NRSCC

esa

HY HJ-1AB CBERS Gaofen Beijing-2 Sentinel-1 Sentinel-2 Sentinel-3 Sentinel-5p Aeolus

2023 DRAGON 5 SYMPOSIUM

Large-Scale Satellite Image Time Series： Learning, Analysis and Application

ID. 58190

Weiwei Guo and Daniela Faur

Sept. 2023

ID.58190

NRSCC

Our team

esa

同濟大學
TONGJI UNIVERSITY


Weiwei Guo

UNIVERSITATEA POLITEHNICA DIN BUCUREŞTI 1818


Prof. Daniela Faur


Prof. Mihai Datcu

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY


Prof. Zenghui Zhang

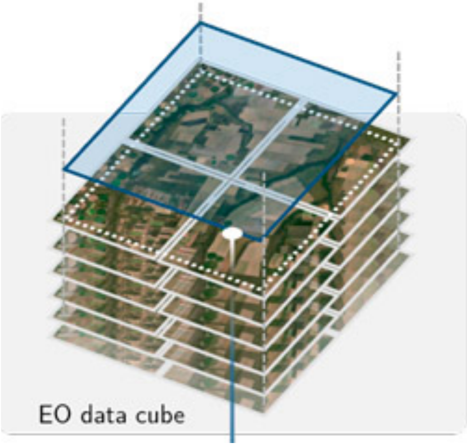

Prof. Wenxian Yu


PhD. Limeng Zhang

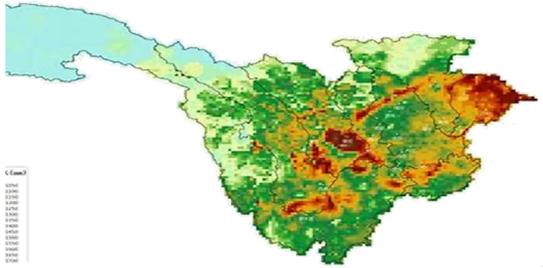

MS. Yan Li

Big EO Data

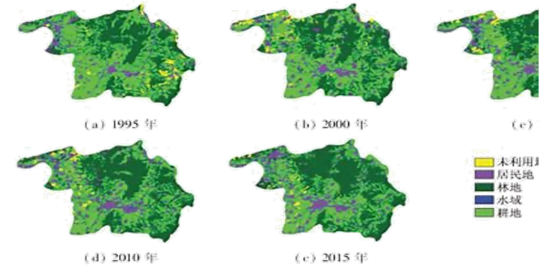
- The Earth is facing unprecedented challenges: environmental, climatic, urbanization, etc.



- Earth Observation data with a broad variety of satellite sensors provide invaluable information to understand our state and changes of large areas and even long periods.



