GeoSET: Generalist Foundation Model for SAR-to-EO Image Translation

Anonymous authors
Paper under double-blind review

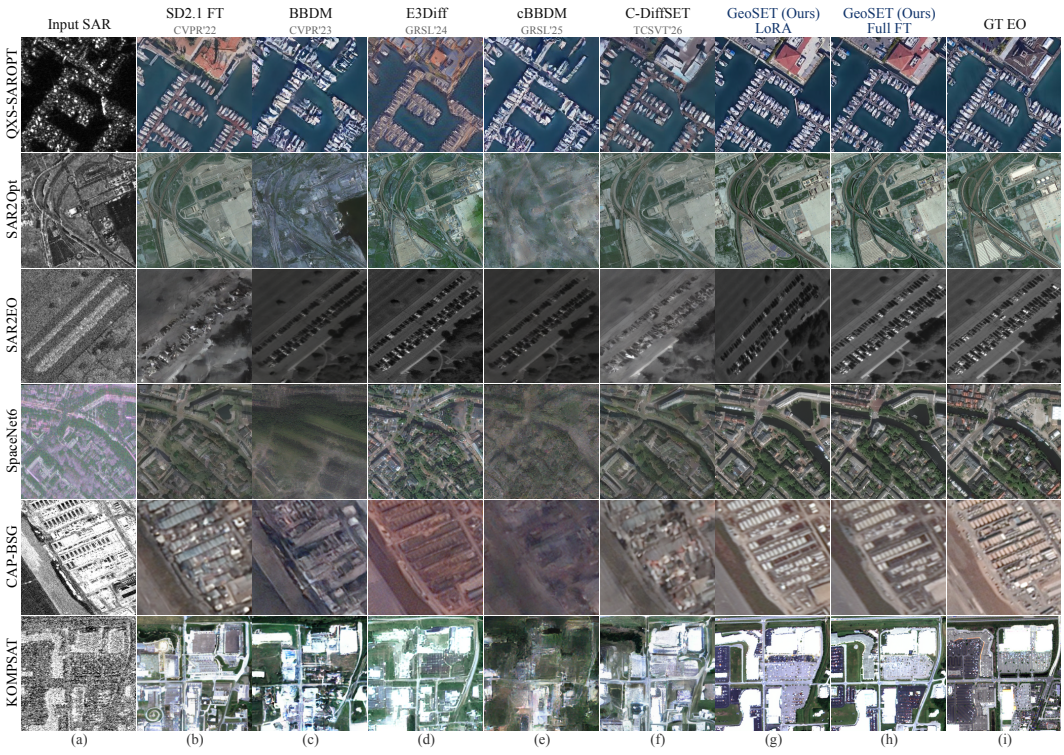

Figure 1: **Qualitative comparison on six SAR-to-EO image translation (SET) benchmarks.** From left to right: (a) input SAR, (b) SD2.1 fine-tuning, (c) BBDM, (d) E3Diff, (e) cBBDM, (f) C-DiffSET, (g) GeoSET with LoRA, (h) GeoSET with full fine-tuning, and (i) ground-truth EO. GeoSET consistently produces coherent EO images while preserving fine-scale scene structure.

Abstract

Paired synthetic aperture radar (SAR) and electro-optical (EO) imagery is increasingly available across sensors, resolutions, and geographic regions. Yet existing SAR-to-EO image translation (SET) methods are typically trained on a single, limited-scale dataset, producing models specialized to particular sensing conditions. We introduce **GeoSET**, *the first generalist model for SET*, built around a single pretrained parent that is adapted to downstream datasets under a common protocol. We curate over 3 million high-quality SAR–EO pairs from a collection of more than 10 million SAR observations, spanning diverse sensors, spatial resolutions, and ground sampling distances. To bridge the modality gap between SAR observations and a pretrained image generator, we develop a speckle-robust SAR encoder and pretrain the conditional generator on this heterogeneous corpus. The resulting parent supports efficient adaptation across downstream datasets through low-rank adaptation (LoRA), updating only 0.60% of the generator parameters and requiring approximately one hour per dataset. Across six downstream benchmarks, GeoSET achieves state-of-the-art results in FID and DISTS with full fine-tuning or LoRA, demonstrating effective transfer across heterogeneous SAR–EO domains. Code and pretrained checkpoints will be released upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Synthetic aperture radar (SAR) enables Earth observation regardless of solar illumination and through most cloud cover (Moreira et al., 2013), making it particularly valuable when electro-optical (EO) observations are unavailable or unreliable. However, speckle and geometry-dependent scattering responses make SAR imagery difficult for humans to interpret and distinguish its appearance from the natural images used to pretrain modern vision models. SAR-to-EO image translation (SET) bridges this gap by mapping SAR observations into intuitive EO-like representations (Zhao et al., 2022; Shermeyer et al., 2020). Such representations facilitate rapid visual analysis and may provide a familiar interface through which large-scale vision models pretrained on natural images can be applied to SAR observations (Liu et al., 2024; Carion et al., 2026). Because SAR and EO capture different physical signals, SET outputs should be interpreted as plausible SAR-conditioned visualizations rather than exact reconstructions of unobserved EO measurements.

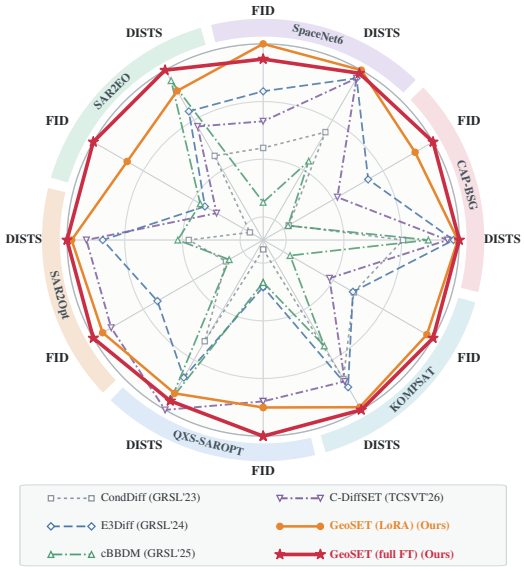

Figure 2: **Cross-dataset comparison of SAR-to-EO image translation methods.** We report FID and DISTS on six benchmarks, independently normalized for each dataset–metric pair as $100 \times \text{best}/\text{value}$.

SET has evolved from adversarial translation (Turnes et al., 2020; Guo et al., 2024; Lee et al., 2023) to diffusion-based generation (Bai et al., 2023; Qin et al., 2024; Kim & Chung, 2025; Do et al., 2026). Despite these advances, existing methods typically develop a separate dataset-specific model using only the SAR–EO training pairs available in each target benchmark. Even when initialized from pretrained image generators, these models do not share a reusable SET prior learned across heterogeneous paired sources. This raises a fundamental question: *Can SET learn a reusable generative prior from heterogeneous SAR–EO sources and adapt it across sensors and datasets under a common protocol?*

We address this question with **GeoSET**, to the best of our knowledge, the first generalist model for SET under a *pretrain-once, adapt-many* framework. We construct a heterogeneous pretraining corpus to exploit scene structures shared across sensors, polarizations, spatial resolutions, ground sampling distances (GSDs), and geographic regions. Naively pooling these sources, however, also aggregates misregistered, low-quality, and invalid pairs. Starting from a collection of over ten million SAR observations, we identify eligible SAR–EO pairs and apply source-aware filtering to obtain over three million high-quality pairs. GeoSET is pretrained once on this corpus and subsequently adapted to every downstream benchmark using a fixed protocol for each adaptation mode. Each downstream model thus inherits a shared cross-source prior rather than learning the SAR-to-EO mapping solely from its target dataset.

Heterogeneous pretraining alone does not resolve the modality mismatch between SAR observations and generators pretrained on natural imagery. Existing latent SET methods typically use a natural-image-pretrained autoencoder to encode both the EO target and the SAR condition (Kim & Chung, 2025; Do et al., 2026; Rombach et al., 2022). Although well suited to natural images, its encoder has not been optimized for multiplicative speckle or sensor-dependent SAR statistics. To obtain stable and informative SAR conditioning, we propose a *speckle-robust SAR encoder*. The encoder learns to recover the original SAR observation from a speckle-perturbed input, encouraging it to preserve scene structures while reducing sensitivity to speckle variations. Meanwhile, keeping the pretrained decoder fixed encourages the learned SAR representation to remain compatible with its latent space, providing a common interface for SAR conditioning and EO generation.

Building on this SAR-compatible representation, GeoSET repurposes a pretrained text-to-image generator (Labs, 2025) for spatial SAR conditioning. We replace its text-conditioning stream with a SAR stream initialized from the corresponding pretrained image-stream parameters, allowing both SAR and EO inputs to be processed as spatial latent tokens. With the SAR encoder and EO autoencoder frozen, we pretrain the resulting conditional generator on the heterogeneous corpus using flow matching (Lipman et al., 2022). The pretrained parent supports both full fine-tuning and parameter-efficient LoRA (Hu et al., 2021), with LoRA adaptation requiring approximately one hour per downstream dataset.

We evaluate GeoSET on six downstream benchmarks spanning diverse sensors, resolutions, GSDs, and geographic settings. Figure 2 summarizes its performance relative to existing SET methods. Across full fine-tuning and LoRA, GeoSET achieves the best reported FID on all six benchmarks and the best DISTS on five. LoRA updates only 0.6% of the generator parameters while remaining competitive with full fine-tuning. Component comparisons further support the value of SAR-specific encoding and speckle augmentation. Together, these results demonstrate the effectiveness of a reusable SET prior across heterogeneous SAR–EO domains.

Our contributions are summarized as follows:

- We introduce **GeoSET**, to the best of our knowledge, *the first generalist SET model under a pretrain-once, adapt-many framework*, pretrained on over three million SAR–EO pairs curated from heterogeneous sources through source-aware filtering.
- We propose a two-stage pretraining strategy that combines decoder-compatible, speckle-robust SAR representation learning with multi-source pretraining of a SAR-conditioned flow model.
- We conduct extensive evaluations and component analyses across six downstream benchmarks. GeoSET achieves state-of-the-art (SOTA) performance on most datasets, while its LoRA variant updates only 0.6% of the generator parameters and remains competitive with full fine-tuning.

