MMO-IG: Multi-Class and Multi-Scale Object Image Generation for Remote Sensing

Chuang Yang, Bingxuan Zhao, Qing Zhou, Qi Wang*, *IEEE Member*

***Abstract*—The rapid advancement of deep generative models (DGMs) has significantly advanced research in computer vision, providing a cost-effective alternative to acquiring vast quantities of expensive imagery. However, existing methods predominantly focus on synthesizing remote sensing (RS) images aligned with real images in a global layout view, which limits their applicability in RS image object detection (RSIOD) research. To address these challenges, we propose a multi-class and multi-scale object image generator based on DGMs, termed *MMO-IG*, designed to generate RS images with supervised object labels from global and local aspects simultaneously. Specifically, from the local view, MMO-IG encodes various RS instances using an iso-spacing instance map (ISIM). During the generation process, it decodes each instance region with iso-spacing value in ISIM—corresponding to both background and foreground instances—to produce RS images through the denoising process of diffusion models. Considering the complex interdependencies among MMOs, we construct a spatial-cross dependency knowledge graph (SCDKG). This ensures a realistic and reliable multidirectional distribution among MMOs for region embedding, thereby reducing the discrepancy between source and target domains. Besides, we propose a structured object distribution instruction (SODI) to guide the generation of synthesized RS image content from a global aspect with SCDKG-based ISIM together. Extensive experimental results demonstrate that our MMO-IG exhibits superior generation capabilities for RS images with dense MMO-supervised labels, and RS detectors pre-trained with MMO-IG show excellent performance on real-world datasets. Code is available at https://github.com/omtcyang/MMO-IG.**

***Index Terms*—Remote sensing, image generation, object detection, diffusion model.**

I. Introduction

OBJECT detection in remote sensing (RS) images [1]–[4] is an important task in earth observation technology, providing essential support for various application scenarios, such as disaster monitoring, military reconnaissance, traffic monitoring, and smart cities. This task aims to determine object categories and locations based on input optical RS images. With the progress of deep learning-based generic detection technology [5]–[8], remote sensing image object detection (RSIOD) research has achieved significant development, where analyzing complex backgrounds and focusing on varied class and scale foreground objects simultaneously is becoming achievable as long as the algorithm that is well-trained based on sufficient and high-quality data. However, the high cost of acquiring satellite RS images and the labor-intensive nature of image annotation limit the availability of data for this research area, making it challenging to adequately train algorithms. This widespread data limitation issue in the field of RS has consistently hindered researchers from making further breakthroughs.

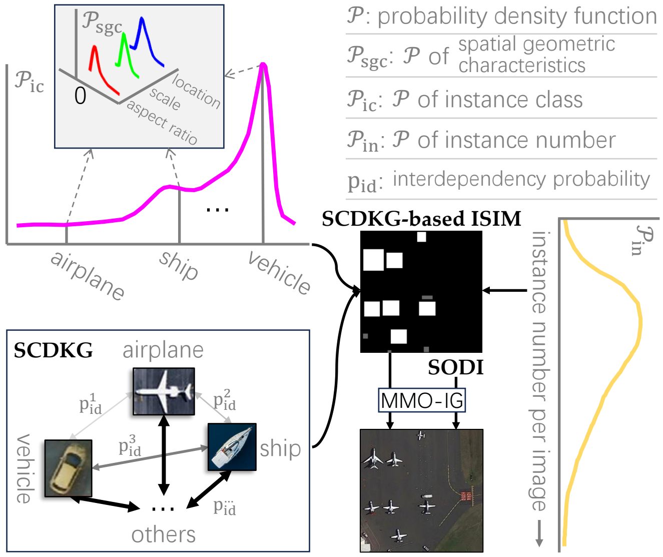
$\mathcal{P}_{sgc}$
0
location
scale
aspect ratio
$\mathcal{P}_{ic}$
...
airplane
ship
vehicle
$\mathcal{P}$: probability density function
$\mathcal{P}_{sgc}$: $\mathcal{P}$ of spatial geometric characteristics
$\mathcal{P}_{ic}$: $\mathcal{P}$ of instance class
$\mathcal{P}_{in}$: $\mathcal{P}$ of instance number
p_{id}: interdependency probability
SCDKG-based ISIM
$\mathcal{P}_{in}$
instance number per image
SODI
MMO-IG
SCDKG
airplane
p_{id}^1
p_{id}^2
p_{id}^3
$p_{id}^{\cdots}$
vehicle
ship
...
others

Fig. 1. Illustration of the generation of RS images containing MMOs by the proposed MMO-IG. Notably, each RS object is modeled by a unique $\mathcal{P}_{\mathrm{sgc}}$ according to the corresponding realistic geometric characteristics.

Recently, the rapidly developed deep generative models (DGMs) [9]–[11] provide a solution for alleviating the above issue. They aim to generate realistic and high-quality images. It enhances the accessibility of data in resource-constrained environments significantly, which makes it possible to build RS generative models (RSGMs) [12], [13] for alleviating the data limitation of RSIOD. While existing RSGMs follow generic DGMs to focus on the achievement of text-to-image or layout-to-image [14]–[17], where the former generates RS images according to text descriptions prompt and the latter is designed to render RS images based on a given layout image prompt (e.g., depth map, sketch, road, edge, etc.). Traditional approaches relying on text or layout prompts face three fundamental limitations when handling Multi-Class and

This work was supported by the National Natural Science Foundation of China under Grant 62471394 and U21B2041.

Chuang Yang and Qi Wang are with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

Bingxuan Zhao and Qing Zhou are with the School of Computer Science, and with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

E-mail: omtcyang@gamil.com, bxuanzhao202@gmail.com, chautsing@gmail.com, crabwq@gmail.com.

Chuang Yang and Bingxuan Zhao contributed equally to this work.

Qi Wang is the corresponding author.

Multi-Scale Objects (MMOs): (1) Layout prompts struggle to encode instance-specific scale variations and precise spatial relationships, often producing rigid object arrangements inconsistent with real RS scenes; (2) Existing methods prioritize global prompt-image alignment while neglecting local interdependencies between adjacent objects; (3) Fixed spatial constraints in layout guidance limit flexible adaptation to diverse RS observation geometries. It means they are hard to provide qualified data with instance-level labels that can model complex interdependencies among different RS objects for the RSIOD task.

Following the above issue, we introduce an instance-aware image generator for providing sufficient and high-quality data with multi-class and multi-scale object (MMO) supervision information for RS detection, namely MMO-IG. To overcome the limitations of traditional guidance signals, we propose the Iso-spacing Instance Map (ISIM) – a novel control mechanism where distinct grayscale regions with unique iso-spacing values encode different object classes. This approach preserves three critical instance attributes: (1) Spatial coordinates through region centroids, (2) Object scales via region areas, and (3) Class identities by grayscale intensities, enabling simultaneous control of multiple object instances while maintaining flexible placement. Specifically, inspired by popular diffusion-based controlled generative models [11], [18]–[20], we formulate the MMO-IG as a controllable generation problem. Different from previous methods that were designed to generate RS images whose content adheres to global layout requirements, the proposed MMO-IG is demanded to focus on each object instance's class, location, and scale in a local view while considering the interdependencies among different instances from a global aspect. Considering that the existing text and layout prompt cannot represent MMOs, an iso-spacing instance map (ISIM) is introduced as the control signal for guidance in our model in generating RS images containing MMPs, where instances with different classes are encoded via various iso-spacing valued grayscale regions. During testing, MMO-IG decodes these regions into instances with different classes according to regions' grayscale values while keeping the location and scale characteristics of the corresponding regions on generated instances.

Existing RSGMs predominantly focus on pairwise object relationships, failing to capture the complex network of dependencies in real RS scenes where objects form hierarchical clusters (e.g., vehicle groups around buildings) and exhibit cross-class interactions (e.g., ships near ports). To address this, we develop the Spatial-Cross Dependency Knowledge Graph (SCDKG) that explicitly models four types of spatial relationships: (1) Intra-class proximity constraints, (2) Cross-class attraction/repulsion forces, (3) Orientation-aware co-occurrence patterns, and (4) Scale compatibility rules.Besides, distinct from previous RSGMs concentrate on pursuing a correct corresponding between the given text or layout prompt and the generated RS image, there exist complex interdependencies among different objects with the same class and also in different classes of objects, which demands our model to ensure a rational interdependent distribution among the MMOs on the generated RS image. Based on this consideration, we construct a spatial-cross dependency knowledge graph (SCDKG) to formulate the complex interdependencies among different RS objects. During testing, we synthesize an ISIM with rational interdependent distribution among the MMOs under the guidance of SCDKG to render the RS image via MMO-IG more accurately reflective of reality. Based on the above proposed ISIM and SCDKG, we construct MMO-IG for generating mass RS images with dense instance-level labels, which provides a solution for alleviating the data limitation problem that exists in the RSIOD field. Furthermore, the hallucination problem in deep generative models (DGMs) can cause remote sensing (RS) objects to appear incorrectly in the background, leading to discrepancies between instance-level labels (i.e., ISIM) and the generated RS images, which negatively impacts downstream tasks. To address this issue, SODI is introduced to guide the image generation process by considering the global perspective of image style and instance characteristics, ensuring both accuracy and control.