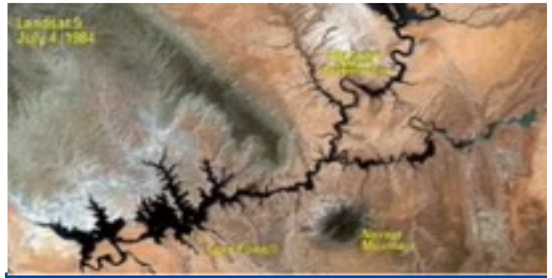
disaster assessment



land use

- The increasing volume of EO data pose challenges to explore these data

urban construction

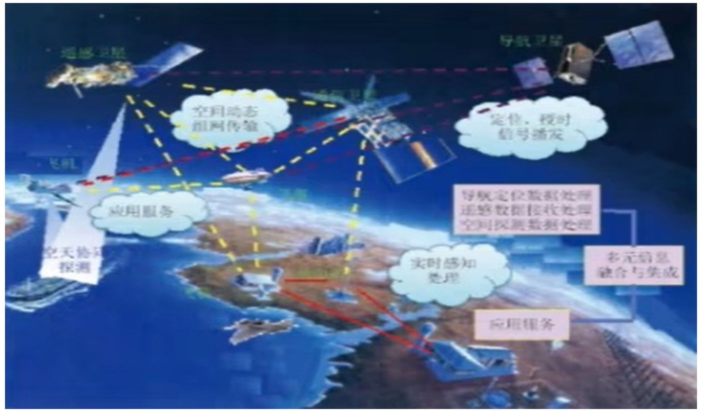


resource exploration

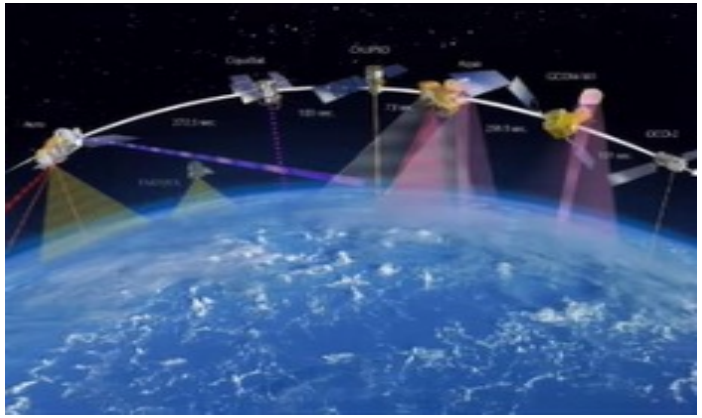
Big EO Data

■ The increasing volume of EO data pose challenges to explore these data

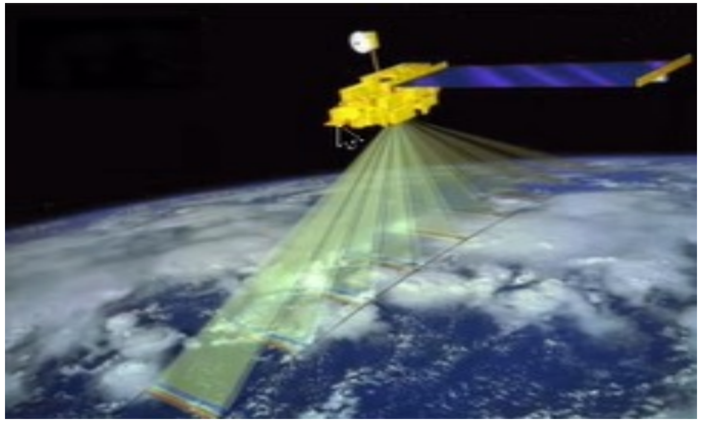
- Big data, but unlabeled
- Multi-modal, heterogeneous data



Multi-platform



Multi-sensor



Multi-perspective

Recent Work 1：Joint SAR and Optical Representation Learning with Vertical Masking

- Mask AutoEncoder (MAE) is a powerful self-supervised learning methods that is designed for nature images.
- But it is suboptimal for multi-modal remote sensing images due to the great domain gap between multi-modal data, such as optical and SAR
- We explore various mask patterns based on MAE and design a novel self-supervised multi-modality pre-trained model with vertical masking to extract complementary information between modals.

