RLI-DM: Robust Layout-Based Iterative Diffusion Model for SAR-to-RGB Image Translation

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

***Abstract*—SAR-to-RGB translation, which transforms Synthetic Aperture Radar (SAR) images into visually interpretable RGB counterparts, is critical for enhancing applications in visual analysis, deep learning, and multi-source data fusion. However, existing methods often fail to preserve both global structural integrity and fine-grained local textures. This deficiency stems from weak feature extraction and the lack of a robust layout framework, leading to outputs with information loss, geometric distortions, and unnatural textures. To overcome these limitations, we propose the Robust Layout-based Iterative Diffusion Model (RLI-DM), a novel three-stage framework for high-fidelity translation. The framework begins with an Optical Reconstruction Module that employs a conditional diffusion model to ensure precise spectral mapping. At its core, the Geometric Robustness Module leverages a Brownian bridge model that we train to derive a noise-resilient layout, overcoming the limitations of conventional edge detection and significantly enhancing global structural fidelity. Finally, this robust layout guides a Customized Multi-Level Refinement Module to iteratively reconstruct local textures, ensuring structural clarity and cross-feature consistency. Extensive experiments on multiple benchmark datasets demonstrate that RLI-DM achieves state-of-the-art performance, significantly outperforming existing methods in both structural integrity and perceptual quality.**

***Index Terms*—Remote sensing, SAR-to-RGB, Robust layout, Iterative refinement, Diffusion model**

I. Introduction

Synthetic Aperture Radar (SAR) has revolutionized remote sensing by providing reliable, all-weather, all-day imaging capabilities. Unlike optical sensors, SAR employs active microwave signals to penetrate atmospheric obscurants like clouds and capture detailed surface information critical for Earth observation, disaster response, and military reconnaissance [1–4]. This operational robustness makes SAR an indispensable tool where optical data is unavailable or unreliable. However, the inherent characteristics of SAR imagery—namely its grayscale format, speckle noise, and complex backscatter patterns—pose significant barriers to direct human interpretation and seamless integration with RGB-based analytical systems. Consequently, SAR-to-RGB image translation has emerged as a critical research area to bridge this modality gap, aiming to synergize SAR's operational advantages with the intuitive visual clarity of RGB images for both human-centric and machine-driven workflows.

Despite promising advancements using deep generative models [5–11], existing translation methods are fundamentally limited in their ability to reconcile global structural integrity with fine-grained local textures. This challenge is exacerbated by the specific failure modes of different model families. Generative Adversarial Networks (GANs)[12, 13], for instance, often prioritize plausible texture synthesis at the expense of global coherence, producing visually appealing but factually unfaithful results. Conversely, while Diffusion Models (DMs) offer more stable training, their direct application to this problem frequently leads to over-smoothing and the loss of critical details, as they lack specialized mechanisms to handle SAR's unique noise profile. A common weakness underpinning these issues is the reliance on noise-sensitive layout extraction techniques, like conventional edge detection, which yield unstable structural priors. This highlights a critical gap in the field: the absence of a framework that explicitly models a robust scene layout to guide the translation process and ensure fidelity.

To overcome these fundamental limitations, we propose the Robust Layout-based Iterative Diffusion Model (RLI-DM), a novel three-stage framework that establishes a new paradigm for high-fidelity SAR-to-RGB translation. Our approach is built upon the principle of decoupling structural representation from texture synthesis, using a robustly derived layout as the guiding backbone for the entire process. The framework begins with an **Optical Reconstruction Module (ORM)**, which utilizes a conditional diffusion model [14–19] to generate a high-quality initial RGB estimate. This stage effectively establishes a strong baseline for plausible color and texture distribution, even if minor structural inaccuracies persist. At the core of our approach, the **Geometric Robustness Module (GRM)** that we train employs a Brownian bridge model, a powerful tool for modeling smooth statistical transitions. By learning a mapping from the noisy SAR domain to a clean layout representation, the GRM inherently suppresses speckle artifacts and produces a structurally coherent and noise-resilient "robust layout," a stark contrast to the fragile outputs of traditional edge detectors. This stable layout then acts as a strong spatial prior, guiding the **Customized Multi-Level Refinement Module (CMRM)**. The CMRM iteratively refines the initial RGB estimate by "painting" realistic textures and high-frequency details directly onto the provided structural canvas, ensuring that local features are generated in a manner that is fully consistent with the global scene context. As

This work was supported in part by the National Natural Science Foundationof China under Grant 62471394, U21B2041, and 62501511

Bingxuan Zhao is with the School of Computer Science, and also with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China.

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

Corresponding author: Qi Wang.