In summary, the contributions of our work are fourfold:

1) An iso-spacing instance map (ISIM) is designed as the control signal for guidance in generating RS images filled with dense MMOs. ISIM is an effective and intuitive encoding method for RS objects with different classes, locations, and scales, which helps the model comprehend objects' geometric characteristics for generating high-quality data.
2) A spatial-cross dependency knowledge graph (SCDKG) is built, which formulates the complex interdependencies among different RS objects for providing support in synthesizing ISIM with a rational spatial distribution for the MMOs. It helps render the RS image more accurately reflective of reality.
3) A SODI is introduced to ensure the alignment between ISIM and the generated RS image. It assists in regulating image content to align with the remote sensing style. Meanwhile, the object statistics description ensures strict correspondences between MMOs in ISIM and the generated remote sensing image, which facilitates an accurate and controllable image generation process.
4) An ISIM and SCDKG-based RS image generation model, termed MMO-IG, is developed to decode synthetic SCDKG-guided ISIM with dense instance regions into RS images through a denoising process. This model provides abundant training data for RSIOD, helping to alleviate existing data limitations such as scarcity and sample imbalance.

The remainder of the paper is structured as follows: Section II reviews the contributions of scholars in related fields, specifically DGMs, and RSGMs, and evaluates the strengths and limitations of existing algorithms, which inform the design of MMO-IG. Section III presents a detailed visualization of the proposed MMO-IG. Section IV discusses the results of ablation studies and comparative analyses. Finally, Section V summarizes the paper's contributions to the data limitation problem of RSIOD.

II. Related Work

The rapid development of DGMs progress RSGMs significantly. In this section, we introduce the related works in the research of DGMs and RSGMs briefly.

A. Remote Sensing Image Object Detection

Object detection is a key topic in computer vision, which provides sufficient support to analyze image context. Inspired by generic object detection approaches [21]–[26], researchers designed RSIOD frameworks according to the specific geometric characteristics of RS objects, which become critical techniques for RS scenario applications. Existing RSIOD methods can be classified into small [3], [27], [28] and oriented [2], [29], [30] object detection methods roughly according to the object characteristics. The former concentrated on designing the feature fusion structure [31]–[33] or introducing attention mechanism into frameworks [34]–[36] to obtain strong expression ability to help the model distinguish small targets from complex backgrounds more effectively. The latter primarily aimed to represent oriented object boundaries [37], [38] accurately to avoid superfluous background interference. Besides, some works [39]–[41] focus on optimizing the whole detection framework to achieve performance balance between accuracy and efficiency. Although these methods achieve competitive performance on multiple public datasets, the limited availability of data in many scenarios hinders their ability to maintain superior performance.

B. Deep Generative Models

Deep generative models (DGMs) are designed for creating realistic synthetic data and enhancing data-driven applications across various fields. Initially, variational autoencoders (VAE) [42]–[44] and generative adversarial networks (GAN) [45]–[47] advanced DGM research considerably and had gained increasing influence in computer vision tasks. However, VAEs often produce blurry outputs due to their reliance on Gaussian assumptions, while GANs can be challenging to train and are prone to mode collapse. In contrast, diffusion models [9]–[11], [48] addressed these limitations by employing a stepwise noise reduction process that yields sharper and more stable outputs, capturing researchers' attention. Specifically, DDPM [9] built on foundational concepts introduced in the original diffusion model [48] by refining the diffusion process to improve sample quality and stability while maintaining flexibility for various tasks. Stable Diffusion [10] introduced a latent space to optimize computational efficiency and accelerate image generation processes. However, these methods lack advanced control mechanisms, limiting their ability to precisely guide image generation. Addressing these issues, ControlNet [11] enhanced generated outputs by integrating additional neural networks that conditionally guide the generation process based on specific input features (e.g., canny edge, human pose, and sketch, etc.), which provided an effective solution for alleviating the data limitation problem exists in remote sensing image analysis.

C. Remote Sensing Generative Models

In remote sensing image analysis research, data acquisition is particularly challenging compared to other computer vision fields, leading to a more severe issue of data scarcity. Based on the above consideration, researchers introduce DGMs into remote sensing image analysis. The authors [13] generate multiple sets of pseudo-labeled samples from real data by SinGAN [49] while a novel quantitative sifting metric is then applied to assess the authenticity and diversity of these samples, enabling the selection of the most optimal pseudo-labeled samples for enhancing model training. To achieve image directional generation, D-SGAN [50] takes a rough segmentation map as input to guide the generation process. RSDiff [51] leverages diffusion models to generate remote sensing images from textual descriptions, enhancing the synthesis quality and semantic alignment between text and imagery. CRS-Diff [12] enables precise manipulation of image attributes through guided compound control conditions, such as the integration of text, depth maps, sketches, and road layouts, thereby enhancing the controllability and specificity of remote sensing image generation. Existing RSGMs pursue text-to-image or layout-to-image based on DDPM [9] or GAN [45] and achieve superior performance. However, a generation model enables provide sufficient data with instance-level labels for the RSIOD task, which is still under-explored.

III. Methodology

RSIOD is an important task for plenty of applications (e.g., disaster monitoring, military reconnaissance, traffic monitoring, and smart cities) in the field of remote sensing image analysis. To alleviate the data limitation problem in RSIOD research, we propose MMO-IG in this paper, the method details will be described in the following paragraphs.

A. Overall Pipeline

An image generation method, namely MMO-IG, tailored for the RSIOD task has been proposed. The overall pipeline of MMO-IG is visualized in Fig. 2. It can be found that the whole generation process consists of three steps: 1) initializing MMOs from SCDKG; 2) synthesizing ISIM and SODI from MMOs; 3) decoding SCDKG-based ISIM to the RS image. On the whole, MMO-IG decodes the regions with different grayscale values in ISIM to instances with different classes while keeping the regions' spatial geometric characteristics (i.e., aspect ratios, scales, and locations). Meanwhile, it ensures the whole image style under the guidance of SODI. Considering the complex spatial-cross dependency relationships among dense objects, SCDKG is designed to ensure a rational instance-level layout that corresponds to realistic scenes. In the following subsections, we will describe SCDKG, ISIM, and SODI in detail respectively.

B. Spatial-Cross Dependency Knowledge Graph

As previously introduced, the proposed MMO-IG aims to decode regions with varying grayscale values in ISIM into RS objects. Unlike generic DGMs that focus on realistic salient

