SGP^2: Coarse-to-Fine Controllable Multi-modal Remote Sensing Image Generation

Pengyu Chen, Xi Yang✉, and Nannan Wang

State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, 710071, China
chenpy@stu.xidian.edu.cn, {yangx,nnwang}@xidian.edu.cn

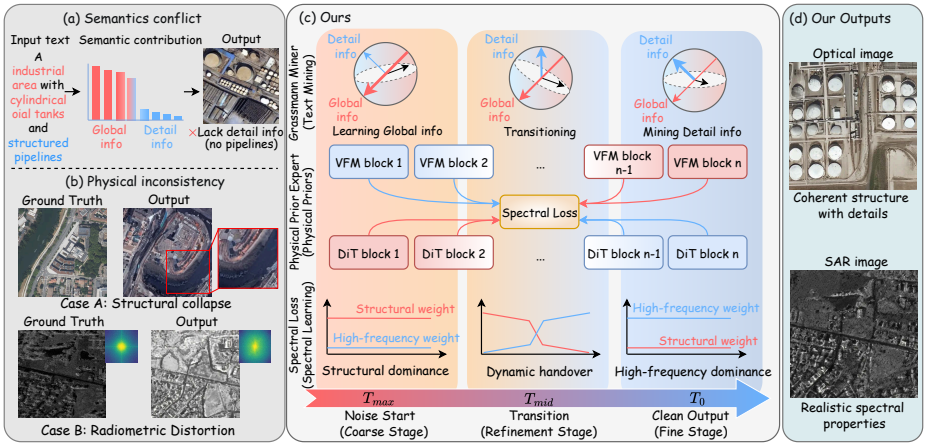

Fig. 1: Existing remote sensing image generation models face challenges regarding semantics conflict and physical inconsistency. Our method achieves coarse-to-fine controllable generation by synergizing geometric and physical priors based on different denoising stages. Compared with existing methods, our model generates high-quality multi-modal (optical and SAR) remote sensing images that are structurally coherent and possess authentic spectral properties.

Abstract. Generating high-fidelity remote sensing images is severely challenged by their specific overhead perspectives, complex structures, and rigorous physical attributes. Although recent controllable RS generation methods have made progress, they are still limited by two key bottlenecks. The first is semantic conflict, where dominant global concepts often overshadow fine-grained details. The second is physical inconsistency, which easily leads to structural collapse and radiometric distortion during the generation process. To address these issues, we propose SGP^2 (**S**ynergizing **G**eometric and **P**hysical **P**riors), a novel coarse-to-fine framework for controllable multimodal RS image generation. Specifically, to

✉Corresponding author

resolve semantic conflicts, we introduce the *Grassmann Miner*, which constructs dynamic geodesic trajectories on the Grassmann Manifold, enabling the model to gradually shift its focus from global backgrounds to fine-grained local details during the denoising process. To ensure physical consistency, we propose *Physical Prior Expert*, a module that dynamically aligns multi-level features from Visual Foundation Models and employs a frequency-aware *Spectral Loss* to enforce both amplitude and phase consistency. Furthermore, to bridge RS data gaps, we present MMEarth-1.5M, a 1.5 million text-image dataset featuring strictly co-registered optical and SAR pairs for robust cross-modal simulation. Our approach, without bells and whistles, achieves favorable performance in controllable multi-modal remote sensing image generation task. The dataset will be available in https://github.com/cpy0029/MMEarth-1.5M.

Keywords: Controllable generation · Remote sensing · Coarse-to-Fine

1 Introduction

Recently, diffusion models [7, 24, 27] have triggered a paradigm shift in generative artificial intelligence by showcasing stunning text-to-image generation capabilities. In Remote Sensing (RS), this capability provides new opportunities for addressing data scarcity and enhancing downstream tasks [29, 51] (*e.g.* object detection and land cover classification) [3, 22, 38]. However, directly transferring natural image generation models to the remote sensing domain poses severe domain adaptation challenges. Compared to natural images, remote sensing images possess unique overhead perspectives, complex structures, and specific physical attributes (*e.g.* scattering and spectral features) [36, 56]. To address this, inspired by ControlNet [57], researchers introduced spatial priors such as segmentation masks and sketches as constraints to achieve controllable generation of RS images.

Although existing controllable RS image generation methods [23, 59] have improved overall visual quality, generating high-fidelity images that align with fine-grained semantic descriptions and possess physical consistency still faces two challenges: (1) **Semantics conflict.** As shown in Fig. 1(a), complex RS scenes contain multi-level semantic information(*e.g.* "industrial area" and common "oil tanks" as global structural information, and "pipelines" as local detail information) [37]. Existing text encoders tend to capture common global concepts while neglecting fine-grained descriptions in text prompts [12, 13], causing semantic incompleteness. (2) **Physical inconsistency.** As shown in Fig. 1(b), RS images possess structural and radiometric properties. Blind generation often results in structural collapse or radiometric distortion. This is because purely data-driven models lack the incorporation of physical information and explicit constraints on frequency domain distribution, which makes it difficult to maintain complex structures and realistic frequency domain energy distributions [31] during the denoising process.

In this paper, we propose a novel framework for controllable multimodal remote sensing generation named SGP^2 (**S**ynergizing **G**eometric and **P**hysical

Priors). Our core perspective is that high-quality RS generation should not rely on a stage-agnostic conditioning strategy, but effectively align with the diffusion process through a coarse-to-fine progression (Fig. 1(c)). (1) To address semantics conflict, we propose the Grassmann Miner. Existing methods typically treat text embeddings as fixed points in Euclidean space [12,13], where dominant global concepts often overshadow fine-grained details due to feature entanglement. To resolve this, we model semantic features from the perspective of subspace geometry on the Grassmann Manifold [10]. As illustrated in the top of Fig. 1(c), we construct a dynamic geodesic trajectory within this manifold. Unlike static representations or simple Euclidean interpolation, this manifold-aware approach ensures geometric integrity, allowing the model to progressively shift its focus from dominant global contexts to submerged fine-grained details across the denoising timesteps. (2) To address the issue of physical consistency, we propose the Physical Prior Expert. Unlike traditional REPA [34, 45, 55] that align only single-level features, our approach simultaneously utilizes both low-level structural features and high-level semantic features from Visual Foundation Models (VFMs) [32] to guide the structural learning of DiT. Furthermore, it dynamically adjusts the contribution weights of different experts according to the evolutionary stages of the denoising process, ranging from coarse contour construction to fine texture filling. Additionally, during feature alignment, we introduce a frequency-domain aware Spectral Loss to simultaneously constrain amplitude and phase consistency within the frequency space. This ensures that the generation images are not only structurally coherent in the spatial domain but also strictly adhere to authentic physical distributions in the frequency domain.