2 Related Work

Conditional image generation. Conditional GANs established a widely used framework for paired image-to-image translation (Isola et al., 2017), while cycle consistency enabled translation between unpaired domains (Zhu et al., 2017). Subsequent methods improved high-resolution synthesis and spatial controllability through coarse-to-fine generation and spatially adaptive normalization (Wang et al., 2018; Park et al., 2019). More recently, diffusion-based approaches have advanced conditional generation through iterative denoising (Saharia et al., 2022), Brownian-bridge formulations (Li et al., 2023), and explicit spatial conditioning of pretrained text-to-image models (Zhang et al., 2023). Latent diffusion reduces computational costs by operating in a compressed latent space (Rombach et al., 2022), diffusion transformers provide scalable generative architectures (Peebles & Xie, 2023), and flow matching offers a framework for learning continuous generative dynamics (Lipman et al., 2022).

SAR-to-EO image translation. Early SET methods predominantly relied on adversarial learning, incorporating multiscale context, scene-level embeddings, or EO-derived semantic supervision (Turnes et al., 2020; Guo et al., 2024; Lee et al., 2023). Recent approaches adopt diffusion and bridge formulations to improve generation quality and efficiency (Bai et al., 2023; Qin et al., 2024; Kim & Chung, 2025; Do et al., 2026). Among these, latent approaches reduce computational costs using autoencoders pretrained on natural imagery, often employing the same encoder for both the SAR condition and the EO target (Kim & Chung, 2025; Do et al., 2026). However, without SAR-specific adaptation, these pretrained encoders may not adequately capture the distinct statistics of coherent radar imagery. More importantly, existing SET methods typically optimize a separate model for each target dataset, without jointly leveraging heterogeneous SAR–EO sources to learn a shared generative prior. GeoSET addresses these limitations through speckle-robust SAR representation learning and multi-source pretraining of a reusable conditional generator.

Large-scale SAR–EO corpora. The growing availability of paired SAR–EO imagery provides new opportunities for learning cross-modal mappings across diverse sensing conditions. SAR-1M (Liu et al., 2026), although introduced for masked SAR pretraining, includes EO counterparts

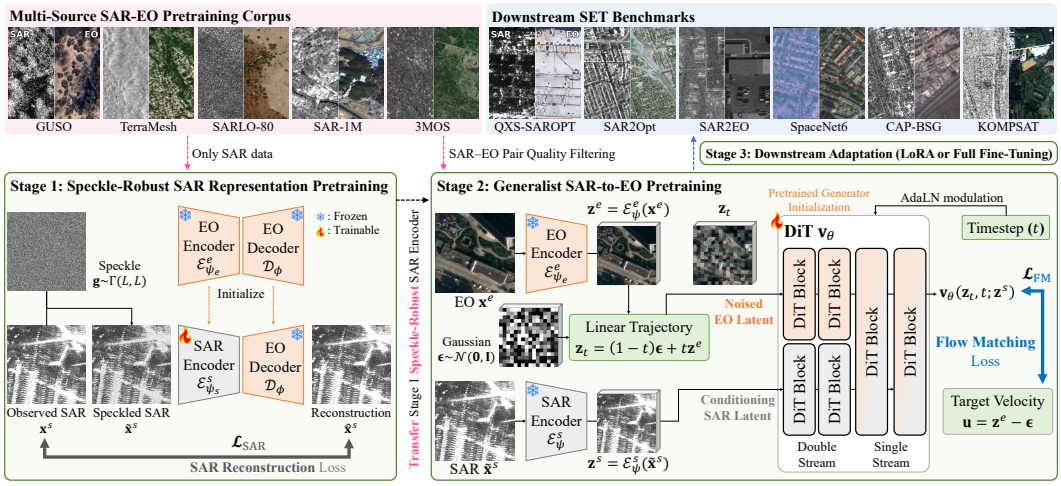

Figure 3: **GeoSET framework.** Stage 1 initializes a SAR encoder from a pretrained autoencoder and trains it to reconstruct the original SAR observation from a domain-aware speckle perturbation while the shared decoder remains frozen. Stage 2 replaces the pretrained generator's text condition with a spatial SAR stream initialized from its image stream and learns a conditional flow from Gaussian noise to the EO latent. Stage 3 adapts the resulting generalist to each downstream benchmark with LoRA or full fine-tuning.

for a substantial portion of its SAR observations. Other recent datasets, including GUSO (Yan et al., 2026), TerraMesh (Blumenstiel et al., 2025), SARLO-80 (Debuysère et al., 2026), and 3MOS (Ye et al., 2025), further expand paired coverage across geographic regions, sensors, spatial resolutions, and ground sampling distances. Their heterogeneity, however, makes direct pooling nontrivial: sources differ in spatial overlap, registration quality, image validity, crop density, and dataset size. GeoSET consolidates these complementary sources into a unified corpus of high-quality pairs for generative pretraining, using explicit pair filtering and source sampling based on crop-equivalent counts to account for differences in data quality, image size, and sampling density.

3 GeoSET

GeoSET implements the *pretrain-once, adapt-many* framework through two pretraining stages followed by downstream adaptation. Figure 3 illustrates the GeoSET framework. First, we initialize a SAR encoder from a pretrained image autoencoder and train it to reconstruct the original SAR observations from speckle-perturbed input. Reconstruction from perturbed inputs encourages speckle robustness, while keeping the decoder fixed encourages compatibility with its pretrained latent interface. Second, we convert a pretrained text-to-image generator (Labs, 2025) into a SAR-conditioned model by replacing its text-conditioning stream with a spatial SAR-conditioning stream initialized from the pretrained image stream. With the SAR encoder and EO autoencoder frozen, the resulting transformer $\mathbf{v}_\theta$ is trained on the filtered multi-source SAR–EO corpus using conditional flow matching (Lipman et al., 2022). Finally, the same generalist checkpoint is adapted to each downstream benchmark through either full fine-tuning or parameter-efficient LoRA (Hu et al., 2021).

3.1 Stage 1: Speckle-Robust SAR Encoder

Speckle-denoising reconstruction. Let $\mathcal{E}^e_{\psi_e}$ and $\mathcal{D}_\phi$ denote the encoder and decoder of a pretrained image autoencoder. We initialize the SAR encoder $\mathcal{E}^s_{\psi_s}$ from $\mathcal{E}^e_{\psi_e}$ and keep $\mathcal{D}_\phi$ frozen. For an observed SAR image $\mathbf{x}^s$ from source r, we generate a perturbed input $\widetilde{\mathbf{x}}^s = \mathcal{A}_{\text{spk}}(\mathbf{x}^s; r)$, where $\mathcal{A}_{\text{spk}}$ is the domain-aware speckle augmentation operator defined below. The reconstruction is

$$\widehat{\mathbf{x}}^s = \mathcal{D}_\phi\left(\mathcal{E}^s_{\psi_s}(\widetilde{\mathbf{x}}^s)\right), \qquad \mathcal{L}_{\text{SAR}} = \mathbb{E}_{\mathbf{x}^s}\left[\|\widehat{\mathbf{x}}^s - \mathbf{x}^s\|_1\right], \tag{1}$$

where only the SAR encoder $\mathcal{E}^s_{\psi_s}$ is optimized. Fixing the decoder $\mathcal{D}_\phi$ prevents $\mathcal{E}^s_{\psi_s}$ and $\mathcal{D}_\phi$ from jointly forming an unconstrained SAR-specific codec, encouraging SAR features that remain com-

patible with the pretrained latent interface of $\mathcal{D}_\phi$. The reconstruction target $\mathbf{x}^s$ is the original SAR observation rather than a speckle-free reference. The objective therefore encourages robustness to the added perturbation without requiring clean SAR supervision. Stage 1 uses only SAR images and requires neither paired EO images nor cross-modal alignment.

Domain-aware speckle augmentation. SAR sources encode radar responses in different numerical domains, making a single multiplicative augmentation inappropriate for all datasets. For each source r, let d_r denote its representation domain—linear intensity, linear amplitude, logarithmic measurement, display imagery, or unknown. We then apply the corresponding routed operator $\mathcal{R}_{d_r}$. Following the standard Gamma model of fully developed speckle (Lee et al., 1994), we sample an i.i.d. multiplicative field $\mathbf{g} \sim \Gamma(L, L)$ using the shape–rate parameterization, where $L \in \{4, 8, 16\}$, so that each element has mean 1 and variance $1/L$. The augmented SAR image $\widetilde{\mathbf{x}}^s$ is defined as

$$\widetilde{\mathbf{x}}^s = \mathcal{A}_{\mathrm{spk}}(\mathbf{x}^s; r) = \mathcal{R}_{d_r}(\mathbf{x}^s; \mathbf{g}). \tag{2}$$

Note that this speckle augmentation operation provides a controlled robustness perturbation rather than simulating a new lower-look acquisition, since the observed input already contains acquisition-dependent speckle. The same augmentation is applied to SAR conditions during generalist pretraining and downstream adaptation, and is disabled at inference; detailed routing rules are provided in the *Supplementary Material*.

3.2 Stage 2: Generalist SAR-to-EO Pretraining

SAR and EO latents. For each retained SAR–EO pair $(\mathbf{x}^s, \mathbf{x}^e)$, we apply the domain-aware speckle augmentation only to the SAR observation $\mathbf{x}^s$, yielding $\widetilde{\mathbf{x}}^s$, while leaving the paired EO target $\mathbf{x}^e$ unchanged. The corresponding latent representations are

$$\mathbf{z}^s = \mathcal{E}^s_{\psi_s}\big(\widetilde{\mathbf{x}}^s\big), \qquad \mathbf{z}^e = \mathcal{E}^e_{\psi_e}\big(\mathbf{x}^e\big), \tag{3}$$

where $\mathcal{E}^s_{\psi_s}$ is the speckle-robust SAR encoder learned in Stage 1 and $\mathcal{E}^e_{\psi_e}$ is the pretrained EO encoder. Both encoders ($\mathcal{E}^s_{\psi_s}, \mathcal{E}^e_{\psi_e}$) and decoder ($\mathcal{D}_\phi$) remain frozen in Stage 2.

Repurposing a pretrained generator for SAR conditioning. We repurpose a pretrained text-to-image flow transformer comprising double-stream and single-stream blocks. The text encoder is removed, and the text-conditioning stream is replaced with a spatial SAR stream, while the image stream processes the noisy EO latent $\mathbf{z}_t$. We initialize the SAR stream by copying the corresponding pretrained image-stream parameters and retain the remaining generator weights. The SAR and EO tokens interact in the double-stream blocks and are jointly processed in the single-stream blocks; only the EO tokens are passed to the final velocity head.

