Geometry-Conditioned Pix2Pix: Leveraging Explicit Conditioning on SAR Projected Local Incidence Angle for SAR-to-EO Translation Quality Improvement

Jinmin Lee, Minkyung Chung, Aisha Javed, Youkyung Han*

Dept. of Civil Engineering, Seoul National University of Science and Technology, Seoul, South Korea – (j1nm2_0212, mkjung4876, javedaisha123, han602)@seoultech.ac.kr

Keywords: SAR-to-EO Translation, Incidence Angle, Conditional Generative Adversarial Networks, Structural Fidelity.

Abstract

Electro-optical (EO) imagery is intuitive but highly dependent on weather and illumination, whereas synthetic aperture radar (SAR) imagery provides reliable all-weather observations yet offers limited spectral information. To complement these modalities, recent studies have applied conditional generative adversarial network (cGAN)-based image-to-image translation to SAR-to-EO translation. However, side-looking SAR introduces spatial distortions such as foreshortening and layover that cause relative misalignment with EO imagery, undermining pixel-wise supervision and yielding structural discrepancies between translated outputs and reference EO imagery. In this study, we propose Geometry-Conditioned Pix2Pix (GC-Pix2Pix), which explicitly incorporates on the projected local incidence angle (PLIA) information derived from SAR imagery to better preserve structure and alignment in translated EO imagery. The method is based on Pix2Pix and comprises a two-branch generator and a PatchGAN discriminator. The generator consists of a main network that processes SAR polarimetric channels (VV, VH) and a conditioning subnetwork that extracts PLIA features. The subnetwork uses multi-layer convolutional blocks to capture local PLIA patterns, and the extracted features are then fused with features from the main branch and emphasized via a spatial attention module. For training and evaluation, we assembled a dataset over South Korea that combines Sentinel-1A/1C GRD VV/VH with PLIA and Sentinel-2B Level-2A RGB imagery. We compared GC-Pix2Pix against representative baselines. Across multiple image quality assessment metrics and complementary qualitative analyses, the proposed approach consistently improved SAR-to-EO translation performance. This study establishes a foundational framework for enhancing SAR data utility by synthesizing high-quality EO imagery as an alternative for continuous Earth observation.

1. Introduction

Electro-optical (EO) satellites provide images that resemble human visual perception, which makes them accessible to both remote sensing experts and non-experts. However, EO imagery is acquired by passive sensors that record solar radiation reflected from surface materials, hence image quality is highly sensitive to weather and illumination. In contrast, synthetic aperture radar (SAR) satellites employ active sensors that transmit microwaves, receive the backscattered signal from surface targets, and reconstruct an image from the returned signals. Thus, SAR can provide reliable observations even during natural disasters and seasons with frequent cloud cover, such as summer (Yamaguchi, 2012). Despite these advantages, SAR imagery remains challenging to interpret visually due to the absence of spectral signatures and the presence of speckle noise.

To exploit the complementary strengths of the two modalities, recent studies have explored SAR-to-EO translation using deep learning-based image-to-image translation methods. A representative approach is conditional generative adversarial networks (cGANs), which learn a mapping from a source domain to a target domain by jointly optimizing a generator and a discriminator through adversarial training (Mirza and Osindero, 2014). Because the discriminator effectively learns realistic texture and high-frequency details, cGANs have been widely adopted for SAR-to-EO translation.

Early approaches for SAR-to-EO image translation simply applied natural scene-based cGANs, such as Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017), DualGAN (Yi et al., 2017), and NICE-GAN (Chen et al., 2020) to the task (Fuentes et al., 2019; Wang et al., 2019; Li et al., 2020). Recently, however, as the utility of SAR-to-EO translation has grown for enhancing downstream applications like cloud removal (Xiang et al., 2024) and wildfire damage mapping (Hu et al., 2023), several methods tailored to SAR and EO modalities have emerged. For instance, Seg-CycleGAN (Zhang et al., 2025) prevents the degradation of contextual information caused by SAR speckle noise by utilizing semantic segmentation results of the translated imagery as a loss function. Furthermore, DOGAN (He et al., 2025) attempts to resolve the domain misalignment between EO and SAR imagery by integrating the representational power of vision foundation models into the generator.