To effectively train our proposed SGP^2 and bridge critical data gaps in the RS field while enabling multi-spatial resolution, controllable, and multi-modal generation, we construct MMEarth-1.5M, a large-scale, multi-modal, and comprehensive dataset with multi-spatial resolution. This dataset comprises 1.5 million text-image pairs that not only support text-to-image pre-training but also feature a curated multi-conditional subset containing co-registered and simultaneous optical and Synthetic Aperture Radar (SAR) images. The core strength of MMEarth-1.5M lies in its cross-modal generation capability, which enables the model to simultaneously depict Earth surface appearances across different sensors and resolutions. This capacity to simulate the response of terrestrial features under various imaging mechanisms provides unprecedented momentum for comprehensive Earth observation, data augmentation for downstream tasks, and the development of general Earth observation foundation models.

In a nutshell, our contributions are summarized as follows:

- We propose SGP^2 for coarse-to-fine controllable multi-modal remote sensing image generation by introducing geometric and physical priors, addressing semantic omission and physical feature distortion in complex scenes.
- We propose Grassmann Miner, leveraging dynamic geodesic trajectories on the Grassmann manifold to resolve Euclidean feature entanglement, enabling progressive attentional shifts from global semantics to fine-grained details.

- We utilize Physical Prior Expert to guide diffusion model training and dynamically adjust the weights of each layer according to different stages of the denoising process. Furthermore, by introducing spectral loss, we enforce joint constraints on amplitude and phase consistency within the frequency domain.
- We construct MMEarth-1.5M, a large-scale, multi-modal, and multi-spatial resolution dataset that supports both text-to-image pre-training and controllable generation fine-tuning. This dataset fills a critical data gap in controllable multi-modal remote sensing image generation.

2 Related Works

2.1 Remote Sensing Image Generation

Early studies utilized Generative Adversarial Networks (GANs) [5] for remote sensing image generation [44], but issues like training instability and mode collapse limited the feasibility of this method in large-scale generation tasks.

As diffusion models [7] have become the mainstream paradigm for image generation, researchers have begun employing them for remote sensing image generation. Works such as SatDM [1], Text2Earth [18], RSDiff [29], and MetaEarth [56] have achieved single-scale or multi-scale remote sensing image generation by fine-tuning diffusion models and super-resolution architectures using remote sensing data, combined with resolution-specific text prompts. Furthermore, leveraging the ControlNet [57], researchers have introduced physical priors such as segmentation masks and sketches as additional conditions to achieve precise control over the generation of complex remote sensing images [22,37,39]. In contrast, CC-Diff [58], AeroGen [38], and OF-Diff [54] focus on conditioning the model on object bounding boxes to guide the generation of remote sensing images containing specific targets. Uni-RS [61] improves spatial faithfulness in remote sensing generation via explicit layout planning. Despite numerous explorations, constrained by the inherent complexity of remote sensing scenes, directly using frozen text encoders often fails to accurately disentangle or capture fine-grained semantic information from global descriptions, leading to critical detailed features being overshadowed by the global context.

2.2 Representation Alignment for Generation

As research into the internal representation mechanisms of diffusion models deepens and self-supervised visual foundation models [4, 32] emerge, researchers have begun exploring the use of external knowledge to guide the representation learning of diffusion models. For instance, REPA [55] proposes aligning the clean image representations extracted by DINOv2 with the projected latent states of DiT, which significantly enhances training efficiency while improving generation quality. Subsequent studies [11,52] have further introduced strategies such as end-to-end fine-tuning [14], semantic feature entanglement [45], and spatial structure

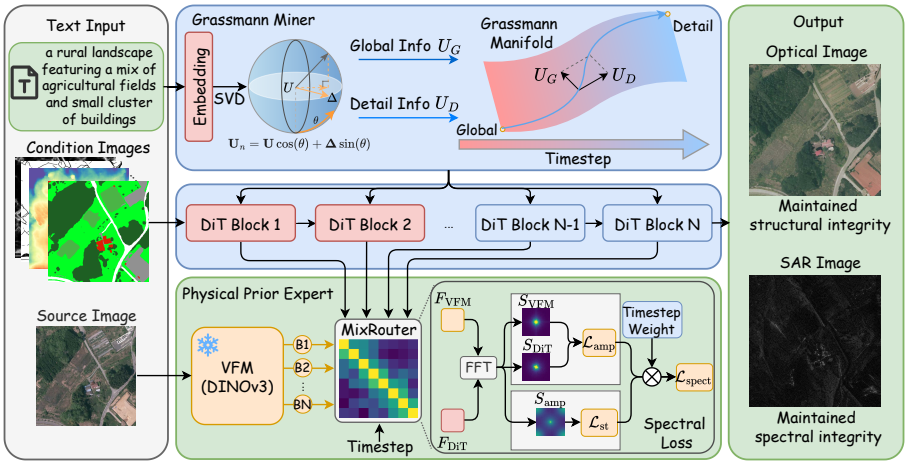

Fig. 2: The framework of SGP^2. Our SGP^2 is based on DiT and consists of Grassmann Miner, Expert of Mixed, and Spectral Loss.

enhancement [34] to optimize model performance. However, existing REPA-like methods mostly utilize only the high-level semantic outputs of vision foundation models for alignment. Given the complex spatial structural features of remote sensing images, these methods based on single-level representations and simple alignment strategies struggle to effectively convey the fine-grained structural and spectral information specific to remote sensing images, even though they introduce physical priors to a certain extent.

3 Method

The overall framework of SGP^2 is drawn in Fig. 2. It first utilizes the Grassmann Miner (GM) to construct a dynamic geodesic trajectory on the Grassmann manifold, guiding text features to achieve a smooth manifold evolution from global semantics to local details. Subsequently, the multimodal conditional latent states are concatenated and fed into the DiT [50]. Through the Physical Prior Expert (PPE), multiscale physical priors are extracted from the visual foundation model [32] to achieve feature alignment with these latent states. On this basis, the Router dynamically allocates alignment weights according to the denoising time steps and provides precise guidance through Spectral Loss, ultimately generating structurally coherent and high quality multimodal RS images.

