SAR2Earth: A SAR-to-EO Translation Dataset for Remote Sensing Applications

Anonymous ICCV submission

Paper ID *****

Abstract

Electro-optical (EO) images are crucial for a wide range of remote sensing applications. However, EO imagery has inherent limitations, including an inability to penetrate cloud cover or capture nighttime images. Synthetic Aperture Radar (SAR) images address these limitations by providing consistent imaging capabilities regardless of weather or lighting conditions. Nevertheless, SAR images are affected by speckle noise, which complicates analysis and limits the direct applicability of EO-based algorithms. To address these challenges, we introduce SAR2Earth, a benchmark dataset specifically designed for SAR-to-EO translation. By translating SAR images into EO-like representations, SAR2Earth enables the application of the extensive range of algorithms initially developed for EO imagery to SAR data. The dataset comprises approximately 100K pairs of spatially aligned SAR and EO images, collected from eight distinct regions covering both urban and rural environments. We provide comprehensive evaluations, detailed model analyses, and extensive experimental results. All code and datasets will be publicly available at https://sar2earth.github.io.

1. Introduction

Remote sensing images provide the capability to observe the Earth on a large scale, making them invaluable for analysis in various applications such as transportation [2], defense [38], natural resource management [10], disaster response [1], and environmental monitoring. However, the vast amount of data generated poses significant challenges for manual analysis due to the time and expertise required. The advent of data-driven models [9, 18, 30] has enabled more efficient and effective analysis of these images. Electro-optical (EO) imagery has been the primary modality for remote sensing applications due to its intuitive representation of the Earth. However, EO imagery has significant limitations: it cannot penetrate cloud cover and is unable to capture images at night, restricting its utility in many scenarios [16, 26]. For instance, during natural disasters like floods—which are often accompanied by heavy cloud cover—EO imagery becomes ineffective for timely disaster assessment and response. To overcome these limitations, synthetic aperture radar (SAR) imagery is employed. SAR sensors can operate independently of daylight and weather conditions, providing consistent imaging capabilities. However, SAR images suffer from speckle noise due to the coherent nature of radar signal processing, which introduces granular interference patterns. This speckle noise makes SAR images challenging to interpret [28, 43], especially for non-experts, and complicates the application of algorithms developed for EO imagery. To bridge this gap, SAR-to-EO translation methods [6, 11, 29, 40] have been proposed, aiming to translate SAR images into EO-like images that are more accessible for analysis using existing EO-based algorithms.

Despite these efforts, there has been a lack of comprehensive analysis of these methods, and they often remain isolated applications without standardized benchmarks. Existing SAR and EO multimodal datasets [15, 17, 24, 27, 33] have several limitations. They frequently have limited geographic diversity and data quantity, restricting the generalizability of model performance across regions. Additionally, many of these datasets feature short temporal intervals—often just one day—failing to represent real-world conditions, including significant temporal discrepancies caused by satellite revisit cycles, cloud cover, or nighttime acquisition. A detailed comparison of existing datasets is provided in the *supplementary material*.

To overcome these challenges, we introduce SAR2Earth, a comprehensive benchmark dataset designed specifically for SAR-to-EO translation. SAR2Earth comprises spatially aligned SAR and EO image pairs collected from 8 distinct regions, covering both urban and rural environments. The dataset also incorporates realistic temporal differences between SAR and EO image acquisitions, better reflecting real-world remote sensing scenarios. All codes and datasets are being made publicly available to support future research in this domain.

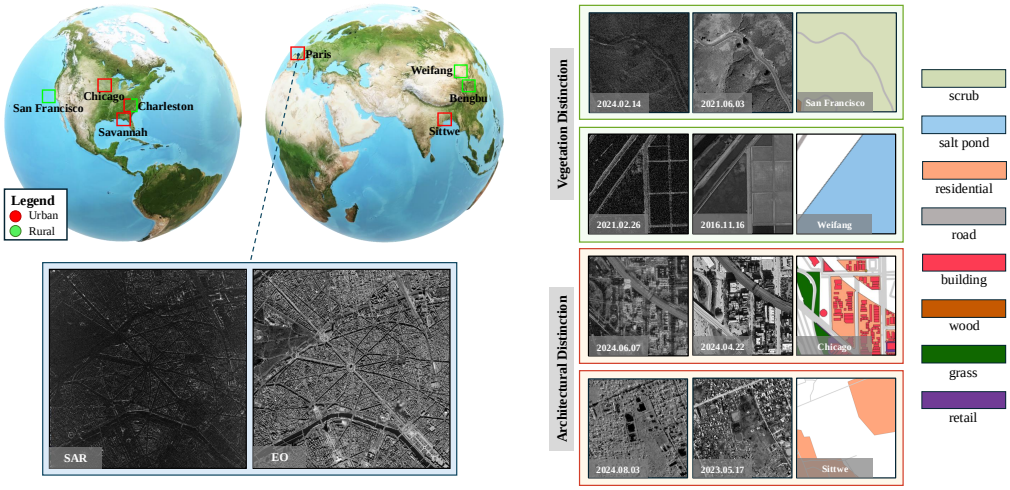


Figure 1. Geographic overview of the SAR2Earth dataset. This dataset highlights the diversity of geographic locations and environments, covering eight distinct regions—including Chicago, San Francisco, Charleston, Savannah, Paris, Bengbu, Weifang, and Sittwe—spanning both urban and rural areas across North America, Europe, and Asia. (As seen on the right, the consecutive columns represent SAR imagery, EO imagery, and OSM-based label maps.)