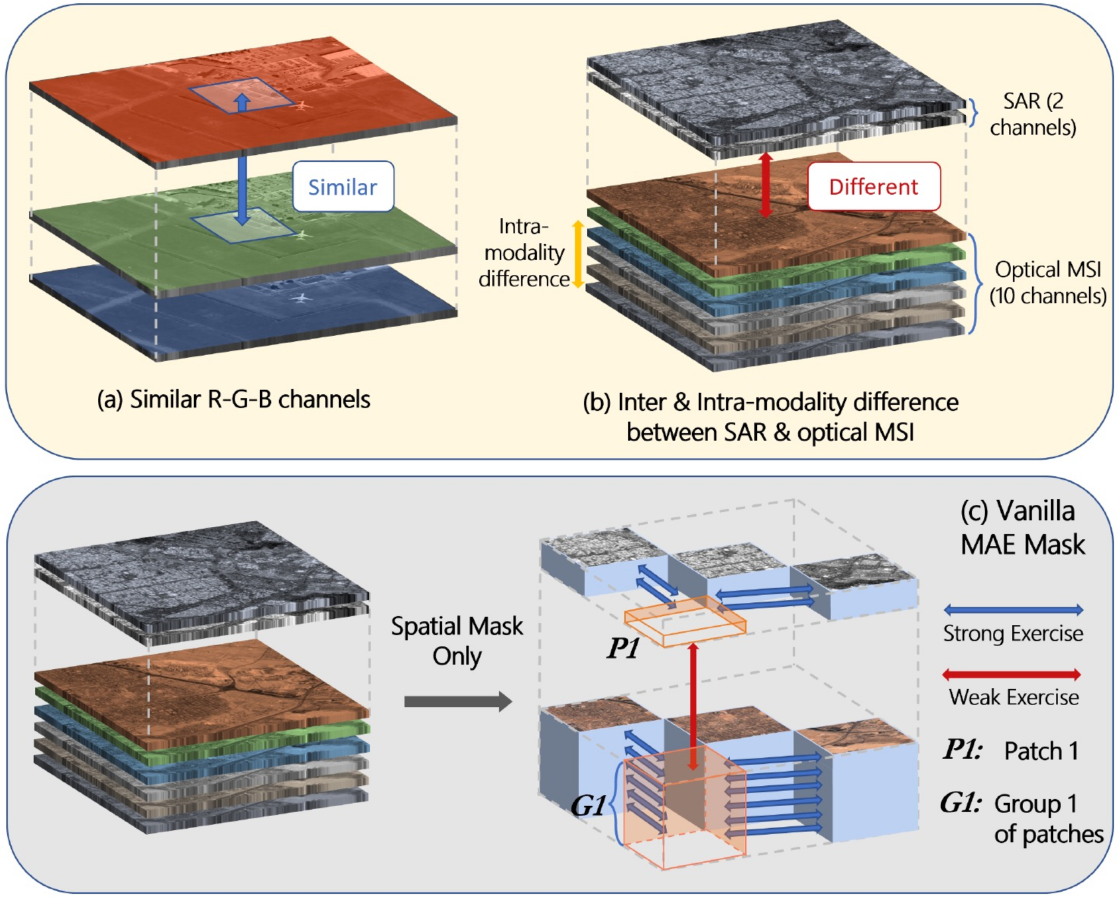
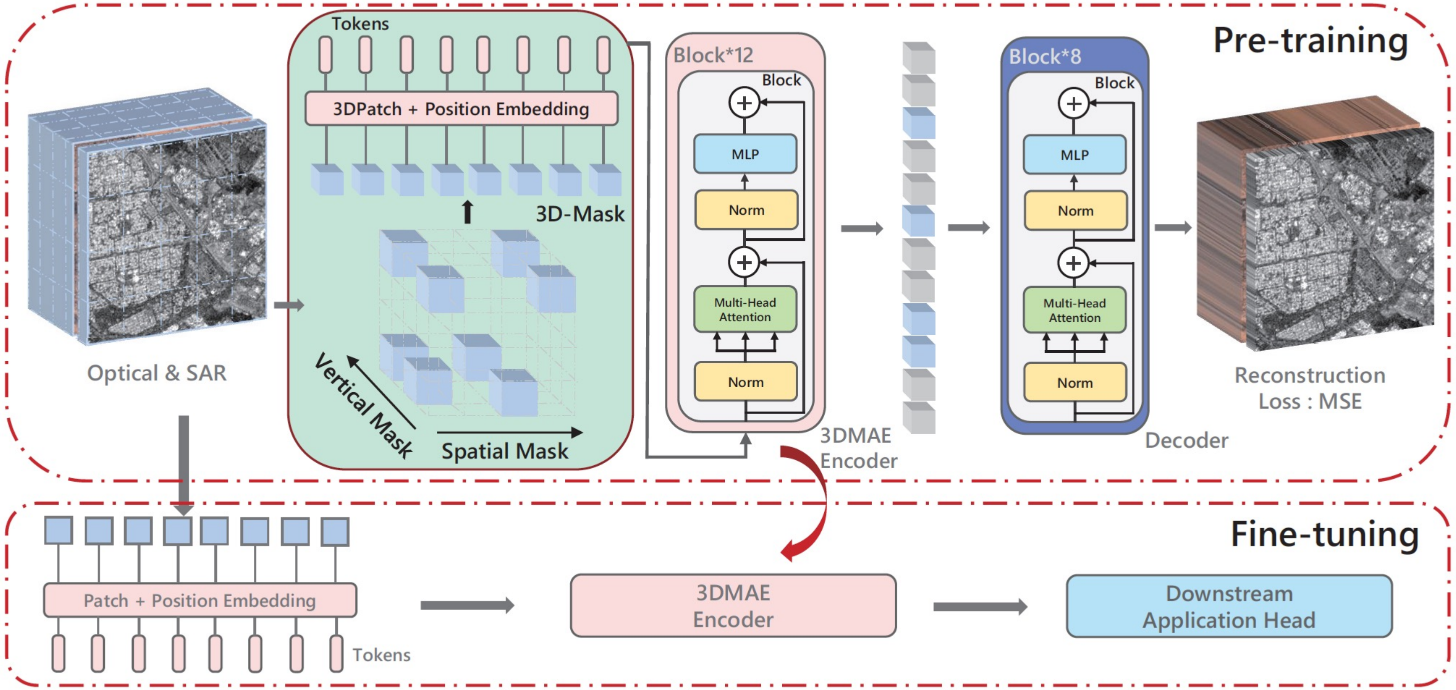


Fig. 1. While the RGB channels of natural images tend to exhibit similar information, there often exist significant disparities between RS images of diverse modalities, such as SAR and optical data.