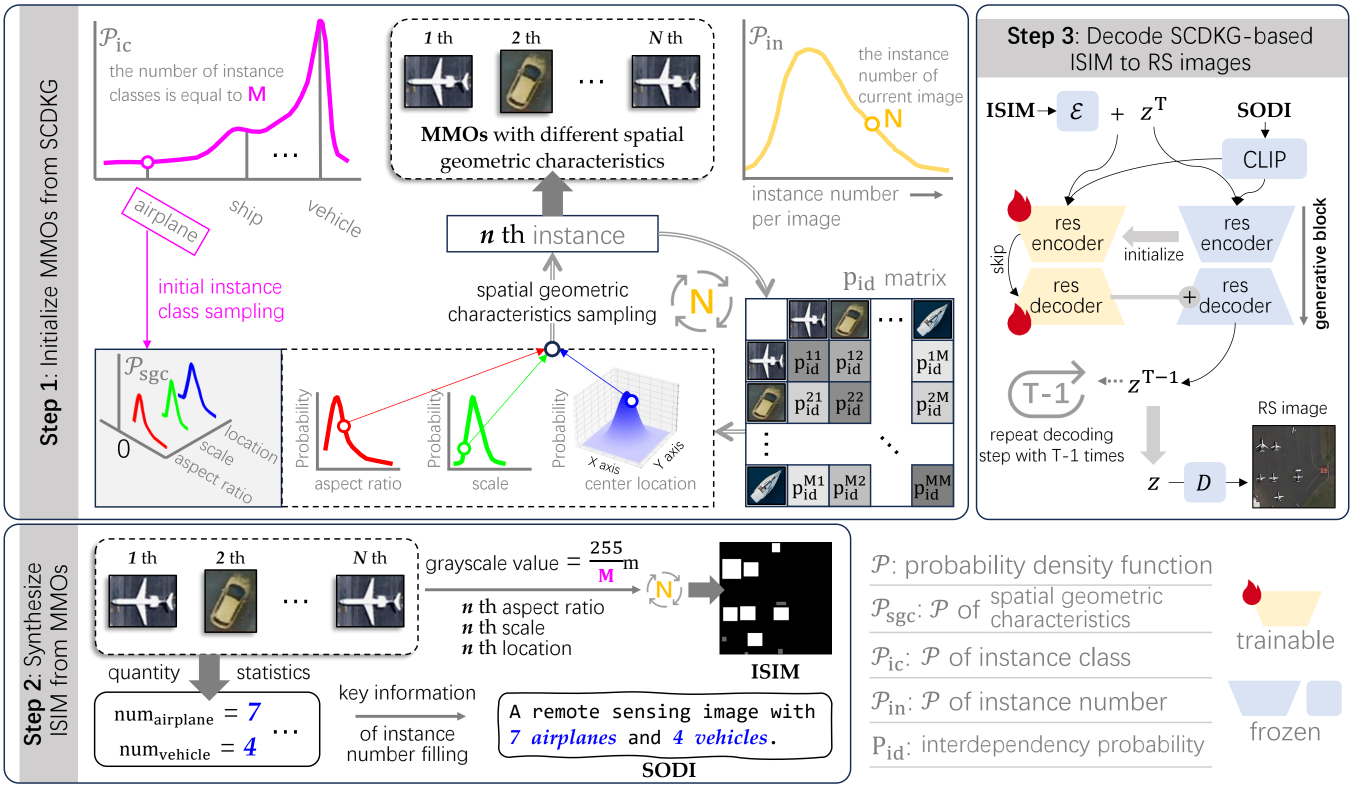
Step 1: Initialize MMOs from SCDKG
$\mathcal{P}_{\mathrm{ic}}$
the number of instance classes is equal to M
...
airplane
ship
vehicle
initial instance class sampling
$\mathcal{P}_{\mathrm{sgc}}$
0
location
scale
aspect ratio
1 th
2 th
N th
MMOs with different spatial geometric characteristics
n th instance
spatial geometric characteristics sampling
N
Probability
aspect ratio
scale
center location
X axis
Y axis
$\mathcal{P}_{\mathrm{in}}$
the instance number of current image
instance number per image
p_{id} matrix
$\mathrm{p}_{\mathrm{id}}^{11}$
$\mathrm{p}_{\mathrm{id}}^{12}$
$\mathrm{p}_{\mathrm{id}}^{1M}$
$\mathrm{p}_{\mathrm{id}}^{21}$
$\mathrm{p}_{\mathrm{id}}^{22}$
$\mathrm{p}_{\mathrm{id}}^{2M}$
$\mathrm{p}_{\mathrm{id}}^{M1}$
$\mathrm{p}_{\mathrm{id}}^{M2}$
$\mathrm{p}_{\mathrm{id}}^{MM}$
Step 3: Decode SCDKG-based ISIM to RS images
ISIM
$\mathcal{E}$
+
z^{T}
SODI
CLIP
res encoder
initialize
res encoder
skip
res decoder
res decoder
generative block
T-1
$z^{\mathrm{T-1}}$
repeat decoding step with T-1 times
RS image
z
D
Step 2: Synthesize ISIM from MMOs
1 th
2 th
N th
grayscale value = $\frac{255}{\mathrm{M}}$m
n th aspect ratio
n th scale
n th location
N
ISIM
quantity
statistics
$\mathrm{num}_{\mathrm{airplane}}$ = 7
$\mathrm{num}_{\mathrm{vehicle}}$ = 4
...
key information of instance number filling
A remote sensing image with 7 airplanes and 4 vehicles.
SODI
$\mathcal{P}$: probability density function
$\mathcal{P}_{\mathrm{sgc}}$: $\mathcal{P}$ of spatial geometric characteristics
$\mathcal{P}_{\mathrm{ic}}$: $\mathcal{P}$ of instance class
$\mathcal{P}_{\mathrm{in}}$: $\mathcal{P}$ of instance number
p_{id}: interdependency probability
trainable
frozen

Fig. 2. Overall pipeline of MMO-IG for generating RS images with dense instance-level bounding box labels. It first synthesizes rational spatial geometric characteristics of MMOs via SCDKG. They then are encoded via the designed ISIM while describing the RS image content through SODI. In the end, following the diffusion model to decode the ISIM to RS image contained MMOs under the guidance of SODI. Notably, each RS object is modeled by a unique $\mathcal{P}_{\mathrm{sgc}}$ according to the corresponding realistic geometric characteristics.

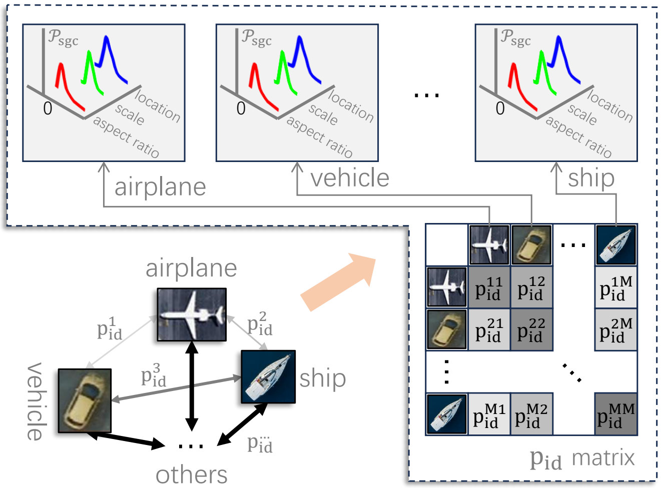
$\mathcal{P}_{\mathrm{sgc}}$
0
location
scale
aspect ratio
...
airplane
vehicle
ship
$\mathrm{p}_{\mathrm{id}}^{11}$
$\mathrm{p}_{\mathrm{id}}^{12}$
$\mathrm{p}_{\mathrm{id}}^{1M}$
$\mathrm{p}_{\mathrm{id}}^{21}$
$\mathrm{p}_{\mathrm{id}}^{22}$
$\mathrm{p}_{\mathrm{id}}^{2M}$
$\mathrm{p}_{\mathrm{id}}^{M1}$
$\mathrm{p}_{\mathrm{id}}^{M2}$
$\mathrm{p}_{\mathrm{id}}^{MM}$
p_{id} matrix
airplane
$\mathrm{p}_{\mathrm{id}}^{1}$
$\mathrm{p}_{\mathrm{id}}^{2}$
$\mathrm{p}_{\mathrm{id}}^{3}$
$\mathrm{p}_{\mathrm{id}}^{\mathrm{ii}}$
vehicle
ship
...
others

Fig. 3. Illustration of the proposed SCDKG, which models complex interdependencies among objects of different classes via p_{id} matrix and their diverse spatial geometric characteristics via $\mathcal{P}_{\mathrm{sgc}}$. Notably, each RS object is modeled by a unique $\mathcal{P}_{\mathrm{sgc}}$ according to the corresponding realistic geometric characteristics.

Algorithm 1 Initialize MMOs from SCDKG

```
Require: The probability density functions $\mathcal{P}_{\mathrm{ic}}$, $\mathcal{P}_{\mathrm{in}}$, and SCDKG ($\mathrm{p}_{\mathrm{id}}$ matrix and $\mathcal{P}_{\mathrm{sgc}}$);
Ensure: object list $\mathrm{L}_{\mathrm{obj}}$ of MMOs for synthesizing ISIM;
  $N \leftarrow$SAMPLE($\mathcal{P}_{\mathrm{in}}$)
  for $n \leftarrow 1$ to $N$ do  //N is instance number per image
      class $\leftarrow$SAMPLE($\mathcal{P}_{\mathrm{ic}}$)
      aspect ratio $\leftarrow$SAMPLE($\mathcal{P}_{\mathrm{sgc}}^{\mathrm{aspect\ ratio}}$, class)
      scale $\leftarrow$SAMPLE($\mathcal{P}_{\mathrm{sgc}}^{\mathrm{scale}}$, class)
      location $\leftarrow$SAMPLE($\mathcal{P}_{\mathrm{sgc}}^{\mathrm{location}}$, class)
      $\mathrm{L}_{\mathrm{obj}} \leftarrow$ (class, aspectratio, scale, location)
      class $\leftarrow$SAMPLE($\mathrm{p}_{\mathrm{id}}$ matrix)
  end for
  return $\mathrm{L}_{\mathrm{obj}}$
```

object details or existing RSGMs that align layout-controlled conditions with corresponding ground truth, MMOs distributed on RS images require MMO-IG to address complex interdependencies among objects of different classes and their diverse spatial geometric characteristics (as illustrated in Fig. 3). This approach ensures a rational and realistic content distribution for the generated RS images. In light of these considerations, SCDKG is designed to model the intricate spatial relationships among RS objects.

SCDKG is tasked with determining MMOs with distinct spatial geometric characteristics for synthesizing ISIM (the corresponding procedures are detailed in Step 1 of Fig.2 and Algorithm 1). Specifically, SCDKG initially samples the RS object class based on the instance class probability density function $\mathcal{P}_{\mathrm{ic}}$, which describes the distribution of RS objects across various types within a given dataset.