2. Related Work

2.1. Applications of SAR Imagery

Numerous applications leverage SAR imagery across various domains. For instance, [14] collected 100K SAR images for object detection, while [22] used aligned SAR and EO images for disaster analysis like floods. Additionally, [16] classified vehicles within SAR imagery. Another key application is cloud removal from EO images using SAR data. Studies such as [25, 36, 37] introduced datasets combining multi-temporal EO and SAR imagery to address cloud cover. However, this approach fails at night when EO data is unavailable and struggles with dynamic objects due to temporal discrepancies. The SAR-to-EO translation task has emerged to address these issues by directly generating EO-like images from SAR data. Despite its benefits, SAR data collection remains expensive and technically challenging due to speckle noise and sensor complexity, limiting widespread availability of standardized datasets.

2.2. SAR-to-EO Translations

To overcome the limitations of SAR datasets, SAR-to-EO translation techniques have been proposed. For instance, [15] introduced a method to utilize SAR images by translating them into EO images. To enhance the performance of SAR-to-EO translation, models such as Pix2Pix [7], Pix2PixHD [32], and CycleGAN [44] have been employed. Recently, diffusion-based methods [13, 23] have been explored to enhance translation quality and applied to tasks like Amazon deforestation monitoring [3]. Despite the numerous SAR-to-EO translation methods proposed, there has not been a rigorous comparison among paired methods, unpaired methods, and diffusion-based approaches. Furthermore, because the pre-processing and post-processing pipelines differ across studies, accurate analysis and benchmarking have been lacking.

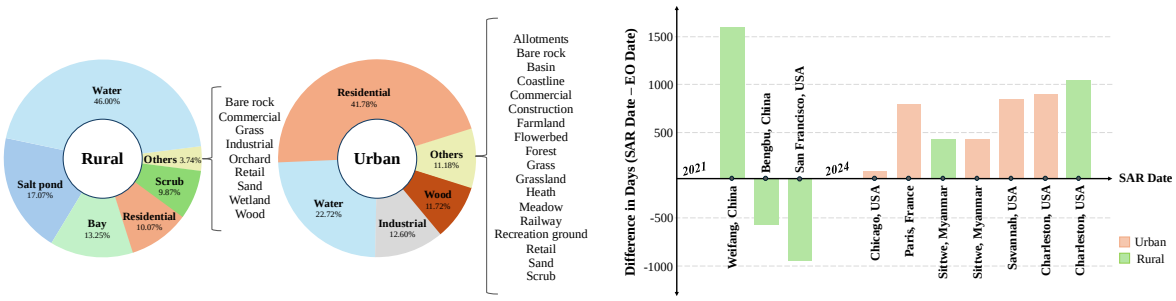
2.3. Remote sensing applications

Recent advancements in large foundation models and generalization models have brought significant benefits to satellite image analysis. GeoChat [9] has demonstrated an EO (Electro-Optical) image-based language model by efficiently fine-tuning large language models. Segment Anything [8] introduced a segmentation model that can be utilized across any domain by training on billion-scale general vision datasets. These technologies have also been applied in the remote sensing domain, being used in various tasks such as change detection [4, 18] and building segmentation [19]. However, as revealed in the study [39], models based on Segment Anything and large language models like GeoChat do not perform effectively on SAR images due to their training on EO images, which have significantly different characteristics. Consequently, in the context of SAR imagery, the benefits of advancements in large foundation models and generalization models have not been fully harnessed.

3. SAR2Earth Dataset

In this section, we provide a detailed description of the SAR2Earth dataset. The SAR2Earth dataset has the following key characteristics:

- **Global Data Collection for Generalization**: To evaluate



(a) Distribution of dataset (b) Time difference

Figure 2. Statistics for the topological distribution and temporal differences in the dataset. (a) Distribution of urban and rural areas by topological elements. (b) Time differences between SAR and EO image captures across regions, indicating the satellite revisit cycles.

generalization performance, the SAR2Earth dataset includes data collected from 8 regions across North America, Europe, and Asia.

- **High Resolution Imagery**: The dataset consists of high resolution images, ranging from 0.15m to 0.6m, offering a diverse mix of spatial resolutions.
- **Consideration of Temporal Shifts**: The dataset accounts for a variety of temporal shifts, ranging from as close as a 1-month difference to as far as a 5-year gap, providing a wide spectrum of temporal scenarios.
- **Structural Diversity**: To address structural shifts, the data is divided into urban and rural categories. The classification is based on the ratio of buildings, amenities, and other structural elements, ensuring a balanced representation of diverse environments.

For sample images and detailed statistics of the dataset, please refer to Figure 1 and Figure 2.