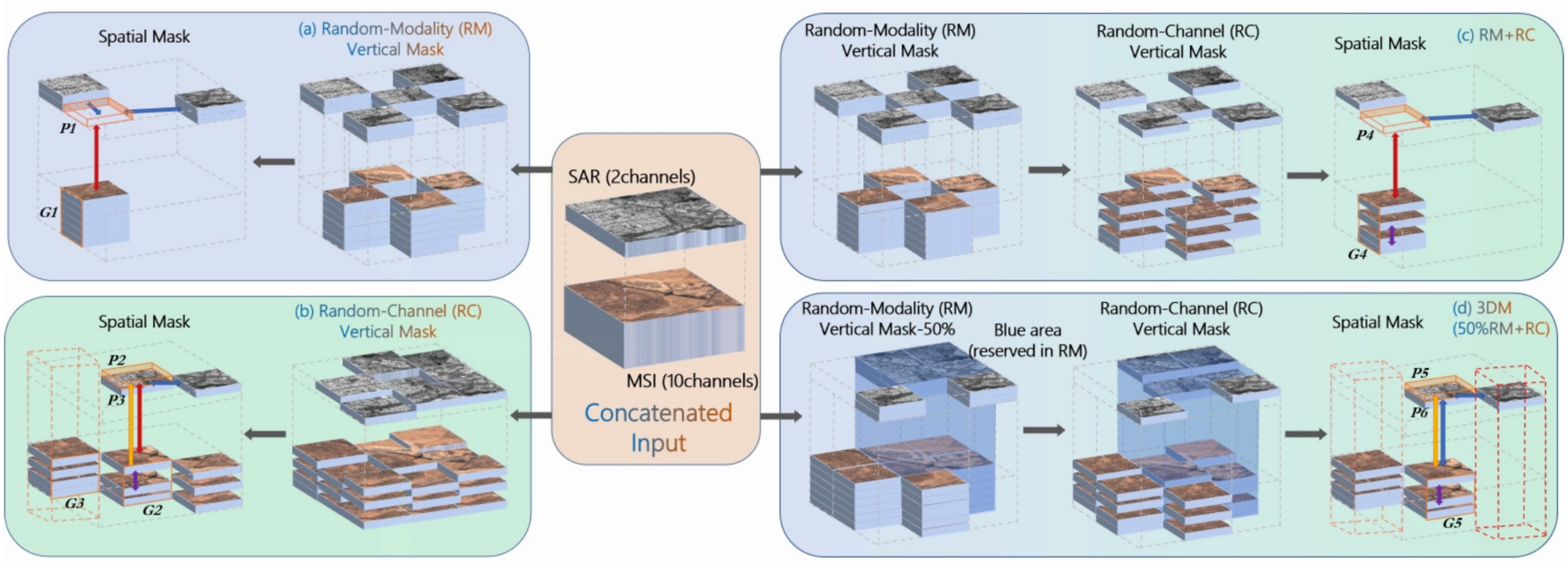
Recent Work 1： Joint SAR and Optical Representation Learning with Vertical Masking

- We propose a 3D-MAE self-supervised learning method that pre-trains on jointly SAR and optical data
- By vertical and spatial masking , the model not only captures the spatial correlation information but channel.
- When finetuning, the pre-trained 3DMAE encoder is utilized to encode the entire image, extracting features that can be applied to various downstream applications.



Recent Work 1: Joint SAR and Optical Representation Learning with Vertical Masking

Different Masking Strategies



For every spatial position->	Only SAR	Only optical	Only partial SAR	Only partial optical	Partial SAR & optical
(a) Random-Modality (RM)	✓	✓	✕	✕	✕
(b) Random-Channel (RC)	✕	✕	✕	✓	✓
(c) RM+RC	✕	✕	✓	✓	✕
(d) 3DM (50%RM+RC, ours)	✕	✕	✓	✓	✓

P1 Short for patch 1
G1 Short for patches group 1
Represent the SAR patch
Represents the optical patch
Strong exercise
Hard exercise (domain gap)
Inter-modality assistance
Intra-modality exercise

Fig. 3. 3D mask exploration. (a) Random-Modality (RM). (b) Random-Channel (RC). (c) Directly RM+RC. (d) 3DM. The input is the concatenation of 2 channels SAR and 10 channels MSI of the same location in channel dimension. The MSI in the figure only shows 6 channels for the convenience of display. The table below gives the various circumstances that may be encountered after the mask.

Recent Work 1： Joint SAR and Optical Representation Learning with Vertical Masking

Experimental Results

■ Multi-label Classification

3DMAE-3DM exhibits superior multi-modality learning capabilities, resulting in not only better performance for S1+S2 compared to S2 only, but also overall higher performance than SatViT. The relatively slight improvement of S1+S2 in comparison to S2 can be attributed to the already elevated performance level of the model, rendering athe attainment of additional improvements challenging.