Upon determining the initial object class, SCDKG proceeds to define object attributes (including aspect ratio, scale, and location) by adhering to the spatial geometric characteristics of realistic RS objects and the initial class. The spatial geometric characteristics are modeled through three core probability functions: $\mathcal{P}_{\rm sgc}$ (Spatial Geometric Characteristics) describes instance attributes through location, scale, and aspect ratio distributions. Specifically, the location function fits center point coordinates, the scale function models object lengths, and the length-to-width ratio function determines size proportions. Meanwhile, $\mathcal{P}_{\rm ic}$ (Probability of Instance Class) governs class occurrence likelihoods, $\mathcal{P}_{\rm in}$ (Probability of Instance Number) regulates per-class instance counts, and $\rm p_{id}$ (Probability of Interdependency) quantifies class co-occurrence relationships.

Concretely, it models each class object through the spatial geometric characteristics probability density function $\mathcal{P}_{\rm sgc}$. For attributes like aspect ratio and scale, SCDKG employs a one-dimensional probability density function after analyzing the relevant geometric features of all instances in the dataset. The location attribute is modeled by decomposing the instance center point coordinates into two dimensions, x and y, and fitting them with a two-dimensional probability density function. Utilizing $\mathcal{P}_{\rm sgc}^{\rm all}$ for all RS objects, SCDKG assigns attributes according to the class-specific $\mathcal{P}_{\rm sgc}$.

After establishing the initial object class and attributes, the selection of the subsequent object class must consider existing interdependencies among different classes. Sampling directly from $\mathcal{P}_{\rm ic}$ would overlook these relationships. For instance, the likelihood of 'ship' and 'harbor' co-occurring is higher than that of 'airplane' and 'harbor'. Thus, ensuring a rational spatial distribution is crucial to bridging the gap between generated RS images and real samples, thereby maintaining the authenticity of synthesized data. SCDKG constructs a bidirectional graph to represent these interdependencies, where each node denotes an instance class present in the RS images. A directed edge from node A to node B signifies the interdependency probability $\rm p_{id}$ of class B objects on class A objects.

The $\rm p_{id}$ values for different object class pairs form an interdependency probability matrix (illustrated in Fig.2 $\rm p_{id}$ matrix). Based on the previous instance class, SCDKG samples the next instance class from this matrix and assigns attributes through the corresponding $\mathcal{P}_{\rm sgc}$. Finally, the model determines a rational instance density (i.e., 'N' in Step 1 of Fig. 2) using the probability density function $\mathcal{P}_{\rm in}$.

C. Iso-Spacing Instance Map

Different from previous methods, the proposed MMO-IG is proposed to focus on the generation of RS images with various MMOs, which demands the control condition to represent all kinds of objects while ensuring the differences among them. Considering the generic control conditions (e.g., depth map, segment map, and edge map) lacks the ability to distinguish objects with different classes, ISIM is introduced in this paper (as shown in Fig. 4).

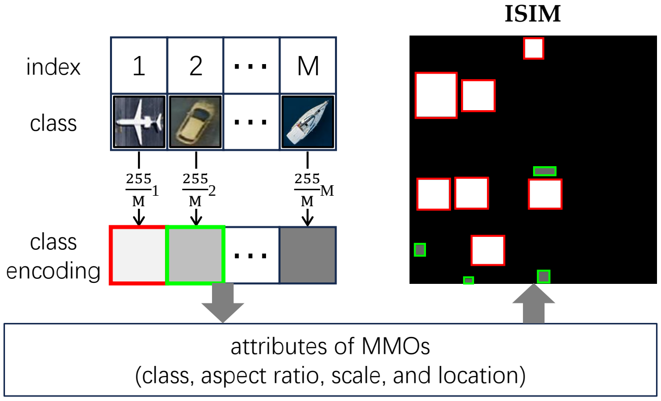

Fig. 4. Illustration of the proposed ISIM encodes instances with different classes according to different grayscale values while keeping the location and scale characteristics of the corresponding regions on generated instances.

Given a remote sensing (RS) dataset containing M classes, the ISIM method initially assigns each class a unique numerical identifier ranging from $[0 \sim M]$. This identifier is then used in an arithmetic interval grayscale assignment strategy. The grayscale value corresponding to each class can be computed as the following equation:

$$v_{gray}(m) = \left\lfloor \frac{255}{M} m \right\rfloor, \tag{1}$$

where m represents the unique numerical identifier for each class instance and v_{gray} is the corresponding grayscale value. $\lfloor \cdot \rfloor$ is floor function (round down to the nearest integer. With the determined class encoding for all kinds of instances in a specific dataset, MMO attributes from Step 1 in Fig. 2 are utilized to synthesize ISIM. Concretely, ISIM encodes each object based on regions corresponding to the class-specific grayscale values and spatial geometric characteristics (i.e., aspect ratio, scale, and location) on a grayscale image.

D. Structured Object Distribution Instruction

As we illustrated before, ISIM is designed to represent the spatial attributes of different RS objects and keep distinguishment among them. However, in the background region of ISIM, some RS objects would occur because of the hallucination problem of DGMs, which leads to a discrepancy between the instance-level labels (ISIM) and the generated RS image and further negatively impacts downstream tasks.

Therefore, SODI is introduced to ensure the alignment between ISIM and the generated RS image. As shown in Fig. 5, with the determined MMOs with different spatial geometric characteristics from Step 1 in Fig. 2, the SODI generation process first counts the number of objects of different classes and fills the information into the statistics template for obtaining a statistics description about MMOs that from SCDKG. Then, the class object with a quantity of zero in the statistics description is filtered out and the filtered description is combined with a structured scene head description (" A remote sensing image with") for generating the SODI. The scene head description assists our model in regulating image content to align with the remote sensing style. Meanwhile, the filtered statistics description ensures that MMO-IG maintains strict alignment between MMOs in ISIM and the generated remote sensing images. This guidance, from

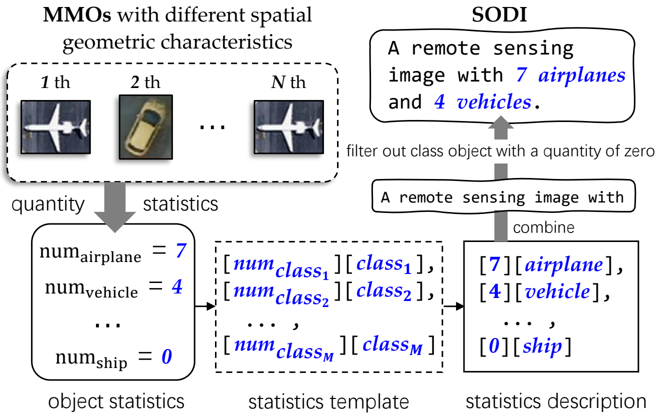

Fig. 5. Illustration of the proposed SODI generation process. It consists of the combination of a structured scene head description (“ A remote sensing image with”) and a statistics description of RS objects.

a global perspective, facilitates an accurate and controllable image generation process.

E. Decoding ISIM to Generate RS Image

The popular DGMs, diffusion models [9]–[11], [48], reconstruct images via a denoising process on noisy counterparts. Considering the diffusion models’ advantages of stable training, high-quality sample generation, tractable likelihood estimation, versatility, and robustness to hyperparameter settings compared with VAE [42] and GAN [45] models, MMO-IG follows the stable diffusion [10] structure to construct the corresponding generative network. As shown in Step 3 of Fig. 2, the generative part consists of two couples of res encoder and res decoder mainly.

In designing our residual encoder (res encoder) and residual decoder (res decoder), we drew inspiration from the architecture of the stable diffusion model. Specifically, the res encoder and res decoder each contain 12 blocks, with the entire model comprising 25 blocks, including the middle block. Among these 25 blocks, 8 blocks are convolutional layers responsible for downsampling or upsampling, while the remaining 17 blocks are the main functional blocks. Each main block consists of 4 residual network layers and 2 Vision Transformers (ViTs). Notably, each ViT incorporates multiple cross-attention and self-attention mechanisms, which further enhance the model’s expressive power and flexibility.

The encoder and decoder are composed of 12 stable diffusion layers respectively and they are connected by a middle layer that of 1 stable diffusion layer. The frozen generative blocks are initialized through the pre-trained weights from stable diffusion [10]. The structures of trainable generative blocks are copied from the right ones and the two block weights are initialized by coping the right block and zero, respectively.

