S2OFormer: High-Performance SAR-Optical Translation Model Based on Transformer

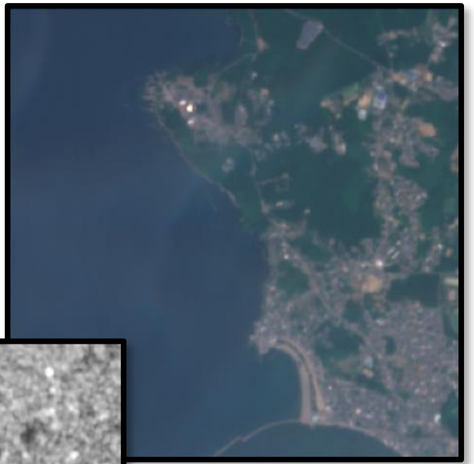
S2OFormer：Transformerに基づく高性能SAR-光学変換モデルの開発

〇富　宣超（東大 /産総研）・神山　徹（産総研）
中村　良介（産総研）・吉川　一朗（東大）

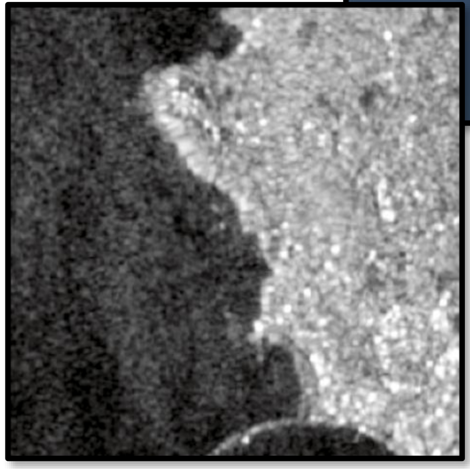
1.1 Advantages and Disadvantages of SAR/OPT

	CAVOK	Overcast	Precipitation	Nighttime
Optical Camera/Sensor	○	×	×	×
SAR	○	○	○	○

Optical Data (OPT)
Visible Light
RGB 3-bands (or More)
Authentic Surface Optical Features
“Vivid / Detailed”



SAR Data (SAR)
Microwave
Multi-Polarization (e.g., VV, VH, HH, HV)
Surface Roughness and Material Properties
“Sparse / Abstracted”



1.2 SAR-to-Optical Image Translation Technology

SAR-to-Optical Image Translation (S2O):

SAR Image

"Highly Challenging Task"

S2O Framework

- **Generating Color**
- **Generating Detail**

"Fake" OPT Image

Current S2O Approaches:

Non-Deep Learning Approaches

- Image Enhancement/Filtering
- Histogram Matching
- Model-Driven

Deep Learning-Based Approaches

- Classic CNNs
- Classic GANs
- **Pix2Pix (SOTA)**

Practical applicability limited by accuracy challenges

1.3 Growing Repository of Large-Scale LULC Datasets

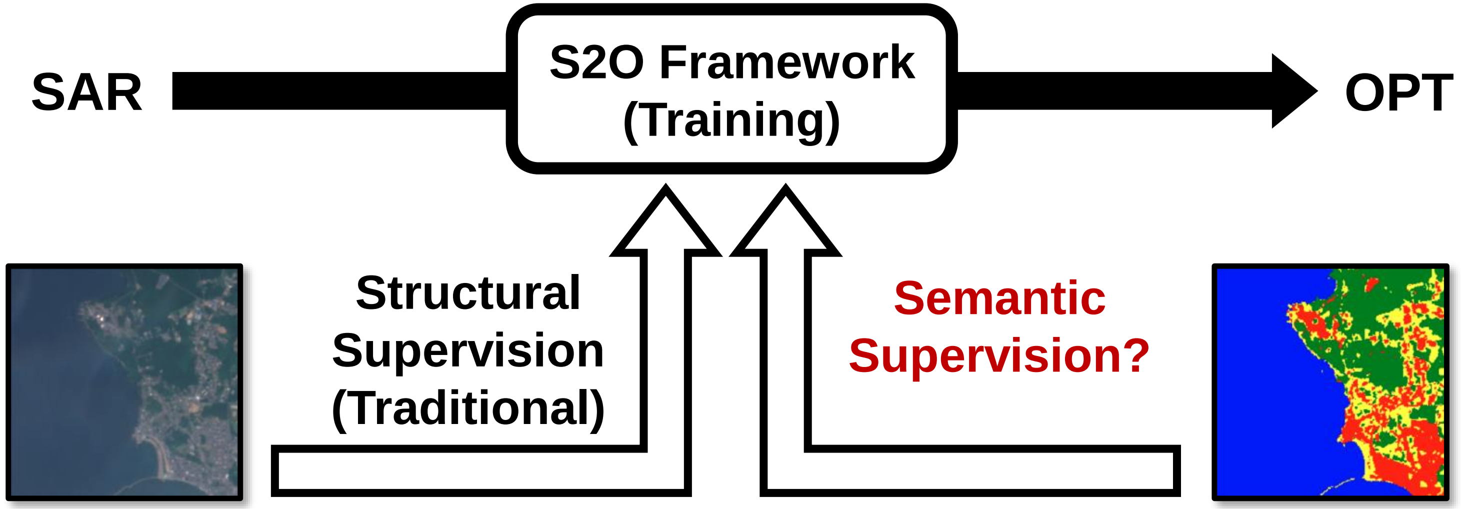
Recent Progress in Large-Scale LULC Datasets:

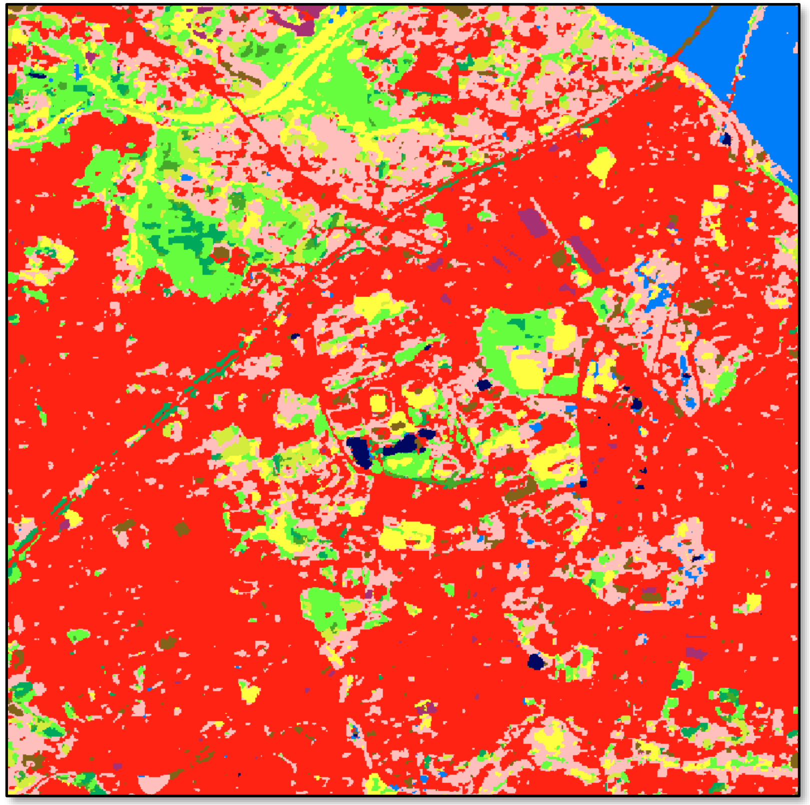
"Land-Use and Land-Cover (LULC)"

- Increasing Dataset Volume
- Higher Resolution (~10m)
- Expanding Coverage Area

Create a Large-Scale High-Resolution SAR-OPT-LULC Dataset?