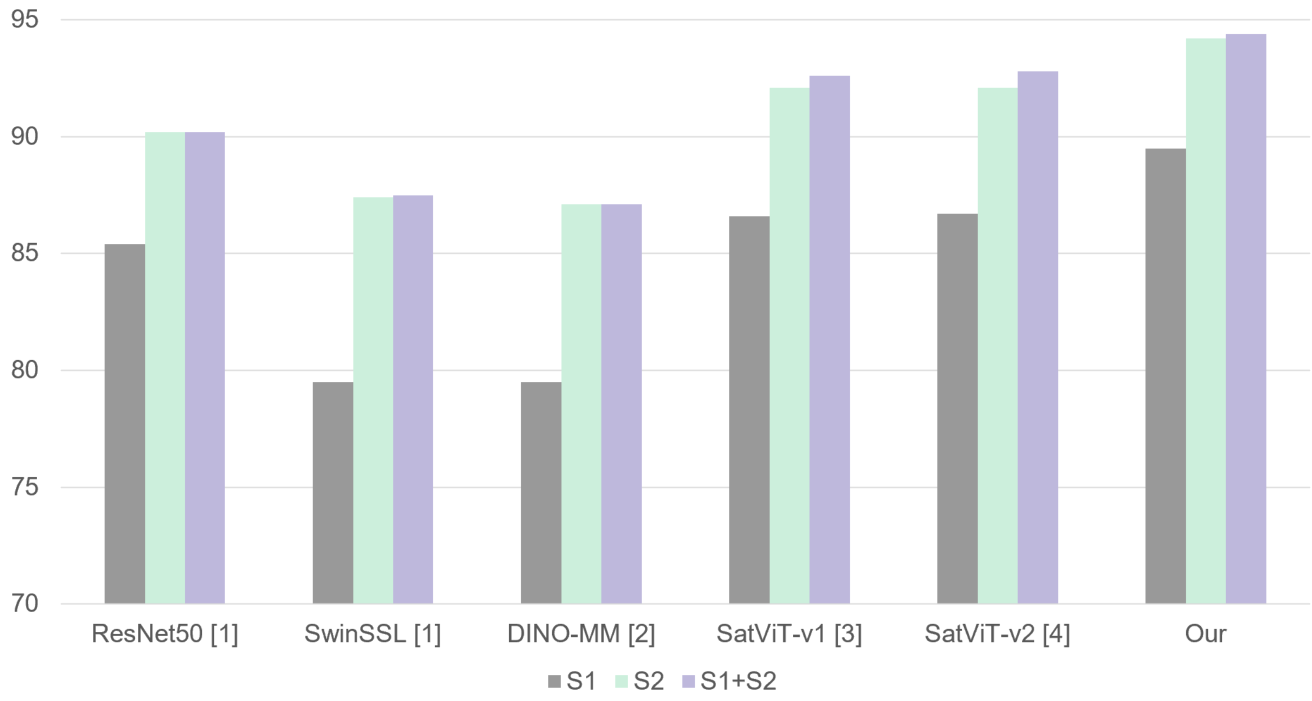


TABLE I

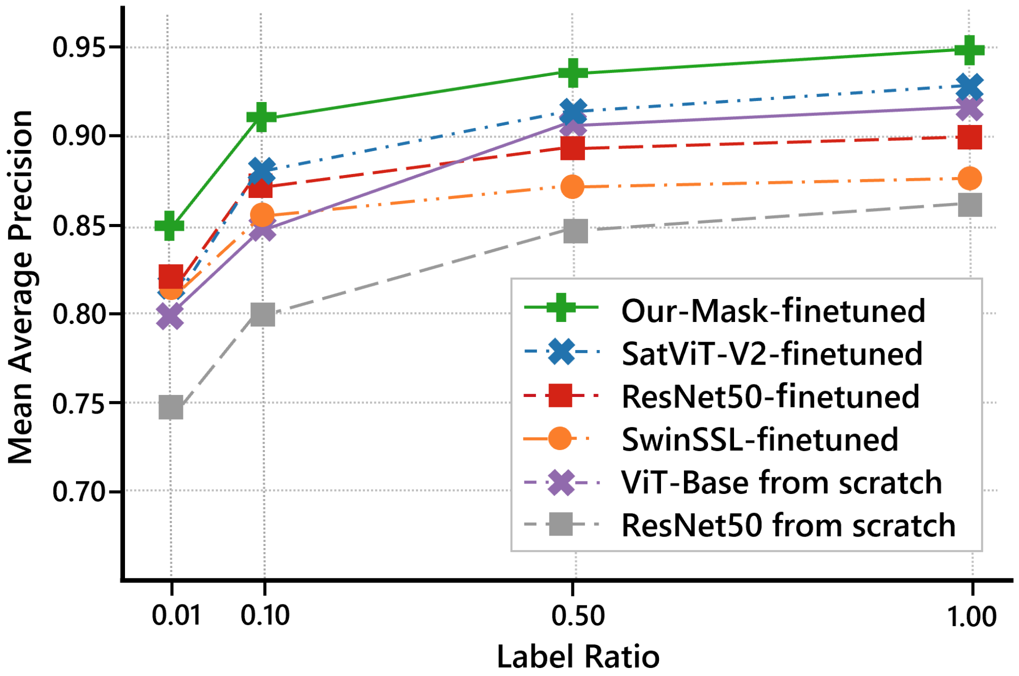
FINE-TUNED RESULTS ON BIGEARTHNET-MM VALIDATION SET. THE EVALUATION METRIC IS MAP. S1 REPRESENTS SAR, S2 REPRESENTS OPTICAL IMAGE, AND + REPRESENT THE CONCATENATION.