3.1. Dataset design

Data acquisition SAR imagery is sourced from the Capella Space Open Data Program, with a resolution ranging from 0.3 to 0.6 meters per pixel. Its capability to capture detailed information irrespective of weather, cloud cover, or lighting makes it reliable for continuous monitoring.

EO imagery is obtained from Google Earth, with resolutions between 0.15 and 0.6 meters per pixel.

SAR Pre-processing SAR images require significant preprocessing to address noise (such as speckle), geometric distortions, and the wide dynamic range of pixel values. One of the critical steps is translating the raw amplitude or intensity values into decibels (dB), which enhances interpretability by compressing the dynamic range and providing a logarithmic representation suitable for further analysis. The conversion to decibels is performed using the following equation:

$$\sigma^0_{\text{dB}} = 10 \log_{10}(S \cdot D^2) \quad (1)$$

where σ^0_{dB} is the backscatter coefficient in decibels, S is a scaling factor specific to the sensor, and D is the calibrated digital number (DN) values in geocoded format. Note that D is typically the square root of the intensity value, as SAR data is often represented in amplitude.

This conversion provides several benefits: it compresses the dynamic range for enhanced visualization, reduces the influence of extreme pixel values, and improves overall data interpretability, which are crucial for subsequent analysis steps.

Co-registration of SAR and EO A significant challenge in SAR-to-EO translation is achieving precise co-registration between the two modalities due to inherent differences in spatial resolution and coordinate systems, and while accurate spatial alignment improves feature correspondence, perfect matching remains elusive. To address these challenges, we experimented with various co-registration methods, including state-of-the-art data-driven methods, as detailed in the *supplementary material*. While such methods showed promising results on local regions, they were insufficient to guarantee consistent alignment across the entire dataset.

Therefore, to ensure global consistency and broad applicability of our dataset, we adopted a reprojection-based method using the World Geodetic System 1984 (WGS84), the most widely adopted geodetic reference framework in remote sensing and geospatial applications. This reprojection guarantees spatial consistency across all patches, enabling accurate overlay and comprehensive analysis of both SAR and EO data.

The co-registration process is performed using QGIS, a robust geographic information system platform. By leveraging the longitude and latitude coordinates inherent to WGS84, we executed image spatial alignment to achieve pixel-level precision. This procedure facilitated the precise synchronization of spatial features across SAR and EO im-

(a) SAR (b) Denoised SAR (c) Synthetic EO (d) Refined EO (e) EO

Figure 3. The results of SAR-to-EO translation at each step. (a) the original SAR image, (b) the denoised SAR, (c) the SAR-to-EO translation result, (d) the output from the refinement model, and (e) the EO image.

agery, thus enabling more effective translation and interpretation between the two data sources.

3.2. Dataset Statistics

To obtain detailed topological information, we utilized OpenStreetMap (OSM), classifying a total of 25 distinct land cover classes across all regions. The entire dataset covers a combined area of 1444.91 km^2. The dataset comprises a total of 99,998 images, each sized 256x256, generated with a stride of 128. For each region, the dataset is divided into training, validation, and test sets in a 7:1:2 ratio, as detailed further in the *supplementary material*. The regions are classified as urban if residential areas cover at least 25% of the total area. Additionally, if non-residential human-made areas, such as commercial, industrial, or retail spaces, occupy at least 5% of the total area, the region is also categorized as urban [5, 21, 30].

As shown in Figure 2-(a), this classification provides an overview of the topological distribution of urban and rural areas. Specifically, rural areas predominantly consist of natural landscapes, such as vegetation and bodies of water, while urban areas are marked by the presence of human-made structures, including residential, commercial, and industrial buildings.

To assess the temporal diversity of our dataset, Figure 2-(b) illustrates the temporal differences between SAR and EO imagery acquisition across various regions. These temporal gaps vary significantly between regions, offering a wide range of temporal shifts. To the best of our knowledge, this makes our dataset the first to incorporate such diverse temporal differences across a broad set of geographic locations. Acquiring temporally aligned SAR-EO pairs without time discrepancies is particularly challenging in real-world settings, making this diversity crucial for practical applications.

4. SAR2EO Pipelines

In this section, we provide a detailed explanation of our proposed SAR-to-EO pipeline. The SAR-to-EO baseline consists of three main stages: first, a de-noising step to remove the speckle noise inherent in SAR images, as shown in Figure 3-(b); second, an image-to-image translation module that translates SAR images into EO images, as illustrated in Figure 3-(c); and finally, a post-processing structure that refines the generated images for enhanced quality, as demonstrated in Figure 3-(d).

4.1. De-noising

SAR images inherently contain speckle noise due to the interference of radar signals interacting with multiple scatterers. This noise has a multiplicative nature and is closely linked to the signal itself. Since speckle noise strongly correlates with neighboring pixels, conventional methods that assume noise and signal independence are less effective in removing it.

To address this, we adopt a blind-spot method, which predicts the clean value of a pixel based on its surrounding pixels rather than the noisy pixel itself. Given the high correlation of speckle noise among neighboring pixels in SAR images, the blind-spot method is particularly effective at distinguishing and removing noise.

This de-noising process enhances image quality for SAR-to-EO translation tasks. In our work, we compare two blind-spot-based de-noising methods: [12] and [42].