Conditional latent flow matching. We learn a conditional flow from an independent Gaussian source to the EO latent $\mathbf{z}^e$. Specifically, we sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \sim \mathcal{U}[0, 1]$, and define the linear probability path and its target velocity as

$$\mathbf{z}_t = (1 - t)\boldsymbol{\epsilon} + t\mathbf{z}^e, \qquad \mathbf{u} = \frac{\partial \mathbf{z}_t}{\partial t} = \mathbf{z}^e - \boldsymbol{\epsilon}. \tag{4}$$

To enable classifier-free guidance (CFG) (Ho & Salimans, 2022), we independently replace the SAR latent with an all-zero latent with probability $p_{\mathrm{drop}} = 0.1$ for each training sample. We train the generator $\mathbf{v}_\theta$ to predict the target velocity $\mathbf{u}$ by minimizing

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\left\|\mathbf{v}_\theta\big(\mathbf{z}_t, t; \mathbf{z}^s\big) - \mathbf{u}\right\|_2^2\right]. \tag{5}$$

Because the dropped condition is represented by an all-zero SAR latent, the generator $\mathbf{v}_\theta$ explicitly learns both conditional and null velocity fields. At inference, we sample $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, combine the conditional and null predictions using two-pass CFG, integrate the resulting velocity field from $t = 0$ to $t = 1$, and decode the endpoint using the frozen decoder $\mathcal{D}_\phi$.

3.3 Stage 3: Downstream Adaptation

For each downstream dataset, we initialize GeoSET from the final exponential moving average (EMA) checkpoint of the Stage 2 generalist model while keeping the SAR and EO encoders

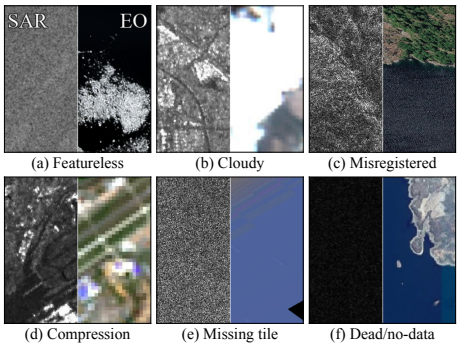

Figure 4: **Pretraining data filtering.**

Table 1: **Stage 2 SAR–EO pretraining corpus.** Drop rates are computed relative to the original SAR sample counts and include source-specific pair eligibility selection and quality filtering. Crop-equivalent counts account for native image size and redundancy and determine the source sampling mixture.

Source	Orig. samples	Kept pairs	Drop (%)	Crop equivalents	Mix (%)
GUSO	589,143	585,424	0.63	2,341,696	38.73
TerraMesh	8,194,048	1,741,434	78.75	1,741,434	28.80
SARLO-80	87,870	77,932	11.31	1,246,912	20.62
SAR-1M	1,130,379	688,458	39.09	688,458	11.39
3MOS	113,074	111,496	1.40	27,874	0.46
Total	10,114,514	3,204,744	68.32	6,046,374	100.00

and the decoder frozen. We consider two adaptation strategies: (i) full fine-tuning (FT) that updates all generator parameters and (ii) low-rank adaptation (LoRA) (Hu et al., 2021) that freezes the pretrained generator weights and introduces low-rank updates to selected linear layers as $\mathbf{W}' = \mathbf{W} + (\alpha_L/r_L)\mathbf{BA}$, where A and B are trainable low-rank matrices, r_L is the adapter rank, and α_L is the scaling factor. The adapters are applied to the attention and MLP projections in both the double-stream and single-stream blocks. Both strategies use the same flow-matching objective and SAR augmentation as in Stage 2.

4 Experiments

4.1 Datasets

Multi-source pretraining corpus. We construct the pretraining corpus from GUSO (Yan et al., 2026), TerraMesh (Blumenstiel et al., 2025), SARLO-80 (Debuysère et al., 2026), SAR-1M (Liu et al., 2026), and 3MOS (Ye et al., 2025). Together, these sources span diverse sensors, polarizations, spatial resolutions, ground sampling distances (GSDs), and geographic regions. GUSO provides globally distributed, ultra-high-resolution pairs at 0.16–0.98 m GSD; TerraMesh provides globally distributed Sentinel-1/2 observations aligned to a 10 m grid; and SARLO-80 contains Umbra spotlight SAR imagery resampled to a 0.8 m slant-range grid. SAR-1M and 3MOS further broaden the corpus with multi-sensor observations covering different frequency bands and spatial resolutions. Source-specific sensor, polarization, GSD, native-resolution, and preprocessing statistics are provided in the *Supplementary Material*.

Pair filtering and sampling. Stage 2 uses only paired SAR–EO observations. We first impose source-specific eligibility criteria. For TerraMesh, we retain pairs with reported zero cloud cover and an acquisition-time difference of at most one day, yielding 1,760,233 eligible pairs (21.5%) from 8,194,048 SAR observations. For SAR-1M, we retain the 731,073 samples (64.7%) with an available EO counterpart from 1,130,379 SAR observations. Together with the remaining sources, these criteria yield 3,281,393 eligible pairs. We then remove pairs exhibiting featureless content (51,515 pairs), cloud contamination (18,722 pairs), SAR–EO misregistration (6,413 pairs), compression artifacts (582 pairs), missing tiles (270 pairs), or dead/no-data regions (161 pairs). It should be noted that because these categories are non-exclusive, their counts do not sum to the number of removed pairs. Filtering the union of these non-exclusive categories leaves 3,204,744 pairs, as summarized in Table 1. Native image sizes and crop densities vary substantially across sources. We determine source sampling probabilities from the number of non-overlapping 256×256 crop equivalents. For 3MOS, we divide the tile count by four to approximately account for the redundancy introduced by half-tile strides along both spatial axes. This correction reduces the overrepresentation of densely overlapping tiles in the pretraining mixture.

Downstream benchmarks. We evaluate GeoSET on six SAR–EO benchmarks spanning satellite and airborne platforms and GSDs from 0.5 to 1.25 m. The train/test splits comprise 16,001/3,999 pairs for QXS-SAROPT (Huang et al., 2021), 1,450/627 for SAR2Opt (Zhao et al., 2022), 68,151/4,000 for SAR2EO (Du et al., 2023), and 2,558/495 for SpaceNet6 (Shermeyer et al., 2020).

Table 2: Quantitative comparison on QXS-SAROPT and SAR2Opt. We retrain (evaluate) all competing methods on the same training (test) splits. **Bold** and underlined values indicate the best and second-best results, respectively.

Method	Venue	QXS-SAROPT							SAR2Opt						
		FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
General image-to-image translation methods															
pix2pix	CVPR'17	174.6	0.373	0.1713	0.275	0.665	0.203	12.33	261.9	0.347	0.2164	0.277	0.657	0.199	13.39
CycleGAN	ICCV'17	115.5	0.362	0.0802	0.283	0.647	0.278	13.25	139.1	0.323	0.0343	0.374	0.642	0.188	12.68
pix2pixHD	CVPR'18	85.7	0.298	0.0492	0.403	0.573	0.358	16.13	146.3	0.283	0.0654	0.475	0.567	0.268	15.95
SPADE	CVPR'19	90.7	0.292	0.0607	0.366	0.599	0.320	14.53	142.5	0.265	0.0518	0.447	0.597	0.234	14.47
DDPM (SR3)	TPAMI'22	43.8	0.311	0.0189	0.425	0.620	0.359	14.04	122.5	0.295	0.0437	0.497	0.610	0.313	13.65
SD2.1 FT	CVPR'22	19.1	0.257	0.0042	0.489	0.561	0.348	15.40	71.8	0.211	0.0094	0.600	0.541	0.293	16.24
BBDM	CVPR'23	76.6	0.270	0.0479	0.414	0.568	0.352	15.34	143.1	0.290	0.0671	0.466	0.590	0.276	15.29
ControlNet	ICCV'23	50.4	0.307	0.0211	0.458	0.604	0.297	13.42	140.5	0.350	0.0480	0.479	0.643	0.217	11.73
HI-Diff	NeurIPS'23	324.3	0.539	0.3269	0.215	0.692	**0.457**	**17.10**	319.8	0.473	0.2357	0.277	0.692	**0.384**	**17.36**
ResShift	NeurIPS'23	140.2	0.334	0.0872	0.295	0.607	0.217	14.20	141.7	0.304	0.0515	0.435	0.597	0.177	14.31
StegoGAN	CVPR'24	106.8	0.384	0.0707	0.261	0.658	0.254	12.96	149.8	0.332	0.0396	0.362	0.652	0.162	12.39
SAR-to-EO image translation (SET) methods															
CondDiff	GRSL'23	88.6	0.355	0.0537	0.310	0.730	0.213	11.55	211.8	0.415	0.1379	0.343	0.686	0.248	12.48
E3Diff	GRSL'24	47.8	0.278	0.0167	0.379	0.530	0.302	16.44	104.7	0.232	0.0306	0.541	**0.529**	0.249	16.09
cBBDM	GRSL'25	50.6	0.246	0.0284	0.492	0.539	0.372	16.02	222.3	0.377	0.1521	0.413	0.571	0.361	17.05
C-DiffSET	TCSVT'26	19.9	**0.233**	0.0055	0.522	**0.526**	0.380	16.92	78.1	0.214	0.0138	0.601	0.529	0.314	16.81
GeoSET (LoRA)	–	19.3	0.254	0.0050	0.544	0.564	0.326	14.67	74.6	0.200	0.0066	**0.626**	0.535	0.278	15.50
GeoSET (full FT)	–	**16.9**	0.244	**0.0032**	**0.558**	0.553	0.332	15.07	**71.1**	**0.196**	**0.0052**	0.614	0.532	0.279	15.66

Table 3: Quantitative comparison on SAR2EO and SpaceNet6. We retrain (evaluate) all competing methods on the same training (test) splits. **Bold** is the best, and underlined is the 2nd best.