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

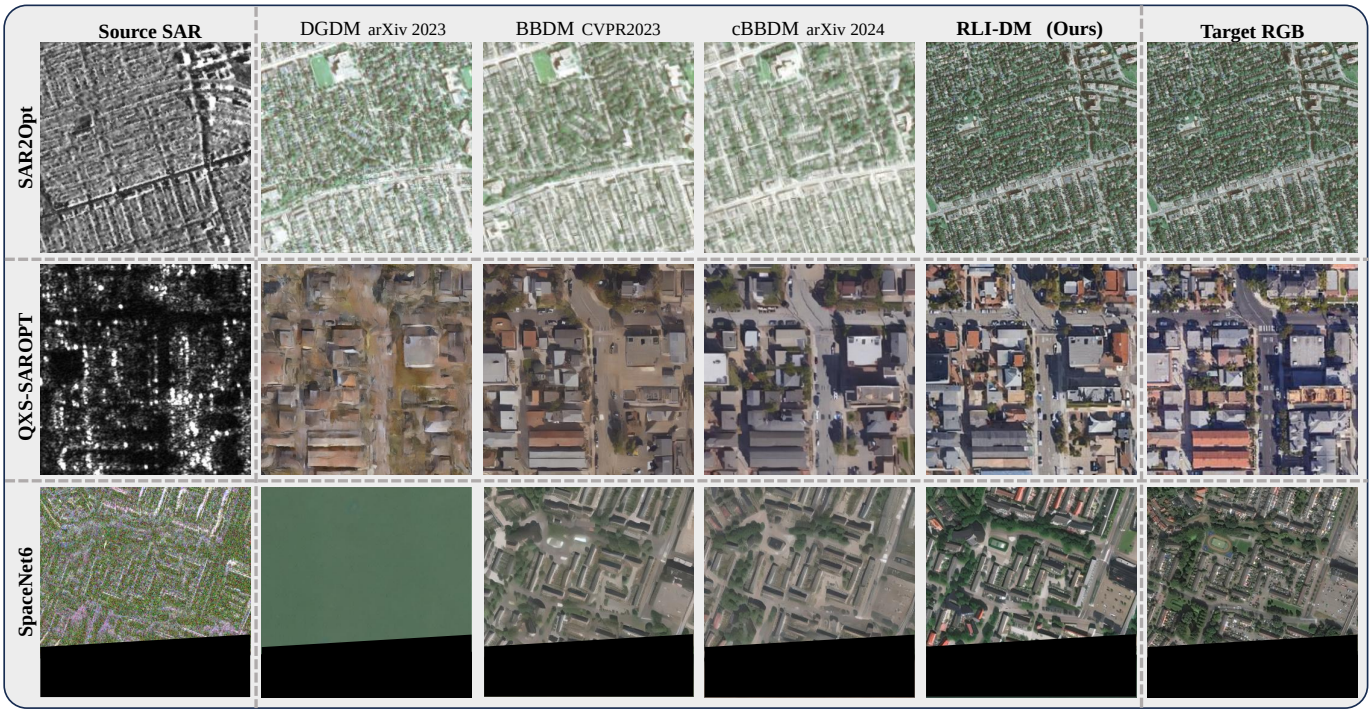

Fig. 1. Qualitative results of SAR-to-RGB image translation using our proposed RLI-DM framework. Rows 1, 2, and 3 show the translation results on the SAR2Opt, QXS-SAROPT, and SpaceNet6 datasets, respectively. Compared to recent methods, RLI-DM exhibits significant advantages in structural accuracy and visual fidelity.

demonstrated in Figure 1, this layout-driven, iterative process allows RLI-DM to deliver state-of-the-art results, achieving superior structural accuracy and visual fidelity.

In summary, the primary contributions of this work are fourfold:

1) We **introduce** a novel SAR-to-RGB translation framework, RLI-DM, which effectively reconciles the trade-off between preserving global structure and synthesizing realistic local textures.
2) We **develop** a Geometric Robustness Module (GRM) that leverages a Brownian bridge model to extract a robust layout, significantly improving the geometric precision and spatial coherence of generated images.
3) We **propose** a Customized Multi-Level Refinement Module (CMRM) that utilizes the robust layout for iterative, multi-level optimization, ensuring high-quality and structurally consistent translation.
4) We **conduct** comprehensive experiments on multiple benchmark datasets, demonstrating that RLI-DM establishes a new state-of-the-art in both quantitative metrics and perceptual quality.

The remainder of this paper is organized as follows. Section II reviews prior art in image generation and SAR-to-RGB translation. Section III details the proposed RLI-DM framework. Section IV presents our experimental setup, comparative results, and ablation studies. Finally, Section V concludes the paper and discusses future directions.

II. Related Work

A. Conditional Image Generation

The field of conditional image synthesis has been redefined by Diffusion Models (DMs) [14, 20], which achieve state-of-the-art fidelity through a progressive denoising process. To enhance computational tractability, Latent Diffusion Models (LDMs) [21] execute this process in a lower-dimensional latent space. The capabilities of these models are significantly expanded by conditioning mechanisms. For instance, text-conditioned models like Stable Diffusion [21] and Imagen [22] leverage pretrained CLIP [23] embeddings to align image generation with textual prompts. Furthermore, a diverse suite of control methodologies has been developed to afford finer user control, enabling spatial guidance via inputs like segmentation masks [24, 25] or personalized generation through few-shot fine-tuning, as exemplified by DreamBooth [26]. While these foundational techniques provide a powerful toolkit, their general-purpose design is not intrinsically tailored to the unique phenomenological and statistical properties of SAR data. Effectively translating these capabilities to the SAR domain thus necessitates a specialized framework.

B. Remote Sensing Image Generation

Building upon these generative tools, remote sensing image generation has emerged as a vibrant research area. Many studies in this domain primarily address *semantic-to-image synthesis*, where the goal is to generate satellite imagery from abstract inputs. Early efforts often adapted Generative Adversarial Networks (GANs) [27] for this purpose; for

example, GeoGAN [28] utilizes land-use labels to generate corresponding images, while MCGAN [29] employs semantic masks for land cover transformation. More recent work has shifted to DMs, with models like Text2Earth [30] generating high-fidelity imagery from natural language descriptions.

However, these methods address a fundamentally different problem than ours. Our work confronts the distinct challenge of *domain-to-domain translation*, where the objective is not to create new scenes but to faithfully translate an existing, structurally-rich source image into a different modality. This task imposes a stringent requirement for preserving the geometric and structural integrity of the input—a constraint far less central to text-driven synthesis. This fundamental distinction in objectives motivates a dedicated analysis of methods designed specifically for the SAR-to-RGB translation problem.

C. SAR-to-RGB Image Translation

The direct translation of SAR to RGB imagery is a long-standing objective for synergistic data interpretation. Early deep learning approaches adapted seminal image-to-image models, such as the GAN-based Pix2pix [31] for paired data and CycleGAN [31] for unpaired scenarios. While capable of producing visually plausible results, their adversarial nature often leads them to prioritize texture realism at the expense of structural fidelity, resulting in geometric distortions and artifacts.

More recently, DMs have emerged as a more powerful alternative due to their stable training and superior synthesis quality [32, 33]. Models like BBDM [32] and cBBDM [33] demonstrate improved consistency over GANs. Nevertheless, a critical bottleneck persists: these models typically lack a robust mechanism to extract and enforce a stable structural layout from the noisy SAR input. Without a reliable structural prior to guide the denoising process, they still struggle with over-smoothing fine details and ensuring precise geometric alignment. Consequently, a fundamental impasse remains in prior work—an inescapable trade-off between structural integrity and textural realism, stemming from the absence of an explicit, noise-resilient layout representation. Our RLI-DM framework is designed to directly break this impasse.

III. Methodology

A. Overall Framework

Transforming SAR imagery into high-fidelity RGB representations poses a unique challenge due to its inherent speckle noise and grayscale nature. This section details the technical implementation of our proposed RLI-DM, a multi-stage framework that synergizes conditional generative modeling, robust geometric feature extraction, and iterative refinement. We formulate the problem using a paired dataset $\mathcal{D} = \{(X^{(i)}, Y^{(i)}) \mid i \in \{1, \ldots, N\}\}$, where $X^{(i)} \in \mathbb{R}^{C_{\text{SAR}} \times H \times W}$ is the SAR input and $Y^{(i)} \in \mathbb{R}^{3 \times H \times W}$ is its ground-truth RGB counterpart. The objective is to learn a mapping $\mathcal{F} : X \to \hat{Y}$ that generates an accurate RGB prediction $\hat{Y}$.