3.1 Grassmann Miner

Existing text encoders tend to capture global concepts while ignoring crucial granular descriptions in text prompts. To address this issue, we break the traditional paradigm of treating text embeddings as fixed points in Euclidean space and propose the Grassmann Miner (GM), transforming text embeddings into

dynamic geometric trajectories during the diffusion denoising process. Given the text embedding $\mathbf{Y} \in \mathbb{R}^{L \times D}$, where L is the sequence length and D is the feature dimension. To disentangle hierarchical semantic responses from the highly entangled text space and extract representation bases with geometric invariance, we perform Singular Value Decomposition (SVD) on this text space:

$$\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^{\top}, \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{L \times L}$ and $\mathbf{V} \in \mathbb{R}^{D \times D}$ represent orthogonal bases, and $\mathbf{S}$ is a diagonal matrix of singular values. To filter out semantic noise and isolate the most informative components, we truncate the basis to the first k dimensions, obtaining $\mathbf{U}_k \in \mathbb{R}^{L \times k}$. We leverage the energy concentration property of Singular Value Decomposition, where the top-k_1 singular vectors $\mathbf{U}_g \in \mathbb{R}^{L \times k_1}$ encapsulate the high-variance semantic anchors (e.g., global scene layouts) that typically dominate the early denoising stage. Conversely, the remaining orthogonal vectors $\mathbf{U}_d \in \mathbb{R}^{L \times (k-k_1)}$ represent the latent fine-grained nuances that are often suppressed in standard Euclidean space. This orthogonal projection on the Grassmann manifold $\mathcal{G}$ ensures that geometric operations on microscopic details do not perturb the established macroscopic context.

After isolating the orthogonal spaces of global and detailed semantics, directly performing linear interpolation or scaling in Euclidean space will inevitably destroy the orthogonality of the matrices, resulting in the collapse and entanglement of semantic features once again. Therefore, we propose evolving the semantic representations along the shortest path, namely the geodesic, on the Grassmann manifold to preserve the geometric integrity of the semantic space. For each subspace $\mathbf{U}_i$ (where $i \in \{g, d\}$), we define a continuous geodesic trajectory parameterized by the angle θ_i:

$$\mathbf{U}_i(\theta_i) = \mathbf{U}_i \cos(\theta_i) + \mathbf{\Delta}_i \sin(\theta_i). \tag{2}$$

To construct an effective tangent vector $\mathbf{\Delta}_i$ that determines the direction of semantic evolution, we introduce a bias term $\mathbf{Z}_i = \mathrm{MLP}(\tau) \cdot \mathrm{AvgPool}(U_i) \in \mathbb{R}^{L \times k_i}$ conditioned on the diffusion timestep, where $\tau = t/T \in [0, 1]$ and T is the total number of denoising steps, and t is the current denoising step. To ensure that the trajectory remains strictly on the manifold, $\mathbf{Z}_i$ is projected into the orthogonal complement space of $\mathbf{U}_i$:

$$\mathbf{\Delta}_i = \frac{\mathbf{Z}_i - \mathbf{U}_i(\mathbf{U}_i^{\top}\mathbf{Z}_i)}{\|\mathbf{Z}_i - \mathbf{U}_i(\mathbf{U}_i^{\top}\mathbf{Z}_i)\|}, \tag{3}$$

this projection yields a unit norm tangent direction $\mathbf{\Delta}_i$. By stepping along it according to the geodesic equation, we safely inject semantic guidance specific to the current step into the basis without destroying its orthogonality, thereby ensuring the mathematical stability of the intervention process.

The generation process of diffusion models is inherently hierarchical: it first establishes coarse global contours and subsequently fills in high frequency local textures. To temporally align the text guidance with this physical denoising

process, we couple the geodesic angles and singular values with the normalized diffusion timestep τ. To this end, we define complementary time scheduling functions: $w_g(\tau) = \sin(\tau\pi/2)$ for the global subspace and $w_d(\tau) = \cos(\tau\pi/2)$ for the detail subspace. The geodesic stepping angles are dynamically scaled as $\theta_g = \alpha_g \cdot w_g(\tau)$ and $\theta_d = \alpha_d \cdot w_d(\tau)$, where α_g and α_d are learnable scaling parameters with an initial value of 1.

Furthermore, given that the singular value matrix $\mathbf{S}$ implicitly encodes the information variance of each orthogonal subspace, we introduce a diagonal operator $\mathbf{\Gamma}(\tau)$ parameterized by the timestep to perform dynamic spectral modulation on it: $\hat{\mathbf{S}} = \mathbf{S} \odot \mathbf{\Gamma}(\tau)$. This enables a smooth transfer of the energy centroid between global and detailed semantics, ensuring that the feature guidance magnitude is strictly isomorphic to the coarse-to-fine denoising process of the model.

Finally, we reconstruct the enhanced manifold representation $\hat{\mathbf{Y}} = \mathbf{U}_{final}\hat{\mathbf{\Sigma}}\mathbf{V}^\top$ utilizing the evolved orthogonal basis $\mathbf{U}_{final} = [\mathbf{U}_g, \mathbf{U}_d]$ and the recalibrated singular values $\hat{\mathbf{\Sigma}}$. To maintain the stability of the original feature distribution, we extract the pure geometric evolution increment $\mathbf{P} = \hat{\mathbf{Y}} - \mathbf{Y}$. Subsequently, we introduce a zero initialized linear transformation operator $\mathbf{W}_p \in \mathbb{R}^{D\times D}$ to realign this increment to the semantic feature space of the diffusion model. The final refined text condition is represented as:

$$\mathbf{Y}_{out} = \mathbf{Y} + \mathbf{P}\mathbf{W}_p. \tag{4}$$

This design shields the pretrained generative prior from uncalibrated feature mutations, facilitating the stable integration of multiscale semantic details from the Grassmann miner while maintaining its original integrity.

3.2 Physical Prior Expert

Unlike traditional methods that solely rely on the rigid single point mapping of high level features from Vision Foundation Models (VFMs), Physical Prior Expert constructs a dynamic multi level collaborative alignment paradigm between the feature pyramids of the VFM and the DiT, achieving a comprehensive prior injection from microscopic physical structures to macroscopic global semantics. Specifically, to follow the denoising evolution regularity of diffusion models, we design MixRouter, a routing mechanism capable of perceiving time steps and dynamically allocating alignment intensity. Given the corresponding time step embedding τ, MixRouter outputs a dense allocation matrix $W(\tau) \in \mathbb{R}^{N_D\times N_V}$:

$$W(\tau) = \text{Softmax}(\text{MLP}(\tau)), \tag{5}$$

where N_V and N_D respectively represent the number of selected DiT feature layers and VFM feature layers. The element $W^{(i,j)}(\tau)$ in the matrix accurately quantifies the alignment weight of the VFM feature at layer j to the DiT feature at layer i on time step τ.