Despite these recent advancements, research focused on preserving the structural fidelity of the real EO image remains significantly limited. SAR imagery is acquired in a side-looking geometry that induces geometric distortions such as layover and foreshortening, which in turn introduce relative misalignment with EO imagery. Such misalignment is a major factor in the deterioration of cGAN performance that rely on pixel-wise losses, leading to structural discrepancies between the real EO images and the translated outputs. Therefore, in this study, we propose a cGAN-based network that enhances the structural fidelity of translated EO imagery by incorporating SAR projected local incidence angle (PLIA) information as an explicit conditioning parameter to learn spatial distortions that cause misalignment between paired SAR and EO imagery.

2. Methodology

In this study, we propose a Geometry-Conditioned Pix2Pix (GC-Pix2Pix) that extends the representative image-to-image network,

* Corresponding author

Pix2Pix, with a dedicated subnetwork for extracting PLIA information and integrating it as explicit conditioning features. Here, PLIA information denotes the incidence angle, defined with respect to the surface normal and projected onto a digital elevation model. GC-Pix2Pix comprises a generator and a conditional discriminator. The generator is factorized into a main network that translates SAR polarimetric channels (i.e., vertical-vertical (VV) and vertical-horizontal (VH) polarimetric bands) and a conditioning subnetwork that extracts and fuses PLIA features. The overall architecture of GC-Pix2Pix is provided in Figure 1.

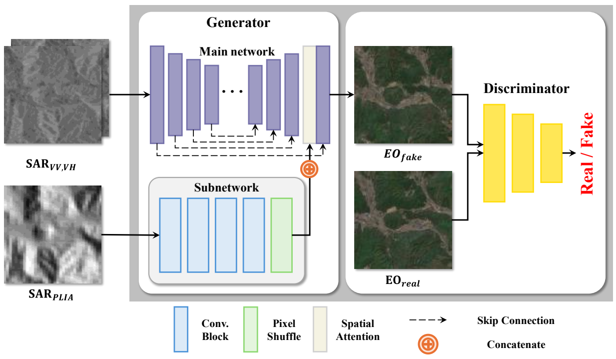

Figure 1. Overview of the GC-Pix2Pix.

2.1 Generator

The proposed generator processes the SAR polarimetric channels ($SAR_{VV,VH}$) through its main network, while the conditioning subnetwork takes the PLIA image (SAR_{PLIA}) as input (Figure 2). The main network adopts a U-Net-like architecture, which preserves the structural content of the $SAR_{VV,VH}$ via skip connections between the encoder and decoder. This design is particularly suitable for SAR-to-EO translation, as the $SAR_{VV,VH}$ contains extensive scattering information that must be retained throughout the reconstruction process, and skip connections effectively prevent the loss of fine structural details.

Meanwhile, the subnetwork comprises stacked convolutional blocks followed by a single pixel-shuffle block. Given that PLIA primarily encodes geometric and angular information rather than intricate texture, a lightweight convolutional structure is sufficient to capture its spatial patterns without introducing unnecessary complexity. Additionally, the inclusion of the pixel-shuffle operation enables efficient resolution recovery, which ensures that the subnetwork features remain precisely aligned with the feature scale of the main network. Accordingly, this subnetwork extracts spatial patterns from SAR_{PLIA} to learn geometric distortions in the $SAR_{VV,VH}$ induced by the incidence angle between the radar propagation direction and the ground.