4.2. Image to image translation

SAR-to-EO image translation poses a complex challenge, requiring the handling of both paired and unpaired settings. Due to changes in ground conditions over time, achieving perfect temporal alignment between SAR and EO images is nearly impossible. For instance, while buildings and fixed structures remain relatively constant, elements like vegetation, moving objects, and lighting conditions vary, complicating precise registration.

Considering these factors, SAR-to-EO translation must effectively address both spatial alignment and temporal misalignment. In this paper, we compare paired and unpaired image-to-image translation approaches. Additionally, we propose a partially-paired image-to-image translation method by incorporating objective functions, such as MSE or MAE loss, into the unpaired setting. Given a SAR

Model	Type	MAE ↓	MSE ↓	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
Pix2Pix [7]	pair	0.172	0.051	13.818	0.085	173.751	0.569
Pix2PixHD [32]	pair	0.151	0.041	15.319	0.162	155.073	0.564
BBDM [13]	pair	0.161	0.047	14.772	0.163	**123.051**	0.477
CycleGAN [44]	unpair	0.244	0.062	12.529	0.101	142.532	0.590
CUT [20]	unpair	0.236	0.086	11.172	0.094	144.312	0.592
StegoGAN [34]	unpair	0.214	0.073	12.041	0.152	158.930	0.595
CycleGAN [44]	pair+unpair	0.189	0.063	13.592	0.109	142.532	0.540
CUT [20]	pair+unpair	**0.132**	**0.039**	**16.500**	**0.199**	140.227	**0.350**
StegoGAN [34]	pair+unpair	0.197	0.059	14.213	0.161	166.325	0.593

Table 1. Results for image-to-image translation baselines on the test set of SAR2Earth. We break down results by training data type: paired training data and unpaired training data. All models are trained on the train set of SAR2Earth.

image I_{sar} and an EO image I_{eo}, the modified loss function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{total}}(G, D_{eo}, I_{sar}, I_{eo}) =& \alpha\mathcal{L}_d(D_{eo}, I_{eo}, G(I_{sar})) \\ &+ \beta\mathcal{L}_g(G, I_{sar}) \\ &+ \gamma\mathcal{L}_{mse}(G(I_{sar}), I_{eo})\end{aligned} \tag{2}$$

Here, $\mathcal{L}_d$ is the discriminator loss, responsible for distinguishing real EO images I_{eo} from generated EO images $G(I_{sar})$. The discriminator D_{eo} learns this differentiation. $\mathcal{L}_g$ is the generator loss, applied to various unpaired image-to-image translation models such as CycleGAN [44] and CUT [20].

The term $\mathcal{L}_{mse}$ represents the MSE or MAE loss, which aims to minimize the reconstruction error between $G(I_{sar})$ and I_{eo}. By leveraging partially-paired data, this loss encourages the generator to produce EO images that closely resemble the real EO data, thereby reducing the differences between the generated and real images.

The terms α, β and γ are all hyperparameters, and in all of our experiments, we set α and β to 1, and γ to 0.5.

4.3. Post-processing

After performing SAR-to-EO translation, the generated images may exhibit blurring or artifacts, especially when the data distribution differs from what is seen during training. However, models such as GeoChat or SAM often struggle to perform well on blurred or artifact-affected objects. Therefore, a refinement process is necessary to eliminate these artifacts.

We adopt Restormer as our refinement model. Let $D(.)$ represent the SAR-to-EO translation model, $G(.)$ the generator, and $R(.)$ the refinement network. The objective of the refinement step is defined as follows:

$$\mathcal{L}_{\text{refinement}} = \mathcal{L}_{\text{mae}}(R(G(D(I_{sar}))), I_{eo}) \tag{3}$$

Model	De-noising	MSE ↓	FID ↓
CUT (pair+unpair)	MedianBlur	0.037	140.530
	GaussianBlur	0.032	140.172
	Noise2Noise [12]	0.029	144.230
	MM-BSN [42]	**0.022**	**136.684**

Table 2. Ablation study on de-noising preprocessing methods.

5. Experiments

In this section, we validate the SAR2Earth dataset using various image-to-image translation methods and experiment with different preprocessing and postprocessing techniques.

5.1. Implementation details

Baselines We selected Pix2Pix [7], Pix2PixHD [32], and the diffusion-based BBDM [13] as paired baselines for image-to-image translation. Additionally, we chose CycleGAN [44], CUT [20], and StegoGAN [34] as unpaired baselines. All hyperparameters strictly followed the default settings of the respective methods [1234]. We refer to the output of SAR-to-EO models as Synthetic EO (*SynEO*), and the approach combining paired and unpaired methods is termed the *hybrid* method.

Experiments settings Table 2 presents results obtained without applying de-noising or post-processing, providing a baseline for comparison. From Table 3 onward, de-noising and post-processing steps are consistently applied, utilizing Hybrid CUT to enhance model performance. This progression demonstrates the impact of these additional steps, ensuring clarity in the experimental setup and the effects of de-noising and post-processing on SAR-to-EO translation performance.