With this routing matrix, we break the rigid single point mapping paradigm in traditional representation alignment, and instead construct a dense multi level

alignment objective based on soft assignment. Specifically, let the DiT feature at layer i be $x_D^{(i)}$, and the corresponding VFM feature at layer j as $x_V^{(j)}$. The initial alignment objective $\mathcal{L}_{align}$ of Physical Prior Expert is formalized as the weighted expectation of the similarity between all feature pairs:

$$\mathcal{L}_{align} = \sum_{i=1}^{N_D} \sum_{j=1}^{N_V} W^{(i,j)}(\tau) \cdot [1 - \text{sim}(x_D^{(i)}, x_V^{(j)})], \tag{6}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity function. Through this routing of time steps, each layer of DiT no longer rigidly fits VFM features of a single dimension, but can adaptively extract and combine the most suitable mixed priors from VFM features according to the physical demands of the current denoising stage.

3.3 Spectral Loss

Although macro level feature alignment was achieved in the previous section, remote sensing images not only possess complex spatial attributes, but their essence is also the physical response of the earth surface to electromagnetic radiation, exhibiting a strict frequency distribution. To fundamentally solve this problem, we introduce a frequency-aware Spectral Loss, jointly constraining the amplitude and phase of feature pairs in the frequency domain.

Given the DiT feature x_D and the VFM feature x_V, we first map them to the frequency domain via Fast Fourier Transform (FFT). In this domain, to anchor the global energy distribution and macroscopic outline of the image, we apply a log smoothed L_1 constraint to the amplitude spectrum of the features:

$$\mathcal{L}_{amp} = ||\log(1 + |\mathcal{F}(x_D)|) - \log(1 + |\mathcal{F}(x_V)|)||, \tag{7}$$

where $\mathcal{F}(\cdot)$ denotes the FFT operation. This term effectively narrows the gap of the low frequency energy distribution between feature pairs, ensuring that the generated images possess authentic macroscopic radiometric properties. Meanwhile, to enforce precise topological constraints on the stringent fine grained structures in remote sensing scenarios, we normalize the absolute error of the real and imaginary parts of the complex signal in the frequency domain relative to the amplitude of the VFM feature:

$$\mathcal{L}_{st} = \mathbb{E}\left[\frac{|Re(\mathcal{F}(x_D)) - Re(\mathcal{F}(x_V))| + |Im(\mathcal{F}(x_D)) - Im(\mathcal{F}(x_V))|}{|\mathcal{F}(x_V)| + \epsilon}\right], \tag{8}$$

where $Re(\cdot)$ and $Im(\cdot)$ respectively denote the real part and imaginary part of the complex signal in the frequency domain. By imposing stringent constraints on the real and imaginary parts respectively, this term not only measures the energy discrepancy of frequencies but also implicitly ensures the strict alignment of phases in the frequency domain.

Finally, we construct a temporally adaptive spectral alignment objective:

$$\mathcal{L}_{\text{sp}} = w_s(\tau)\mathcal{L}_{\text{amp}} + (1 - w_s(\tau))\,\mathcal{L}_{\text{st}}, \tag{9}$$

where $w_s(\tau) = \sin(\pi\tau/2)$ and $\tau = t/T \in [0, 1]$ denotes the normalized diffusion timestep. Since τ is large at the early denoising stage, $w_s(\tau)$ is close to 1, and the optimization is mainly dominated by $\mathcal{L}_{\text{amp}}$, which encourages global amplitude and radiometric distribution consistency. As denoising approaches the final stage ($\tau \to 0$), $w_s(\tau)$ decreases, so the effective weight of the high-frequency structural term, $1 - w_s(\tau)$, increases smoothly. Therefore, the spectral objective gradually shifts its emphasis from global frequency energy alignment to fine-grained structural and phase consistency through $\mathcal{L}_{\text{st}}$.

3.4 Training Objectives

To ensure the model possesses stringent physical consistency while generating high quality RS images, we construct a multi dimensional joint training objective:

$$\mathcal{L}_{total} = \mathcal{L}_{DiT} + \mathcal{L}_{align} + \mathcal{L}_{sp}, \tag{10}$$

where $\mathcal{L}_{DiT}$ is the original denoising loss of DiT [50], $\mathcal{L}_{align}$ is the alignment loss described in § 3.2, and $\mathcal{L}_{sp}$ is the spectral loss described in § 3.3.

4 MMEarth-1.5M Construction

To support controllable multi-modal remote sensing image generation, we need to construct a large scale text-image dataset of registered multi-modal images and various conditional images. The dataset construction pipeline includes three main steps: collecting data, extracting conditional images and text annotation.

Collecting data. To construct the MMEarth-1.5M, we integrate public high quality image sources, including Git-10M [18], HRSC2016 [60], HRRSD [62], RSICD trainset [19], WHU-OPT-SAR [16], DFC25 trainset [46], DFC23 [25], FUSAR-Map [30], 3MOS [53], M4-SAR [40], QXS-SAROPT [9], SAR2Opt [65], EarthMiss [67], and YESeg-OPT-SAR [43]. Among these, Git-10M images are randomly sampled based on different spatial resolution ratios for text-image pre-training. HRSC2016, HRRSD and RSICD trainset are applied for controllable generation of optical images. The remaining datasets are used for multi-modal controllable generation. To ensure consistency in data scale and fairness in comparison, we uniformly apply Real-ESRGAN [41] for super resolution processing on all images with a resolution below 512×512. Tab. 1 details the composition of MMEarth-1.5M.

Extracting conditional images. To achieve controllable generation, we extract various conditional for all applicable datasets. Specifically, we utilize the HED [48], LETR [49], DepthAnythingv3 [17], LtS [33], SGCN [66], and SegEarth-OV3 [15], to extract HED (Holistically-nested Edge Detection), MLSD (Multiscale Line Segment Detection), depth maps, sketch, road maps, and segmentation masks. In multi-modal controllable generation, we further utilize optical and SAR images as equivalent conditions for cross modal generation.

Table 1: Property comparisons with popular remote sensing datasets. Gen. and Ctrl. represent generation and controllable.

Datasets	Resolution	Modality	Text	Gen.	Ctrl.
HRSC2016 [60]	0.5M-2M	Optical	✗	✗	✗
RSICD [19]	0.5M-8M	Optical	✓	✓	✗
RS5M [63]	Multiscale	Optical	✓	✓	✗
WHU-OPT-SAR [16]	5M	Optical, SAR	✗	✗	✗
Git-10M [18]	0.5M-128M	Optical	✓	✓	✗
MMEarth-1.5M (Ours)	0.15M-128M	Optical, SAR	✓	✓	✓

Text annotation. Considering that manual annotation is infeasible for large-scale images, we design an automated text annotation pipeline to address this challenge. This pipeline utilizes Qwen3-VL-32B [2] for initial annotation. Subsequently, the text is segmented and fed back into another Qwen3-VL for hallucination checking. The pipeline outputs the final text annotation only when all segments are confirmed to be free of hallucinations. In addition, regular manual sampling was conducted to evaluate the accuracy. Errors identified during the review process were fed back into the annotation pipeline, prompting rapid improvements and reprocessing of erroneous samples. This automated annotation pipeline successfully generated high-quality, semantically rich, and contextually accurate text descriptions for every image in the dataset.