At the outermost stage of the main network’s decoder, the feature maps produced by the subnetwork are concatenated with those of the main branch. Subsequently, the concatenated feature maps are fed into a spatial attention module (Woo et al., 2018) to fuse features and emphasize the effects of PLIA at the pixel level. The spatial attention module compresses the pixel-wise channels of input feature maps F_{input} into a single-channel attention map. It then conducts an element-wise product to generate the refined feature maps F_{refine}. The process of spatial attention is expressed as follows:

$$F_{refine} = \sigma(Conv_{7\times7}(F_{input})) \otimes F_{input}, \qquad (1)$$

where σ = sigmoid function
$Conv_{7\times7}$ = 7 × 7 convolution operation
$\otimes$ = element-wise product

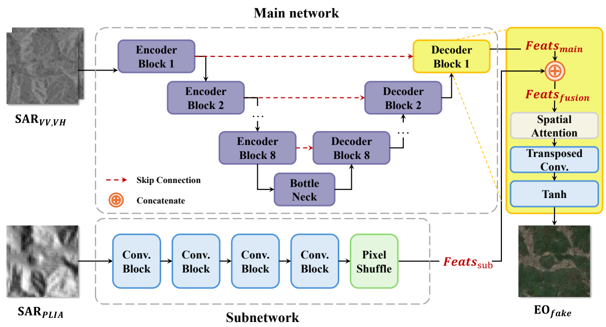

Figure 2. Architecture of the proposed generator.

After emphasizing the positional cues of the fused feature maps, the final upsampling stage resizes the feature maps and produces the fake EO image (EO_{fake}).

2.2 Discriminator

Next, the generated EO_{fake} and the corresponding real EO image (EO_{real}) are fed into the PatchGAN discriminator proposed by Isola et al. (2017). Unlike conventional GANs (Goodfellow et al., 2014) that models unconditional data distribution, the PatchGAN discriminator operates as a conditional discriminator, which takes both source and target domain to estimate the target distribution. Accordingly, both EO_{fake} and EO_{real} are concatenated with the $SAR_{VV,VH}$, respectively, and provided as inputs to the discriminator. As depicted in Figure 3, The PatchGAN discriminator applies convolution operations with a fixed receptive field to produce a patch-wise realism score map. Through this patch-wise formulation, the PatchGAN discriminator encourages the network to focus on local details rather than relying on a single global decision, which distinguishes it from conventional GAN discriminators. In this work, we adopt a 70 × 70 size PatchGAN discriminator to balance local detail preservation and global consistency.

Afterward, the resulting patch-wise score map is used to compute the adversarial loss, which is then utilized to optimize the discriminator. Since the discriminator fundamentally functions as a binary classifier, cross-entropy (CE) loss, which is commonly used in classification tasks, has traditionally been the standard choice for adversarial training. However, CE loss suffers from instability during training—particularly in the early stages when the generated images are of low quality, allowing the discriminator to converge too quickly and resulting in vanishing gradients in the generator (Mao et al., 2017).

To address this issue, we replace the conventional CE-based loss with the hinge loss proposed by Lim and Ye (2017), with the aim of improving training stability and mitigating gradient saturation. Compared with CE loss, the hinge loss has been shown in prior studies to more effectively penalize margin violations at the decision boundary, thereby promoting more stable generator-discriminator interaction (Wang et al., 2019; Bai and Xu, 2024). Based on this, the adversarial loss functions for the discriminator $\mathcal{L}^{D}_{Adv,real}$ and $\mathcal{L}^{D}_{Adv,fake}$ are defined as:

$$\mathcal{L}^{D}_{Adv,real} = \mathbb{E}_{y\sim p_{Y|X}}\left[max\left(0, 1 - D(y)\right)\right], \qquad (2)$$

$$\mathcal{L}^{D}_{Adv,fake} = \mathbb{E}_{x\sim p_X}\left[max\left(0, 1 + D(G(x))\right)\right], \qquad (3)$$

where $D(\cdot)$= output of the discriminator
$G(\cdot)$= output of the generator
$x \sim p_X$= input $SAR_{VV,VH}$ samples
$y \sim p_{Y|X}$= EO_{real} samples conditioned on x
$\mathbb{E}[\cdot]$ = expectation
$max(0,\ \cdot)$ = hinge function with margin 1

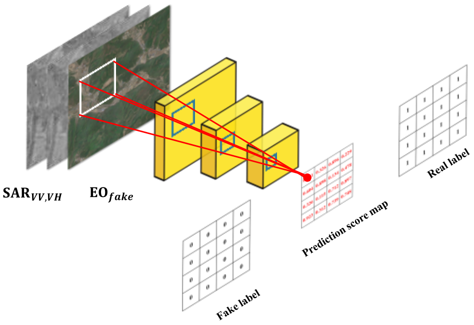

Figure 3. Illustration of the PatchGAN discriminator process.