Method	SAR2EO							SpaceNet6						
	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
General image-to-image translation methods														
pix2pix	169.9	0.315	0.1551	0.400	0.573	0.455	18.06	166.7	0.242	0.1091	0.589	0.419	0.459	17.07
CycleGAN	360.7	0.522	0.3903	0.369	0.683	0.300	15.17	119.3	0.219	0.0436	0.660	0.373	0.478	16.88
pix2pixHD	106.0	0.254	0.0638	0.543	0.439	0.619	21.68	175.1	0.246	0.1136	0.682	0.361	0.525	19.21
SPADE	98.5	0.260	0.0646	0.504	0.501	0.562	19.81	179.7	0.242	0.1228	0.662	0.385	0.489	17.77
DDPM (SR3)	85.0	0.272	0.0390	0.529	0.471	0.479	13.49	219.1	0.333	0.1616	0.419	0.622	0.111	12.45
SD2.1 FT	67.7	0.270	0.0282	0.511	0.487	0.459	15.46	124.7	0.196	0.0537	0.612	0.348	0.513	19.29
BBDM	63.7	0.231	0.0245	0.613	0.418	0.601	19.08	230.2	0.381	0.1647	0.470	0.468	0.449	18.03
ControlNet	87.8	0.315	0.0424	0.509	0.523	0.418	12.93	131.3	0.279	0.0635	0.615	0.413	0.448	14.45
HI-Diff	315.5	0.556	0.3007	0.155	0.647	**0.672**	21.68	261.2	0.300	0.2036	0.499	0.403	**0.589**	**20.61**
ResShift	130.6	0.273	0.0768	0.418	0.510	0.504	19.29	155.4	0.230	0.0858	0.540	0.367	0.421	18.58
StegoGAN	384.7	0.550	0.4343	0.387	0.698	0.180	12.49	113.1	0.225	0.0361	0.649	0.388	0.457	16.04
SAR-to-EO image translation (SET) methods														
CondDiff	113.2	0.317	0.0639	0.461	0.551	0.334	11.12	160.1	0.274	0.1012	0.514	0.517	0.198	14.68
E3Diff	55.5	0.228	0.0148	0.550	0.446	0.514	20.58	110.6	0.197	0.0440	0.696	0.361	0.462	18.32
cBBDM	52.7	0.191	0.0214	0.675	**0.369**	0.650	**21.73**	280.4	0.347	0.2620	0.450	0.419	0.447	19.72
Seg-CycleGAN	–	–	–	–	–	–	–	131.4	0.229	0.0611	0.649	0.386	0.471	16.62
C-DiffSET	64.0	0.252	0.0257	0.527	0.462	0.512	17.09	132.3	0.197	0.0614	0.611	0.346	0.520	19.43
GeoSET (LoRA)	29.5	0.202	0.0042	0.691	0.440	0.547	19.19	**87.9**	**0.189**	**0.0162**	**0.733**	**0.338**	0.513	17.95
GeoSET (full FT)	**24.5**	**0.181**	**0.0032**	**0.730**	0.399	0.564	19.72	94.1	0.192	0.0183	0.724	0.342	0.515	17.99

We additionally evaluate on two private datasets, CAP-BSG and KOMPSAT, using train/test splits of 7,331/2,024 and 6,167/1,680 pairs, respectively. Complete sensor, resolution, and split-construction details are provided in the *Supplementary Material*.

4.2 Implementation Details and Evaluation Protocol

Training details. GeoSET is trained on 256×256 crops. Stages 1 and 2 use four NVIDIA B200 GPUs. Stage 1 is trained for 50,000 updates with a global batch size of 432, using AdamW with a learning rate of 10^{-5} and a 500-update linear warmup. Stage 2 optimizes all 3.852B generator parameters for 500,000 updates with a global batch size of 256, using AdamW with a peak learning rate of 10^{-4} and a 1,000-update linear warmup. Conditioning dropout is applied with probability 0.1. Stages 1 and 2 require 14.8 and 129.3 hours of training, respectively. Each downstream GeoSET model is adapted for 20,000 updates with a global batch size of 16. Full fine-tuning uses a learning rate of 2×10^{-5}, whereas LoRA uses 10^{-4} with rank $r_L = 16$ and scaling factor $\alpha_L = 16$. We clip the gradient norm to 1.0 in all stages and maintain an exponential moving average (EMA) with decay 0.9999 during Stages 2 and 3. At inference, we use 50 integration steps with two-pass classifier-free guidance (CFG) and a guidance weight of 2.0.

Baselines. We retrain a broad set of baselines on the same training splits, using official implementations when available, and evaluate all outputs on identical test pairs. General image-to-image trans-

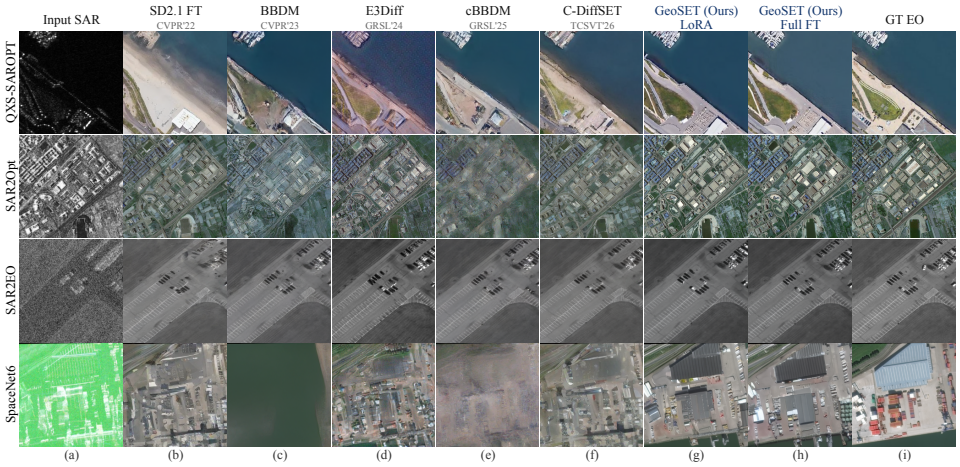

Figure 5: **Qualitative comparison on SAR-to-EO image translation (SET) benchmarks.** All methods receive the same SAR observation, and the paired EO image is shown as a reference.

lation baselines include pix2pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017), pix2pixHD (Wang et al., 2018), SPADE (Park et al., 2019), and StegoGAN (Wu et al., 2024). Restoration and conditional-generation baselines include DDPM (SR3) (Saharia et al., 2022), SD2.1 fine-tuning (FT) (Rombach et al., 2022), BBDM (Li et al., 2023), ControlNet (Zhang et al., 2023), HI-Diff (Chen et al., 2023), and ResShift (Yue et al., 2023). SET-specific baselines comprise CondDiff (Bai et al., 2023), E3Diff (Qin et al., 2024), cBBDM (Kim & Chung, 2025), C-DiffSET (Do et al., 2026), and Seg-CycleGAN (Zhang et al., 2025). We additionally compare FLUX.2 LoRA (Labs, 2025) with the proposed configurations in the component analysis.

Metrics. We use FID (Heusel et al., 2017) and DISTS (Ding et al., 2020) as our primary measures of distributional realism and perceptual fidelity, respectively. We additionally report KID (Bińkowski et al., 2018), DINO feature similarity computed with a frozen DINOv3-SAT ViT-L/16 (Siméoni et al., 2025), LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), and PSNR.

4.3 Comparison with the State of the Art

Quantitative comparison. Tables 2 and 3 compare GeoSET with existing methods on four public benchmarks; results on CAP-BSG and KOMPSAT are provided in the *Supplementary Material*. Across its two adaptation modes, GeoSET achieves the best FID on all four public benchmarks and the best DISTS on three, with gains also observed in KID and DINO similarity. On SAR2EO, GeoSET (full fine-tuning) reduces FID by over 50% relative to the strongest competing baseline, cBBDM. On SpaceNet6, GeoSET (LoRA) achieves the strongest FID and DISTS, demonstrating effective transfer with limited parameter updates. C-DiffSET retains the best DISTS on QXS-SAROPT, while competing methods often achieve stronger LPIPS, PSNR, or SSIM. PSNR and SSIM measure agreement with a specific EO reference; however, SAR backscatter does not uniquely determine optical color and texture, and residual misregistration can further penalize otherwise plausible outputs. We therefore emphasize FID for distributional realism and DISTS for paired structural and textural similarity (Ding et al., 2020), while retaining LPIPS, PSNR, and SSIM as complementary fidelity measures. GeoSET's gains thus reflect improved distributional quality and competitive perceptual fidelity rather than uniformly better pixel-aligned reconstruction.

Qualitative comparison. Figures 1 and 5 compare fixed, score-independent examples using identical inputs and spatial extents for all methods. We examine whether each method preserves large-scale scene layout while producing locally coherent building boundaries, road structures, vegetation patterns, and water regions. This comparison complements the scalar metrics by revealing whether gains in FID and DISTS correspond to plausible local structure rather than only changes in global color or contrast. Additional qualitative results are provided in the *Supplementary Material*.

Table 4: **Stage 1 encoder diagnostics and Stage 3 adaptation efficiency.** (a) Reconstruction PSNR on original and perturbed SAR inputs before and after encoder training. EO reports reconstruction with the frozen pretrained codec. (b) Computational costs of full fine-tuning and LoRA.

(a) Stage 1 Encoder Reconstruction

Source	Original SAR			Perturbed SAR			EO
	Before	After	Δ	Before	After	Δ	
GUSO	27.47	29.18	+1.71	23.54	26.96	+3.42	31.02
TerraMesh	20.11	32.04	+11.93	19.12	30.48	+11.37	33.95
SARLO-80	17.39	19.21	+1.82	16.71	18.49	+1.78	27.81
SAR-1M	29.95	31.25	+1.30	22.01	26.99	+4.98	33.59
3MOS	35.15	36.07	+0.92	27.30	31.92	+4.61	34.62
Average	26.01	29.55	+3.54	21.74	26.97	+5.23	32.20

(b) Downstream Adaptation Efficiency

GeoSET	Full FT	LoRA
Trainable parameters	3.852B	23.10M
Trainable fraction	100%	0.5997%
Per-dataset storage	30.8 GB	185 MB
Peak memory	96.9 GB	49.1 GB
Batch size	16	16
Throughput	67 img/s	94 img/s
Wall time/dataset	1.34 h	0.98 h

4.4 Ablation Studies and Analysis

Stage 1 encoder reconstruction. Table 4(a) evaluates the SAR encoder before and after Stage 1 with the decoder fixed. Training improves reconstruction across all sources, increasing average PSNR by 3.54 dB on original inputs and 5.23 dB on perturbed inputs. The larger gain under perturbation supports speckle robustness, while improvements on original inputs indicate better adaptation to SAR statistics. The pretrained codec already provides strong EO reconstruction, motivating us to preserve the EO pathway and focus representation adaptation on SAR.

Parameter-efficient adaptation. Table 4(b) shows that LoRA adapts the shared parent by updating only 0.6% of the generator parameters, requiring approximately one hour per dataset on one GPU. It substantially reduces per-dataset storage and peak training memory while increasing training throughput. LoRA remains competitive with full fine-tuning and achieves stronger FID, DISTS, and LPIPS on SpaceNet6 under the evaluated settings. This makes a shared parent with compact dataset-specific adapters a practical option for extending GeoSET to multiple downstream datasets.