In the generation process, the generative network takes ISIM, SODI, and z^T as input to achieve the RS images, where ISIM and SODI are embedding into feature vectors via a lightweight network ε and CLIP [52] respectively, where ε consists of four convolution layers with 4×4 kernels and 2×2 strides $\mathrm{Conv}_{4,2}^{16}, \mathrm{Conv}_{4,2}^{32}, \mathrm{Conv}_{4,2}^{64}, \mathrm{and Conv}_{4,2}^{128}$. z^T is the feature vector of noisy image that can be efficiently obtained by conducting the well-trained ε on the noisy image. After T-1 times iterative denoising through res blocks, MMO-IG decoding each region in ISIM into the corresponding object with the decoded image feature vector z, and the final RS image reconstructed by conducting D (that proposed in [53]) on z.

Given the clean input image z_0, diffusion models gradually corrupt the image through t successive noising steps, producing the noisy latent z_t where $t \in \{1, ..., T\}$ denotes the timestep index. Conditioned on the timestep t, structured object distribution instruction (SODI) c_t, and iso-spacing instance map (ISIM) c_f, the model learns a denoising network ϵ_θ to estimate the injected noise in z_t. This process is formalized as:

$$\mathcal{L} = \mathbb{E}_{z_0,t,c_t,c_f,\epsilon\sim\mathcal{N}(0,\mathbf{I})} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)\|_2^2 \right] \tag{2}$$

where $\mathcal{L}$ denotes the loss function for training the diffusion model.

IV. Experiments

In this section, we detail the dataset configuration used for our experiments, present the experimental results of our MMO-IG model, and conduct an ablation study. Additionally, we design downstream experiments to demonstrate the effectiveness of the generated remote sensing multi-object detection data using this method, achieving results comparable to those obtained with real remote sensing images.

A. Datasets and Experimental Setup

Datasets. We utilized the DIOR and DIOR-R [54] dataset to train and evaluate our model. This dataset comprises 23,463 high-quality remote sensing images and 190,288 meticulously annotated object instances, resulting in a total of 192,472 axis-aligned object annotations. Each image is 800×800 pixels, with spatial resolutions ranging from 0.5 to 30 meters. The dataset is divided into a training and validation set (11,725 images) and a test set (11,738 images). DIOR serves as a comprehensive benchmark for object detection in optical remote sensing images, encompassing 20 object classes, including **A**irplane, **A**irport, **B**aseball, **B**asketball, **B**ridge, **C**himney, **D**am, **E**xpressway service area, **E**xpressway toll station, **G**olf field, **G**round track, **H**arbor, **O**verpass, **S**hip, **S**tadium, **S**torage tank, **T**ennis court, **T**rain station, **V**ehicle, **W**indmill.

Experimental Setup. Our model is trained using the Adam optimizer with a learning rate of 1e-5. Training is conducted on four NVIDIA Tesla A100 GPUs, each with 80 GB of memory, and takes approximately two days to complete. During the sampling phase, six samples are generated, with a control parameter (scale) set to 2.5 to enhance sampling quality and refine the model’s outputs. This configuration allows for efficient utilization of computational resources, achieving high performance in both the training and sampling stages.

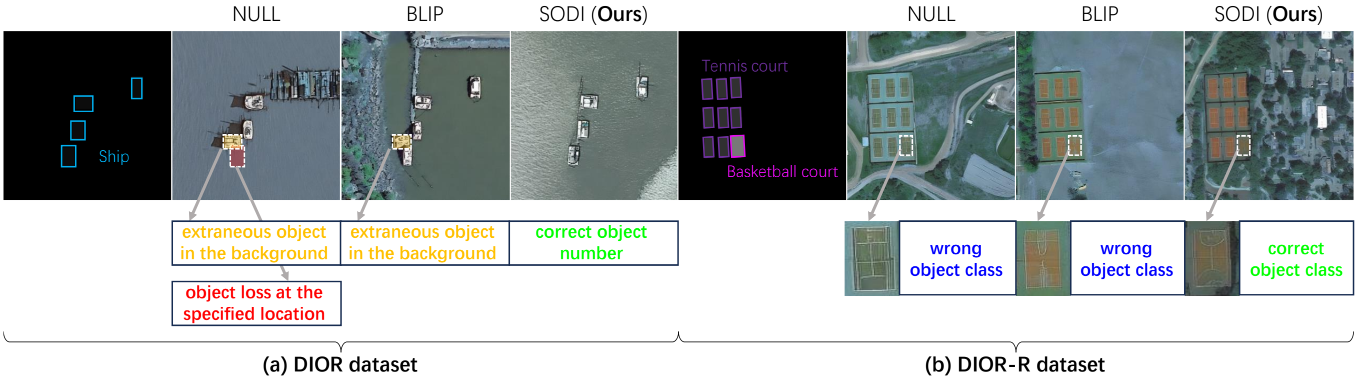

Fig. 6. Visualization comparison of the generated image by our MMO-IG with different text conditions on the DIOR dataset. Considering the grayscale values of some objects are too low and difficult to discern, we enhance their visibility by marking the object regions with colored bounding boxes in the ISIM.

TABLE I

RESULTS ON MMO-IG WITH DIFFERENT TEXT CONDITIONS. 'NULL' AND 'BLIP' DENOTE THE TEXT CONDITION IS NULL AND AN IMAGE DESCRIPTION GENERATED BY BLIP [55]. $\mathrm{Acc_c}$ AND $\mathrm{Acc_n}$ ARE THE ACCURACY OF THE OBJECT CLASS AND NUMBER THAT ARE EVALUATED ON THE GENERATED RS IMAGES.

dataset	text condition	$\mathrm{Acc_c}$ ↑	$\mathrm{Acc_n}$ ↑	FID ↓	CAS↑
DIOR	NULL	92.8	93.9	128.4	27.1
	BLIP	92.3	94.5	75.6	38.4
	SODI (**Ours**)	**97.9**(5.6)	**98.7**(4.2)	**65.6**(10.0)	**45.8**(7.4)
DIOR-R	NULL	91.7	93.7	134.6	28.3
	BLIP	93.4	93.1	69.8	36.9
	SODI (**Ours**)	**98.2**(4.8)	**97.2**(4.1)	**64.8**(5.0)	**47.2**(10.3)

B. Ablation Study

Effectiveness of SODI. Considering the hallucination problem in deep generative models (DGMs) can cause remote sensing (RS) objects to be inaccurately represented in the background, resulting in inconsistencies between instance-level labels (ISIM) and the generated RS images (as shown in Fig. 6). When low-quality ISIM is used as a control condition for generating images, it leads to a mismatch between the generated images and the provided SODI. This mismatch negatively affects downstream tasks, such as object detection, since the labels for the generated detection data are derived from the SODI. If the generated images do not align with the SODI, the resulting data becomes unreliable, which can significantly impair the model's performance during training. This discrepancy adversely affects downstream tasks. Therefore, SODI is implemented to guide the image generation process by incorporating a global perspective on image style and instance characteristics, thereby ensuring precision and control. To verify the effectiveness of SODI in ensuring the alignment between ISIM and the generated RS images, we show the testing quantitative metrics of our model with different text conditions in Table I, where 'NULL' and 'BLIP' denote the text condition is null and an image description generated by BLIP [55]. It can be found that MMO-IG with 'NULL' text condition makes it hard to control the accuracy of object class and number on the generated RS images, where the $\mathrm{Acc_c}$ and $\mathrm{Acc_n}$ are 92.8 and 93.9 respectively. It means that ISIM, the instance-level label, enjoys low-quality correspondence with the generated RS images, further influencing the following downstream task. $\mathrm{Acc_c}$ evaluates category alignment between generated instances and SODI-defined classes, while $\mathrm{Acc_n}$ measures numerical consistency in instance counts relative to SODI specifications.

As for the 'BLIP' text condition, it can describe image content in detail, which provides a more controlled ability to achieve a realistic image style. It can be found from Table I, that the 'BLIP' text condition brings more than 50 and 8 improvements in FID and CAS respectively, which means there is a significant gain in the similarity between generated images and real images. However, in generating object detection data, the 'BLIP' text condition often presents challenges. Specifically, it typically emphasizes the overall features of the image, such as color, background, or scene, while paying insufficient attention to the categories and quantities of instances within the image. For instance, BLIP might generate a descriptive text like "A beautiful aerial view of a city with roads and buildings," but it rarely accurately identifies specific object categories (such as "cars," "tennis courts," etc.) or the number of objects in the image. Consequently, the text generated by BLIP may lack the precision necessary, especially when such information is crucial for training object detection models. It leads to the problem of low-quality correspondence between ISIM and the generated RS image is still under-alleviated. The above phenomenon can be verified by the results in Table I, compared with 'NULL' text condition, the $\mathrm{Acc_c}$ and $\mathrm{Acc_n}$ of the model with 'BLIP' text condition not achieve significant improvements, and is even slightly degraded.

