex objects (for example, vehicles and ships), enable OBB-based detection, and improve fidelity to specific platforms and environments.

5. Conclusions

In this study, we introduced Text2AIRS, a diffusion-based framework for generating realistic, fine-grained airplane imagery from natural language in remote sensing. To our knowledge, this is the first work to explicitly incorporate ground sample distance (GSD) as a structural prior in text-to-image generation for RSI. By injecting GSD at both the prompt level and the feature level—via resolution-aware captions and attention-guided modulation within the denoising U-Net—our approach calibrates the generative process to the target spatial resolution and improves intra-image scale consistency. Beyond perceptual quality, Text2AIRS also proves practically useful: when applied as targeted augmentation on Fair1M under severe class imbalance, the synthesized instances yield measurable gains in detection accuracy while keeping the validation and test sets unchanged. The contributions of this work lie in turning GSD from passive metadata into an active constraint that governs scale, offering a principled remedy to the long-standing issue of inconsistent object sizes within single images. Architecturally, the dual-branch conditioning integrates smoothly with a standard diffusion backbone, making resolution-aware generation attainable without heavyweight redesign. Practically, the generated images transfer into downstream benefits under data scarcity, demonstrating that physically grounded constraints can improve both realism and utility. We also clarify the evaluation protocol and implementation details to facilitate transparency and reproducibility. Nevertheless, several limitations remain. For clarity and reproducibility of augmentation, our detection experiments adopt HBBs (horizontal bounding boxes), which simplifies instance insertion but under-represents orientation cues that matter in remote sensing. Our current evaluation focuses on aircraft; broader taxonomies and more cluttered backgrounds require further validation. The approach assumes access to reliable GSD, and performance may degrade when metadata are

noisy or missing. Extremely low-texture or highly specular scenes also remain challenging and can introduce artifacts. Looking ahead, we plan to extend Text2AIRS to OBB-based augmentation and evaluation so that orientation and rotation consistency are explicitly modeled, to introduce stronger layout and geometric controls (such as instance count, spacing, and pose priors) for improved scene realism, and to broaden the evaluation to multi-class, multi-sensor benchmarks. We also aim to develop uncertainty-aware GSD inference when metadata are unavailable and to study robustness under GSD noise, while releasing standardized protocols and code to support fair comparisons and reproducible research. Overall, by enforcing a physically grounded resolution prior through GSD, Text2AIRS advances the fidelity of synthesized fine-grained aircraft and the effectiveness of downstream detection, pointing toward scalable, resolution-aware data generation in remote sensing.

Author Contributions: Conceptualization, Y.Y., Y.X. and Y.Z.; methodology, Y.Y., Y.X., and Y.Z.; validation, Y.Y., Y.C., and J.H.; formal analysis, Y.Y.; investigation, Y.Y., Y.C. and J.H.; data curation, Y.Y.; writing—original draft preparation, Y.Y.; writing—review and editing, Y.X. and Y.Z.; visualization, Y.Y.; supervision, Y.Z.; project administration, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62471415.

Data Availability Statement: The data presented in this study are available within the article. The image generation and object detection models were trained and evaluated using the publicly available FAIR1M dataset. The synthetic images generated by the proposed Text2AIRS method are derived from this dataset based on the methodology described. No entirely new standalone dataset was created and released as part of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, Z.; Ma, Y.; Zhang, Y. Review of pixel-level remote sensing image fusion based on deep learning. *Inf. Fusion* **2023**, *90*, 36–58. [CrossRef]
2. Xia, Y.; Wu, Q.; Li, W.; Chan, A.B.; Stilla, U. A Lightweight and Detector-Free 3D Single Object Tracker on Point Clouds. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 5543–5554. [CrossRef]
3. Hao, X.; Liu, L.; Yang, R.; Yin, L.; Zhang, L.; Li, X. A Review of Data Augmentation Methods of Remote Sensing Image Target Recognition. *Remote Sens.* **2023**, *15*, 827. [CrossRef]
4. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5612822. [CrossRef]
5. Li, W.; Wei, W.; Zhang, L. GSDet: Object detection in aerial images based on scale reasoning. *IEEE Trans. Image Process.* **2021**, *30*, 4599–4609. [CrossRef]
6. Yang, Y.; Wang, C.; Cai, Z.; Song, P.; Huang, G.; Cheng, M.; Zang, Y. GSDDet: Ground Sample Distance Guided Object Detection for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5626012. [CrossRef]
7. Zhang, F.; Du, B.; Zhang, L. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2175–2184. [CrossRef]
8. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2183–2195. [CrossRef]
9. Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; Wang, Z. NWPU-captions dataset and MLCA-net for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5629419. [CrossRef]
10. Xu, Y.; Yu, W.; Ghamisi, P.; Kopp, M.; Hochreiter, S. Txt2Img-MHN: Remote sensing image generation from text using modern Hopfield networks. *IEEE Trans. Image Process.* **2023**, *32*, 5737–5750. [CrossRef] [PubMed]
11. Ramsauer, H.; Schäfl, B.; Lehner, J.; Seidl, P.; Widrich, M.; Adler, T.; Gruber, L.; Holzleitner, M.; Pavlović, M.; Sandve, G.K.; et al. Hopfield networks is all you need. *arXiv* **2020**, arXiv:2008.02217.
12. Dieste, Á.G.; Argüello, F.; Heras, D.B. ResBaGAN: A Residual Balancing GAN with Data Augmentation for Forest Mapping. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 6428–6447. [CrossRef]