Table 5: **Cumulative component comparison on SAR2EO and SpaceNet6.**

Configuration	SAR2EO			SpaceNet6		
	FID↓	DISTS↓	LPIPS↓	FID↓	DISTS↓	LPIPS↓
SD2.1 Full FT	67.7	0.270	0.487	124.7	0.196	0.348
FLUX.2 LoRA (Backbone)	71.2	0.275	0.481	117.4	0.201	0.361
+ Pretraining corpus	50.5	0.250	0.470	107.9	0.194	0.351
+ SAR encoder	39.8	0.220	0.448	94.2	0.193	0.347
+ Speckle aug. (**GeoSET**)	**29.5**	**0.202**	**0.440**	**87.9**	**0.189**	**0.338**

Effect of pretraining, SAR encoding, and speckle augmentation. Table 5 compares the cumulative effects of corpus pretraining, SAR encoding, and speckle augmentation. All FLUX.2 configurations use matched training lengths for Stage 3 LoRA tuning and, where applicable, Stage 2 corpus pretraining and Stage 1 SAR encoder training. The FLUX.2 LoRA baseline shows mixed results relative to SD2.1 Full FT, with neither configuration consistently outperforming the other. Adding corpus pretraining improves all three metrics on both datasets, reducing FID from 71.2 to 50.5 on SAR2EO and from 117.4 to 107.9 on SpaceNet6. Introducing the SAR encoder further improves all metrics, supporting the benefit of SAR-specific conditioning representations. Speckle augmentation provides additional gains across all metrics, yielding the best results with the complete GeoSET (LoRA) configuration. Together, these cumulative comparisons support the complementary benefits of corpus pretraining, SAR-specific encoding, and speckle augmentation.

5 Conclusion

We introduced GeoSET, the first generalist model for SAR-to-EO image translation under a *pretrain-once, adapt-many* framework. By combining speckle-robust SAR encoding with generative pretraining on over three million curated SAR–EO pairs, GeoSET learns a reusable SET prior that transfers across heterogeneous sensing conditions. Across six downstream benchmarks, GeoSET achieves state-of-the-art FID and DISTS results, while LoRA enables adaptation in approximately one hour per dataset by updating only 0.60% of the generator parameters. These results demonstrate the potential of a shared SET prior to advance SAR-to-EO translation beyond dataset-specific models.

AI USE STATEMENT

We used generative AI tools to assist with translation, English-language editing, manuscript organization, and review of experimental reporting and interpretation. The authors reviewed and revised the AI-assisted content, checking technical descriptions against the implementation and empirical statements against the reported results. The authors take full responsibility for the final submission, including its text, methodology, results, claims, and references.

ETHICS STATEMENT

This study uses existing public and private SAR–EO datasets, which remain subject to their original access, usage, and redistribution conditions. GeoSET produces plausible SAR-conditioned EO-like images rather than direct optical measurements; generated colors, textures, or structures may be inaccurate and should not be treated as verified observations. Geospatial imagery may also reveal sensitive information about locations, infrastructure, and human activities, raising concerns about privacy, data governance, and misuse, including unauthorized monitoring. Benchmark performance does not establish reliability or equitable performance across geographic regions, sensors, or acquisition conditions. Consequential applications, such as disaster assessment, require independent validation and corroboration with acquired observations.

REPRODUCIBILITY STATEMENT

Section 3 describes the GeoSET framework and learning objectives, while Section 4 reports corpus construction, training budgets, downstream adaptation settings, and the evaluation protocol. Appendix A provides additional results, SAR-conditioning diagnostics, and limitations. Appendix B summarizes dataset provenance and evaluation splits; Appendix C details speckle augmentation, SAR input/output stems, and the backbone and autoencoder; and Appendix D describes the evaluation metrics. Code and pretrained checkpoints will be released upon acceptance, subject to applicable licenses.

REFERENCES

Xinyu Bai, Xinyang Pu, and Feng Xu. Conditional diffusion for sar to optical image translation. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2023.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Benedikt Blumenstiel, Paolo Fraccaro, Valerio Marsocci, Johannes Jakubik, Stefano Maurogiovanni, Mikolaj Czerkawski, Rocco Sedona, Gabriele Cavallaro, Thomas Brunschwiler, Juan Bernabe Moreno, et al. Terramesh: A planetary mosaic of multimodal earth observation data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2419–2427, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. *Advances in neural information processing systems*, 36:29114–29125, 2023.

Solène Debuysère, Nicolas Trouvé, Nathan Letheule, Elise Colin, and Georgia Channing. Sarlo-80: Worldwide slant sar language optic dataset 80cm. *arXiv preprint arXiv:2606.20523*, 2026.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Jeonghyeok Do, Jaehyup Lee, Seungchul Lee, and Munchurl Kim. C-diffset: Leveraging latent diffusion for sar-to-eo image translation with confidence-guided reliable object generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Shenshen Du, Jun Yu, Guochen Xie, Renjie Lu, Pengwei Li, Zhongpeng Cai, and Keda Lu. Sar2eo: A high-resolution image translation framework with denoising enhancement. In *Australasian Joint Conference on Artificial Intelligence*, pp. 91–102. Springer, 2023.

Zhe Guo, Rui Luo, Qinglin Cai, Jiayi Liu, Zhibo Zhang, and Shaohui Mei. Scene-embedded generative adversarial networks for semi-supervised sar-to-optical image translation. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Meiyu Huang, Yao Xu, Lixin Qian, Weili Shi, Yaqin Zhang, Wei Bao, Nan Wang, Xuejiao Liu, and Xueshuang Xiang. The qxs-saropt dataset for deep learning in sar-optical data fusion. *arXiv preprint arXiv:2103.08259*, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5967–5976. Ieee, 2017.

Seon-Hoon Kim and Daewon Chung. Conditional brownian bridge diffusion model for vhr sar to optical image translation. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Jaehyup Lee, Hyun-Ho Kim, Doochun Seo, and Munchurl Kim. Segmentation-guided context learning using eo object labels for stable sar-to-eo translation. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2023.

Jong-Sen Lee, L Jurkevich, Piet Dewaele, Patrick Wambacq, and André Oosterlinck. Speckle filtering of synthetic aperture radar images: A review. *Remote sensing reviews*, 8(4):313–340, 1994.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1952–1961. IEEE, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Danxu Liu, Di Wang, Hebaixu Wang, Haoyang Chen, Wentao Jiang, Yilin Cheng, Haonan Guo, Wei Cui, and Jing Zhang. Sarmae: Masked autoencoder for sar representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6496–6507, 2026.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1):6–43, 2013.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2332–2341. IEEE, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Jiang Qin, Bin Zou, Haolin Li, and Lamei Zhang. Efficient end-to-end diffusion model for one-step sar-to-optical translation. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hänsch, Alexei Bastidas, Scott Soenen, Todd Bacastow, et al. Spacenet 6: Multi-sensor all weather mapping dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 768–777. IEEE, 2020.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Javier Noa Turnes, Jose David Bermudez Castro, Daliana Lobo Torres, Pedro Juan Soto Vega, Raul Queiroz Feitosa, and Patrick N Happ. Atrous cgan for sar to optical image translation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2020.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8798–8807. Ieee, 2018.

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

Sidi Wu, Yizi Chen, Samuel Mermet, Lorenz Hurni, Konrad Schindler, Nicolas Gonthier, and Loic Landrieu. Stegogan: Leveraging steganography for non-bijective image-to-image translation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7922–7931. IEEE, 2024.

Heng Yan, Ailong Ma, Hong Shu, Yuting Wan, Liangpei Zhang, and Yanfei Zhong. Ultra-high-resolution sar and optical image registration: From global benchmark dataset to frequency-guided registration method. *ISPRS Journal of Photogrammetry and Remote Sensing*, 235:190–210, 2026.

Yibin Ye, Xichao Teng, Hongrui Yang, Shuo Chen, Yuli Sun, Yijie Bian, Tao Tan, Zhang Li, and Qifeng Yu. 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.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in neural information processing systems*, 36: 13294–13307, 2023.

Hannuo Zhang, Huihui Li, Jiarui Lin, Yujie Zhang, Jianghua Fan, Hang Liu, and Kun Liu. Segcyclegan: Sar-to-optical image translation guided by a downstream task. *IEEE Geoscience and Remote Sensing Letters*, 22:1–5, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yitao Zhao, Turgay Celik, Nanqing Liu, and Heng-Chao Li. A comparative analysis of gan-based methods for sar-to-optical image translation. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5, 2022.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 2242–2251. Ieee, 2017.

APPENDIX

This *Supplementary Material* provides additional results and implementation details for GeoSET. Section A presents quantitative and qualitative comparisons, SAR conditioning diagnostics, and limitations. Section B summarizes the pretraining sources and downstream evaluation protocols. Section C describes domain-aware speckle augmentation, SAR input/output stems, and the backbone and autoencoder. Section D explains the evaluation metrics.

Table 6: Overview of the *Supplementary Material*.

Section	Contents
Sec. A	Additional results and discussion
Sec. B	Datasets and evaluation protocols
Sec. C	Implementation details
Sec. D	Evaluation metrics

A ADDITIONAL RESULTS AND DISCUSSIONS

Quantitative comparison on private datasets. Table 8 reports results on CAP-BSG and KOMPSAT. GeoSET achieves the best FID, DISTS, KID, and DINO similarity on both datasets, extending the improvements observed on public benchmarks to additional sensing conditions. GeoSET (full fine-tuning) provides the strongest FID and DISTS, while GeoSET (LoRA) remains competitive and matches its DINO similarity at the reported precision. Competing methods retain advantages in LPIPS, PSNR, and SSIM, consistent with the different metric strengths discussed in the main paper.

Additional qualitative comparison. Figures 6, 7, 8, 9, and 10 present additional examples on QXS-SAROPT, SAR2Opt, SAR2EO, SpaceNet6, and CAP-BSG, respectively, using identical SAR inputs and paired EO references. These examples complement the quantitative results by enabling closer inspection of scene layout, local structures, and texture consistency across methods.

Does GeoSET use the SAR condition? We evaluate the Stage 2 parent after 500,000 updates using true, shuffled, and null SAR conditions. Each setting uses the same 192 examples per source, 50 integration steps, and random seed. As shown in Table 7, true conditioning consistently improves PSNR and LPIPS over both alternatives across all three sources. This indicates that GeoSET uses the input SAR observation to guide generation, rather than relying solely on its learned EO prior. These diagnostics are conducted on pretraining sources.

Table 7: **SAR conditioning intervention on pretraining sources.** True denotes the corresponding SAR observation, shuffled uses a mismatched SAR observation, and null uses an all-zero SAR latent. The null condition is included during training through conditioning dropout.