Different from the 'BLIP' text condition, SODI begins with "a remote sensing image with," followed by a detailed description based on the actual instance categories and quantities present in the image. For example, if there are 37 ships and 3 harbors in the image, the generated prompt would be a remote sensing image with 37 ships, and 3 harbors, if there are 4 bridges and 3 ground track fields in the image, the generated prompt would be: a remote sensing image with 4 bridges, 3 ground track fields. It ensures that the generated text not only accurately describes the overall features of the image but

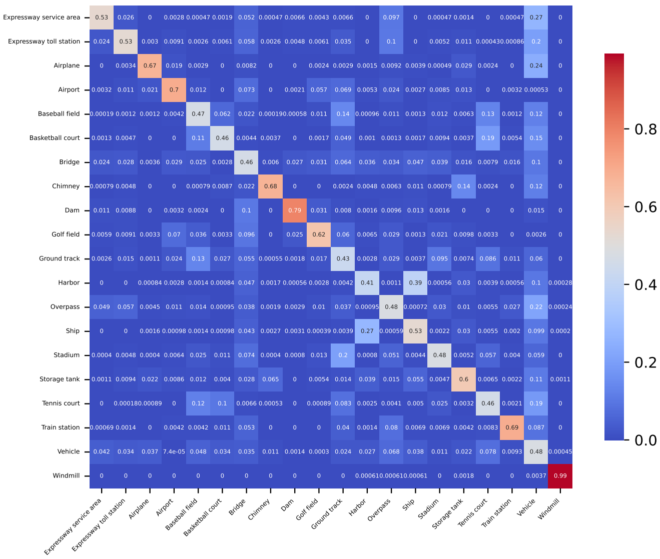

Fig. 7. Visualization of the $\mathrm{p_{id}}$ matrix of SCDKG on DIOR dataset. There are 20 classes in the dataset and the value of each element represents the interdependencies between the corresponding two classes.

also provides a detailed account of the specific object types and quantities contained within the image, thereby offering more precise training data for the model. Models trained using this method can effectively eliminate errors related to instance quantity and category during the generation process. During the image generation process, the model no longer makes errors due to inaccurate descriptions of object types or quantities. This method resolves instance recognition errors in the generation of RS images with MMOs, ensuring that the strict correspondences between ISIM and generated images better meet the practical requirements of downstream object detection tasks. These advantages help our MMO-IG achieve 97.9 $\mathrm{Acc_c}$ and 98.7 $\mathrm{Acc_n}$ on the DIOR dataset and 98.2 $\mathrm{Acc_c}$ and 97.2 $\mathrm{Acc_n}$ on the DIOR-R dataset, which enhances the correspondence between ISIM and the generated RS image significantly. Moreover, our model surpasses the 'BLIP' text condition a lot in FID and CAS, which means the SODI's ability to improve the reality of the generated image. The superiority also has been verified by the qualified generated RS image samples in Fig. 6. The above results show that SODI can improve the generation effect, making the image content real, and the alignment between ISIM and generated RS images.

Effectiveness of SCDKG. To address complex interdependencies among objects of different classes and their diverse spatial geometric characteristics (as illustrated in Fig. 3). SCDKG is proposed to ensure a rational and realistic content distribution for the generated RS images.

SCDKG consists of $\mathrm{p_{id}}$ matrix that models interdependencies among all objects of different classes (the $\mathrm{p_{id}}$ matrix of DIOR dataset is shown in Fig. 7) and $\mathcal{P}_{\mathrm{sgc}}$ that formulates diverse spatial geometric characteristics of each object class individually. Notably, $\mathcal{P}_{\mathrm{sgc}}$ models the aspect ratio, scale, and location characteristics of RS objects separately. It effectively avoids influence brought by irrational characteristics (e.g., aspect ratio and positional distribution that do not conform to real-world object spatial geometric characteristics visualized in Fig. 8). To verify the effectiveness of SCDKG, we test the performance enhancement brought by each factor in SCDKG and visualize the results in Table II. For $\mathcal{P}_{\mathrm{sgc}}$, It can be found that all of the factors can bring performance gains for the generated RS images. Specifically, the modeling of the complex interdependencies among different RS objects makes a more rational object distribution, which helps our model achieve 3.8 and 1.1 improvements in $\mathrm{FID_{zs}}$ and CAS on DIOR dataset and 6.8 and 1.0 gains in $\mathrm{FID_{zs}}$ and CAS on DIOR-R dataset. Notably, $\mathrm{FID_{zs}}$ and $\mathrm{CAS_{zs}}$ denote the zero-shot FID and zero-shot CAS, which is introduced to evaluate the quality of the generated images that not in the test set. The improvements brought by $\mathrm{p_{id}}$ matrix demonstrate the effectiveness of the interdependencies modeling.

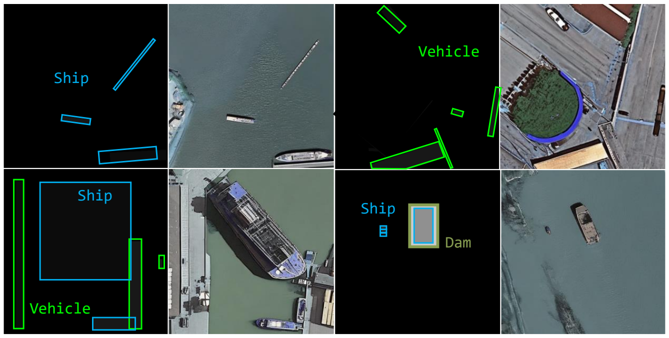

Fig. 8. Visualization of some samples with irrational aspect ratio, scale, and locations without the guidance of SCDKG. Considering the grayscale values of some objects are too low and difficult to discern, we enhance their visibility by marking the object regions with colored bounding boxes in the ISIM.

TABLE II
RESULTS OF MODELS WITH DIFFERENT SCDKG SETTINGS ON DIOR DATASET. $\mathrm{FID_{zs}}$ AND $\mathrm{CAS_{zs}}$ DENOTE THE ZERO-SHOT FID AND CAS METRICS.

#	aspect ratio	scale	location	$\mathrm{p_{id}}$ matrix	DIOR		DIOR-R	
					$\mathrm{FID_{zs}}\downarrow$	$\mathrm{CAS_{zs}}\uparrow$	$\mathrm{FID_{zs}}\downarrow$	$\mathrm{CAS_{zs}}\uparrow$
1					92.3	28.2	94.2	27.7
2				✔	88.5	29.3	87.4	28.7
3	✔			✔	82.7	30.7	83.9	29.8
4		✔		✔	74.5	32.4	73.5	31.9
5			✔	✔	79.7	34.9	78.2	35.3
6	✔	✔		✔	70.2	38.2	71.2	37.9
7		✔	✔	✔	78.9	40.1	74.5	42.3
8	✔		✔	✔	68.7	44.6	67.2	46.2
9	✔	✔	✔	✔	**65.6**	**45.8**	**64.8**	**47.2**

Besides, the rational characteristics of the object aspect ratio, scale, and location bring 8.8, 5.8, and 14.0 improvements in $\mathrm{FID_{zs}}$ and 1.4, 3.1, and 5.6 gains in $\mathrm{CAS_{zs}}$. This is mainly because the synthesized images enjoy more rational and realistic object spatial characteristics than real RS images, which makes it easier for our model to learn the feature distribution and bring significant performance improvements. Meanwhile, the combination of the three characteristics also can bring further improvements in both $\mathrm{FID_{zs}}$ and $\mathrm{CAS_{zs}}$. Specifically, the model equipped with SCDKG gains 26.7 and 17.6 in $\mathrm{FID_{zs}}$ and $\mathrm{CAS_{zs}}$ on DIOR dataset and 29.4 and

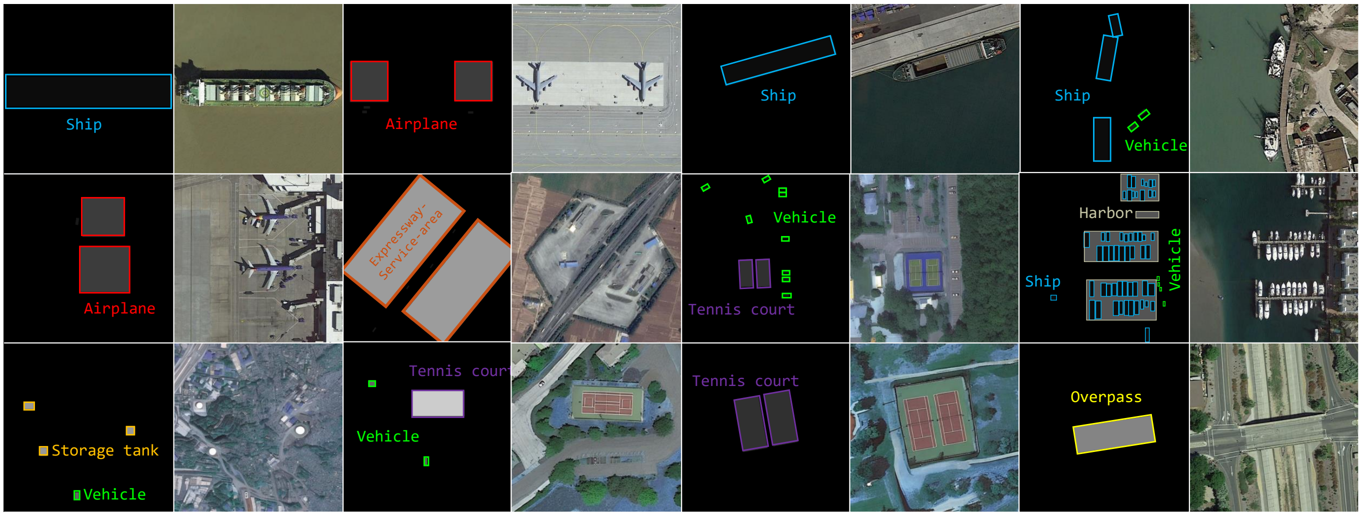

Fig. 9. Visualization of some qualified generated RS images by our MMO-IG. Considering the grayscale values of some objects are too low and difficult to discern, we enhance their visibility by marking the object regions with colored bounding boxes in the ISIM.