As illustrated in Figure 2, our framework orchestrates a three-module pipeline. The process commences with the **Optical Reconstruction Module (ORM)**, where the SAR input X is first encoded into latent features $\mathcal{H}_{\text{I}} = \mathcal{F}_\theta(X)$ by a trainable image encoder $\mathcal{F}_\theta$. Simultaneously, a fixed prompt, "SAR image to RGB image modal transformation," is encoded by a pretrained CLIP text encoder $\mathcal{E}_T$ to produce a task-specific semantic prior $\mathcal{H}_{\text{T}}$. Both features condition a pretrained denoising U-Net $\mathcal{G}$ to generate an initial RGB estimate, Y_1:

$$Y_1 = \mathcal{G}(\mathcal{H}_{\text{T}}, \mathcal{H}_{\text{I}}). \tag{1}$$

The primary training objective at this stage is to optimize the image encoder's parameters θ by minimizing the reconstruction loss $\mathcal{L}(\theta) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\|Y_1 - Y\|_2^2]$, while the U-Net $\mathcal{G}$ remains frozen.

However, the structural integrity of this initial estimate Y_1 is often compromised by SAR's noise characteristics. To rectify this, the **Geometric Robustness Module (GRM)**, the cornerstone of our approach, extracts a noise-resilient structural representation, $L_{\text{robust}} = \text{GRM}(X)$. This robust layout serves as a stable geometric prior for the subsequent refinement process, which is handled by the **Customized Multi-Level Refinement Module (CMRM)**. The CMRM takes the output from the previous step, Y_i, and the robust layout, L_{robust}, to produce a refined version, Y_{i+1}. This process is applied iteratively, starting with Y_1:

$$Y_{i+1} = \text{CMRM}(Y_i, L_{\text{robust}}), \quad \text{for } i = 1, 2, \ldots, K-1, \tag{2}$$

where K is the total number of refinement iterations. This framework ensures high-quality translation by explicitly decoupling and then reconciling structural representation with textural synthesis.

B. Geometric Robustness Module (GRM)

The GRM is designed to overcome the limitations of conventional structural extraction techniques, such as Canny detection [34] or HED [35, 36]. These methods are notoriously sensitive to the high noise levels in SAR imagery and often produce fragmented or unstable edge maps, failing to capture the geometric configurations essential for RGB reconstruction. To overcome this, our module introduces the *robust layout*, a structural representation specifically designed to enhance clarity and resilience under SAR's adverse conditions.

The generation of this robust layout is a two-phase process. First, for each training pair, we create a high-quality supervision signal by extracting a target layout from the ground-truth RGB image Y. This is achieved using **Algorithm 1**, which leverages filtering and morphological operations to produce a clean, binary structural map. Second, we train a **Brownian Bridge Diffusion Model (BBDM)** [32] to learn the direct mapping from a noisy SAR input X to its corresponding target layout. BBDM is an image-to-image diffusion technique particularly well-suited for this task, as it models a smooth, statistical transition from a noisy source to a clean target, making it inherently robust to speckle noise.

The intermediate state z_t in BBDM at time step $t \in [0, T]$ is defined by the distribution:

$$z_t \sim \mathcal{N}\left(\left(1 - \frac{t}{T}\right) z_0 + \frac{t}{T} z_T, \sigma_t^2 \mathbf{I}\right), \tag{3}$$

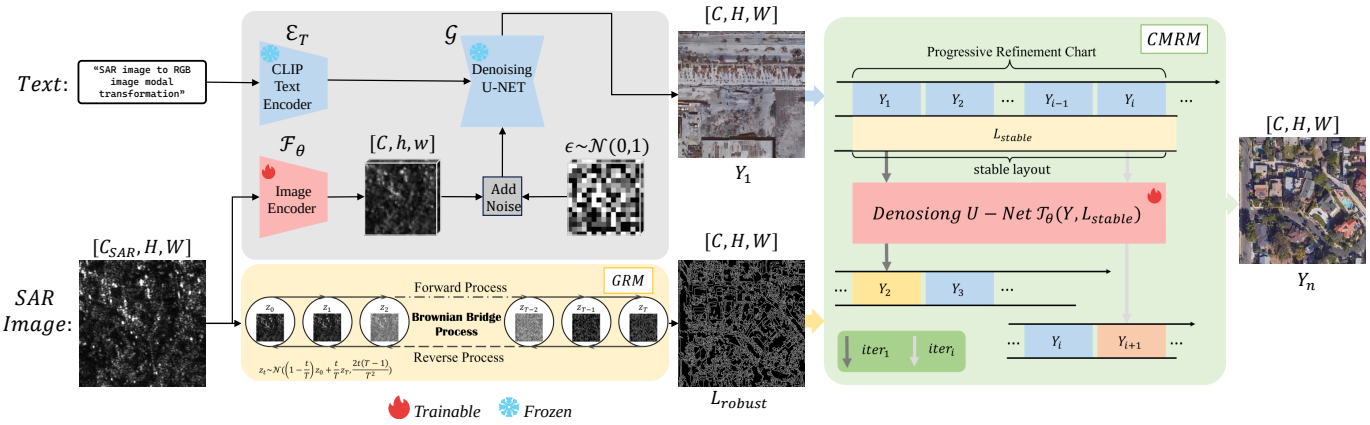

Fig. 2. Overall framework of our Robust Layout-based Iterative Diffusion Model (RLI-DM) for SAR-to-RGB translation. The framework operates via a three-stage pipeline. First, an **Optical Reconstruction Module (ORM)** generates an initial RGB estimate (Y_1) by conditioning a frozen denoising U-Net ($\mathcal{G}$) on latent features from the SAR image (X_{SAR}) and a text prompt. In parallel, a **Geometric Robustness Module (GRM)** employs a Brownian Bridge Process to derive a noise-resilient robust layout (L_{robust}) from the SAR input. Finally, the **Customized Multi-Level Refinement Module (CMRM)** leverages this layout as an explicit structural prior to iteratively refine the image from Y_1 to the final, high-fidelity output (Y_n). Trainable components are marked with a flame icon.

where z_0 is the SAR image, z_T is the target layout, $\sigma_t^2 = \frac{2t(T-t)}{T^2}$ is the noise variance, and T is the total number of time steps. This formulation ensures gradual noise suppression while preserving continuous geometric structures. Once trained, the GRM can take any new SAR image and generate a reliable robust layout, L_{robust}, which serves as a high-quality conditioning signal for the CMRM, providing a dependable structural basis for the final RGB reconstruction.