5 Experiment

5.1 Experimental setup

Implementation details. Following JoDI [50], SGP^2 adopts a DiT backbone with a hidden size of 2240 and 20 attention heads. Training is conducted in two stages rather than on all raw MMEarth-1.5M images as controllable pairs. We first perform text-to-image pre-training for 2 epochs on 1,198,845 text-image samples, and then fine-tune the model for 5 epochs on 66,124 controllable multimodal pairs. The global batch size is 32 in both stages, resulting in roughly 85K optimization steps in total. We use CAME-8bit [20] with learning rates of 4×10^{-5} for pre-training and 1×10^{-5} for fine-tuning, and train with bf16 precision. During fine-tuning, all available conditions are jointly used. At inference, we employ Flow-DPM-Solver [47] adapted for rectified flow with 20 denoising steps, and set the classifier-free guidance [8] scale to 4.5.

Evaluation metrics. Similar to [37, 51], we report the Inception Score (IS) [28], the Fr'echet Inception Distance (FID) [6], the CLIP Score (CLIP$_s$) [26] and the Overall Accuracy (OA) [51] for zero-shot classification for text-to-optical image generation at 512×512. For text-to-SAR generation, considering that the FID utilizes an RGB pre-trained InceptionV3 [35] for feature extraction and is thus unsuitable for SAR imagery, we employ Structural Similarity Index Measure (SSIM) and Peak Signal-to-Noise Ratio (PSNR) for performance evaluation. In addition, we report SAR-specific physical metrics, including Equivalent Number of

Table 2: Quantitative comparisons on RSICD testset [19]. † indicates images are generated and evaluated using official code and weights. FT represents fine-tuning on our MMEarth-1.5M.

Methods	IS↑	FID↓	CLIP$_s$↑	OA↑
CRS-Diff [37][TGRS 24]	18.39	50.72	20.33	69.21
InstancedDiff [42][CVPR 24]	-	138.60	**24.70**	-
Text2Earth† [18][GRSM 25]	18.17	53.15	22.62	68.27
AeroGen$_{(FT)}$ [38][CVPR 25]	16.83	58.12	22.31	66.72
JoDI [50][ArXiv 25]	4.28	181.36	20.23	22.74
JoDI$_{(FT)}$ [50][ArXiv 25]	18.03	57.24	21.46	67.59
SGP^2(Ours)	**21.23**	**44.17**	24.58	**86.27**

Looks (ENL) and Radiometric Resolution (RaRes). For controllable optical image generation, we use SSIM and mIoU to evaluate performance. For controllable SAR image generation, we report SSIM.

5.2 Experimental Results

Text-to-Image generation. In Tab. 2, we present the optical image generation results on the RSICD testset. SGP^2 consistently surpasses other remote sensing image generation methods. Compared to the latest JoDI [50], our method obtains a higher IS (+17.7%), CLIP score (+14.5%), OA (+27.6%), and lower FID (+22.8%) on the RSICD testset. Although InstancedDiff [42] achieves a higher CLIP Score, its generated images are not from a RS perspective, thus resulting in a worse FID. Furthermore, SGP^2 significantly outperforms some well-known diffusion-based methods [18,37,38]. Experimental results verify that our proposed grassmann miner can effectively mine local details with global context preserved, thus producing high-quality remote sensing images.

SGP^2 demonstrates a deep understanding of image semantics and structural information. Therefore, it can be used to generate multi-modal remote sensing images. In Tab. 3, we present the SAR image generation results at 512×512 resolution on the DFC25 valset [46]. Similar to the optical image generation results, on the text-to-SAR image generation task, our method also outperforms JoDI$_{(FT)}$ on the SSIM (+10.9%) and PSNR (+3.16%) metrics, which once again confirms the effectiveness of grassmann miner. Additionally, we observed that JoDI, after being fine-tuned with the MMEarth-1.5M dataset, demonstrates SAR image generation performance that significantly surpasses the original model. Recognizing that SAR and optical are different physical manifestations of the same geographic location, Physical Prior Expert utilizes multi-condition joint training to transfer structural and geometric priors from optically pre-trained VFMs, effectively generalizing to SAR generation with enhanced physical consistency.

Single-condition image generation. SGP^2 is not limited to text conditions and supports a broad range of control signals to enhance the fidelity of image generation. We conduct a comparative evaluation of SGP^2 against state-of-the-art models, including JoDI and CRS-Diff, on the RSICD (optical) and DFC25 (SAR)

Table 3: Quantitative comparisons on DFC25 valset [46].

Methods	SSIM↑	PSNR↑	ENL↑	RaRes↓
Text2Earth† [18][GRSM 25]	0.340	21.52	-	-
JoDI [50][ArXiv 25]	0.012	12.21	5.01	1.88
$\text{JoDI}_{(\text{FT})}$ [50][ArXiv 25]	0.513	22.43	5.12	1.80
SGP^2(Ours)	**0.569**	**23.14**	**5.31**	**1.76**

Table 4: Comparison of model performance using single conditions. M., Dep., Skh., Opt., and Seg. represent generated image modality, depth map, sketch, optical, and segmentation mask. For optical image generation, mIoU is reported for Seg. and Road; otherwise, SSIM is reported. The same applies below.

Methods	M.	HED	MLSD	Dep.	Skh.	Seg.	Road
		SSIM				mIoU/SSIM	
CtrlNet [57]	Opt.	0.43	0.83	0.80	0.28	0.38	0.34
U-CtrlNet [64]		0.44	0.73	0.87	0.52	0.30	0.33
CRS-Diff [37]		0.45	0.74	0.87	0.56	0.29	0.37
$\text{JoDI}_{(\text{FT})}$ [50]		0.51	0.76	0.80	0.57	0.31	0.30
SGP^2(Ours)		**0.59**	**0.84**	**0.91**	**0.62**	**0.42**	**0.40**
$\text{JoDI}_{(\text{FT})}$ [50]	SAR	0.49	0.51	0.48	0.49	0.48	0.45
SGP^2(Ours)		**0.55**	**0.56**	**0.53**	**0.55**	**0.51**	**0.49**

datasets at a resolution of 512×512. After acquiring diverse inputs through specific condition extraction methods, the experimental results in Tab. 4 demonstrate that our method significantly outperforms JoDI and CRS-Diff for single condition optical image generation. These findings indicate that the advantage of SGP^2 stems from the Spectral Loss constraint during the joint training stage, which enables the model to learn effective spectral features for inferring target structures and ensures superior generation performance.