2.3 Loss Functions

The objective of the proposed network is to optimize the generator and discriminator through an adversarial min-max game. The discriminator is trained to distinguish EO_{real} from generated EO_{fake} by minimizing the hinge-based adversarial loss, which enforces a margin between real and fake samples. In contrast, the generator is optimized to maximize the discriminator's response for generated samples, aiming to increase the discriminator score for generated outputs so that they are indistinguishable from real data. To this end, the hinge-based adversarial loss for the generator is defined as follows:

$$\mathcal{L}_{Adv}^{G} = \ -\mathbb{E}_{x \sim p_X}[D(G(x))], \tag{4}$$

where $D(\cdot)$= output of discriminator
$G(\cdot)$= output of generator
$x \sim p_X$ = input random noise distribution
$\mathbb{E}[\cdot]$ = expectation

In addition to the adversarial objective, we incorporate multiple reconstruction losses to preserve structural and perceptual fidelity. The L1 loss enforces pixel-level consistency between generated EO_{fake} and reference EO_{real}, while the Laplacian-based edge loss enhances boundary preservation by emphasizing high-frequency components. Furthermore, a VGG19-based perceptual loss is employed to measure high-level feature similarity, encouraging the generated images to align with human visual perception. Taken together, the final discriminator loss ($\mathcal{L}_D$) and generator loss ($\mathcal{L}_G$) are formulated as weighted combinations of terms, where each loss weight is empirically set to ensure a balanced optimization between structural accuracy and perceptual realism.

$$\mathcal{L}_D \ = \ 0.5\, \mathcal{L}_{Adv,real}^{D} + 0.5\, \mathcal{L}_{Adv,fake}^{D}, \tag{5}$$
$$\mathcal{L}_G \ = \ 2\, \mathcal{L}_{Adv}^{G} + 10\, \mathcal{L}_{L1} + 10\, \mathcal{L}_{Edge} \ + 10\, \mathcal{L}_{Percep}, \tag{6}$$

Scene number	EO		SAR	
	Satellite	Acquisition date	Satellite	Acquisition date
Scene 1	S-2B	25.04.26	S-1A	25.04.26
Scene 2	S-2B	24.05.31	S-1A	24.05.30
Scene 3	S-2B	25.04.26	S-1C	25.04.25

Table 1. Summary of EO and SAR satellite data acquisition per scene (S indicates Sentinel satellite).

where $\mathcal{L}_{L1}$= L1 reconstruction loss
$\mathcal{L}_{Edge}$= Laplacian-based edge loss
$\mathcal{L}_{Percep}$ = VGG19-based perceptual loss

3. Experimental Dataset and Implementation Details

3.1 Experimental Dataset

For network training and evaluation, we constructed a three paired dataset over South Korea by collecting Sentinel-1A/1C ground range detected (GRD) SAR imagery together with Sentinel-2B Level-2A EO imagery, as illustrated in Figure 4, Table 1, and Table 2. The EO data utilized in this study consisted of 10,980 × 10,980-pixel images with red, green, and blue bands at a spatial resolution of 10 m. To process the SAR imagery, the GRD preprocessing workflow proposed by Filipponi (2019) was employed to generate $SAR_{VV,VH}$ and SAR_{PLIA}.

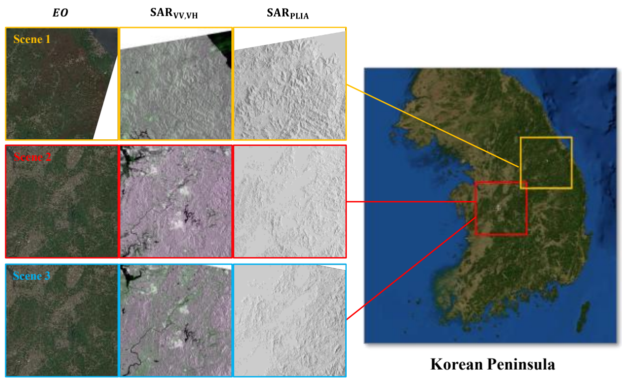

Figure 4. Experimental sites and corresponding satellite images used in this study.