Potential for Leveraging LULC Datasets in S2O:





High-Resolution LULC Map of the Vicinity of Kashiwanoha (JAXA)

1.4 Advanced S2O Techniques Using High-Resolution LULC Map

Main Objective:

Developing a High-Accuracy Framework for SAR-to-Optical Image Translation by Incorporating Land-Use and Land-Cover Maps During the Training Phase

Technological objectives:

Build a Large-Scale "SAR-OPT-LULC Map" Dataset

[1] Xuanchao Fu et al., 2024, IGARSS

Trinity of Objectives

Develop a Novel S2O Framework for Concurrent Structural and Semantic Supervision

[1] Xuanchao Fu et al., 2024, IGARSS

Develop a Novel Generative Model for Complex Feature Transformation in S2O Tasks

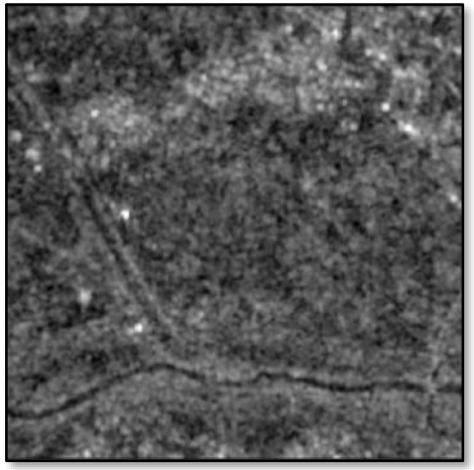
This Presentation

2. Data and Methods

2.1 Construction of SEN12JHL Dataset

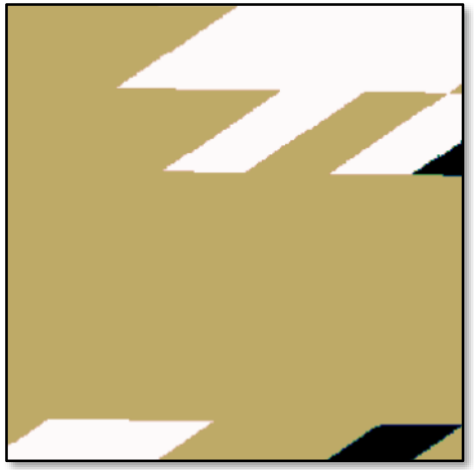
SEN12MS Dataset:

- 180662 triplets of Sentinel-1 (SAR-VV/VH) image patches, Sentinel-2 (OPT-R/G/B) image patches, and MODIS land cover maps **(11 categories at 500m resolution)**
- All images resampled to **10m resolution**
- Includes images from multiple countries worldwide **(Japan approx. 3%)**



SAR

OPT



MODIS

→ **Insufficient Resolution**

2. Data and Methods

2.1 Construction of SEN12JHL Dataset

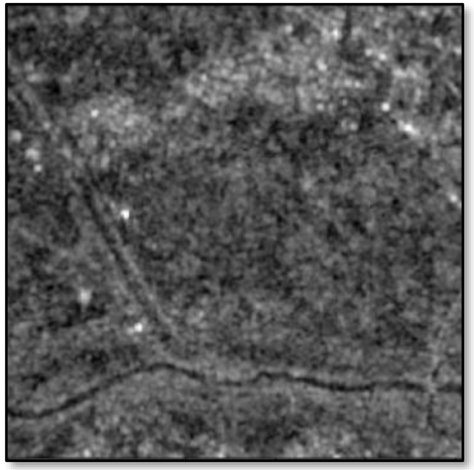
JAXA's High-Resolution Land-Use and Land-Cover Map (JHL):

- **10m** high-resolution LULC dataset **covering entire Japan**
- **12 land cover categories** based on Japanese context
- The **most accurate** LULC map of Japan to date
  (Derived from multiple satellite data sources)

#1: Water bodies
#2: Built-up
#3: Paddy field
#4: Cropland
#5: Grassland
#6: Forest (DBF)
#7: Forest (DNF)
#8: Forest (EBF)
#9: Forest (ENF)
#10: Bare
#11: Bamboo forest
#12: Solar panel