Multi-condition image generation. To comprehensively validate model performance, we tested eight different condition pairing schemes (such as HED+Sketch and MLSD+HED) for optical and SAR image generation tasks, while recording relevant metrics including SSIM and mIoU. The results in Tab. 5 demonstrate that SGP^2 maintains leading performance even in complex scenarios involving multiple condition inputs. Using the generation of optical images with the Sketch+Depth maps combination as an example, the SSIM of our method is 4.7% higher than that of CRS-Diff and 10.1% higher than JoDI. Notably, the method remains robust even when certain condition combinations may cause potential spatial conflicts. This is primarily attributed to our proposed Physical Prior Expert, which enables the model to capture discriminative physical consistency information in the feature space, thereby achieving effective generalization for generating diverse modalities.

Table 5: Results comparison of SGP^2 and other methods on multi-condition image generation. † indicates images are generated and evaluated using official code and weights. Abbreviations for condition groups are defined as follows: G1: HED+Skh., G2: MLSD+HED, G3: Skh.+MLSD, G4: Skh.+Dep., G5: HED+Seg., G6: Road+Seg., G7: Skh.+Seg., and G8: Dep.+Seg.

Methods	M.	G1	G2	G3	G4	G5	G6	G7	G8
		SSIM				mIoU (Opt.) / SSIM (SAR)			
CRS-Diff† [37]	Opt.	0.503	0.531	0.528	0.750	0.314	0.305	0.318	0.304
JoDI$_{(FT)}$ [50]	Opt.	0.492	0.541	0.537	0.738	0.329	0.307	0.322	0.315
SGP^2(Ours)	Opt.	**0.568**	**0.579**	**0.604**	**0.813**	**0.389**	**0.364**	**0.413**	**0.386**
JoDI$_{(FT)}$ [50]	SAR	0.490	0.527	0.491	0.502	0.491	0.478	0.509	0.486
SGP^2(Ours)	SAR	**0.557**	**0.569**	**0.556**	**0.538**	**0.547**	**0.512**	**0.557**	**0.541**

Table 6: Ablation study of components on FID. "GM", "PPE", and "$\mathcal{L}_{sp}$" represent grassmann miner, physical prior expert, and spectral loss.

Components			HED	MLSD	Dep.	Skh.	Seg.	Road
GM	PPE	$\mathcal{L}_{sp}$						
✗	✗	✗	64.7	66.5	65.1	60.6	64.7	69.3
✓	✗	✗	61.7	66.2	63.9	59.3	62.4	66.1
✗	✓	✗	55.8	57.4	56.2	56.8	60.9	61.3
✗	✓	✓	55.1	54.7	53.8	53.2	57.9	59.6
✓	✓	✗	47.6	49.7	50.2	54.9	52.1	51.4
✓	✓	✓	**42.4**	**43.7**	**42.1**	**41.8**	**42.6**	**43.2**

Table 7: Experiments on different variants of GM. FID is calculated as the mean FID values.

Method	Avg. FID↓
ToRA [12]	45.6
DATE [21]	43.6
Ours(GM)	**42.6**
w/o SVD	45.8
w/o $\mathbf{U}_g$ & $\mathbf{U}_d$	44.3
w/o geodesic trajectory	43.6
w/o $\mathbf{\Gamma}(\tau)$	43.1

5.3 Ablation Study

We conduct a comprehensive ablation study on the single-condition optical image generation to validate the contributions of our proposed components.

Analysis of critical modules. To validate the effectiveness of the three core components, Grassmann Miner (GM), Physical Prior Expert (PPE), and Spectral Loss ($\mathcal{L}_{sp}$), we conduct comprehensive ablation studies in Tab. 6. The experimental results indicate that our proposed method achieves significant performance gains of 31.0% and 35.3% on FID under sketch and depth map conditions respectively when compared to the baseline [50]. Further analysis reveals that removing Spectral Loss while retaining only PPE makes it difficult for the model to effectively learn consistent spectral information, which consequently leads to degraded generation quality. These findings demonstrate that these three core components work in synergy to optimize latent features and collectively drive a substantial improvement in generative performance.

Analysis of grassmann miner. To evaluate the internal design of the Grassmann Miner (GM) module and its specific contributions to the overall model performance, Tab. 7 presents the results of the ablation experiments. Compared to the baseline method ToRA [12] which follows the traditional text processing paradigm, the full GM achieves the best performance with an FID of 42.6. The

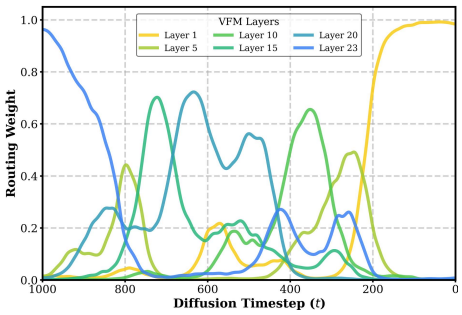

Fig. 3: Visualization of MixRouter weights at different time steps

Fig. 4: Disentanglement evaluation of $\mathbf{U}_g$ & $\mathbf{U}_d$

removal of SVD causes the FID to increase to 45.8, which demonstrates that SVD is crucial for decoupling highly entangled text feature spaces and extracting bases with geometric invariance. If the global and detail subspaces are not explicitly partitioned (w/o $\mathbf{U}_g$ & $\mathbf{U}_d$), local details are easily overwhelmed by the global background (FID 44.3). Furthermore, removing the geodesic trajectory or the dynamic spectral modulation $\mathbf{\Gamma}(\tau)$ results in the FID degrading to 43.6 and 43.1 respectively. These results confirm the necessity of maintaining manifold geometric integrity and matching guidance intensity with denoising stages.

Analysis of mixrouter. Fig. 3 illustrates the dynamic weight evolution logic of MixRouter during the denoising process. At the early denoising stage when t is large, the model assigns the highest weights to the deep layers of the VFM such as Layer 23 to extract macroscopic semantic priors and establish the global image layout. As the denoising process progresses toward the later stages when t decreases, the weights dynamically transition to the shallow layers of the VFM such as Layer 1, utilizing low level features to guide the generation of microscopic physical structures and high frequency textures. This observation aligns with our previous analysis, ensuring that the generated remote sensing images possess precise physical details while maintaining global coherence.