Specification	Sentinel-2B	Sentinel-1A/1C
Sensor type	Multispectral Imager	C-band SAR
Processing level	Level-2A	GRD
Operate mode	-	Interferometric wide swath
Band/Polarization	B02 (Blue), B03 (Green), B04 (Red)	VV, VH
Ground sample distance	10 m × 10 m	10 m × 10 m

Table 2. Detailed specifications of Sentinel-1 and Sentinel-2 sensors used in this study

Consequently, the overlapping regions of the SAR and EO images were partitioned into 256 × 256-pixel patches. During this process, any images containing cloud cover or "NoData" values were removed through rigorous visual inspection to ensure the integrity of the dataset. The final dataset comprised 6,310 SAR–EO pairs, which were split into training and evaluation sets at a 70:30 ratio.

3.2 Implementation Details

To validate the proposed network, we compared SAR-to-EO translation results from representative image-to-image translation networks—Pix2Pix, CycleGAN, and StegoGAN (Wu et al., 2024) —as well as MT_GAN (Wang et al., 2025), a task-specific baseline, and our GC-Pix2Pix. The training process was

Model	PSNR (↑)	SSIM (↑)	FSIM (↑)	SAM (↓)	LPIPS (↓)	FID (↓)
Pix2Pix	<u>28.6945</u>	<u>0.7107</u>	<u>0.8054</u>	<u>0.1337</u>	<u>0.3695</u>	2.4881
CycleGAN	25.3282	0.6165	0.7477	0.1804	0.4427	0.7459
StegoGAN	26.5475	0.6379	0.7707	0.1598	0.4183	<u>0.5963</u>
MT_GAN	27.9110	0.6845	0.7870	0.1449	0.3878	1.6204
GC-Pix2Pix (ours)	**29.0362**	**0.7121**	**0.8070**	**0.1117**	**0.3462**	**0.4304**

Table 3. Image quality assessment of generated EO images by different networks. The best and second-best results for each metric are highlighted in bold and underline, respectively.