C. Customized Multi-level Refinement Module (CMRM)

The CMRM iteratively refines the generated RGB image, Y_i, by leveraging the high-quality 'robust layout' provided by the GRM. To effectively fuse the content from the previous image estimate with the structural guidance from the layout, the CMRM adopts a dual-path architecture, as illustrated in Figure 3. This design masterfully balances the stability of pretrained features with adaptive learning for the new structural condition.

The first path, the *content path*, is based on a frozen pretrained image encoder ($\mathcal{E}_I$, e.g., from CLIP). It takes the output from the previous iteration, Y_i, and encodes it into multi-resolution feature maps: $\mathcal{H}_{Y_i} = \mathcal{E}_I(Y_i)$. This ensures that the rich, pretrained knowledge of natural image textures is preserved and carried forward.

The second path, the *structure path*, is trainable and designed to interpret the robust layout. It utilizes a parallel encoder architecture to process the layout L_{robust} into structural features $\mathcal{H}_{L_{\text{robust}}}$. These structural features are then fused with the content features via element-wise addition to form a combined representation, $\mathcal{H}_{\text{combined}}$. This combined feature map is processed by a series of zero-initialized convolutional layers to produce a control signal, $\mathcal{H}_{\text{control}}$.

The critical step is the fusion of this control signal back into the main generation process. The $\mathcal{H}_{\text{control}}$ tensor is injected into the content path's U-Net decoder ($\mathcal{T}_{\text{left}}$) at multiple resolutions using a **cross-attention mechanism**. This allows the robust layout to provide explicit spatial guidance, dictating *where*

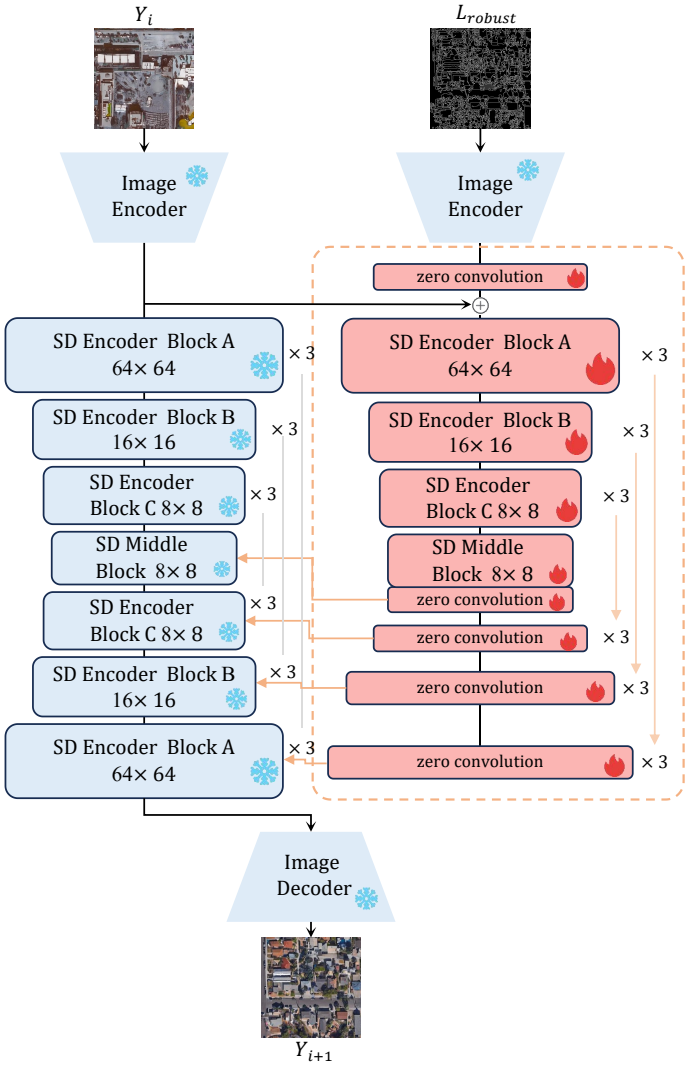

Fig. 3. Enhanced denoising U-Net $\mathcal{T}$ architecture in CMRM: a dual-branch framework with frozen and trainable paths, copy-based parameter initialization, and zero convolutions for cross-modal feature learning.

Algorithm 1 Enhanced Robust Layout Extraction for Remote Sensing Image

Require: Input image $I \in \mathbb{R}^{H \times W}$
Ensure: Enhanced binary layout map $S \in \{0,1\}^{H \times W}$

1: **Preprocessing:** Apply bilateral filtering to reduce noise:

$$\tilde{I} \leftarrow \text{BilateralFilter}(I, d = 5, \sigma_s = 50, \sigma_r = 0.1)$$

2: **Gradient Computation:** Use Scharr operators for edge detection:

$$K_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix}, \quad K_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix}$$

3: **for** each pixel $(i, j) \in [1, H] \times [1, W]$ **do**
4: $\quad G_x^{(i,j)} \leftarrow (\tilde{I} * K_x)[i, j]$
5: $\quad G_y^{(i,j)} \leftarrow (\tilde{I} * K_y)[i, j]$
6: $\quad E_{i,j} \leftarrow \sqrt{(G_x^{(i,j)})^2 + (G_y^{(i,j)})^2}$
7: **end for**
8: **Adaptive Thresholding:** Compute initial threshold via Otsu's method:

$$\tau_{\text{init}} \leftarrow \text{OtsuThreshold}(E)$$

9: $S \leftarrow \mathbb{I}(E > \tau_{\text{init}})$
10: **Edge Refinement:** Apply multi-scale non-maximum suppression:

$$S \leftarrow \text{MultiScaleNonMaxSuppression}(S)$$

11: **Threshold Linking:** Perform hysteresis thresholding:

$$S \leftarrow \text{HysteresisThreshold}$$

12: **Morphological Enhancement:** Close small gaps with a disk kernel:

$$S \leftarrow \text{MorphologicalClose}(S, \text{kernel} = \text{Disk}(3))$$

13: **return** $L_{\text{robust}} = S$

details should be sharpened and structures should be enforced. The refined image for the next iteration is then generated as:

$$Y_{i+1} = \mathcal{T}_{\text{left}}(\mathcal{H}_{Y_i}, \mathcal{H}_{\text{control}}). \tag{4}$$

The CMRM is trained with the following supervised loss to ensure alignment with ground-truth structure and detail:

$$\mathcal{L}_{\text{CMRM}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{H}_{\text{control}})\|_2^2 \right], \tag{5}$$

where $\mathbf{x}_0$ is the ground-truth image, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the noise added at timestep t, $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$, and $\boldsymbol{\epsilon}_\theta$ is the noise predictor conditioned on the control signal $\mathcal{H}_{\text{control}}$.