Analysis of semantic disentanglement. Fig. 4 provides a rigorous evaluation of the decoupling between global basis vectors $\mathbf{U}_g$ and detail basis vectors $\mathbf{U}_d$, demonstrating the effectiveness of the Grassmann Miner (GM) module in separating macro-layouts from micro-details in remote sensing imagery. Experimental results show that fixing $\mathbf{U}_g$ ensures the structural stability of global scenes, such as farmland or sports fields, while $\mathbf{U}_d$ is responsible for accurately reconstructing fine-grained features, such as the meandering of rivers or running track textures. This decoupling mechanism validates that the model resolves "semantic conflict" within the manifold space, enabling the generated imagery to maintain global coherence while achieving high-fidelity physical details.

Qualitative comparison. As shown in Fig. 5, our method incorporates three enhancements including Grassmann Miner, Expert of Mixed, and Spectral Loss to generate high-quality RS images with consistent color, structure, and spectral information. This approach effectively prevents spatial inconsistencies and the

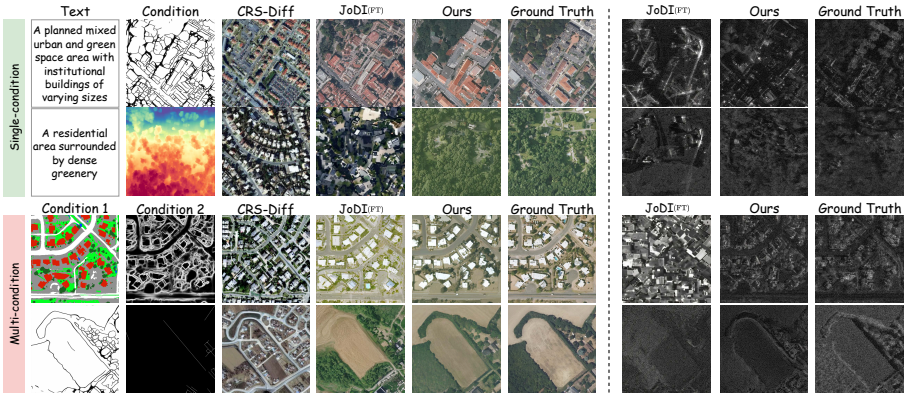

Fig. 5: Qualitative results of the proposed method for single-condition image generation and multi-condition image generation.

loss of detail seen in the second image. Notably, even in regions where feature conditions are not directly covered, our method still demonstrates superior generalization capabilities and generates more reasonable and complete scene content.

6 Conclusion

To address the semantic conflicts and physical inconsistencies in existing controllable remote sensing image generation, we propose a novel coarse-to-fine framework, SGP^2. We transcend traditional Euclidean paradigms by introducing Grassmann Miner, which constructs dynamic geodesic trajectories to smoothly transition the semantic focus from global backgrounds to fine-grained details. To fundamentally ensure physical consistency, our Physical Prior Expert dynamically aligns multi-level features across different denoising stages. Furthermore, a frequency-aware Spectral Loss is introduced to enforce strict amplitude and phase coherence. Finally, we contribute MMEarth-1.5M, a large-scale multi-modal dataset, to bridge critical data gaps in this domain.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62372348, in part by the Key Research and Development Program of Shaanxi under Grant 2024GXZDCYL-02-10, in part by Scientific and Technological Innovation Teams in Shaanxi Province under Grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11, in part by the Establishment Fund of the State Key Laboratory of Wakefullness/Sleep and Cognition(Chongqing Institute for Brain and Intelligence), in part by the Fundamental Research Funds for the Central Universities under Grant QTZX26138.

References

1. Baghirli, O., Askarov, H., Ibrahimli, I., Bakhishov, I., Nabiyev, N.: Satdm: Synthesizing realistic satellite image with semantic layout conditioning using diffusion models. arXiv preprint arXiv:2309.16812 (2023)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Chen, P., Nie, X., Ning, Y., Zhang, Y.: Learning efficient and adaptive cross-channel dependencies for weakly-supervised object detection. IEEE Transactions on Multimedia **27**, 8954–8966 (2025)
4. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
5. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (2017)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (2020)
8. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
9. Huang, M., Xu, Y., Qian, L., Shi, W., Zhang, Y., Bao, W., Wang, N., Liu, X., Xiang, X.: The qxs-saropt dataset for deep learning in sar-optical data fusion. arXiv preprint arXiv:2103.08259 (2021)
10. Huang, Z., Wu, J., Van Gool, L.: Building deep networks on grassmann manifolds. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
11. Jiang, D., Wang, M., Li, L., Zhang, L., Wang, H., Wei, W., Dai, G., Zhang, Y., Wang, J.: No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. arXiv preprint arXiv:2505.02831 (2025)
12. Kang, S., Han, W., Ju, D., Hwang, S.J.: Rare text semantics were always there in your diffusion transformer. arXiv preprint arXiv:2510.03886 (2025)
13. Kong, Z., Zhang, Y., Yang, T., Wang, T., Zhang, K., Wu, B., Chen, G., Liu, W., Luo, W.: Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In: European Conference on Computer Vision (2024)
14. Leng, X., Singh, J., Hou, Y., Xing, Z., Xie, S., Zheng, L.: Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. arXiv preprint arXiv:2504.10483 (2025)
15. Li, K., Zhang, S., Deng, Y., Wang, Z., Meng, D., Cao, X.: Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images. arXiv preprint arXiv:2512.08730 (2025)
16. Li, X., Zhang, G., Cui, H., Hou, S., Wang, S., Li, X., Chen, Y., Li, Z., Zhang, L.: Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. International Journal of Applied Earth Observation and Geoinformation **106**, 102638 (2022)
17. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: International Conference on Learning Representations (2026)