	PSNR↑			LPIPS↓		
Source	True	Shuffled	Null	True	Shuffled	Null
GUSO	**13.36**	10.24	10.58	**0.548**	0.678	0.690
SARLO-80	**15.88**	11.96	11.65	**0.514**	0.678	0.693
TerraMesh	**13.69**	10.94	10.80	**0.491**	0.670	0.695

SAR and EO reconstruction. Figure 11 visualizes reconstructions from the adapted SAR encoder and the frozen EO encoder using their corresponding decoder paths. Both retain scene structures across diverse pretraining sources. For speckle-perturbed SAR inputs ($L = 4$), reconstruction suppresses the added perturbation and improves PSNR relative to the original SAR observation in every example. These results support robustness to additional speckle while preserving the observed SAR structure.

Limitations. Despite heterogeneous pretraining, GeoSET does not consistently achieve strong zero-shot performance across unseen SAR–EO domains. Differences in sensors, acquisition geometry, resolution, and radiometric processing remain challenges for direct transfer. Its generalist capability therefore lies in a reusable SET prior that enables efficient downstream adaptation, rather

Table 8: Quantitative comparison on CAP-BSG and KOMPSAT. We retrain (evaluate) all competing methods on the same training (test) splits. **Bold** is the best, and underlined is the 2nd best.

Method	CAP-BSG							KOMPSAT						
	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑	FID↓	DISTS↓	KID↓	DINO↑	LPIPS↓	SSIM↑	PSNR↑
General image-to-image translation methods														
pix2pix	245.8	0.402	0.2086	0.254	0.626	0.251	13.77	255.6	0.346	0.2892	0.298	0.669	0.116	10.33
CycleGAN	100.0	0.327	0.0485	0.365	0.569	0.353	16.84	160.8	0.359	0.1105	0.299	0.653	0.139	10.14
pix2pixHD	212.7	0.335	0.1673	0.301	0.570	0.367	17.23	223.2	0.379	0.2004	0.305	0.649	0.128	11.24
SPADE	187.5	0.354	0.1407	0.287	0.593	0.348	16.20	226.3	0.377	0.2106	0.303	0.670	0.157	11.47
DDPM (SR3)	132.0	0.403	0.0815	0.297	0.646	0.373	12.41	119.8	0.344	0.0695	0.365	0.645	0.121	10.21
SD2.1 FT	98.6	0.348	0.0406	0.321	0.587	0.380	15.86	100.7	0.333	0.0535	0.359	0.635	0.150	11.23
BBDM	145.2	0.346	0.0789	0.305	0.577	0.360	16.16	154.3	0.343	0.1101	0.322	0.656	0.145	9.74
ControlNet	120.5	0.426	0.0466	0.290	0.660	0.303	13.10	115.5	0.358	0.0672	0.357	0.671	0.137	9.63
HI-Diff	200.0	0.496	0.1328	0.234	0.697	**0.463**	16.75	277.8	0.523	0.2664	0.223	0.756	**0.226**	12.39
ResShift	201.2	0.378	0.1282	0.280	0.611	0.212	15.11	247.9	0.348	0.2193	0.306	0.652	0.115	11.12
StegoGAN	97.3	0.319	0.0464	0.367	0.567	0.345	16.89	325.4	0.497	0.3565	0.222	0.713	0.092	**12.70**
SAR-to-EO image translation (SET) methods														
CondDiff	212.1	0.418	0.1347	0.246	0.734	0.234	10.50	115.6	0.349	0.0672	0.352	0.692	0.082	8.70
E3Diff	87.2	0.324	0.0337	0.346	**0.543**	0.332	16.79	116.5	0.333	0.0762	0.388	**0.600**	0.147	11.74
cBBDM	214.6	0.365	0.1580	0.280	0.578	0.437	**17.91**	245.3	0.434	0.2300	0.279	0.666	0.189	11.92
C-DiffSET	113.0	0.333	0.0517	0.310	0.573	0.382	17.03	144.9	0.344	0.0943	0.315	0.640	0.144	11.17
GeoSET (LoRA)	64.7	0.318	0.0154	**0.402**	0.566	0.343	16.85	72.0	0.299	0.0294	0.446	0.625	0.152	10.64
GeoSET (full FT)	**58.9**	**0.316**	**0.0109**	**0.402**	0.565	0.352	16.60	**69.8**	**0.295**	**0.0292**	**0.447**	0.620	0.153	10.78

than universal zero-shot translation. LoRA makes this adaptation practical, although it still requires paired data from the target domain.

B Datasets and Evaluation Protocols

Table 9 summarizes the pretraining sources, their sensor provenance, and the SAR representations used by GeoSET. Table 10 details the downstream datasets, training and evaluation set sizes, and split protocols. All competing methods are evaluated on the same test pairs with identical crops.

C Implementation Details

Domain-aware speckle augmentation. The SAR input representations summarized in Table 9 determine the routing of $\mathcal{R}_{d_r}$ in Eq. 2. TerraMesh uses the decibel branch, SARLO-80 uses the amplitude branch, and GUSO, SAR-1M, and both 3MOS streams use the display branch. The complete operator is

$$\mathcal{R}_{d_r}(\mathbf{x}^s;\mathbf{g}) = \begin{cases} \mathbf{I}\odot\mathbf{g}, & \text{linear intensity},\\ \mathbf{A}\odot\sqrt{\mathbf{g}}, & \text{linear amplitude},\\ \mathbf{x}^s + 10\log_{10}\mathbf{g}, & \text{decibel-valued input},\\ \mathbf{x}^s\odot[1+\kappa(\mathbf{g}-1)], & \text{display imagery},\\ \mathbf{x}^s, & \text{unknown representation}, \end{cases} \tag{6}$$

where $\mathbf{I}$ and $\mathbf{A}$ denote linear intensity and amplitude, respectively, and $\odot$ denotes elementwise multiplication. The amplitude and decibel branches follow from $\mathbf{I} = \mathbf{A}^2$ and the logarithmic intensity representation, respectively. For display imagery, whose rendering may alter the original radiometric relationship, we use an attenuated multiplicative perturbation with $\kappa = 0.5$. Inputs with unknown representations are left unchanged. Figure 12 illustrates the augmentation across the six pretraining streams for $L \in \{16, 8, 4\}$. Smaller L produces stronger perturbations, reflected in lower PSNR relative to the original SAR observations, while the large-scale scene layout remains recognizable.

SAR input and output stems. We use channel-specific input and output stems around shared encoder and decoder trunks to accommodate single-channel SAR and dual-polarization VV/VH inputs. Single-channel SAR uses the pretrained three-channel path with grayscale folding, while VV/VH inputs use a dedicated two-channel path that preserves separate polarization channels. During Stage 1, we train the encoder trunk, input stems, and the new two-channel output projection, while keeping the decoder trunk and pretrained three-channel output projection frozen.

Table 9: **Pretraining corpus and representations.** Five dataset families form six streams, with 3MOS divided into `mr` (mid-resolution) and `hr` (high-resolution). Spatial scales denote reported resolutions or sampling intervals. Tile sizes are in pixels; display denotes rendered SAR imagery.

Source	SAR source	EO source	Spatial scale	Tile size	SAR input
GUSO (Yan et al., 2026)	ICEYE, Capella, Umbra	Supplied optical imagery	0.16–0.98m	512^2	1-ch display
TerraMesh (Blumenstiel et al., 2025)	Sentinel-1 RTC, VV/VH	Sentinel-2 L2A RGB	10 m grid	264^2	2-ch dB
SARLO-80 (Debuysère et al., 2026)	Umbra spotlight SLC	Google satellite imagery	0.8m slant-range grid	1024^2	1-ch amplitude
SAR-1M (Liu et al., 2026)	Multi-source paired subset	Paired optical counterparts	Source-dependent	256^2	1-ch display
3MOS `mr` (Ye et al., 2025)	Sentinel-1, ALOS-family	Google Earth	10 / 12.5m	256^2	1-ch display
3MOS `hr` (Ye et al., 2025)	GF-3, RADARSAT-2, RCM	Google Earth	3.5 / 6.3 / 12.5m	256^2	1-ch display

Table 10: **Downstream datasets and evaluation protocols.** GSDs refer to the SAR observations where specified; – denotes an unspecified value. SAR2EO uses the fixed first 4,000 samples of the 21,260-pair E3Diff test split. SAR2Opt uses a deterministic center crop.

Dataset	SAR source	EO source	GSD	Eval. size	Split protocol
QXS-SAROPT (Huang et al., 2021)	GF-3	Google Earth	1m	256^2	C-DiffSET splits (Do et al., 2026)
SAR2Opt (Zhao et al., 2022)	TerraSAR-X	Google Earth	1m	$600^2 \rightarrow 512^2$	Official train/test folders
SAR2EO (Du et al., 2023)	Airborne UNICORN SAR	Monochrome WAMI	–	256^2	E3Diff split (Qin et al., 2024)
SpaceNet6 (Shermeyer et al., 2020)	Airborne Capella X-band	WorldView-2	0.5m	256^2	UTM-easting blocks; 450 m guard band
CAP-BSG	Capella X-band GEC	BlackSky	0.9–1.25m	256^2	Location-group split
KOMPSAT	X-band SAR	KOMPSAT-3 RGB	–	256^2	Acquisition-group split

Backbone and autoencoder. GeoSET builds on FLUX.2 4B Base (Labs, 2025), with five double-stream and twenty single-stream blocks, a hidden dimension of 3,072, and 24 attention heads. The SAR stream is initialized from the corresponding pretrained image-stream projections, attention layers, MLPs, and modulation parameters. Both block types retain SAR and EO tokens and apply full self-attention over their combined sequence, using shared spatial rotary embeddings. Double-stream blocks use separate projection, MLP, and modulation parameters for the two modalities, whereas single-stream blocks process the concatenated tokens with shared parameters. Only EO tokens are passed to the final velocity head. The pretrained VAE produces 32-channel posterior-mean latents at $1/8$ spatial resolution, which are packed into 128 channels at $1/16$ resolution and normalized using fixed BatchNorm statistics. Decoding reverses normalization and packing.

D Evaluation Metrics

We evaluate GeoSET using seven complementary metrics, with FID and DISTS as the primary measures of distributional realism and paired perceptual fidelity. All methods are evaluated on the same test pairs and spatial extents specified in Table 10. The symbols $\downarrow$ and $\uparrow$ indicate that lower and higher values are better, respectively.

FID. Fréchet Inception Distance (Heusel et al., 2017) measures the distance between Gaussian approximations to the generated and reference EO feature distributions. It evaluates distributional similarity rather than correspondence between individual SAR inputs and generated outputs.

DISTS. Deep Image Structure and Texture Similarity (Ding et al., 2020) compares generated and reference images through structural and textural statistics of deep features. Its tolerance to texture resampling complements the distributional assessment provided by FID.

KID. Kernel Inception Distance (Bińkowski et al., 2018) estimates the squared maximum mean discrepancy between generated and reference Inception features using a polynomial kernel. It provides a complementary distributional comparison to FID.

