SAR-to-Optical Image Translation Using Wavelet Based GAN Networks

Artificial Intelligence Methods Lab

Kürşat Kömürcü, Vytautas Paura, Valentas Gružauskas, Linas Petkevičius

Artificial Intelligence Methods Lab
Institute of Computer Science, Vilnius University
Corresponding author: vytautas.paura@mif.vu.lt

Follow us: aiml.lt

Abstract

Translating synthetic aperture radar (SAR) images into optical-like images makes all-weather observations easier to interpret, but unpaired generative models tend to trade geometric fidelity for plausible texture. We propose a modified generative adversarial network (called SWIFT-GAN) for synthetic aperture radar (SAR)-to-optical image translation that uses a single-pass generator based on Haar-wavelets and Fourier phase locking. A collection of Sentinel-1/2 imagery over 150 areas of the Baltic states (consisting of more than 30,000 SAR and 120,000 optical images) was gathered to create training and testing data. The newly created GAN network with the single-pass generator has about 63% fewer parameters than the baselines: $3.5M$ vs $7.9-13.1M$ parameters compared to existing models. This new proposed method greatly reduces inference times compared to baseline models.

Problem

Synthetic aperture radar (SAR) images the Earth's surface through clouds and at night, which makes it the most reliable source of observations for flood mapping, disaster response and continuous land monitoring. Its images are, however, hard to interpret: SAR is an active, side-looking microwave sensor, so its intensities are governed by surface roughness and backscattering geometry rather than by reflectance, and the imagery is corrupted by coherent speckle and by geometric effects such as layover and shadowing. Translating SAR into optical-like imagery is therefore an attractive way to make SAR usable by analysts and by models trained on optical data.

Dataset

We selected 181 areas of 20×20 km across Lithuania, Latvia and Estonia, covering urban, rural, agricultural, forest and water surfaces (Figure 1), and exported for each area Sentinel-1 and Sentinel-2 imagery through Google Earth Engine between January 2024 and October 2025. Sentinel-1 data are Ground Range Detected products in Interferometric Wide swath mode with VV and VH polarisations. Sentinel-2 data are Level-2A surface reflectance; scenes with more than 50% cloud cover were discarded, and in some areas the acquisition window was limited to March-October to avoid snow.

Dataset

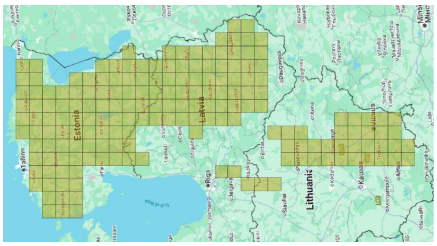

Figure 1: The 181 study areas in Lithuania, Latvia and Estonia, chosen to cover urban, rural, agricultural, forest and water surfaces.

Generator

The generator (Figure 2) consists of an encoder with two Haar-wavelet downsampling stages, a bottleneck operating at 64×64 with 256 channels, and a decoder:

- **Wavelet downsampling.** Each downsampling stage applies a one-level Haar transform to every channel, concatenates the four sub-bands and projects them with a 1×1 convolution.
- **Bottleneck.** The bottleneck interleaves three transformer blocks with two spectral transform blocks (STB).
- **Spectral transform block.** For an input feature map h the STB computes

$$\mathcal{F}(h) = A\, e^{i\Phi}, \qquad A = |\mathcal{F}(h)|, \quad \Phi = \angle\mathcal{F}(h).$$

- **Decoder.** Two transposed convolutions (3×3, stride 2) with instance normalisation and ReLU restore the input resolution, followed by a 7×7 convolution with reflection padding and a tanh output. The generator has 3.54 million parameters; without the STBs it has 2.88 million.
- **Losses. Wavelet loss** - reconstructions are also compared in the Haar domain; **Color loss** - To match color statistics generated and real optical images are converted to CIELAB; **Phase consistency** - cannot phase angles directly, because angles wrap around, instead compare unit phasors.

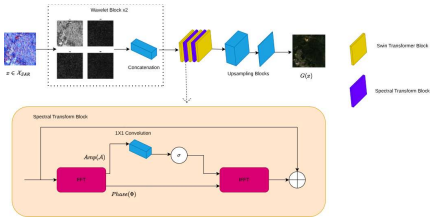

Figure 2: Architecture of the SWIFT-GAN generator.

Discriminator

Both domains use a PatchGAN discriminator[2] with four 4×4 convolutional stages (64, 128, 256 and 512 channels, instance normalization, leaky ReLU) and a final 4×4 convolution producing a map of real/fake scores.

Results

Table 1: Results on 2,000 co-registered Sentinel-1/2 patch pairs ($|\Delta t| \leq 3$ days, 149 areas). Fidelity metrics are computed against the aligned optical image.

Method	SSIM↑	PSNR↑	MAE↓	LPIPS↓	GSSIM↑	FID↓	KID↓
Identity (SAR)	0.006	5.07	0.493	0.862	0.057	393.1	0.454
Random optical	0.307	–	0.144	0.568	0.177	0.0	0.000
CycleGAN[5]	0.270	15.34	0.153	0.558	**0.261**	**104.6**	**0.071**
CUT[4]	0.041	7.14	0.407	0.677	0.093	255.7	0.272
SwinCUT	**0.337**	**16.49**	0.150	0.553	0.246	207.5	0.207
QS-Attn[1]	0.198	15.31	0.140	0.630	0.163	284.0	0.310
UNSB[3]	0.029	6.83	0.402	0.814	0.069	368.8	0.410
SWIFT-GAN	0.290	16.32	**0.134**	**0.521**	0.204	121.7	0.102

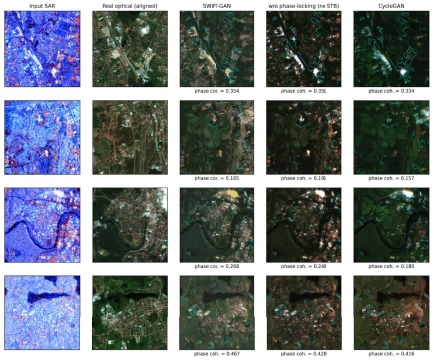

Figure 3: Columns: input SAR, real optical image acquired within three days, SWIFT-GAN, SWIFT-GAN (no phase-locking), and CycleGAN. Rows: cloud-free test patches with the largest GSSIM margin of SWIFT-GAN over the better of the other two models.

Conclusions

- Model without phase-locking produces images that are less realistic and whose edge structure agrees less with the optical reference.
- Training a SWIFT-GAN model on the Baltic data takes about 2.5 h per epoch on one NVIDIA V100. One epoch contains 23,888 optical patches.
- Generators built only from residual convolution blocks are faster (9–10 ms), whereas UNSB[3], which samples in five steps. is 2.8 times slower.
- Removing the color loss gives the lowest distortion of all models (LPIPS 0.491, MAE 0.108, SSIM 0.403) at a higher FID (131.9), whereas removing the wavelet loss lowers FID to 107.8 at a slightly higher LPIPS.

Bibliography

[1] X. Hu, X. Zhou, Q. Huang, Z. Shi, L. Sun, and Q. Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18291–18300, 2022.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[3] B. Kim, G. Kwon, K. Kim, and J. C. Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *International Conference on Learning Representations*, volume 2024, pages 19312–19331, 2024.

[4] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision – ECCV 2020*, volume 12354 of *Lecture Notes in Computer Science*. pages 319–345, Cham, 2020. Springer.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.