[1]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix
[2]https://github.com/taesungp/contrastive-unpaired-translation
[3]https://github.com/xuekt98/BBDM
[4]https://github.com/sian-wusidi/StegoGAN

Experiment Setting	Region	MAE ↓	MSE ↓	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
In-Domain (Single region)	Charleston-U	0.108	0.030	17.235	0.230	130.582	0.320
	Chicago	0.112	0.033	16.983	0.225	132.467	0.327
	Paris	**0.105**	**0.029**	**17.301**	**0.235**	**128.430**	**0.315**
	Savannah	0.115	0.034	16.875	0.222	135.098	0.330
	Sittwe-U	0.109	0.031	17.102	0.229	131.744	0.322
	Bengbu	0.098	0.025	18.512	0.240	120.320	0.300
	Charleston-R	0.101	0.027	18.301	0.238	123.982	0.308
	San Francisco	0.097	0.024	18.734	0.242	118.567	0.295
	Sittwe-R	0.099	0.026	18.589	0.239	121.765	0.305
	Weifang	**0.096**	**0.023**	**18.852**	**0.245**	**117.231**	**0.292**
In-Domain	Urban→Urban	0.106	0.028	17.478	0.240	125.345	0.310
	Rural→Rural	0.097	0.024	18.715	0.241	115.984	0.298
Cross-Domain	Urban→Rural	0.135	0.043	16.253	0.210	145.450	0.360
	Rural→Urban	0.132	0.041	16.438	0.218	143.890	0.355

Table 3. Results for regional test set when trained with 10 regions or the entire urban (Charleston-U, Chicago, Paris, Savannah, Sittwe-U) and rural regions (Bengbu, Charleston-R, San Francisco, Sittwe-R, Weifang).

We use the official codes for OpenEarthMap [35] and GeoChat, where the UnetFormer [31] model are used for land cover segmentation, and the 7B model are used for GeoChat. For further details on the experimental setup of land cover segmentation, please refer to *supplementary material*. We strictly followed all the hyperparameters and settings from the original code.

Evaluation metrics To evaluate the performance of the SAR-to-EO image translation task, we use MAE (Mean Absolute Error), MSE (Mean Squared Error), PSNR (Peak Signal-to-Noise Ratio), and SSIM (Structural Similarity Index Measure) to measure pixel-level accuracy and structural similarity. These metrics capture the absolute and squared differences between the generated and real EO images, assess image quality in terms of noise (PSNR), and ensure structural consistency (SSIM), which are crucial for maintaining fidelity in pixel values and structures in SAR-to-EO translation.

Additionally, we use FID (Fréchet Inception Distance) and LPIPS (Learned Perceptual Image Patch Similarity) to evaluate the perceptual quality and realism of the generated EO images. FID assesses the similarity in feature distributions between the generated and real EO images, while LPIPS focuses on perceptual differences based on deep feature representations, ensuring that the generated images visually resemble real EO data.

5.2. Comparison of baseline

Table 1 presents the results of comparing image-to-image translation methods on the SAR2Earth dataset. As observed in the comparison table, methods under the *paired* setting achieved high accuracy results (MSE, MAE). In contrast, methods under the *unpaired* setting showed lower accuracy (MSE, MAE) but attained higher perceptual scores (FID).

The SAR2Earth task aims to accurately *predict* the correct EO image rather than simply *generate* plausible images. Therefore, metrics such as perceptual scores and MSE, MAE are both important. Accordingly, we combined unpaired baselines that achieved high perceptual scores with paired methods that obtained high MSE and MAE performance. We conducted experiments by applying Eq. 2 on the paired images using existing unpaired methods such as CycleGAN, CUT, and StegoGAN.

Experimental results showed that the hybrid CUT in Table 1 achieved the highest performance. This is because the SAR2Earth dataset is spatially aligned but temporally unaligned. As a result, objects like buildings are in a paired setting, while moving objects are in an unpaired setting. Therefore, a baseline that considers both settings achieved the best performance.

5.3. Comparison of processing

Comparison of de-noising SAR images contain a large amount of speckle noise. This noise appears as granular interference, obscuring important features and textures in the image. It complicates the feature extraction process in data-driven models by introducing high-frequency artifacts, making it challenging to learn accurate mappings between SAR and EO images. To address this issue, de-noising methods have been applied, but because elements in SAR images that appear as noise can actually be important signals, de-noising methods need to be applied carefully. Table 2 shows the performance variations of SAR-to-EO translation according to different de-noising methods.

The results in Table 2 demonstrate that as the de-noising methods become more advanced, performance improves. These experimental results indicate that in the SAR-to-EO translation task, employing more advanced de-noising methods positively impacts performance.