Model	S1	S2	S1+S2
ResNet50 [2]	85.4	90.2	90.2
SwinSSL [2]	79.5	87.4	87.5
DINO-MM [3]	79.5	87.1	87.1
SatViT-v1 [5]	86.6	92.1	92.6
SatViT-v2 [6]	86.7	92.1	92.8
3DMAE-RM	87.5	93.7	93.7
3DMAE-RC	89.2	94.0	94.1
3DMAE-RM+RC	88.3	93.9	93.9
3DMAE-3DM (ours)	**89.5**	**94.2**	**94.4**

Experimental Results

■ Small-scale Data Study

- 3DMAE outperforms other self-supervised pre-trained models when applied to small-scale data.
- Furthermore, it shows significantly better performance compared to models that underwent supervised training from scratch on small datasets of BigEarthNet-MM.



Recent Work 1: Joint SAR and Optical Representation Learning with Vertical Masking

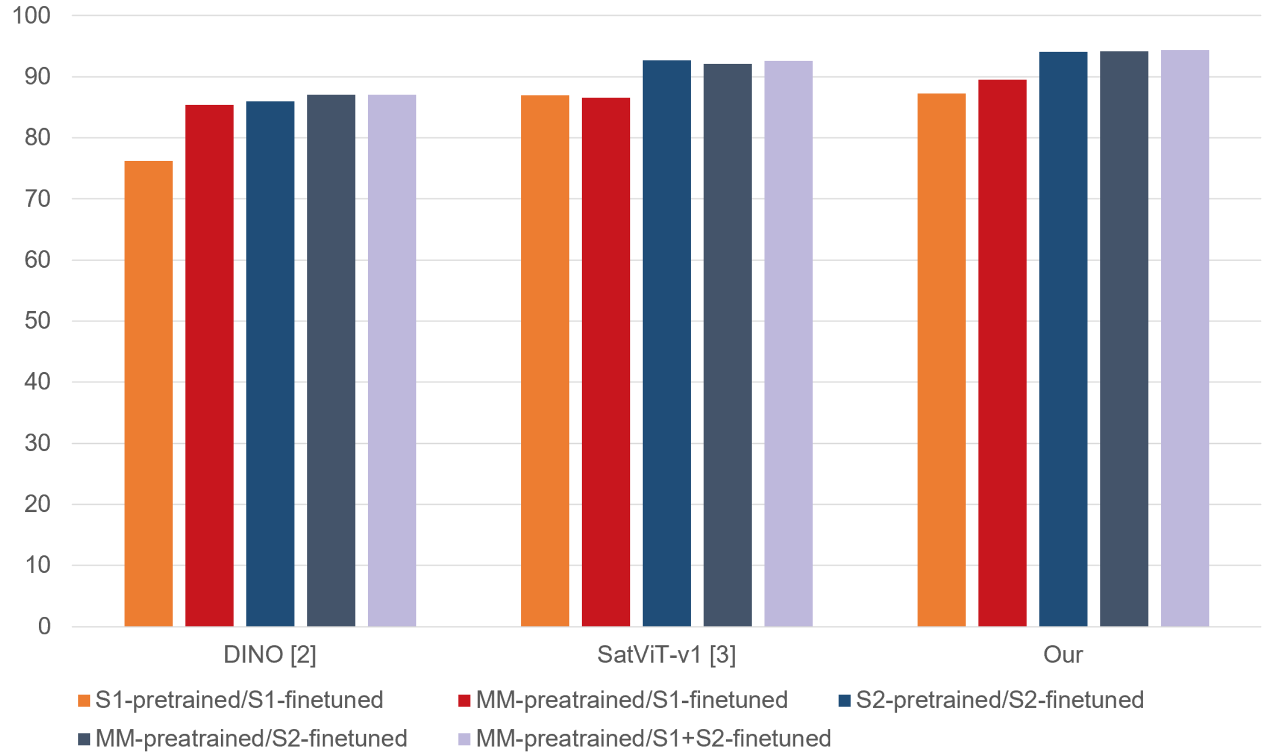
Experimental Results

■ Data Ablation Study

The data for the model pre-trained and fine-tuned varies.

TABLE II
FINE-TUNED RESULTS ON THE SINGLE-MODALITY BIGEARTHNET-MM VALIDATION SET USING MAP AS THE EVALUATION METRIC. S1 REPRESENTS SAR, S2 REPRESENTS THE OPTICAL IMAGE.

Model	S1	S2	S1+S2
DINO-S1/S2 [3]	76.2	86.0	-
MoCo-v2-S1/S2 [1]	82.8	89.3	-
SatViT-v1-S1/S2 [5]	87.0	92.7	-
3DMAE-3DM-S1/S2	**87.3**	**94.1**	-
DINO-MM [3]	79.5	87.1	87.1
SatViT-v1-MM [5]	86.6	92.1	92.6
3DMAE-3DM-MM	**89.5**	**94.2**	**94.4**