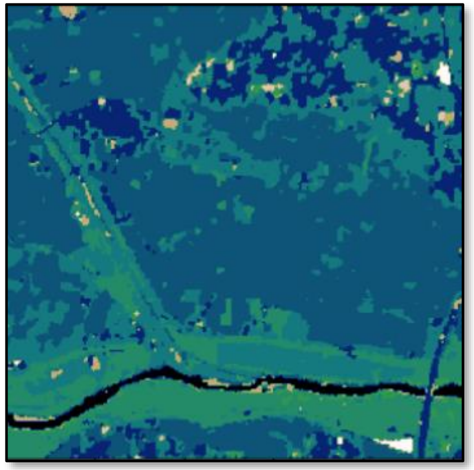
OA: 97.67%

[2] Sota Hirayama et al., 2022, RSSJ



SAR

OPT

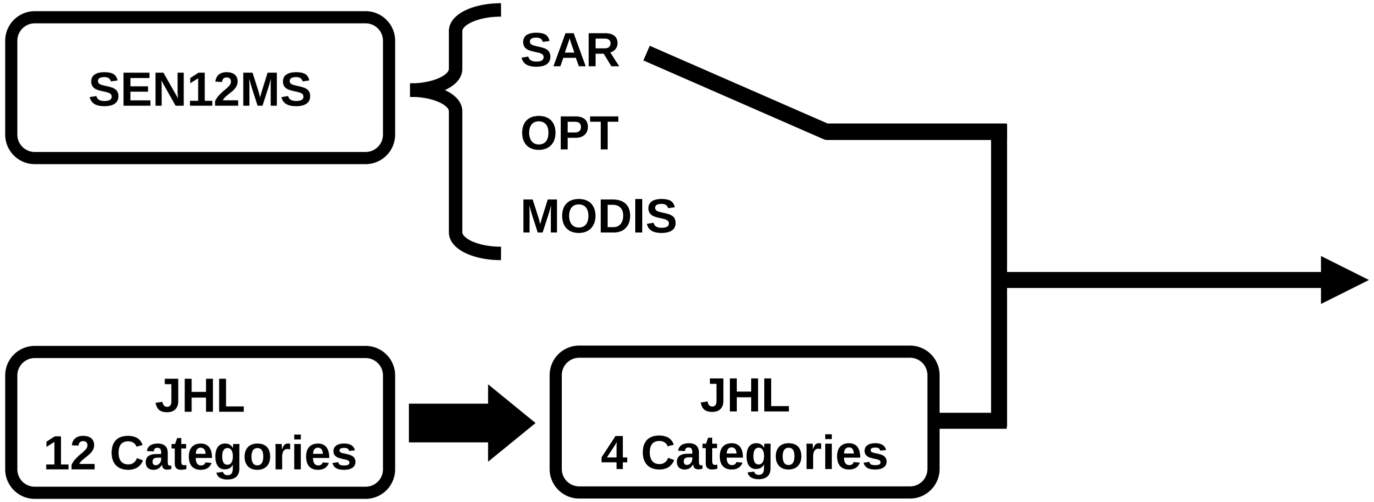


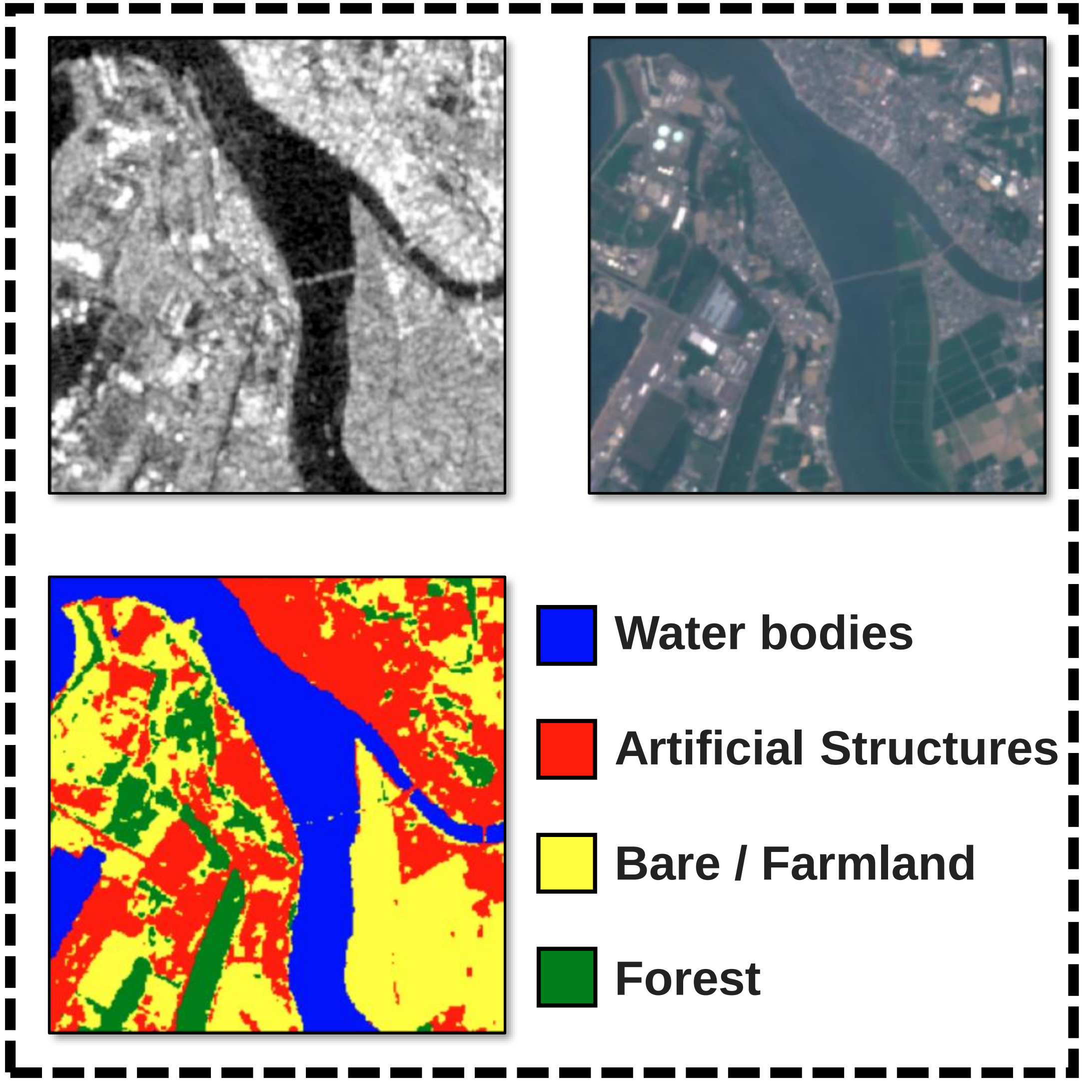
JHL
(12 Categories)

2. Data and Methods

2.1 Construction of SEN12JHL Dataset

Construction of SEN12JHL Dataset Based on SEN12MS and JHL:





- 5917 triplets of SAR, OPT, and LULC Maps
  (256x256 pixels/ 10m resolution / 4 categories)

2.1 Construction of SEN12JHL Dataset

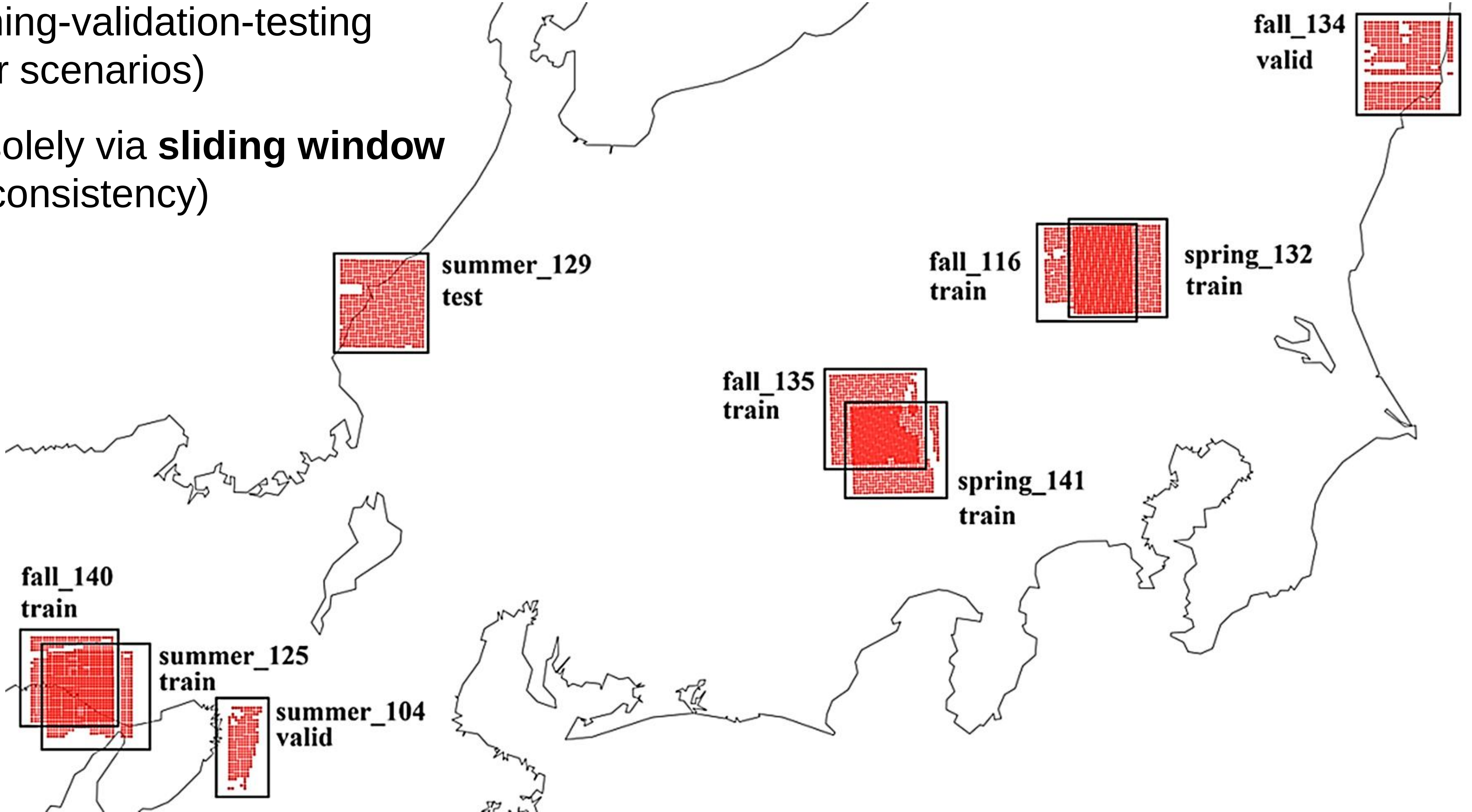
Partitioning of the Experimental Dataset:

- **Cross-regional** training-validation-testing (Emulating unfamiliar scenarios)
- Data augmentation solely via **sliding window** (Ensuring semantic consistency)
- Dataset Size

  Training Set
  - 4033

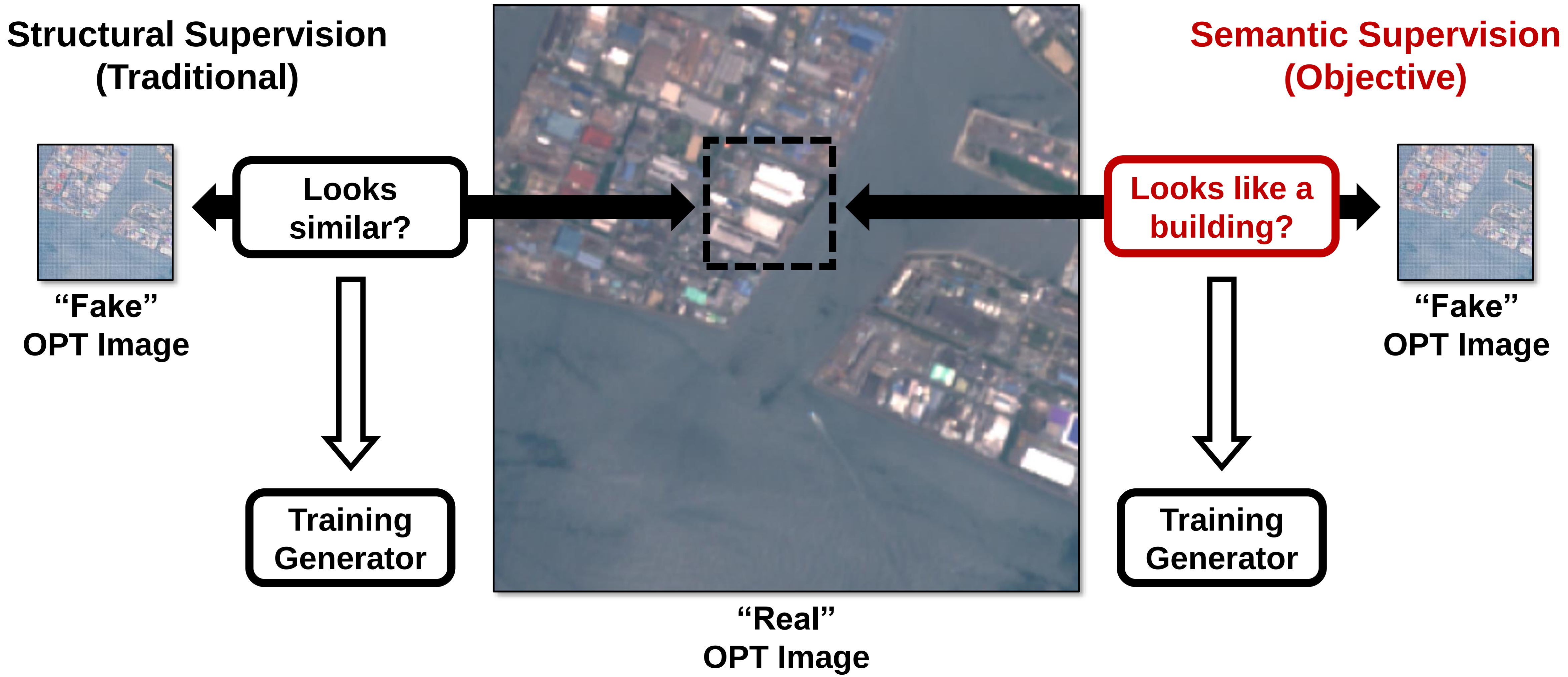
  Validation Set
  - 738

  Test Set
  - 589



2.2 S2O Framework Based on DSD Module

Supervising Generator Training with Semantic Information:



2.2 S2O Framework Based on DSD Module

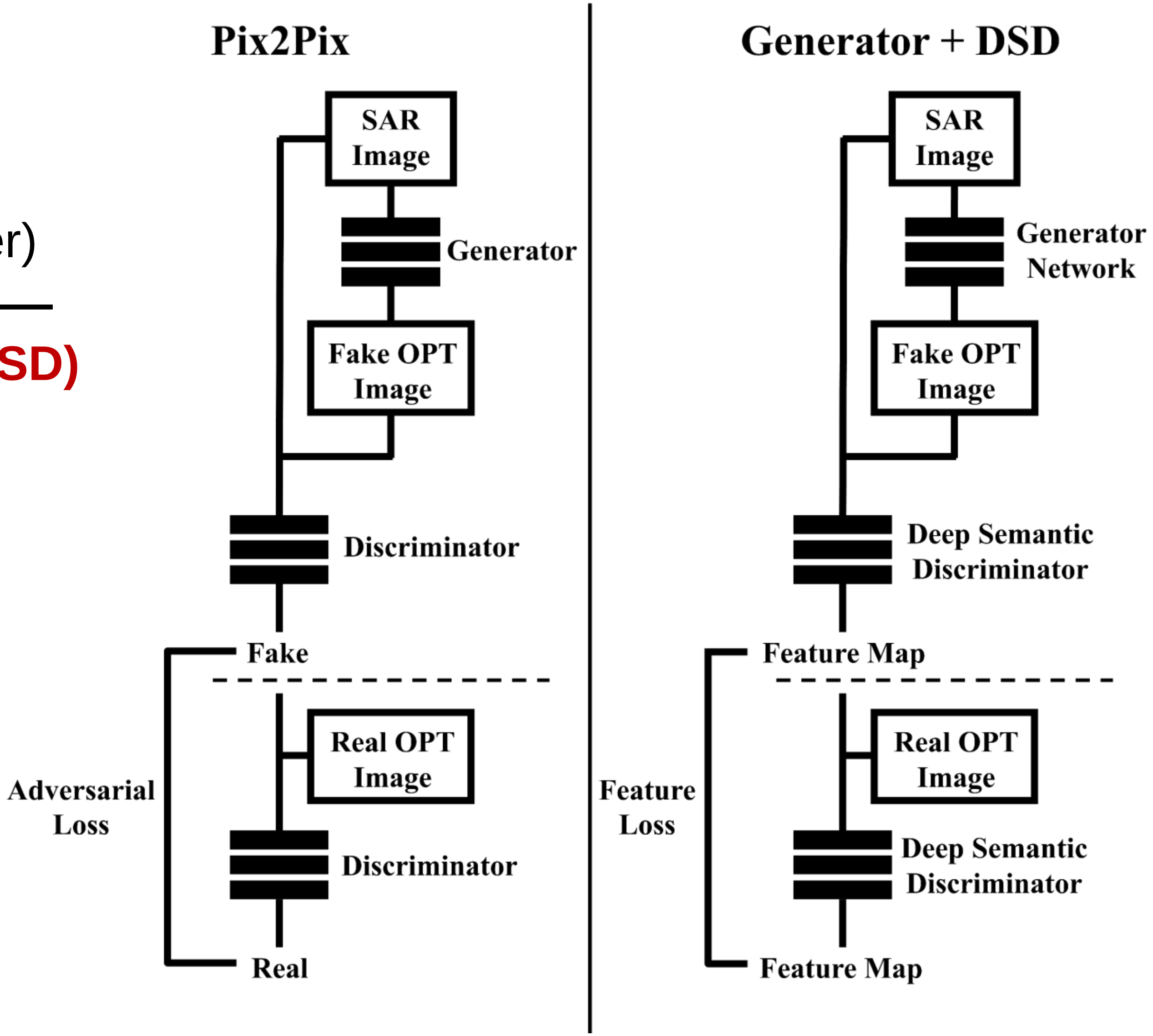
Structure of the Pix2Pix Framework:

- Generator - **U-Net**
- Discriminator - **PatchGAN** (Image patch classifier)

Our Proposal: Deep Semantic Discriminator (DSD)

- An arbitrary **semantic segmentation model** with **sufficient depth**
- Dual-modal architecture accommodating SAR and OPT (For **feature fusion**)
- **Feature Loss:**
  Calculate MSE Loss between Real-/Fake-OPT at the **deepest feature map** in the DSD

Structural Supervision + **Semantic Supervision**

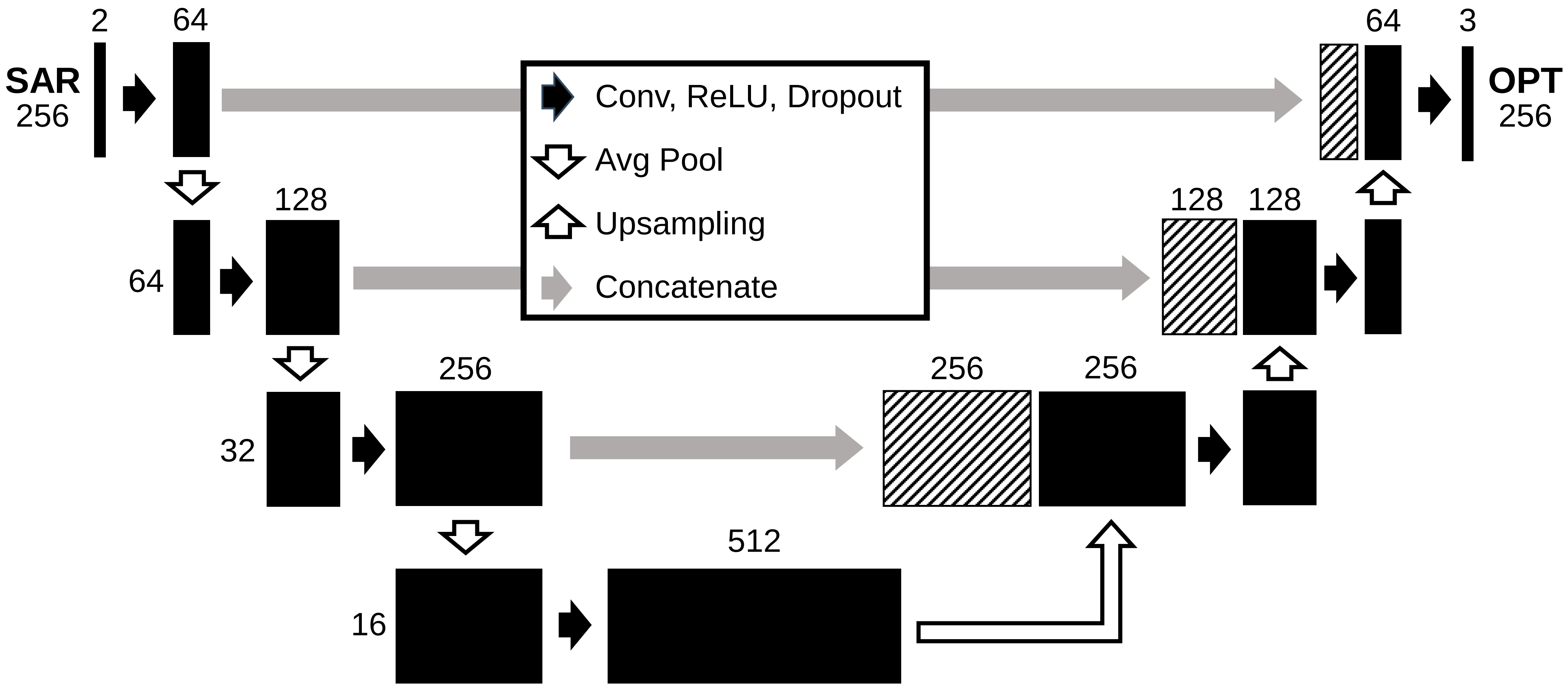


[1] Xuanchao Fu et al., 2024, IGARSS

2.3 S2O Former Based on Swin Transformer Architecture

Bottleneck of Pix2Pix Generator + DSD Framework :

- The Pix2Pix generator is **fully CNN-based**, making **complex feature transformations** challenging