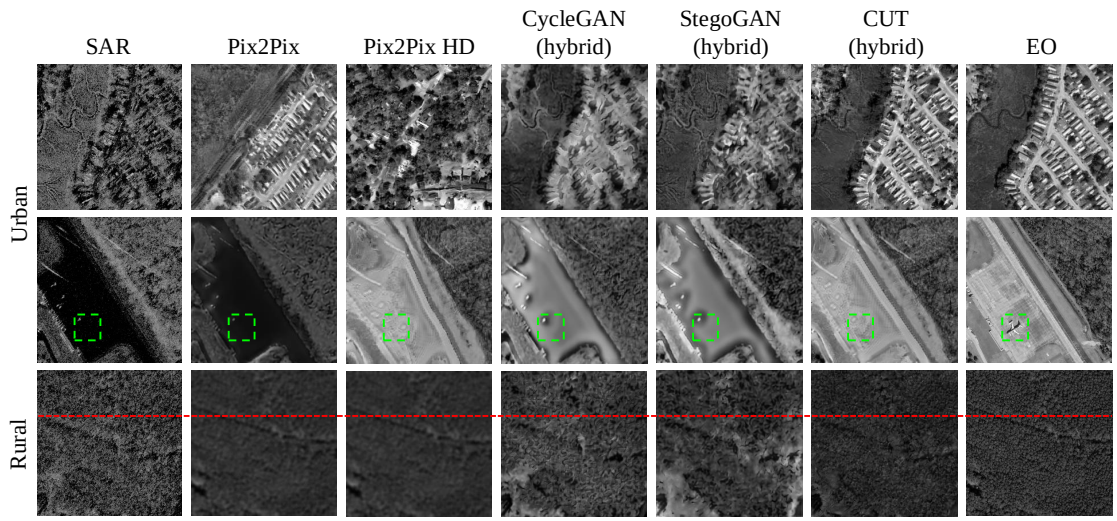


Figure 4. Qualitative comparison of various image-to-image translation methods for SAR-to-EO translation in rural and urban cases.

Comparison of refinement We compared the performance of SAR-to-EO translation with respect to post-processing. For post-processing, we used [41], and during training, we aimed for refinement by adding random deformations (affine transforms, random Gaussian noise) to the EO images. After that, we applied a refinement model to the images translated from SAR-to-EO. We observed that the FID score decreased from **136** to **128**, indicating an improvement in perceptual quality, while the other scores did not change significantly. As observed in the results, we confirmed that the performance improved slightly. Figure 3 illustrates (a) the original SAR, (b) the denoised SAR, (c) the synthetic EO, (d) the refined EO, and (e) the ground truth EO. As shown in Figure 3, we confirmed that the artifacts present in (c) disappeared in (d) through refinement. These experimental results indicate the cause of the performance improvement due to refinement.

5.4. Model Generalization evaluation

The characteristics of SAR images vary significantly by region due to radar backscatter, making it difficult to distinguish between surfaces with similar structures, like oceans and flat areas. As a result, domain gaps in SAR data are often larger than in EO imagery. To evaluate this, we conduct in-domain experiments by training and testing models within the same region.

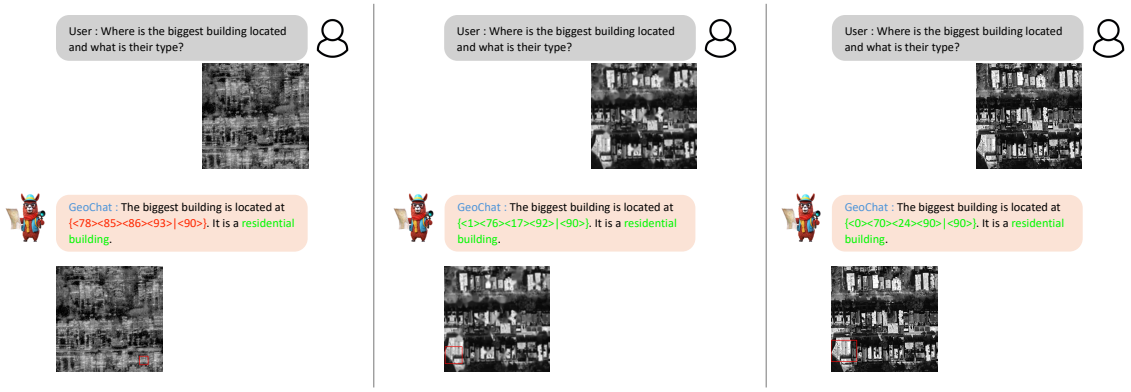
Urban areas, with their complex structures, present larger domain gaps compared to rural areas, which tend to have more uniform natural features. As shown in Table 3, rural regions generally outperform urban areas in in-domain evaluations across all metrics. Notably, training on combined urban regions often yields better results than training on a single region, likely due to increased data diversity. However, for rural regions, training on individual regions produces better results, suggesting that localized models perform better for natural features.

In cross-domain experiments (Urban $\rightarrow$ Rural and Rural $\rightarrow$ Urban), we observe significant performance drops, emphasizing the large differences between these domains. Thus, for practical applications, collecting and training data tailored to specific regional characteristics is more beneficial than simply expanding the dataset without considering regional uniqueness.

5.5. Qualitative results

Figure 4 qualitatively compares the results of SAR-to-EO translation across different baselines. As shown in the figure, CUT (hybrid) produces the most visually plausible results. Specifically, in the second row, indicated by the green dotted box, the SAR image does not contain an airplane signal, and all baselines succeed to generate an airplane in their corresponding SAR-to-EO translation outputs. This experiment demonstrates that, despite the temporally unaligned nature of the SAR-to-EO setting, combining paired and unpaired training approaches effectively mitigates this challenge.