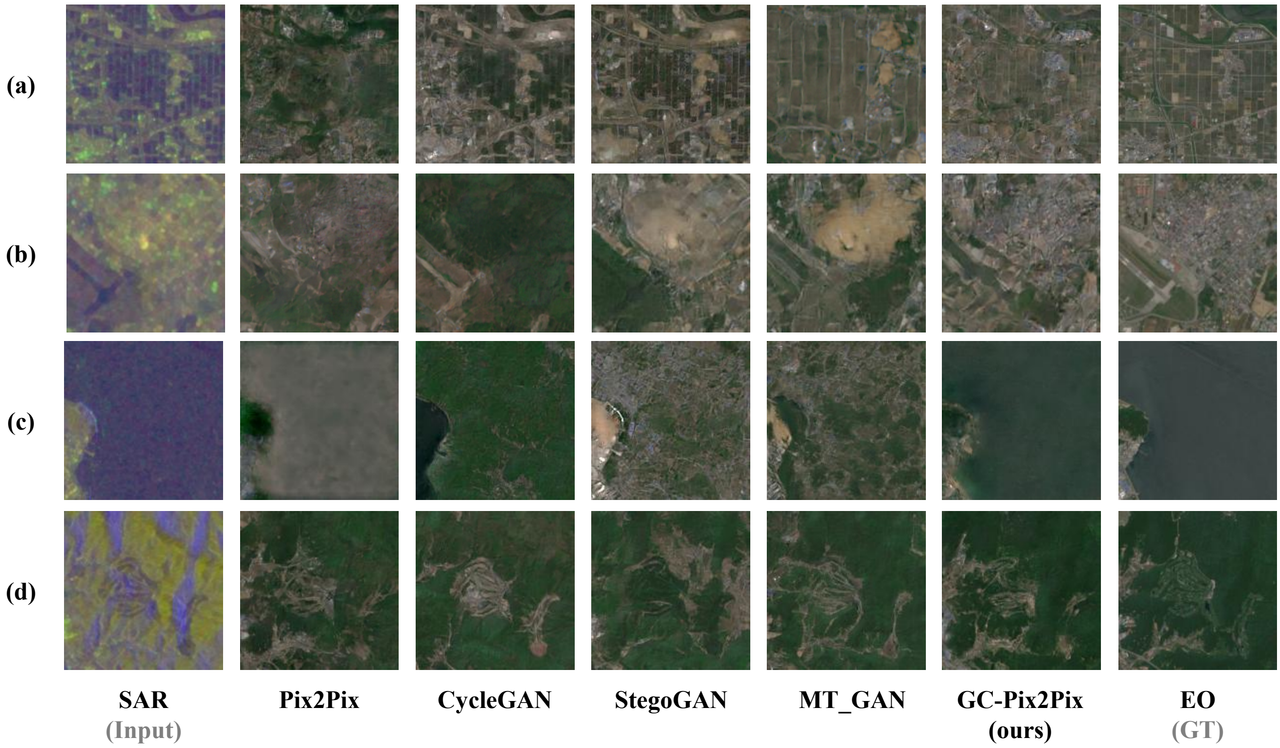

Figure 5. Visual comparison of SAR-to-EO translation results generated by different networks: (a) Agricultural area, (b) Residual area, (c) Water body area, and (d) Mountainous area.

conducted on a system equipped with an Intel Xeon Gold 5220 18-Core Processor, 512 GB of RAM, and an NVIDIA GeForce RTX A6000 GPU. Training spanned 200 epochs, utilizing a batch size of 8, a learning rate of 1×10^{-4}, and the Adam optimizer.

For the existing networks, $SAR_{VV,VH}$ and SAR_{PLIA} images were simply stacked as 3-band input, whereas our method uses $SAR_{VV,VH}$ as main 2-band input and incorporates SAR_{PLIA} features through explicit conditioning within the generator. To ensure a fair and consistent comparison, all networks—including those originally designed for unpaired image translation—were trained in a paired setting on SAR and EO imagery from the same locations.

4. Experimental Results and Analysis

To assess the performance of the proposed network, the outputs were evaluated quantitatively using multiple image quality assessment (IQA) metrics and qualitatively through visual assessment. For the quantitative analysis, structural and spectral similarities were evaluated using representative IQA metrics, specifically peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) (Wang et al., 2004). To further examine the consistency of robust features such as edges, we employed the feature similarity index measurement (FSIM), while the spectral angle mapper (SAM) was utilized to determine spectral fidelity (Zhang et al., 2011).

Furthermore, to evaluate the perceptual quality of the generated imagery, we incorporated metrics that quantify perceptual discrepancies based on pre-trained deep features. Specifically, we employed learned perceptual image patch similarity (LPIPS) and Fréchet inception distance (FID), which utilize the feature representations of VGG16 and Inception-v3, respectively, to ensure that the translated results align with human visual perception (Zhang et al., 2018; Heusel et al., 2017).

According to the quantitative comparison (Table 3), the proposed GC-Pix2Pix achieved superior performance across all metrics that represent structural and spectral fidelity, specifically PSNR (29.0362), SSIM (0.7121), FSIM (0.8070), and SAM (0.1117). Despite the geometric distortions inherent in SAR imagery, GC-Pix2Pix attained higher SSIM and FSIM scores compared to the baseline Pix2Pix. This improvement suggests that the explicit conditioning on PLIA information effectively preserves physical land-cover characteristics by accounting for the SAR acquisition geometry. Furthermore, the lowest SAM value indicates that the model enables the synthesis of EO imagery with colors that closely resemble the ground truth.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-163-2026

In terms of perceptual similarity and the quality of the generated data distribution, GC-Pix2Pix significantly outperformed competing models, reaching the lowest scores for LPIPS (0.3462) and FID (0.4304). Notably, the FID score showed a substantial improvement, being approximately five times lower than that of the Pix2Pix baseline (2.4881). This indicates that the proposed method effectively generates the realistic textures and fine-scale details of EO imagery in a statistically consistent manner. While StegoGAN was the second-best performer in terms of FID, GC-Pix2Pix demonstrated a superior overall balance, validating its robust performance across all evaluated IQA metrics.