To further enhance feature adaptation, we strategically employ deformable and dilated convolutions within the CMRM. This sophisticated dual-path architecture, guided by cross-attention, enables the model to meticulously "paint" realistic textures onto the structurally accurate canvas provided by the GRM, achieving state-of-the-art results.

IV. Experiments

A. Datasets

We conducted a comprehensive evaluation of the RLI-DM framework using three publicly available SAR-to-RGB remote sensing datasets: QXS-SAROPT [37], SAR2Opt [38], and SpaceNet6 [39]. These datasets offer substantial diversity in satellite platforms, ground sampling distance (GSD), and SAR polarization modes, facilitating a robust assessment of RLI-DM's adaptability and generalization across varied real-world scenarios.

QXS-SAROPT Dataset [37] comprises 20,000 paired SAR (Gaofen-3) and RGB (Google Earth) images. The single-polarization SAR images and 1-meter GSD RGB patches primarily capture intricate maritime environments across various port cities.

SAR2Opt Dataset [38] includes 2,076 SAR (TerraSAR-X) and RGB (Google Earth) image pairs. The data, captured over diverse Asian cities with a 1-meter GSD, is particularly well-suited for urban analysis and infrastructure monitoring.

SpaceNet6 Dataset [39] features 3,401 SAR (Capella Space) and RGB (Maxar WorldView-2) image pairs. Its full-polarization SAR data and 0.5-meter GSD RGB imagery focus on urban landscapes, enabling detailed structural analysis for tasks like building detection.

B. Experimental Setup

All experiments were conducted using the PyTorch framework on four NVIDIA A100 GPUs in a distributed setup. We utilized Stable Diffusion v2.1 as the pretrained LDM backbone, keeping its CLIP-ViT-H/14 text encoder frozen during training. For all datasets, original images were randomly cropped to generate 512×512-pixel patches. We employed standard data augmentation techniques, including random horizontal and vertical flips and 90-degree rotations, to enhance model generalization. The datasets were partitioned into an 80% training set and a 20% testing set.

Framework performance was assessed using a comprehensive suite of five quantitative metrics. We employed Fréchet Inception Distance (FID) [47] to measure the realism of generated images and Learned Perceptual Image Patch Similarity (LPIPS) [48] to capture perceptual differences. To evaluate spatial and structural fidelity, we used Spatial Correlation Coefficient (SCC) [49] and Structural Similarity Index (SSIM) [50]. Finally, Peak Signal-to-Noise Ratio (PSNR) [51] was used to evaluate pixel-level reconstruction quality.

C. Comparisons with Existing Methods

a) Qualitative Comparison.: As illustrated in **Figure 1**, a qualitative comparison visually substantiates the superiority of our proposed RLI-DM framework. Conventional **GAN-based methods**, such as Pix2Pix and CycleGAN, largely fail to bridge the significant modality gap, resulting in outputs with severe blurriness, color imbalance, and a near-total loss of structural integrity. While recent **DM-based** methods like BBDM and cBBDM produce sharper textures, they still struggle with geometric fidelity, introducing noticeable artifacts

TABLE I
QUANTITATIVE COMPARISON OF IMAGE-TO-IMAGE TRANSLATION METHODS AND SET METHODS ON SAR2OPT AND SPACENET6 DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE IN EACH METRIC.

Types	Methods	Publications	SAR2Opt Dataset					SpaceNet6 Dataset				
			FID↓	LPIPS↓	SCC↑	SSIM↑	PSNR↑	FID↓	LPIPS↓	SCC↑	SSIM↑	PSNR↑
GANs	Pix2Pix [40]	CVPR 2017	196.87	0.426	0.0006	0.216	15.422	124.55	0.256	0.0102	0.522	19.357
	CycleGAN [31]	ICCV 2017	139.72	0.425	0.0022	0.224	14.931	114.81	0.274	0.0097	0.493	17.798
	NiceGAN [41]	CVPR 2020	141.13	0.435	0.0006	0.232	15.128	113.04	0.265	0.0091	0.443	17.022
	CRAN [42]	Sci China Inf Sci 2021	128.25	0.412	0.0009	0.236	15.113	113.31	0.273	0.0089	0.461	16.917
	CFCA-SET [43]	TGRS 2023	152.27	0.430	0.0009	0.223	15.183	164.78	0.279	0.0097	0.498	18.297
	StegoGAN [44]	CVPR 2024	144.54	0.398	0.0034	0.237	15.624	75.12	0.244	0.0106	0.516	18.958
DMs	BBDM [32]	CVPR 2023	94.72	0.473	0.0005	0.234	15.131	81.86	0.302	0.0019	0.217	17.678
	S2O-CSD [45]	IGARSS 2024	114.34	0.422	0.0002	0.213	14.946	114.34	0.422	0.0072	0.213	14.946
	DGDM [46]	ECCV 2024	156.12	0.541	0.0004	0.273	15.568	238.37	0.438	0.0015	0.253	17.124
	cBBDM [33]	arXiv 2024	97.64	0.394	0.0022	0.285	16.591	72.77	0.243	0.0079	0.254	19.033
	RLI-DM (Ours)	–	**87.81**	**0.386**	**0.0030**	**0.286**	**16.715**	**67.44**	**0.192**	**0.0175**	**0.577**	**22.312**

TABLE II
QUANTITATIVE COMPARISON OF IMAGE-TO-IMAGE TRANSLATION METHODS ON QXS-SAROPT DATASET. **BOLD** INDICATES THE BEST PERFORMANCE IN EACH METRIC.

Types	Methods	Publications	FID↓	LPIPS↓	SCC↑	SSIM↑	PSNR↑	Inference Time↓
GANs	Pix2Pix [40]	CVPR 2017	196.89	0.454	0.0000	0.247	14.924	45 ms
	CycleGAN [31]	ICCV 2017	195.38	0.455	0.0001	0.251	14.977	52 ms
	NiceGAN [41]	CVPR 2020	198.28	0.465	0.0002	0.242	15.128	62ms
	CRAN [42]	Sci China Inf Sci 2021	211.42	0.471	0.0001	0.263	15.317	65 ms
	CFCA-SET [43]	TGRS 2023	79.06	0.406	0.0006	0.273	15.094	69 ms
	StegoGAN [44]	CVPR 2024	85.60	0.391	0.0019	0.280	15.580	57 ms
DMs	S2O-CSD [45]	IGARSS 2024	114.34	0.422	0.0002	0.213	14.946	1.8 s
	BBDM [32]	CVPR 2023	65.15	0.522	0.0004	0.238	13.946	1.5 s
	DGDM [46]	ECCV 2024	147.23	0.634	0.0001	0.288	11.564	2.2 s
	cBBDM [33]	arXiv 2024	69.47	0.420	0.0023	0.304	16.248	1.4 s
	RLI-DM (Ours)	–	**59.81**	**0.301**	**0.0031**	**0.316**	**16.782**	1.2 s

such as local distortions (BBDM) and global misalignments (cBBDM). This demonstrates their inability to derive a stable structural layout from the noisy SAR input.