In the rural example (third row), all baselines produce more plausible images compared to their urban counterparts. However, as highlighted by the red dotted line, fully paired methods like pix2pix and pix2pixHD tend to distort features. This is due to the differing imaging angles between SAR and EO data, where SAR images are often captured from a perspective distinct from that of EO imagery. As a result, the paired models attempt to generate EO-like angles, even for features not present in the original SAR image, creating non-existent structures in the SynEO out-



(a) SAR Image (b) Synthetic EO Image (c) EO Image

Figure 5. Comparison of visual grounding tasks using SAR, EO, and SynEO.

put. In contrast, baselines that combine paired and unpaired approaches do not exhibit this distortion tendency, maintaining consistency with the original SAR imagery. These results suggest that if the goal is to generate EO-like angles from SAR data, a paired setting is optimal. However, if the aim is to faithfully replicate the appearance of SAR imagery, a combined paired and unpaired training approach is more effective.

5.6. Application

GeoChat Figure 5 illustrates the results of testing SAR images, SynEO images obtained through SAR-to-EO translation, and actual EO images using the GeoChat large language model (LLM). As shown in the figure, when a SAR image is input into GeoChat, the responses from the model contain entirely incorrect content. This indicates a failure to interpret the SAR data accurately, primarily because SAR images are excessively noisy and differ significantly from the EO or RGB images on which LLMs are predominantly trained. In contrast, when the SynEO and EO images are provided as input, GeoChat generates correct answers, demonstrating its ability to understand and analyze these images effectively.

Land Cover Segmentation As shown in Figure 6, the land cover segmentation results show that SynEO images lead to higher accuracy than SAR images, particularly for artificial classes such as buildings and roads. This indicates that using SynEO as input produces outputs more similar to those from EO images, compared to directly using SAR images. Since most existing land cover segmentation models are trained on EO images, applying them directly to SAR data often results in suboptimal performance. Furthermore, our results highlight the potential of leveraging SAR-to-EO translation to expand the applicability of EO-trained models to SAR data, enabling land cover segmentation across diverse classes despite the inherent differences between SAR and EO imagery.

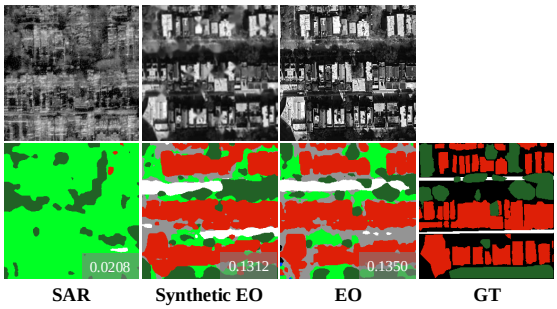


Figure 6. Inference results of SAR, SynEO, and EO images using UnerFormer trained on grayscale OpenEarthMap. The bottom-right corner of each prediction shows the mIoU score.

6. Conclusion

In this paper, we present SAR2Earth, a public benchmark dataset for SAR-to-EO translation designed to support diverse remote sensing applications. We evaluate SAR2Earth using state-of-the-art image-to-image translation models, provide benchmark results, and perform ablation studies on data pre-processing and model architecture. Additionally, experiments on remote sensing applications such as GeoChat and Land Cover Segmentation demonstrate the potential of SAR-to-EO translation in enhancing data accessibility and utility. Our dataset and code are publicly available to encourage future research in applications such as disaster response and AI for social good.

References

[1] Zahraa Tarik AlAli and Salah Abdulghani Alabady. A survey of disaster management and sar operations using sensors and supporting techniques. *International Journal of Disaster Risk Reduction*, 82:103295, 2022. 1

[2] John E Ball, Derek T Anderson, and Chee Seng Chan. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of applied remote sensing*, 11(4):042609–042609, 2017. 1

[3] Miriam Cha, Gregory Angelides, Mark Hamilton, Andy Soszynski, Brandon Swenson, Nathaniel Maidel, Phillip Isola, Taylor Perron, and Bill Freeman. Multiearth 2023–multimodal learning for earth and environment workshop and challenge. *arXiv preprint arXiv:2306.04738*, 2023. 2

[4] Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, and Lorenzo Bruzzone. Adapting segment anything model for change detection in vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[5] Thomas Esch, Wieke Heldens, Andreas Hirner, Manfred Keil, Mattia Marconcini, Achim Roth, Julian Zeidler, Stefan Dech, and Emanuele Strano. Breaking new ground in mapping human settlements from space–the global urban footprint. *ISPRS Journal of Photogrammetry and Remote Sensing*, 134:30–42, 2017. 4

[6] Mario Fuentes Reyes, Stefan Auer, Nina Merkle, Corentin Henry, and Michael Schmitt. Sar-to-optical image translation based on conditional generative adversarial networks—optimization, opportunities and limits. *Remote Sensing*, 11(17):2067, 2019. 1

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 5

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[9] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 1, 2

[10] Navneet Kumar, SS Yamaç, and A Velmurugan. Applications of remote sensing and gis in natural resource management. *Journal of the Andaman Science Association*, 20(1): 1–6, 2015. 1