The visual comparison presented in Figure 5 further substantiates the superiority of the proposed GC-Pix2Pix architecture. In the agricultural region (Figure 5a), baseline models such as Pix2Pix and CycleGAN incorrectly reconstruct the farmland with textures resembling dense forests. Meanwhile, StegoGAN successfully delineates the agricultural parcels but exhibits significant spectral degradation. Although MT_GAN achieves both parcel delineation and spectral recovery, it erroneously synthesizes the residential area in the upper right corner as bare soil. In contrast, the proposed method accurately delineates the parcel boundaries while successfully restoring the spatial footprint of the residential complex.

Furthermore, in the complex residential zone (Figure 5b), Pix2Pix and CycleGAN largely fail to reconstruct the underlying topography. Both StegoGAN and MT_GAN show partial success but incorrectly generate bare land over the dense housing areas at the top of the image. Conversely, GC-Pix2Pix effectively preserves the high-frequency structural outlines of the buildings and reconstructs the residential areas with a high degree of similarity to the ground truth.

The robustness of the proposed method extends to diverse natural land covers, such as water bodies and mountainous terrain. For the water body shown in Figure 5(c), the comparative models produce severe artifacts that resemble forest or urban textures. The proposed method, however, successfully estimates the homogeneous, smooth texture and accurate color of the water surface while properly restoring the adjacent vegetation.

Finally, in the mountainous forest region depicted in Figure 5(d), all evaluated networks generate outputs that generally resemble the real EO imagery, although some confusion between bare soil and vegetation remained. Among the compared methods, the proposed method synthesizes the most accurate land-cover representation, which closely matches the complex textural distribution of the actual EO reference. Overall, these qualitative results confirm that explicit PLIA conditioning enables GC-Pix2Pix to excel in preserving fine-scale structural details and preventing the hallucination of incorrect land covers across diverse landscapes.

5. Conclusion

In this study, we proposed GC-Pix2Pix, a conditional generative adversarial network that explicitly incorporates SAR PLIA information to improve structural fidelity of SAR-to-EO image translation. Unlike conventional cGAN-based approaches that rely primarily on stacked multi-band inputs, the proposed method introduces a dedicated conditioning subnetwork to model geometry-induced distortions inherent in side-looking SAR imagery and integrates PLIA features into the generator through feature fusion and spatial attention.

Experimental results on a paired Sentinel-1A/1C and Sentinel-2B dataset over South Korea demonstrated that GC-Pix2Pix consistently outperformed representative baselines, including Pix2Pix, CycleGAN, StegoGAN, and MT_GAN, across multiple quantitative IQA metrics. Qualitative comparisons further confirmed that GC-Pix2Pix better preserves land-cover structures and fine-scale details of both natural and human-made objects, which can be attributed to the explicit modeling of geometry-related effects via PLIA conditioning.

These results suggest that incorporating physically meaningful SAR acquisition parameters as explicit conditioning signals is an effective strategy for mitigating geometric misalignment between SAR and EO imagery. Future work will focus on extending the proposed framework to multi-temporal and multi-sensor settings, as well as investigating its impact on downstream remote sensing tasks such as change detection, object detection, and damage assessment.

Acknowledgements

This study was conducted with the support of the “Development of Application Support System for Satellite Information Big Data (RS-2022-00165154)” of the Korea Aerospace Administration.

References

Bai, X., & Xu, F. (2024). Accelerating diffusion for SAR-to-Optical image translation via adversarial consistency distillation. *arXiv prepr.*, arXiv:2407.06095.

Chen, R., Huang, W., Huang, B., Sun, F., Fang, B., 2020. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 8168-8177.