[3] P. Isola et al., 2017, CVPR

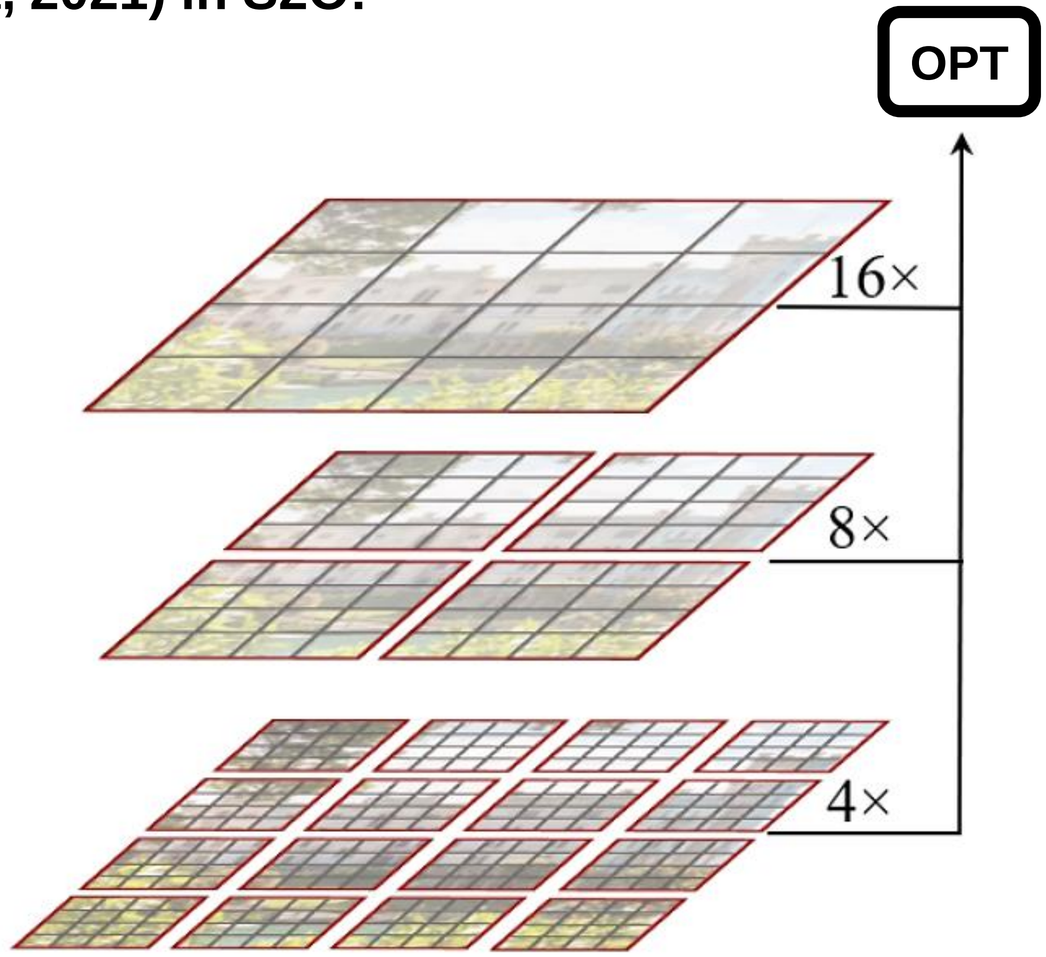
2.3 S2O Former Based on Swin-Transformer Architecture

Potential of Swin-Transformer Architecture (Microsoft, 2021) in S2O:

- **Local Window Self-Attention:**
  Employs **self-attention mechanism** to capture global dependencies, thus enhancing understanding and transformation of **scenes in complex SAR imagery**

- **Hierarchical Feature Maps:**
  Enables efficient processing of high-resolution images, essential for **preserving details in SAR** imagery during the S2O process

- **Adaptability:**
  Adapts to varying scales of image content, automatically adjusting its focus to meticulously process **multi-scale features in S2O**

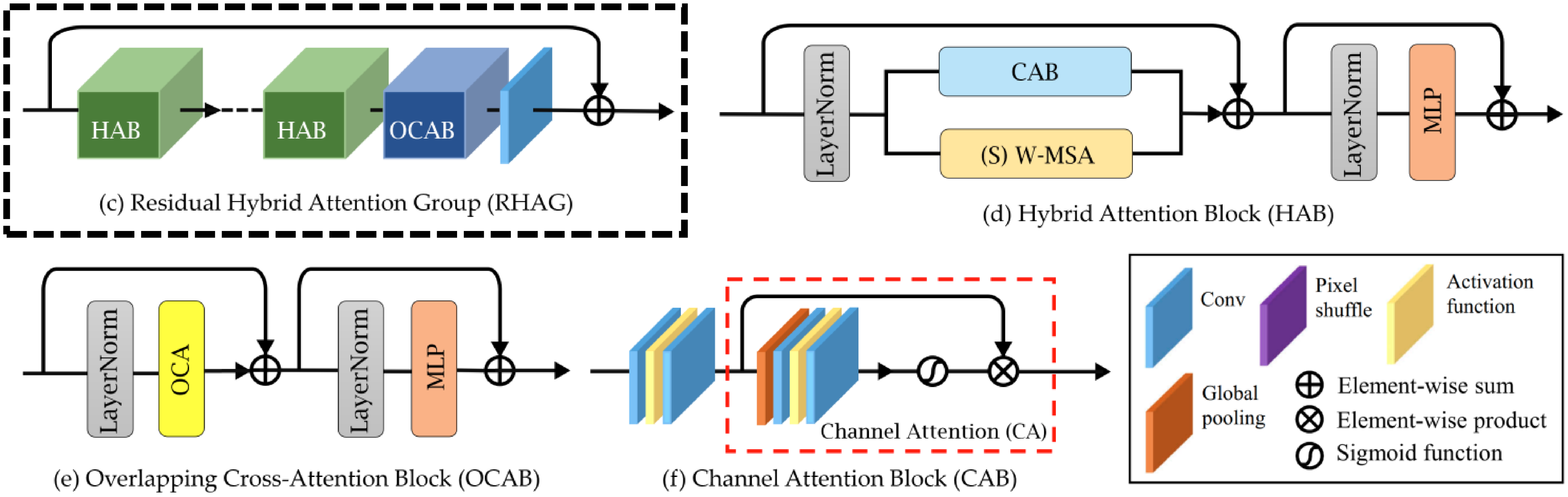


[4] Z. Liu et al., 2021, ICCV

2.3 S2O Former Based on Swin Transformer Architecture

SOTA Swin Transformer Module - RHAG (XPixel, 2022):