In stark contrast, our **RLI-DM** excels by overcoming these limitations. Echoing the design of our framework, the **Geometric Robustness Module** first establishes a noise-resilient layout, ensuring the global geometry—like the road network—is correctly preserved. Subsequently, this robust layout guides the **Customized Multi-Level Refinement Module** to iteratively reconstruct local textures with high precision. This results in an image that is not only perceptually sharp but also structurally sound, free from the distortions that plague other state-of-the-art methods and establishing a new benchmark in both structural integrity and perceptual quality. For a more comprehensive visual analysis, additional qualitative results are provided in the **Appendix A**.

b) Quantitative Analysis.: As shown in Tables I and II, GAN-based methods consistently exhibit suboptimal performance. Their high FID and LPIPS scores, coupled with low SSIM and PSNR values (e.g., Pix2Pix on SAR2Opt), quantitatively confirm their tendency to produce geometrically distorted and texturally inconsistent results. This aligns with the qualitative evidence in the Appendix, which reveals artifacts such as blurriness and color imbalance in their outputs.

While DM-based approaches generally outperform GANs, they still face significant challenges. For instance, BBDM and cBBDM show improved FID scores but suffer from high LPIPS values (e.g., 0.473 for BBDM on SAR2Opt), reflecting persistent pixel-level misalignments. DGDM, despite its deterministic initialization, yields inferior SSIM and PSNR scores, indicating a struggle with fine-grained texture synthesis.

In stark contrast, our proposed RLI-DM establishes a new state-of-the-art across all datasets and nearly all metrics. On SpaceNet6, for example, it reduces the best prior FID score from 72.77 (cBBDM) to **67.44** and LPIPS from 0.243 (cBBDM) to **0.192**. Simultaneously, it achieves the highest SSIM (0.577) and PSNR (22.312). This demonstrates RLI-DM's unique ability to concurrently enhance perceptual realism (lower FID/LPIPS) and preserve precise structural integrity (higher SSIM/PSNR), effectively breaking the trade-off that limits prior work. This superior performance is directly attributable to our three-stage design, where the robust layout from the GRM provides accurate structural guidance for the iterative refinement in the CMRM.

D. Ablation Study and Component Analysis

To validate the effectiveness of our proposed modules, we conducted a series of ablation studies on the large-scale QXS-SAROPT dataset.

a) Effectiveness of the Robust Layout.: To verify the superiority of our *robust layout*, we compared its performance as a control condition within the CMRM against several traditional edge-detection techniques. As detailed in Table III,

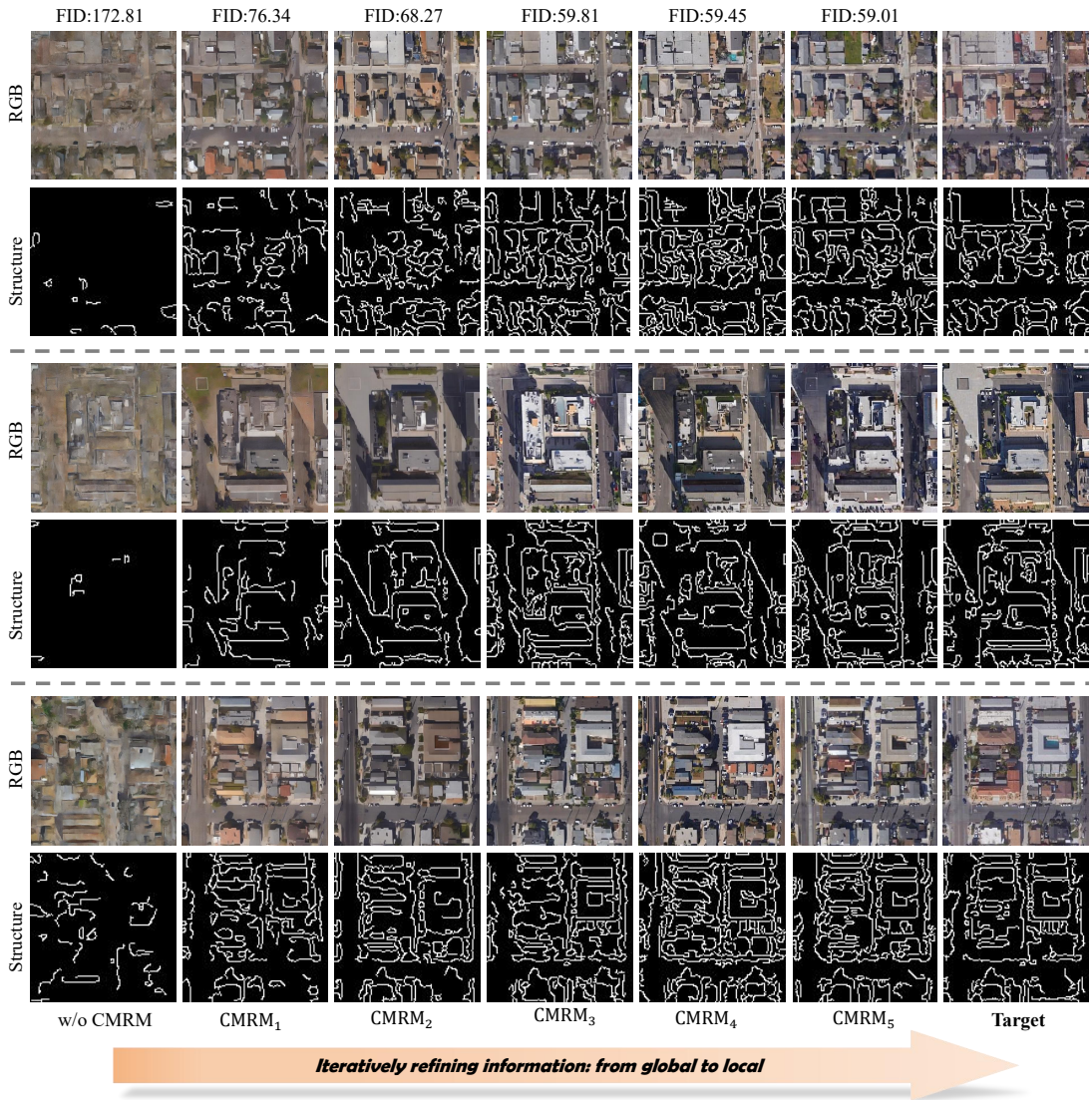

Fig. 4. Iterative Refinement Process of Generated Images. This figure illustrates the progressive refinement of three selected images (separated by dashed lines). The first row of each section depicts the RGB images evolving from an initial state (without CMRM) to the target RGB output, with FID scores decreasing from left to right, indicating improved fidelity. The second row reveals the structural evolution at each iteration, showcasing the transition from coarse global outlines to refined local details.