[11] Jaehyup Lee, Hyebin Cho, Doochun Seo, Hyun-Ho Kim, Jaeheon Jeong, and Munchurl 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*, 2023. 1

[12] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. *Proceedings of the 35th International Conference on Machine Learning*, 80:2965–2974, 2018. 4, 5

[13] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023. 2, 5

[14] Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *arXiv preprint arXiv:2403.06534*, 2024. 2

[15] Spencer Low, Oliver Nina, Angel D Sappa, Erik Blasch, and Nathan Inkawhich. Multi-modal aerial view image challenge: Translation from synthetic aperture radar to electro-optical domain results-pbvs 2023. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 515–523, 2023. 1, 2

[16] Spencer Low, Oliver Nina, Angel D Sappa, Erik Blasch, and Nathan Inkawhich. Multi-modal aerial view object classification challenge results-pbvs 2023. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 412–421, 2023. 1, 2

[17] Spencer Low, Oliver Nina, Dylan Bowald, Angel D Sappa, Nathan Inkawhich, and Peter Bruns. Multi-modal aerial view image challenge: Sensor domain translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3096–3104, 2024. 1

[18] Youngtack Oh, Minseok Seo, Doyi Ki, and Junghoon Seo. Prototype-oriented unsupervised change detection for disaster management. *arXiv preprint arXiv:2310.09759*, 2023. 1, 2

[19] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato Junior. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124:103540, 2023. 2

[20] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345, 2020. 5

[21] Martino Pesaresi, Guo Huadong, Xavier Blaes, Daniele Ehrlich, Stefano Ferri, Lionel Gueguen, Matina Halkia, Mayeul Kauffmann, Thomas Kemper, Linlin Lu, et al. A global human settlement layer from optical hr/vhr rs data: Concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5): 2102–2131, 2013. 4

[22] Clément Rambour, Nicolas Audebert, Elise Koeniguer, Bertrand Le Saux, Michel Crucianu, and Mihai Datcu. Sen12-flood : a sar and multispectral dataset for flood detection. *IEEE Dataport*, 2020. 2

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[24] Michael Schmitt, Lloyd Haydn Hughes, and Xiao Xiang Zhu. The sen1-2 dataset for deep learning in sar-optical data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 141–146, 2018. 1

[25] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019. 2

[26] Minseok Seo, Youngtack Oh, Doyi Kim, Dongmin Kang, and Yeji Choi. Improved flood insights: Diffusion-based sar to eo image translation. *arXiv preprint arXiv:2307.07123*, 2023. 1

[27] Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hansch, Alexei Bastidas, Scott Soenen, Todd Bacastow, et al. Spacenet 6: Multi-sensor all weather mapping dataset. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 196–197, 2020. 1

[28] Marc Spigai, Céline Tison, and Jean-Claude Souyris. Time-frequency analysis in high-resolution sar imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 49(7): 2699–2711, 2011. 1

[29] Haixia Wang, Zhigang Zhang, Zhanyi Hu, and Qiulei Dong. Sar-to-optical image translation with hierarchical latent features. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022. 1

[30] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021. 1, 4

[31] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. 6

[32] 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. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2, 5

[33] Yuanyuan Wang and Xiao Xiang Zhu. The sarptical dataset for joint analysis of sar and optical image in dense urban area. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 6840–6843, 2018. 1

[34] 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7922–7931, 2024. 5

[35] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 6

[36] Yu Xia, Wei He, Qi Huang, Guoying Yin, Wenbin Liu, and Hongyan Zhang. Crformer: Multi-modal data fusion to reconstruct cloud-free optical imagery. *International Journal of Applied Earth Observation and Geoinformation*, 128: 103793, 2024. 2

[37] Fang Xu, Yilei Shi, Patrick Ebel, Wen Yang, and Xiao Xiang Zhu. Multimodal and multiresolution data fusion for high-resolution cloud removal: A novel baseline and benchmark. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15, 2023. 2

[38] Hang Xu, Sylvain Barbot, and Teng Wang. Remote sensing through the fog of war: Infrastructure damage and environmental change during the russian-ukrainian conflict revealed by open-access data. *Natural Hazards Research*, 4(1):1–7, 2024. 1

[39] Zhiyuan Yan, Junxi Li, Xuexue Li, Ruixue Zhou, Wenkai Zhang, Yingchao Feng, Wenhui Diao, Kun Fu, and Xian Sun. Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 2

[40] Xi Yang, Jingyi Zhao, Ziyu Wei, Nannan Wang, and Xinbo Gao. Sar-to-optical image translation based on improved cgan. *Pattern Recognition*, 121:108208, 2022. 1

[41] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 7

[42] Dan Zhang, Fangfang Zhou, Yuwen Jiang, and Zhengming Fu. Mm-bsn: Self-supervised image denoising for real-world with multi-mask based on blind-spot network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4189–4198, 2023. 4, 5

[43] Yueting Zhang, Chibiao Ding, Xiaolan Qiu, and Fangfang Li. The characteristics of the multipath scattering and the application for geometry extraction in high-resolution sar images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(8):4687–4699, 2015. 1

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2, 5