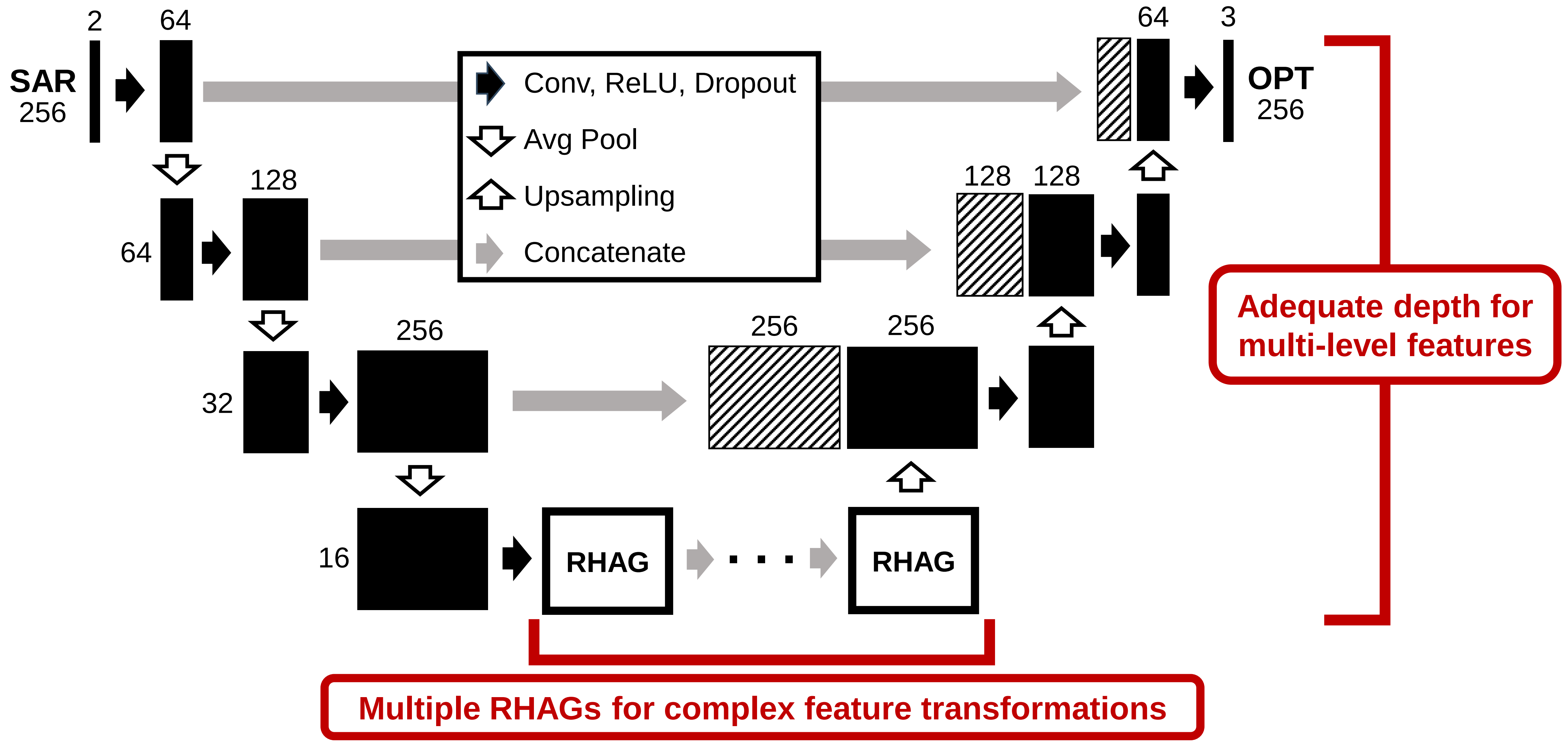
- **Residual Hybrid Attention Groups (RHAG),** proposed as the core module of the SOTA Super-Resolution model **HAT**



[5] X. Chen et al., 2023, CVPR

2.3 S2O Former Based on Swin Transformer Architecture

Our Proposal: High-Performance Generator Specialized for S2O Task - S2O Former



3.1 Experimental Setup and Parameters

Experimental Setup:

	Pix2Pix-DSD	HAT-DSD	S2OFormer-DSD (Ours)
Generator	**U-Net**	**HAT[1]**	**S2OFormer[1]**
Discriminator	**DSD[2]**	**DSD[2]**	**DSD[2]**

1*: RHAG parameters in the HAT: depth = [4,4,4,4], num heads = [4,4,4,4], window size = 8
2*: SNUNet-CD

Experimental Setup:

- PSNR (Peak Signal-to-Noise Ratio)
- SSIM (Structural Similarity Index Measure)
- Visual quality

Training Parameters:

- Batch size/Epochs: **16/128**
- Optimizer/Learning Rate: **Adam/0.0003**
- Equipment: **NVIDIA A100**

3.2 Experimental Results

Experimental Results (PSNR/SSIM):

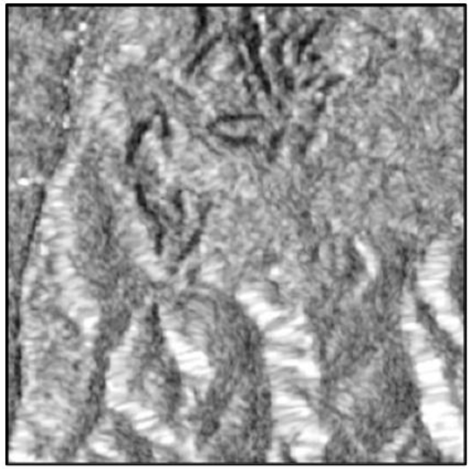
	U-Net-DSD	HAT-DSD	S2OFormer-DSD (Ours)
PSNR	19.50	19.92	**20.14**
SSIM	0.565	0.575	**0.642**

*Only generator is subjected to testing, with discriminator not being involved in the inference phase

3.3 Experimental Results

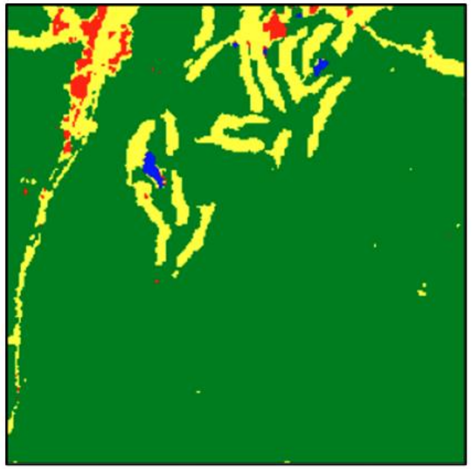
(LULC is NOT used in inference phase)

Land Region (Kaga, Ishikawa)

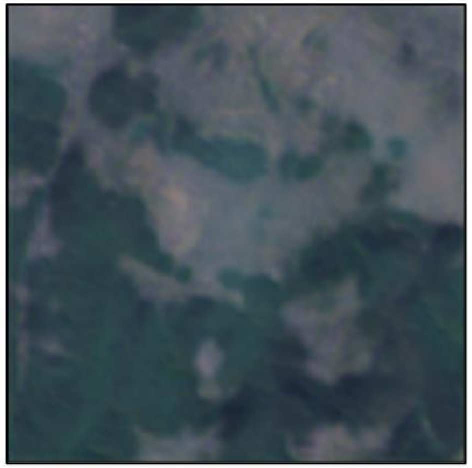


SAR

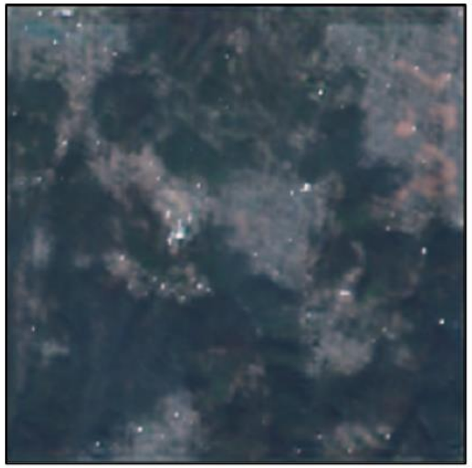
OPT



LULC



U-Net



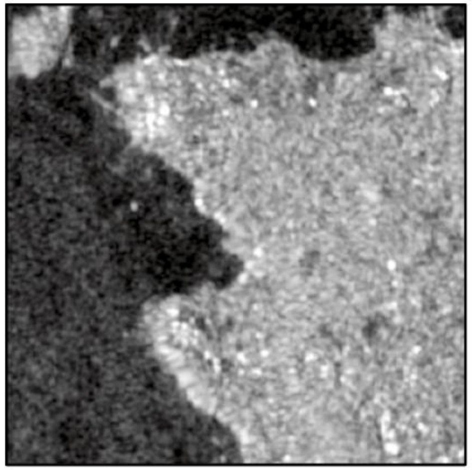
HAT

S2OFormer (ours)