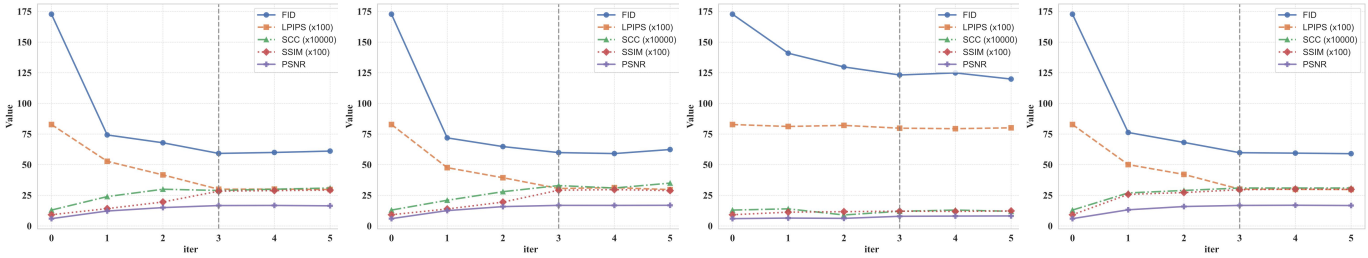

Fig. 5. Performance Curves of Different Architectures on the QXS-SAROPT Dataset, Showing Metric Variations Across Iterations. The four architectures—original structure, w/o 16×16 blocks, w/o both 16×16 and 32×32 blocks, and w/o 32×32 blocks—are presented sequentially. All performance metrics stabilize after the third iteration.

TABLE III
PERFORMANCE COMPARISON AFTER CMRM WITH DIFFERENT CONTROL CONDITIONS. **BOLD** INDICATES THE BEST PERFORMANCE IN EACH METRIC.

Control Condition	FID↓	LPIPS↓	SCC↑	SSIM↑	PSNR↑
Laplace edge	73.18	0.351	0.0022	0.273	13.145
Sobel edge	72.04	0.345	0.0026	0.225	12.145
Canny edge	64.40	0.302	0.0021	0.234	16.210
robust layout	**59.81**	**0.301**	**0.0031**	**0.316**	**16.782**

TABLE IV
STRUCTURAL REFINEMENTS FOR EFFICIENT CMRM. **BOLD** INDICATES THE BEST AND SECOND-BEST VALUES. "F" DENOTES THE FULL MODEL, "W/O 16" REFERS TO REMOVAL OF 16×16 BLOCKS, "W/O 16&32" INDICATES REMOVAL OF BOTH 16×16 AND 32×32 BLOCKS, AND "W/O 32" REPRESENTS OUR VARIANT EXCLUDING 32×32 BLOCKS. GD DENOTES GPU DAYS.

Variant	FID↓	LPIPS↓	SCC↑	SSIM↑	PSNR↑	GD↓
F	**59.23**	**0.300**	0.0029	0.284	16.689	3.5
w/o 16	59.89	0.306	**0.0033**	**0.292**	**16.832**	2.4
w/o 16&32	123.21	0.798	0.0012	0.121	7.921	**1.0**
w/o 32 (Ours)	**59.81**	**0.301**	**0.0031**	**0.296**	**16.782**	**1.6**

using Laplace, Sobel, or Canny edges as structural priors leads to significantly worse results across all metrics. For instance, the Canny-edge-guided model achieves an FID of 64.40, whereas our robust-layout-guided model lowers it to **59.81**. This experiment provides compelling evidence that the noise-resilient layout generated by our GRM is critical for achieving high-fidelity translation and is far more effective than conventional, noise-sensitive structural representations.

b) Efficacy of Iterative Refinement and CMRM Architecture: To validate the design of our **Customized Multi-level Refinement Module (CMRM)**, we first determined the optimal number of refinement iterations. The effectiveness of this iterative process is quantitatively analyzed in Figure 4, which plots key performance metrics as a function of the refinement step. The performance curves clearly show that all metrics, such as FID and LPIPS, improve sharply during the initial steps. Crucially, these gains diminish and the curves saturate around the third iteration, indicating that further steps yield minimal improvement. This trend confirms that an efficient yet highly effective result is achieved at three iterations. The detailed numerical data for these curves is provided in the **Appendix B**.

Based on this finding, we established three iterations as the standard for ensuring a fair and efficient comparison in our subsequent architectural ablation study. The results of this study, presented in Table IV, confirm that our final configuration ("w/o 32 (Ours)") achieves a superior trade-off between performance and computational cost (1.6 GPU-Days). It decisively outperforms both more computationally intensive baselines and heavily pruned variants. Collectively, this two-stage analysis validates that our deliberate design of the CMRM—adopting a three-step iterative process within an efficiently pruned architecture—is fundamental to achieving state-of-the-art performance cost-effectively.

V. CONCLUSION

In conclusion, we introduced RLI-DM, a state-of-the-art framework that significantly advances SAR-to-RGB image translation. By uniquely decoupling structure extraction from texture synthesis via a Brownian Bridge Diffusion Model and an iterative refinement strategy, our method successfully generates images with both high structural fidelity and rich perceptual details. This approach is validated by extensive experiments where RLI-DM outperforms all baseline methods on a comprehensive suite of metrics. Looking forward, our efforts will be directed at enhancing color accuracy through contextual data integration and developing lightweight architectures to boost efficiency. These future steps will unlock the potential of RLI-DM for critical real-world applications, including real-time disaster monitoring and urban planning.