Filipponi, F., 2019. Sentinel-1 GRD preprocessing workflow. *Proc.*, *18*(1), 11. https://doi.org/10.3390/ECRS-3-06201

Fuentes Reyes, M., Auer, S., Merkle, N., Henry, C., & Schmitt, M., 2019. SAR-to-optical image translation based on conditional generative adversarial networks—Optimization, opportunities and limits. *Remote Sens.*, *11*(17), 2067. https://doi.org-/10.3390/rs11172067

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. *Adv. neural inf. process. syst.*, 27.

He, J., Chen, L., Shi, H., Chen, Y., Yang, J., Li, W., 2025. DOGAN: DINO-based optical-prior-driven GAN for SAR-to-optical image translation. *IEEE Trans. Geosci. Remote Sens.*, *63*, 1-16. https://doi.org/10.1109/TGRS.2025.3606979

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. neural inf. process. syst.*, *30*.

Hu, X., Zhang, P., Ban, Y., Rahnemoonfar, M., 2023. GAN-based SAR and optical image translation for wildfire impact assessment using multi-source remote sensing data. *Remote Sens. Environ.*, 289, 113522. doi.org/10.1016/j.rse.2023.113522

Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,* 1125-1134.

Li, Y., Fu, R., Meng, X., Jin, W., Shao, F., 2020. A SAR-to-optical image translation method based on conditional generation adversarial network (cGAN). *IEEE Access*, *8*, 60338-60343. https://doi.org/10.1109/ACCESS.2019.2939649

Lim, J. H., Ye, J. C., 2017. Geometric GAN. *arXiv prepr.*, arXiv:1705.02894.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2794-2802.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. *arXiv prepr.*, arXiv:1411.1784.

Wang, L., Xu, X., Yu, Y., Yang, R., Gui, R., Xu, Z., Pu, F., 2019. SAR-to-optical image translation using supervised cycle-consistent adversarial networks. *IEEE Access*, *7*, 129136-129149. https://doi.org/10.1109/ACCESS.2019.2939649

Wang, P., Chen, Y., Huang, B., Zhu, D., Lu, T., Dalla Mura, M., Chanussot, J, 2025. MT_GAN: A SAR-to-optical image translation method for cloud removal. *ISPRS J. Photogramm. Remote Sens.*, 225, 180-195. https://doi.org/10.1016/j.isprsjprs.2025.04.011

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, *13*(4), 600-612. https://doi.org/10.1109/TIP.2003.819861

Woo, S., Park, J., Lee, J. Y., Kweon, I. S., 2018. CBAM: Convolutional block attention module. In *Proc. Eur. Conf. Comput. Vis.*, 3-19.

Wu, S., Chen, Y., Mermet, S., Hurni, L., Schindler, K., Gonthier, N., Landrieu, L., 2024. StegoGAN: Leveraging steganography for non-bijective image-to-image translation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 7922-7931.

Xiang, X., Tan, Y., & Yan, L., 2024. Cloud-guided fusion with SAR-to-optical translation for thick cloud removal. *IEEE Trans. Geosci. Remote Sens.*, *62*, 1-15. https://doi.org/10.1109-/TGRS.2024.3431556

Yamaguchi, Y., 2012. Disaster monitoring by fully polarimetric SAR data acquired with ALOS-PALSAR. *Proc. IEEE*, 100(10), 2851-2860. https://doi.org/10.1109/JPROC.2012.2195469

Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. DualGAN: Unsupervised dual learning for image-to-image translation. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2849-2857.

Zhang, H., Li, H., Lin, J., Zhang, Y., Fan, J., Liu, H., Liu, K., 2025. Seg-CycleGAN: SAR-to-optical image translation guided by a downstream task. *IEEE Geosci. Remote Sens. Lett.*, *22*, 1-5. https://doi.org/10.1109/LGRS.2025.3538868

Zhang, L., Zhang, L., Mou, X., Zhang, D., 2011. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, *20*(8), 2378-2386. https://doi.org/10.1109/TIP.2011.2109730

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 586-595.

Zhu, J. Y., Park, T., Isola, P., Efros, A. A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2223-2232.