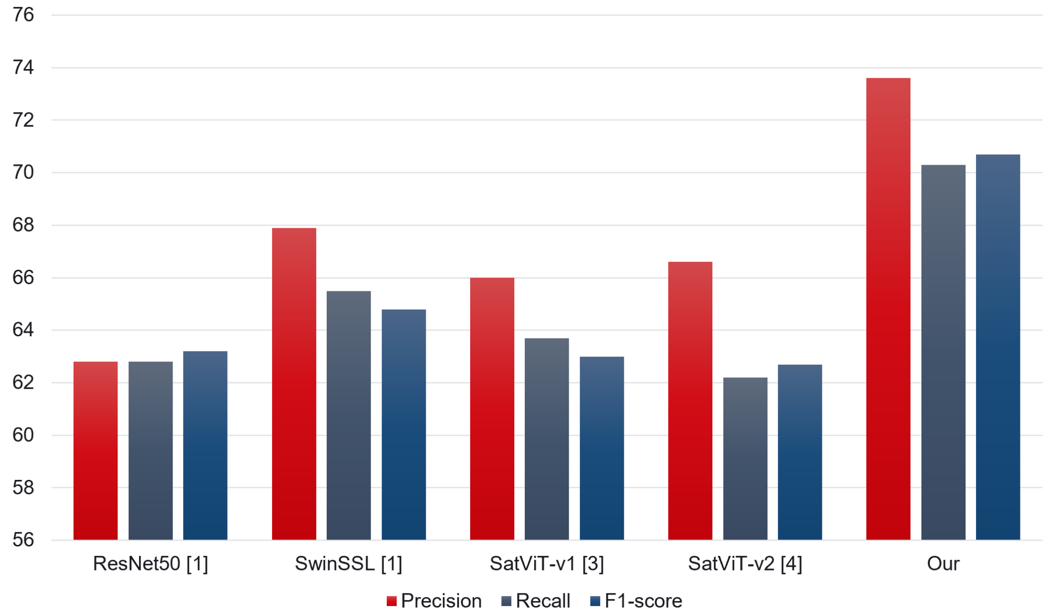
Experimental Results

■ Generalization Study

The model was pretrained on the BigEarthNet dataset, and fine-tuned on SEN12MS dataset, which consists merely of European landscapes, leading to a significant difference in data distribution between the two datasets.

TABLE III
FINE-TUNED RESULTS ON SEN12MS VALIDATION SET.

Model	Precision	Recall	F1-score	Pre-trained Dataset
ResNet50 [2]	62.8	62.8	63.2	SEN12MS(Global)
SwinSSL [2]	67.9	65.5	64.8	SEN12MS(Global)
SatViT-v1 [5]	66.0	63.7	63.0	130MILLION(Global)
SatViT-v2 [6]	66.6	62.2	62.7	130MILLION(Global)
3DMAE-3DM	**73.6**	**70.3**	**70.7**	BigEarthNet-MM(European)



Conclusions and Discussion:

- The proposed method enhances the model’s performance in multi-modality situations by effectively exploiting complementary information, achieving state-of-the-art (SOTA) performance..

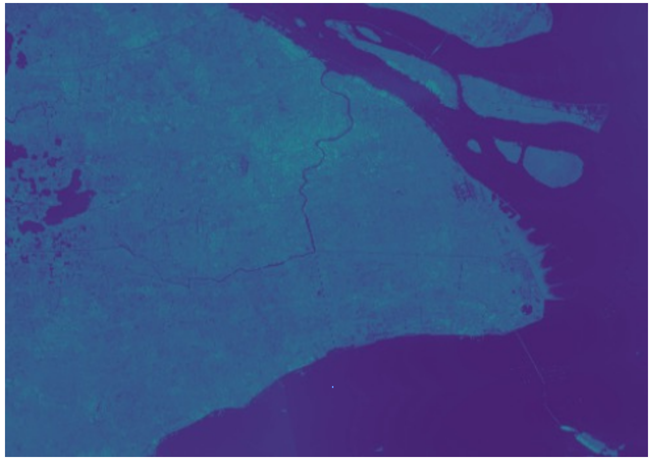
- The proposed model can handle a variety of downstream applications in both single and multi-modality scenario. Substantially, it enhances performance in single-modality situations where only SAR images are available, such as in emergency scenarios.

- However, due to the limitation of computing resource overhead, there is still room for further exploration and improvement of mask ratio.

Recent Work 2： Spatial-Temporal Masked Image Modeling for Satellite Image Time Series

- EO data not only attain the spatial and spectral information but also contain the temporal dimension, exhibiting a big 4D tensor.

- We intend to explore the temporal information for the self-supervised learning on image time series.