3.3 Experimental Results

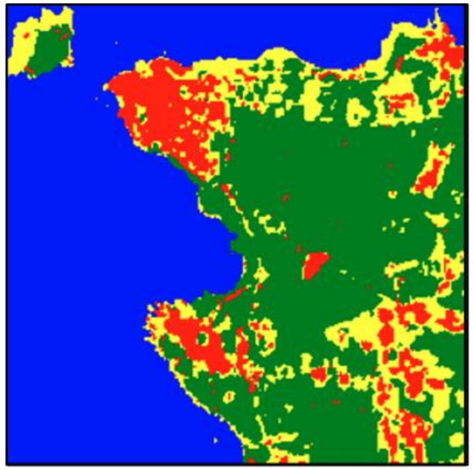
(LULC is NOT used in inference phase)

Coastal Region (Mikuni, Fukui)

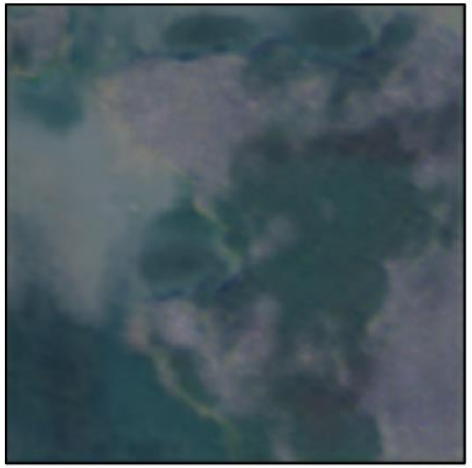


SAR

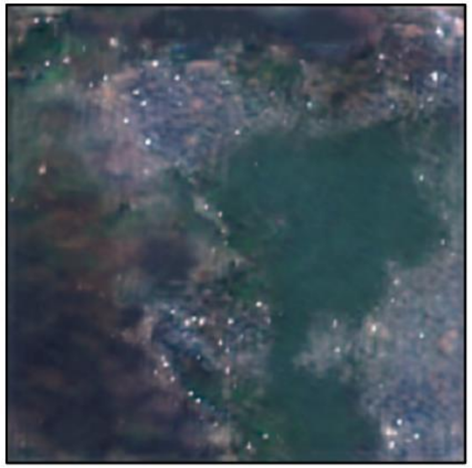
OPT



LULC



U-Net



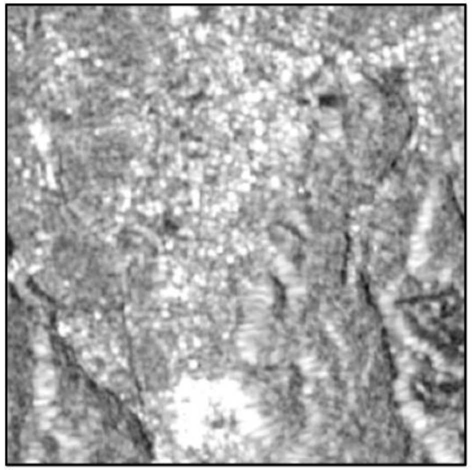
HAT

S2OFormer (ours)

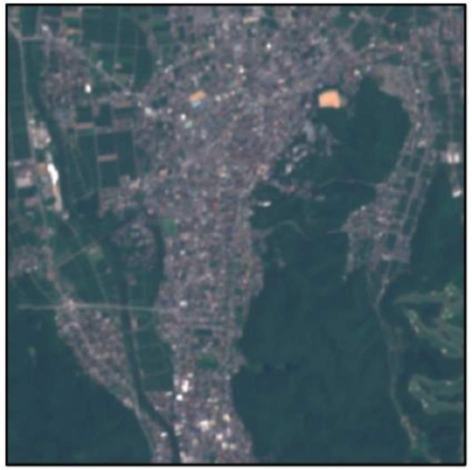
3.3 Experimental Results

(LULC is NOT used in inference phase)

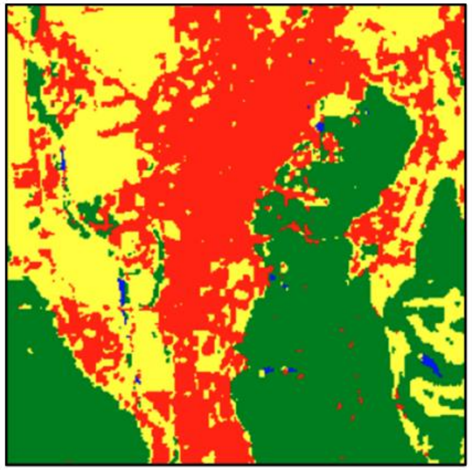
Urban Area (Kaga, Ishikawa)



SAR



OPT



LULC

U-Net

HAT

S2OFormer (ours)

4. Conclusions and Future Work

4.1 Conclusions

- S2OFormer **outperforms traditional CNN and baseline Transformer models in** SAR-to-optical translation, achieving superior **PSNR (20.14) and SSIM (0.642) metrics**

- S2OFormer delivers **superior visual results across various terrains**, **effectively reducing noise and enhancing feature clarity** compared to U-Net and HAT models

- The integration of Swin-Transformer-based RHAG modules in S2OFormer demonstrates promise for **extended applications**, including performance **under complex conditions and real-time processing**

4.2 Future Work

Developing Framework Specialized for S2O Tasks:

- We are pioneering a framework that **transcends traditional GAN structures**, creating a novel approach based on **Diffusion and Foundation techniques** to more effectively utilize LULC Maps
- Our novel framework also attempts to integrate the LULC Maps as an **essential intermediary step** for **fully semantic guidance** of the S2O process

Enhanced SAR-OPT-LULC Dataset with Expanded Coverage / Higher Resolution:

- Incorporate data from **additional countries / regions**
- Attempted to create datasets utilizing newly emerged **ultra-high-resolution** SAR data

5. References

[1] **<u>Fu, X</u>.**, Kouyama, T., Seki, S., Nakamura, R., & Yoshikawa, I. (2024). Advanced SAR-to-Optical Image Translation Techniques Using JAXA's High-Resolution Land-Use and Land-Cover Map. In IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. IEEE. （受理済み）

[2] 平山颯太, 田殿武雄, 大木真人, 水上陽誠, 奈佐原(西田)顕郎, 今村功一, ... & 山之口勤. (2022). JAXA 高解像度土地利用土地被覆図日本域 21.(HRLULC-Japan v21. 11) の作成. 日本リモートセンシング学会誌, 42(3), 199-216.

[3] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134). IEEE.

[4] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).

[5] Chen, X., Wang, X., Zhou, J., Qiao, Y., & Dong, C. (2023). Activating more pixels in image super-resolution transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 22367-22377).