DINO similarity. We compare generated and paired reference EO images using features from a frozen DINOv3-SAT ViT-L/16 (Siméoni et al., 2025). This metric assesses correspondence in a satellite-image representation space, complementing the Inception-based distributional measures.

LPIPS. Learned Perceptual Image Patch Similarity (Zhang et al., 2018) measures differences between normalized deep features using learned perceptual weights. It evaluates the perceptual distance between each generated image and its paired EO reference.

SSIM. Structural Similarity (Wang et al., 2004) compares local luminance, contrast, and structure between generated and reference images. It measures spatially corresponding image fidelity and is sensitive to residual misregistration.

PSNR. Peak Signal-to-Noise Ratio expresses pixelwise mean squared error relative to the squared image intensity range on a logarithmic scale. It measures radiometric reconstruction fidelity to the paired EO reference and complements the perceptual and distributional metrics.

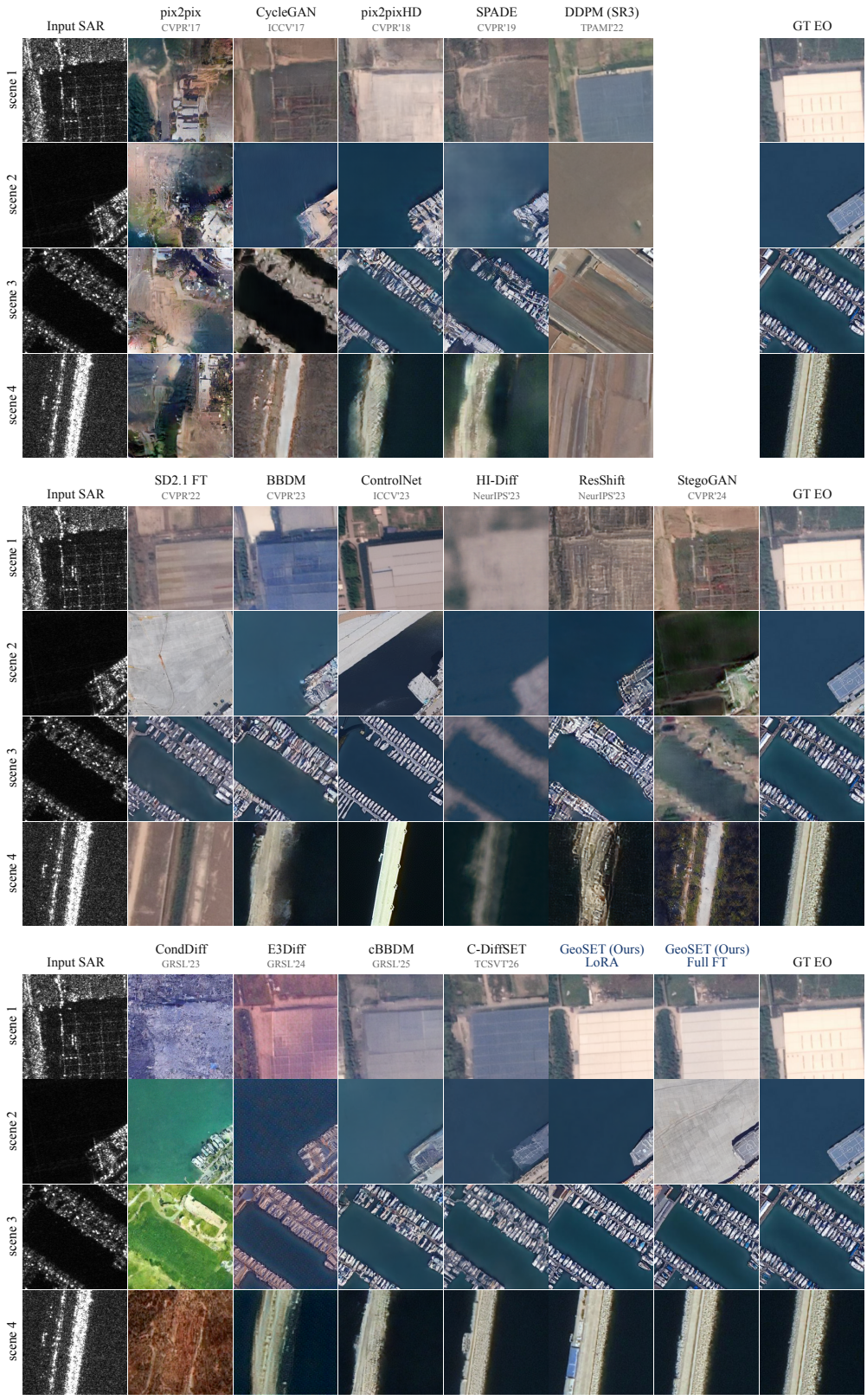

Figure 6: **Qualitative comparison on the QXS-SAROPT dataset.** All methods receive identical SAR inputs, with paired EO images shown as references.

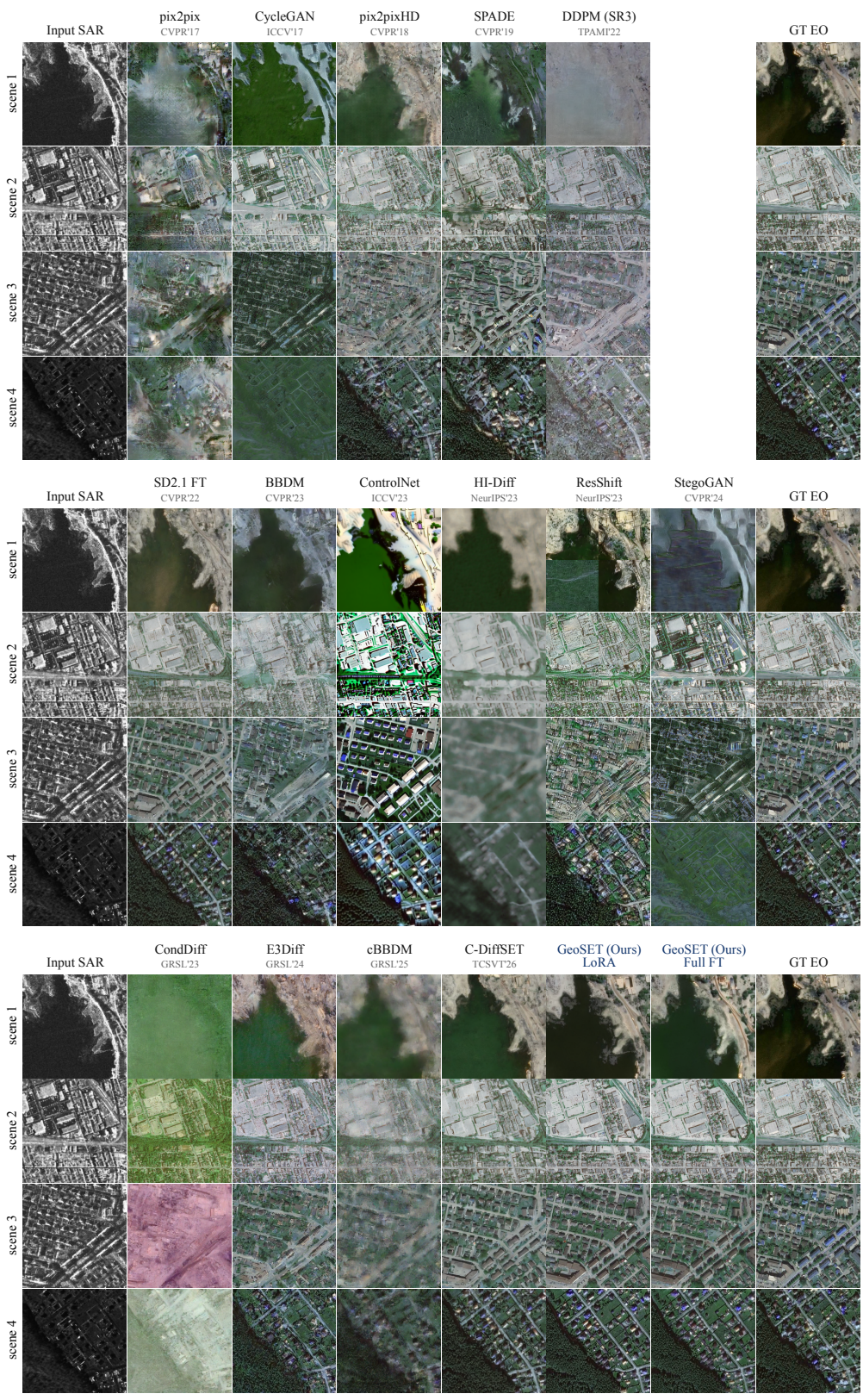

Figure 7: **Qualitative comparison on the SAR2Opt dataset.** All methods receive identical SAR inputs, with paired EO images shown as references.