13. Sebaq, A.; ElHelw, M. RSDiff: Remote Sensing Image Generation from Text Using Diffusion Model. *arXiv* **2023**, arXiv:2309.02455. [CrossRef]
14. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
15. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
16. Sun, X.; Wang, P.; Yan, Z.; Diao, W.; Lu, X.; Yang, Z.; Zhang, Y.; Xiang, D.; Yan, C.; Guo, J.; et al. Automated high-resolution earth observation image interpretation: Outcome of the 2020 Gaofen challenge. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8922–8940. [CrossRef]
17. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
18. Guo, B.; Zhang, R.; Guo, H.; Yang, W.; Yu, H.; Zhang, P.; Zou, T. Fine-Grained Ship Detection in High-Resolution Satellite Images With Shape-Aware Feature Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1914–1926. [CrossRef]
19. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. In Proceedings of the International Conference on Machine Learning (ICML 2016), New York, NY, USA, 20–22 June 2016; pp. 1060–1069.
20. Reed, S.E.; Akata, Z.; Mohan, S.; Tenka, S.; Schiele, B.; Lee, H. Learning what and where to draw. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 217–225.
21. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.
22. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML 2021), Virtual, 18–24 July 2021; pp. 8748–8763.
23. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv* **2022**, arXiv:2204.06125. [CrossRef]
24. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
25. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv* **2023**, arXiv:2303.12712. [CrossRef]
26. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
27. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494.
28. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
29. Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023, pp. 22500–22510.
30. Quattrochi, D.A.; Ridd, M.K. Analysis of vegetation within a semi-arid urban environment using high spatial resolution airborne thermal infrared remote sensing data. *Atmos. Environ.* **1998**, *32*, 19–33. [CrossRef]
31. Atkinson, P.M.; Curran, P.J. Choosing an appropriate spatial resolution for remote sensing investigations. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 1345–1351.
32. Benjamin, S.; Gaydos, L. Spatial resolution requirements for automated cartographic road extraction. *Photogramm. Eng. Remote Sens.* **1990**, *56*, 93–100.
33. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
34. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In Proceedings of the 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujscie, Poland, 9–12 May 2018; IEEE: New York, NY, USA, 2018; pp. 117–122.
35. Li, B.; Hou, Y.; Che, W. Data augmentation approaches in natural language processing: A survey. *AI Open* **2022**, *3*, 71–90. [CrossRef]
36. Paschali, M.; Simson, W.; Roy, A.G.; Göbl, R.; Wachinger, C.; Navab, N. Manifold exploring data augmentation with geometric transformations for increased performance and robustness. In *Proceedings of the Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, 2–7 June 2019*; Proceedings 26; Springer: Cham, Switzerland, 2019; pp. 517–529.

37. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
38. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
39. He, Q.; Sun, X.; Yan, Z.; Li, B.; Fu, K. Multi-object tracking in satellite videos with graph-based multitask modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5619513. [CrossRef]
40. Wang, J.; Yang, W.; Li, H.C.; Zhang, H.; Xia, G.S. Learning center probability map for detecting objects in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4307–4323. [CrossRef]
41. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
42. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
43. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
44. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
45. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef]
46. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 3163–3171.
47. Han, J.; Ding, J.; Xue, N.; Xia, G.S. Redet: A rotation-equivariant detector for aerial object detection. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021; pp. 2786–2795.
48. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605415. [CrossRef]
49. Li, C.; Cheng, G.; Wang, G.; Zhou, P.; Han, J. Instance-aware distillation for efficient object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5602011. [CrossRef]
50. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
51. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
52. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 783–792.
53. Qin, C.Z.; Zhan, L.J.; Zhu, A.X. How to apply the geospatial data abstraction library (GDAL) properly to parallel geospatial raster I/O? *Trans. GIS* **2014**, *18*, 950–957. [CrossRef]
54. Zeng, L.; Guo, H.; Yang, W.; Yu, H.; Yu, L.; Zhang, P.; Zou, T. Instance Switching-Based Contrastive Learning for Fine-Grained Airplane Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5633416. [CrossRef]
55. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
56. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
57. Qiao, S.; Chen, L.C.; Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10213–10224.
58. Xie, X.; Lang, C.; Miao, S.; Cheng, G.; Li, K.; Han, J. Mutual-assistance learning for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 15171–15184. [CrossRef] [PubMed]
59. Zhang, S.; Wang, X.; Wang, J.; Pang, J.; Lyu, C.; Zhang, W.; Luo, P.; Chen, K. Dense Distinct Query for End-to-End Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7329–7338.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.