REFERENCES

[1] F. Tosti, V. Gagliardi, F. D'Amico, and A. M. Alani, "Transport infrastructure monitoring by data fusion of gpr and sar imagery information," *Transportation Research Procedia*, vol. 45, pp. 771–778, 2020.

[2] L. White, B. Brisco, M. Dabboor, A. Schmitt, and A. Pratt, "A collection of sar methodologies for monitoring wetlands," *Remote sensing*, vol. 7, no. 6, pp. 7615–7645, 2015.

[3] Y. Yamaguchi, "Disaster monitoring by fully polarimetric sar data acquired with alos-palsar," *Proceedings of the IEEE*, vol. 100, no. 10, pp. 2851–2860, 2012.

[4] Z. Zhao, K. Ji, X. Xing, H. Zou, and S. Zhou, "Ship surveillance by integration of space-borne sar and ais–review of current research," *The Journal of Navigation*, vol. 67, no. 1, pp. 177–189, 2014.

[5] C. Yang, B. Zhao, Q. Zhou, and Q. Wang, "Mmo-ig: Multi-class and multi-scale object image generation for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[6] C. Yang, K. Zhuang, M. Chen, H. Ma, X. Han, T. Han, C. Guo, H. Han, B. Zhao, and Q. Wang, "Traffic sign interpretation via natural language description," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[7] C. Yang, X. Han, T. Han, H. Han, B. Zhao, and Q. Wang, "Edge approximation text detector," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[8] 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.

[9] J. Zhang, J. Zhou, and X. Lu, "Feature-guided sar-to-optical image translation," *Ieee Access*, vol. 8, pp. 70 925–70 937, 2020.

[10] J. Zhang, Y. Liu, G. Ding, B. Tang, and Y. Chen, "Adaptive decomposition and extraction network of individual fingerprint features for specific emitter identification," *IEEE Transactions on Information Forensics and Security*, 2024.

[11] Y. Li, L. Wang, T. Wang, X. Yang, J. Luo, Q. Wang, Y. Deng, W. Wang, X. Sun, H. Li *et al.*, "Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 47, no. 3, pp. 1832–1849, 2025.

[12] Z. Han, Z. Zhang, S. Zhang, G. Zhang, and S. Mei, "Aerial visible-to-infrared image translation: Dataset, evaluation, and baseline," *Journal of remote sensing*, vol. 3, p. 0096, 2023.

[13] Z. Han, S. Zhang, Y. Su, X. Chen, and S. Mei, "Dr-avit: Toward diverse and realistic aerial visible-to-infrared image translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[14] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[15] J. O. Cross-Zamirski, P. Anand, G. Williams, E. Mouchet, Y. Wang, and C.-B. Schönlieb, "Class-guided image-to-image diffusion: Cell painting from brightfield images with class labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3800–3809.

[16] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[17] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[18] 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.

[19] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2302.08453

[20] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[21] 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.

[22] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022. [Online]. Available: https://arxiv.org/abs/2205.11487

[23] 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," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[24] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in *European Conference on Computer Vision*. Springer, 2022, pp. 89–106.

[25] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin, "Spatext: Spatio-textual representation for controllable image generation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023, p. 18370–18380. [Online]. Available: http://dx.doi.org/10.1109/CVPR52729.2023.01762

[26] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2023. [Online]. Available: https://arxiv.org/abs/2208.12242

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[28] S. Ganguli, P. Garzon, and N. Glaser, "Geogan: A conditional gan with reconstruction and style loss to generate standard layer of maps from satellite images," 2019. [Online]. Available: https://arxiv.org/abs/1902.05611

[29] G. Ji, Z. Wang, L. Zhou, Y. Xia, S. Zhong, and S. Gong, "Sar image colorization using multidomain cycle-consistency generative adversarial network," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 2, pp. 296–300, 2020.

[30] C. Liu, K. Chen, R. Zhao, Z. Zou, and Z. Shi, "Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model," 2025. [Online]. Available: https://arxiv.org/abs/2501.00895

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020. [Online]. Available: https://arxiv.org/abs/1703.10593

[32] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Bbdm: Image-to-image translation with brownian bridge diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2205.07680

[33] S.-H. Kim and D.-w. Chung, "Conditional brownian bridge diffusion model for vhr sar to optical image translation," *arXiv preprint arXiv:2408.07947*, 2024.

[34] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3828–3837.

[35] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1395–1403.

[36] R. Grompone von Gioi and G. Randall, "A brief analysis of the holistically-nested edge detector," *Image Process-*

ing On Line, vol. 12, pp. 369–377, 2022.
[37] M. Huang, Y. Xu, L. Qian, W. Shi, Y. Zhang, W. Bao, N. Wang, X. Liu, and X. Xiang, "The qxs-saropt dataset for deep learning in sar-optical data fusion," *arXiv preprint arXiv:2103.08259*, 2021.
[38] Y. Zhao, T. Celik, N. Liu, and H.-C. Li, "A comparative analysis of gan-based methods for sar-to-optical image translation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
[39] J. Shermeyer, D. Hogan, J. Brown, A. Van Etten, N. Weir, F. Pacifici, R. Hansch, A. Bastidas, S. Soenen, T. Bacastow *et al.*, "Spacenet 6: Multi-sensor all weather mapping dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 196–197.
[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018. [Online]. Available: https://arxiv.org/abs/1611.07004
[41] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8168–8177.
[42] S. Fu, F. Xu, and Y.-Q. Jin, "Reciprocal translation between sar and optical remote sensing images with cascaded-residual adversarial networks," *Science China Information Sciences*, vol. 64, pp. 1–15, 2021.
[43] J. Lee, H. Cho, D. Seo, H.-H. Kim, J. Jeong, and M. Kim, "Cfca-set: Coarse-to-fine context-aware sar-to-eo translation with auxiliary learning of sar-to-nir translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.
[44] S. Wu, Y. Chen, S. Mermet, L. Hurni, K. Schindler, N. Gonthier, and L. Landrieu, "Stegogan: Leveraging steganography for non-bijective image-to-image translation," 2024. [Online]. Available: https://arxiv.org/abs/2403.20142
[45] X. Bai and F. Xu, "Sar to optical image translation with color supervised diffusion model," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 963–966.
[46] D. Yoon, M. Seo, D. Kim, Y. Choi, and D. Cho, "Deterministic guidance diffusion model for probabilistic weather forecasting," 2023. [Online]. Available: https://arxiv.org/abs/2312.02819
[47] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6626–6637, 2017.
[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.
[49] R. Dosselmann and X. D. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image and Video Processing*, vol. 2, no. 2, pp. 81–88, 2008.
[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[51] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.

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.

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.

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/).