TABLE III
COMPARISONS WITH EXISTING LAYOUT-TO-IMAGE METHODS ON DIOR AND DIOR-R DATASETS.

dataset	method	FID ↓	CAS↑
DIOR	LostGAN [17]	57.10	46.02
	Layout Diffusion [15]	45.31	56.98
	ReCo [16]	42.56	55.42
	GLIGEN [14]	41.31	63.50
	MMO-IG (**Ours**)	**34.48**(6.83)	**78.64**(15.14)
DIOR-R	GLIGEN [14]	48.43	58.89
	MMO-IG (**Ours**)	**35.07**(13.36)	**76.13**(17.24)

19.5 in FID_{zs} and CAS_{zs} on DIOR-R dataset. Significant performance improvements verify the effectiveness of the proposed SCDKG in the generation task of RS images containing MMOs. Some qualified samples are shown in Fig. 9, it can be found that SCDKG enables ensuring rational spatial geometric characteristics for objects with different classes to enhance the reality of generated images.

C. Comparisons with Existing DGMs

Different from existing layout-to-image DGMs, our MMO-IG enables us to focus on the complex interdependencies among different RS objects, which helps ensure the rational distribution of the generated RS images. Meanwhile, the SODI can suppress the appearance of objects in the background. The above advantages facilitate MMO-IG's superior performance in the generation process. In this section, we compare our MMO-IG with existing layout-to-image DGMs to show the superiority of our method. As shown in Table III, for horizontal labeled samples (DIOR data), our method achieves 6.83 and 15.14 improvements in FID and CAS evaluation metrics respectively. It is mainly because of the more direct encoding strategy (ISIM) for RS objects with different classes and the more effective constraints on the object quantity and class brought by SODI. The former helps our model decode ISIM regions into instances following the simple correspondence between the classes and grayscale values, which is more intuitive than the way to represent instance class through text description. The latter ensures a strict correspondence of RS objects between ISIM and generated RS images. The above advantages either facilitate MMO-IG to surpass GLIGEN [14] 13.36 and 17.24 in the aspect of FID and CAS respectively on rotated labeled data (DIOR-R). The results demonstrate the superiority of our MMO-IG and the effectiveness of the introduced ISIM and SODI on the generation task of RS images that contain MMOs.

D. Downstream Task

As we mentioned before, RSIOD [1]–[4] is an important task in the research of remote sensing. However, the high cost of acquiring satellite RS images and the labor-intensive nature of image annotation limit the availability of data for this research area, making it challenging to adequately train algorithms. Based on the above consideration, MMO-IG is proposed to alleviate the data limitation problem.

In this section, we demonstrate the effectiveness of the generated remote sensing (RS) images using our MMO-IG model. We generated 20,000 images, which were integrated with the original DIOR dataset to train various models, including R-CNN [7], Faster R-CNN [5], YOLO [21], PANet [23], and CornerNet [25]. These models were then evaluated on the DIOR test set. The results, presented in Table IV and Fig. 10 (for better observing the gains of all kinds of objects), compare models trained solely on the DIOR training set with those trained on the augmented dataset. The first row for each method indicates the accuracy of models trained exclusively on the DIOR dataset, while the second row shows the results with the combined generated images and DIOR dataset.

The findings reveal that the augmented data from our MMO-IG model significantly enhances performance for the vast majority of objects compared to models trained only on the DIOR dataset. For instance, R-CNN, Faster R-CNN, YOLO,

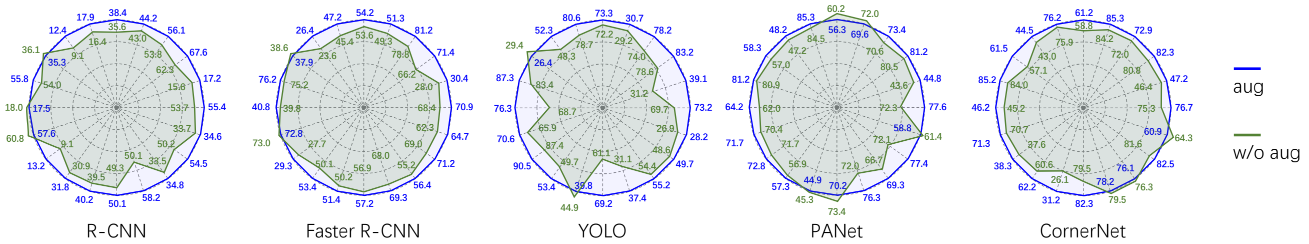

Fig. 10. Comparison of different detectors' accuracy across each category on the DIOR dataset under the setting of data augmentation with 20k generation images (aug) and without augmentation (w/o aug). There are 20 classes of objects (i.e. Airplane, Airport, Baseball, Basketball, Bridge, Chimney, Dam, Expressway service area, Expressway toll station, Golf field Ground track, Harbor, Overpass, Ship, Stadium, Storage tank, Tennis court, Train station, Vehicle, Windmill) in DIOR dataset. The performance metrics for these classes are arranged clockwise from the 12 o'clock position in the diagram.

TABLE IV

RESULTS OF DIFFERENT DETECTORS' ACCURACY ACROSS EACH CATEGORY ON THE DIOR DATASET UNDER THE SETTING OF DATA AUGMENTATION WITH 20K GENERATION IMAGES (AUG) AND WITHOUT AUGMENTATION (W/O AUG). EXPRESSWAY S-A AND EXPRESSWAY T-S DENOTE EXPRESSWAY SERVICE AREA AND EXPRESSWAY TOLL STATION.

methods	aug	Airplane	Airport	Baseball	Basketball	Bridge	Chimney	Dam	Expressway s-a	Expressway t-s	Golf field
R-CNN [7]	✘	35.6	43.0	53.8	62.3	15.6	53.7	33.7	50.2	33.5	50.1
	✔	**38.4**	**44.2**	**56.1**	**67.6**	**17.2**	**55.4**	**34.6**	**54.5**	**34.8**	**58.2**
Faster R-CNN [5]	✘	53.6	49.3	78.8	66.2	28.0	68.4	62.3	69.0	55.2	68.0
	✔	**54.2**	**51.3**	**81.2**	**71.4**	**30.4**	**70.9**	**64.7**	**71.2**	**56.4**	**69.3**
YOLO [21]	✘	72.2	29.2	74.0	78.6	31.2	69.7	26.9	48.6	54.4	31.1
	✔	**73.3**	**30.7**	**78.2**	**83.2**	**39.1**	**73.2**	**28.2**	**49.7**	**55.2**	**37.4**
PANet [23]	✘	60.2	72.0	70.6	80.5	43.6	72.3	61.4	72.1	66.7	72.0
	✔	**56.3**	**69.6**	**73.4**	**81.2**	**44.8**	**77.6**	**58.8**	**77.4**	**69.3**	**76.3**
CornerNet [25]	✘	58.8	84.2	72.0	80.8	46.4	75.3	64.3	81.6	76.3	79.5
	✔	**61.2**	**85.3**	**72.9**	**82.3**	**47.2**	**76.7**	**60.9**	**82.5**	**76.1**	**78.2**

methods	aug	Ground track	Harbor	Overpass	Ship	Stadium	Storage tank	Tennis court	Train station	Vehicle	Windmill	Avg
R-CNN [7]	✘	49.3	39.5	30.9	9.1	60.8	18.0	54.0	36.1	9.1	16.4	37.1
	✔	**50.1**	**40.2**	**31.8**	**13.2**	**57.6**	**17.5**	**55.8**	**35.3**	**12.4**	**17.9**	**37.7**
Faster R-CNN [5]	✘	56.9	50.2	50.1	27.7	73.0	39.8	75.2	38.6	23.6	45.4	53.4
	✔	**57.2**	**51.4**	**53.4**	**29.3**	**72.8**	**40.8**	**76.2**	**37.9**	**26.4**	**47.2**	**54.0**
YOLO [21]	✘	61.1	44.9	49.7	87.4	65.9	68.7	83.4	29.4	48.3	78.7	55.3
	✔	**69.2**	**39.8**	**53.4**	**90.5**	**70.6**	**76.3**	**87.3**	**26.4**	**52.3**	**80.6**	**55.7**
PANet [23]	✘	73.4	45.3	56.9	71.7	70.4	62.0	80.9	57.0	47.2	84.5	62.9
	✔	**70.2**	**44.9**	**57.3**	**72.8**	**71.7**	**64.2**	**81.2**	**58.3**	**48.2**	**85.3**	**63.2**
CornerNet [25]	✘	79.5	26.1	60.6	37.6	70.7	45.2	84.0	57.1	43.0	75.9	62.4
	✔	**82.3**	**31.2**	**62.2**	**38.3**	**71.3**	**46.2**	**85.2**	**61.5**	**44.5**	**76.2**	**62.5**

PANet, and CornerNet exhibit performance gains of up to 8.1, 5.2, 8.1, 5.3, and 5.1 percentage points in detection accuracy, respectively. This underscores the effectiveness of the generated RS images in improving the performance of remote sensing image object detection (RSIOD) models.