18. Liu, C., Chen, K., Zhao, R., Zou, Z., Shi, Z.: Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. IEEE Geoscience and Remote Sensing Magazine **13**(3), 238–259 (2025)
19. Lu, X., Wang, B., Zheng, X., Li, X.: Exploring models and data for remote sensing image caption generation. IEEE Transactions on Geoscience and Remote Sensing **56**(4), 2183–2195 (2017)
20. Luo, Y., Ren, X., Zheng, Z., Jiang, Z., Jiang, X., You, Y.: Came: Confidence-guided adaptive memory efficient optimization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (2023)
21. Na, B., Park, M., Sim, G., Shin, D., Bae, H., Kang, M., Kwon, S.J., Kang, W., Moon, I.C.: Diffusion adaptive text embedding for text-to-image diffusion models. In: Proceedings of the Neural Information Processing Systems (2026)
22. Pan, J., Lei, S., Fu, Y., Li, J., Liu, Y., Sun, Y., He, X., Peng, L., Huang, X., Zhao, B.: Earthsynth: Generating informative earth observation with diffusion models. arXiv preprint arXiv:2505.12108 (2025)
23. Pang, L., Cao, X., Tang, D., Xu, S., Bai, X., Zhou, F., Meng, D.: Hsigene: A foundation model for hyperspectral image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(1), 730–746 (2026)
24. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2023)
25. Persello, C., Hänsch, R., Vivone, G., Chen, K., Yan, Z., Tang, D., Huang, H., Schmitt, M., Sun, X.: 2023 ieee grss data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction [technical committees]. IEEE Geoscience and Remote Sensing Magazine **11**(1), 94–97 (2023)
26. 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: International Conference on Machine Learning (2021)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2022)
28. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems (2016)
29. Sebaq, A., ElHelw, M.: Rsdiff: Remote sensing image generation from text using diffusion model. Neural Computing and Applications **36**(36), 23103–23111 (2024)
30. Shi, X., Fu, S., Chen, J., Wang, F., Xu, F.: Object-level semantic segmentation on the high-resolution gaofen-3 fusar-map dataset. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 3107–3119 (2021)
31. Sigillo, L., He, S., Comminiello, D.: Latent wavelet diffusion for ultra-high-resolution image synthesis. arXiv preprint arXiv:2506.00433 (2025)
32. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
33. Simo-Serra, E., Iizuka, S., Sasaki, K., Ishikawa, H.: Learning to simplify: fully convolutional networks for rough sketch cleanup. ACM Transctions on Graphics **35**(4), 1–11 (2016)
34. Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., Xie, S.: What matters for representation alignment: Global information or spatial structure? arXiv preprint arXiv:2512.10794 (2025)

35. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
36. Tang, C., Powell, J., Koch, D., Mullins, R.D., Weddell, A.S., Chauhan, J.: Physwin: An efficient and physically-informed foundation model for multispectral earth observation. In: Neural Information Processing Systems (2025)
37. Tang, D., Cao, X., Hou, X., Jiang, Z., Liu, J., Meng, D.: Crs-diff: Controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
38. Tang, D., Cao, X., Wu, X., Li, J., Yao, J., Bai, X., Jiang, D., Li, Y., Meng, D.: Aerogen: Enhancing remote sensing object detection with diffusion-driven data generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
39. Toker, A., Eisenberger, M., Cremers, D., Leal-Taixé, L.: Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2024)
40. Wang, C., Lu, W., Li, X., Yang, J., Luo, L.: M4-sar: A multi-resolution, multi-polarization, multi-scene, multi-source dataset and benchmark for optical-sar fusion object detection. arXiv preprint arXiv:2505.10931 (2025)
41. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
42. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2024)
43. Wei, K., Dai, J., Hong, D., Ye, Y.: Mgfnet: An mlp-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images. International Journal of Applied Earth Observation and Geoinformation **135**, 104241 (2024)
44. Wu, C., Du, B., Zhang, L.: Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9774–9788 (2023)
45. Wu, G., Zhang, S., Shi, R., Gao, S., Chen, Z., Wang, L., Chen, Z., Gao, H., Tang, Y., Yang, J., et al.: Representation entanglement for generation: Training diffusion transformers is much easier than you think. In: Neural Information Processing Systems (2025)
46. Xia, J., Chen, H., Broni-Bediako, C., Wei, Y., Song, J., Yokoya, N.: Openearthmap-sar: A benchmark synthetic aperture radar dataset for global high-resolution land cover mapping. arXiv preprint arXiv:2501.10891 (2025)
47. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In: International Conference on Learning Representations (2025)
48. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2015)
49. Xu, Y., Xu, W., Cheung, D., Tu, Z.: Line segment detection using transformers without edges. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2021)
50. Xu, Y., He, Z., Kan, M., Shan, S., Chen, X.: Jodi: Unification of visual generation and understanding via joint modeling. arXiv preprint arXiv:2505.19084 (2025)

51. Xu, Y., Yu, W., Ghamisi, P., Kopp, M., Hochreiter, S.: Txt2img-mhn: Remote sensing image generation from text using modern hopfield networks. IEEE Transactions on Image Processing **32**, 5737–5750 (2023)
52. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (2025)
53. Ye, Y., Teng, X., Yang, H., Chen, S., Sun, Y., Bian, Y., Tan, T., Li, Z., Yu, Q.: 3mos: a multi-source, multi-resolution, and multi-scene optical-sar dataset with insights for multi-modal image matching. Visual Intelligence **3**(1), 19 (2025)
54. Ye, Z., Ma, S., Yang, J., Yang, X., Gong, Z., Yang, X., Wang, H.: Object fidelity diffusion for remote sensing image generation. In: International Conference on Learning Representations (2026)
55. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. In: International Conference on Learning Representations (2025)
56. Yu, Z., Liu, C., Liu, L., Shi, Z., Zou, Z.: Metaearth: A generative foundation model for global-scale remote sensing image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 1764–1781 (2025)
57. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
58. Zhang, M., Liu, Y., Liu, Y., Zhao, Y., Ye, Q.: Cc-diff: enhancing contextual coherence in remote sensing image synthesis. arXiv preprint arXiv:2412.08464 (2024)
59. Zhang, M., Liu, Y., Liu, Y., Zhao, Y., Ye, Q.: Cc-diff++: Spatially controllable text-to-image synthesis for remote sensing with enhanced contextual coherence. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–16 (2025)
60. Zhang, R., Yao, J., Zhang, K., Feng, C., Zhang, J.: S-cnn-based ship detection from high-resolution remote sensing images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **41**, 423–430 (2016)
61. Zhang, W., Hu, Y., Li, Y., Liu, Y.: Uni-rs: A spatially faithful unified understanding and generation model for remote sensing. arXiv preprint arXiv:2601.17673 (2026)
62. Zhang, Y., Yuan, Y., Feng, Y., Lu, X.: Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. IEEE Transactions on Geoscience and Remote Sensing **57**(8), 5535–5548 (2019)
63. Zhang, Z., Zhao, T., Guo, Y., Yin, J.: Rs5m and georsclip: A large-scale vision-language dataset and a large vision-language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–23 (2024)
64. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. In: Advances in Neural Information Processing Systems (2023)
65. Zhao, Y., Celik, T., Liu, N., Li, H.C.: A comparative analysis of gan-based methods for sar-to-optical image translation. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022)
66. Zhou, G., Chen, W., Gui, Q., Li, X., Wang, L.: Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2021)
67. Zhou, Y., Ma, A., Wang, J., Chen, Z., Zhong, Y.: Remote sensing meta modal representation for missing modality land cover mapping: From earthmiss dataset to metars method. Remote Sensing of Environment **333**, 115132 (2026)