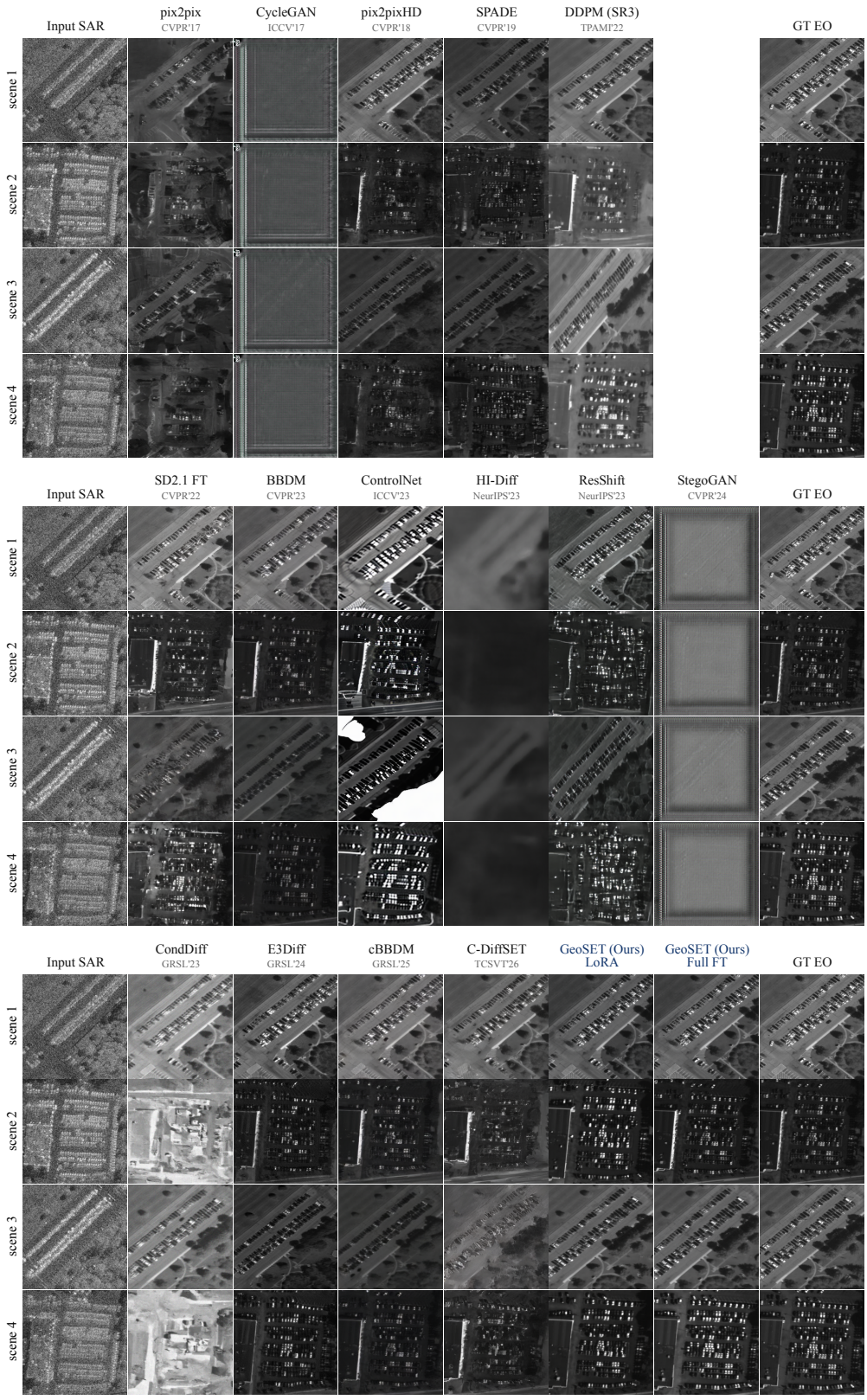

Figure 8: **Qualitative comparison on the SAR2EO dataset.** All methods receive identical SAR inputs, with paired EO images shown as references.



Figure 9: **Qualitative comparison on the SpaceNet6 dataset.** All methods receive identical SAR inputs, with paired EO images shown as references.

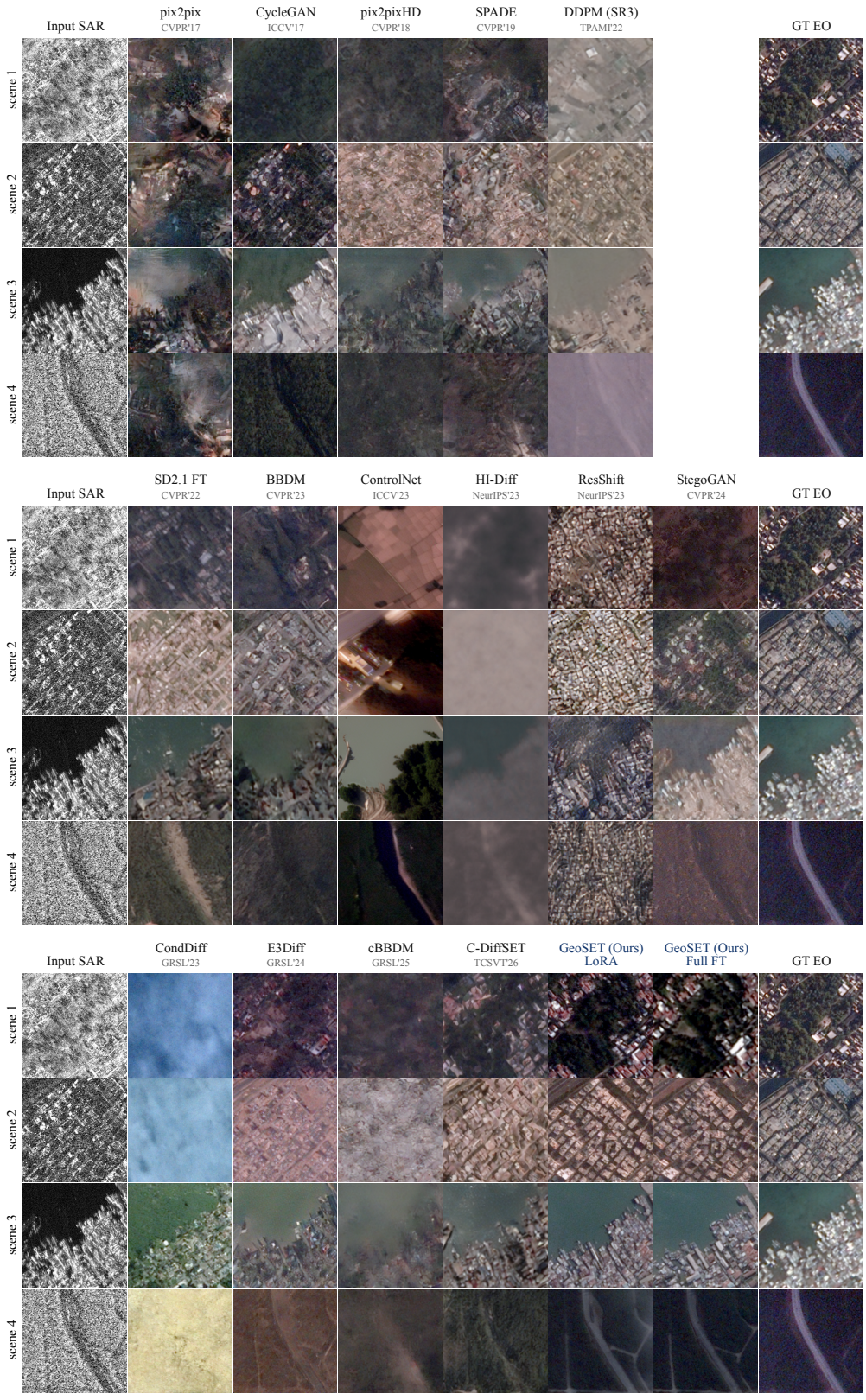

Figure 10: **Qualitative comparison on the CAP-BSG dataset.** All methods receive identical SAR inputs, with paired EO images shown as references.

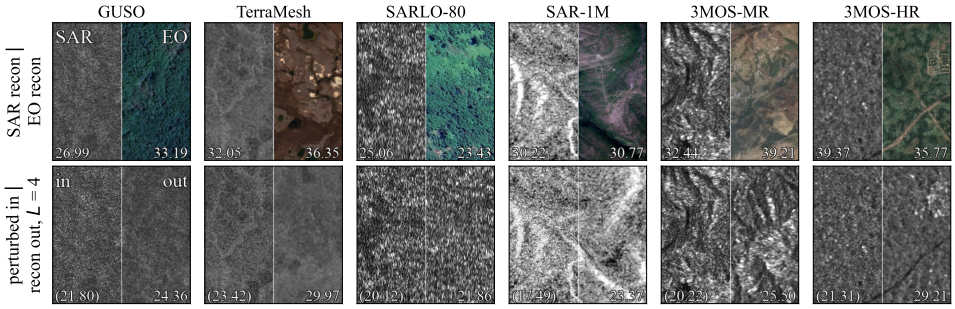

Figure 11: **SAR and EO reconstruction across pretraining sources.** Top: reconstructions of original SAR and EO inputs. Bottom: speckle-perturbed SAR inputs ($L = 4$, left) and their reconstructions (right). Values indicate PSNR (dB) relative to the corresponding original observations; parenthesized values refer to perturbed inputs.

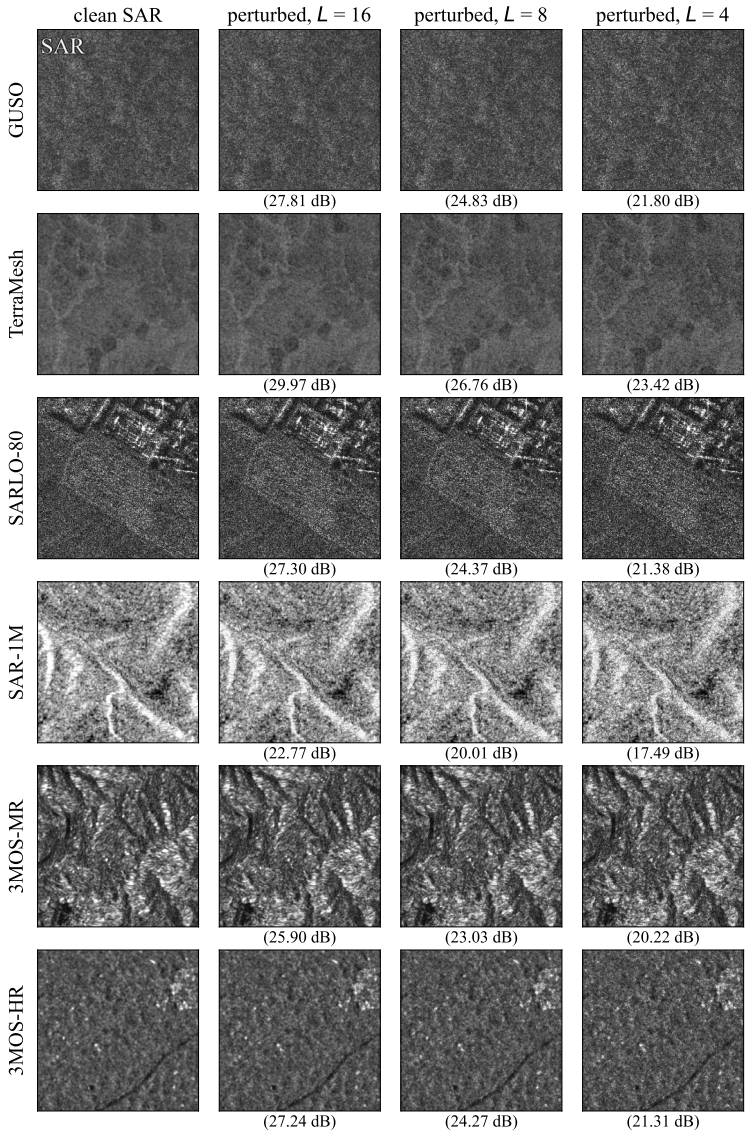

Figure 12: **Domain-aware speckle augmentation across pretraining sources.** From left to right: original SAR inputs and perturbed inputs with $L = 16$, 8, and 4. Smaller L yields stronger perturbations. Values indicate PSNR (dB) relative to the original SAR observations.