Although there is a decline in performance for a small number of object classes (averaging 15% per method), the RS data generated by MMO-IG still provides substantial benefits for these detectors. Additionally, we visualize the detection performance in Fig. 6 to better illustrate the improvements from synthesized data across different object classes. These results demonstrate the capability of MMO-IG to alleviate data limitations in RSIOD.

V. CONCLUSION

In this paper, we present a generative framework, MMO-IG, designed to synthesize remote sensing (RS) images with multi-modal objects (MMOs) and provide corresponding instance-level labels for Remote Sensing Image Object Detection (RSIOD). We introduce ISIM and SODI as control conditions to encode MMO and image content, and propose SCDKG to model interdependencies between object classes and their spatial characteristics. Experimental results show that MMO-IG effectively generates high-quality RS images and improves detection performance.

However, MMO-IG has some limitations. When handling rare instances, the generated images may fail to accurately represent them. Function fitting sometimes oversimplifies instance attributes, and the method's performance can degrade with large numbers of targets due to increased complexity. Additionally, while MMO-IG performs well on remote sensing images, it may struggle with natural scenes that involve more complex backgrounds, lighting variations, and occlusions. Finally, the computational requirements of the image generation process limit its application in resource-constrained environments. We aim to address these challenges in future work by improving the framework's adaptability, reducing its computational demands, and enhancing its handling of rare instances and large target counts.

References

[1] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[2] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[3] T. Gao, Q. Niu, J. Zhang, T. Chen, S. Mei, and A. Jubair, "Global to local: A scale-aware network for remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[4] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 794–16 805.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[6] C. Yang, M. Chen, Z. Xiong, Y. Yuan, and Q. Wang, "Cm-net: Concentric mask based arbitrary-shaped text detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2864–2877, 2022.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[8] C. Yang, M. Chen, Y. Yuan, and Q. Wang, "Text growing on leaf," *IEEE Transactions on Multimedia*, vol. 25, pp. 9029–9043, 2023.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[11] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[12] D. Tang, X. Cao, X. Hou, Z. Jiang, J. Liu, and D. Meng, "Crs-diff: Controllable remote sensing image generation with diffusion model," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[13] X. Qian, X. Chen, W. Yue, X. Liu, J. Guo, Z. Li, Y. Li, and W. Wang, "Generating and sifting pseudolabeled samples for improving the performance of remote sensing image scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4925–4933, 2020.

[14] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 511–22 521.

[15] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "Layoutdiffusion: Controllable diffusion model for layout-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 490–22 499.

[16] Z. Yang, J. Wang, Z. Gan, L. Li, K. Lin, C. Wu, N. Duan, Z. Liu, C. Liu, M. Zeng *et al.*, "Reco: Region-controlled text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 246–14 255.

[17] W. Sun and T. Wu, "Image synthesis from reconfigurable layout and style," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 531–10 540.

[18] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen, "Controlnet++: Improving conditional controls with efficient consistency feedback," in *European Conference on Computer Vision*. Springer, 2025, pp. 129–147.

[20] D. Zavadski, J.-F. Feiden, and C. Rother, "Controlnet-xs: Designing an efficient and effective architecture for controlling text-to-image diffusion models," *arXiv preprint arXiv:2312.06573*, 2023.

[21] J. Redmon, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[22] C. Yang, M. Chen, Y. Yuan, and Q. Wang, "Bip-net: Bidirectional perspective strategy based arbitrary-shaped text detection network," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2255–2259.

[23] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[24] Z. Xiao, Z. Mai, Z. Xu, Y. Cui, and J. Li, "Corporate event predictions using large language models," in *2023 10th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, 2023, pp. 193–197.

[25] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.

[26] C. Yang, H. Ma, and Q. Wang, "Instance mask growing on leaf." in *BMVC*, 2023, pp. 4–6.

[27] Y. Li, Q. Huang, X. Pei, Y. Chen, L. Jiao, and R. Shang, "Cross-layer attention network for small object detection in remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2148–2161, 2020.

[28] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[29] X. Qian, B. Wu, G. Cheng, X. Yao, W. Wang, and J. Han, "Building a bridge of bounding box regression between oriented and horizontal object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–9, 2023.

[30] Z. Li, E. Li, T. Xu, A. Samat, and W. Liu, "Feature alignment fpn for oriented object detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[31] Z. Li, Y. Wang, D. Xu, Y. Gao, and T. Zhao, "Tbnet: A texture and boundary-aware network for small weak object detection in remote-sensing imagery," *Pattern Recognition*, vol. 158, p. 110976, 2025.

[32] F. Xiaolin, H. Fan, Y. Ming, Z. Tongxin, B. Ran, Z. Zenghui, and G. Zhiyuan, "Small object detection in remote sensing images based on super-resolution," *Pattern Recognition Letters*, vol. 153, pp. 107–112, 2022.

[33] W. Ma, N. Li, H. Zhu, L. Jiao, X. Tang, Y. Guo, and B. Hou, "Feature split–merge–enhancement network for remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[34] Y. Ye, X. Ren, B. Zhu, T. Tang, X. Tan, Y. Gui, and Q. Yao, "An adaptive attention fusion mechanism convolutional network for object detection in remote sensing images," *Remote Sensing*, vol. 14, no. 3, p. 516, 2022.

[35] X. Dong, Y. Qin, Y. Gao, R. Fu, S. Liu, and Y. Ye, "Attention-based multi-level feature fusion for object detection in remote sensing images," *Remote Sensing*, vol. 14, no. 15, p. 3735, 2022.

[36] H. Gong, T. Mu, Q. Li, H. Dai, C. Li, Z. He, W. Wang, F. Han, A. Tuniyazi, H. Li *et al.*, "Swin-transformer-enabled yolov5 with attention mechanism for small object detection on satellite images," *Remote Sensing*, vol. 14, no. 12, p. 2861, 2022.

[37] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 161, pp. 294–308, 2020.

[38] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 819–15 829.

[39] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[40] J. Lin, Y. Zhao, S. Wang, and Y. Tang, "YOLO-DA: an efficient yolo-based detector for remote sensing object detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[41] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "Fast tiny object detection in large-scale remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5512–5524, 2019.

[42] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014,*

Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2014.

[43] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 2352–2360.

[44] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 1945–1954.

[45] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 2672–2680.

[46] M. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 469–477.

[47] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2813–2821.

[48] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[49] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4570–4580.

[50] N. Lv, H. Ma, C. Chen, Q. Pei, Y. Zhou, F. Xiao, and J. Li, "Remote sensing data augmentation through adversarial training," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 9318–9333, 2021.

[51] A. Sebaq and M. ElHelw, "Rsdiff: Remote sensing image generation from text using diffusion model," *Neural Computing and Applications*, pp. 1–9, 2024.

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., vol. 139, 2021, pp. 8748–8763.

[53] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition 2021, virtual, June 19-25, 2021*, 2021, pp. 12 873–12 883.

[54] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[55] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

Chuang Yang received the B.E. degree in automation and the M.E. degree in control engineering from Civil Aviation University of China, Tianjin, China, in 2017 and 2020 respectively. He is currently working toward the Ph.D. degree in the School of Computer Science and School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, embodied AI, and intelligent transportation.

Bingxuan Zhao received the B.Sc. degree in Information and Computing Science from Northwestern Polytechnical University, Xi'an, China, in 2024. He is currently working toward the Ph.D. degree in the School of Computer Science and School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and machine learning.

Qing Zhou is currently pursuing the Ph.D degree in computer science and technology with the school of Artificial Intelligence, Optics and Electronics (iOPEN). His research interests include computer vision and pattern recognition.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.For more information, visit the link (http://crabwq.github.io/).