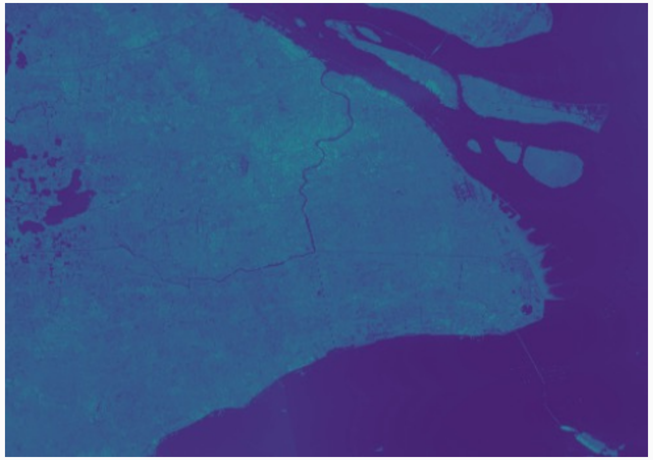
Sentinel-1 SAR
8X (2015~2022)

Sentinel-2 RGB
6X (2017~2022)

Recent Work 2： Spatial-Temporal Masked Image Modeling for Satellite Image Time Series

Data Preparation

data

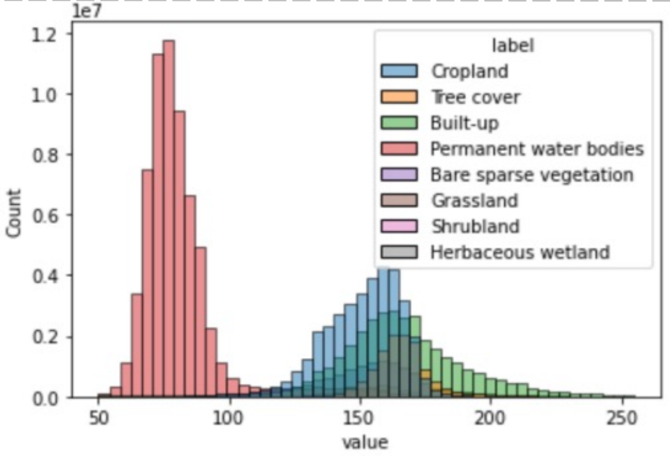


Sentinel-1 SAR
8X (2015~2022)

Sentinel-2 RGB
6X (2017~2022)

label

Ground truth (2020)



Category Statistics

Preprocessing

missing value	outlier
data type	split

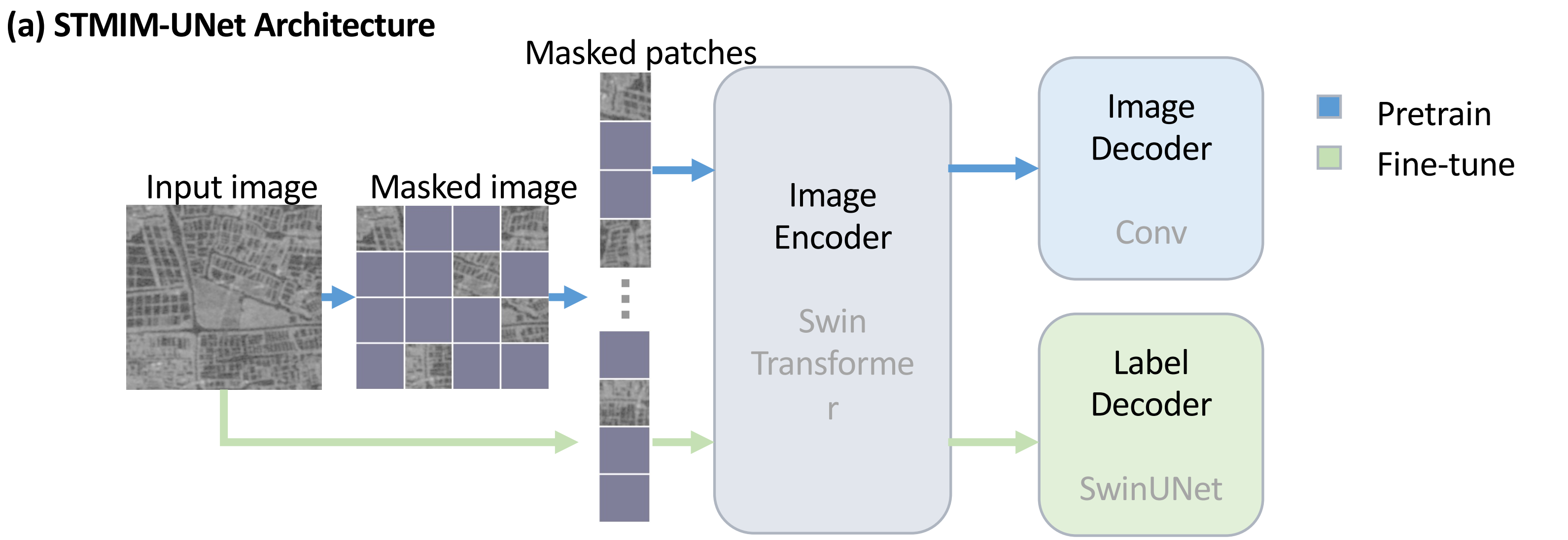
$$x = \begin{cases} 0 & x < \mu - 3\sigma \\ \frac{x-(\mu-3\sigma)}{6\sigma} \times 255 & \mu - 3\sigma < x < \mu + 3\sigma \\ 255 & x > \mu + 3\sigma \end{cases}$$

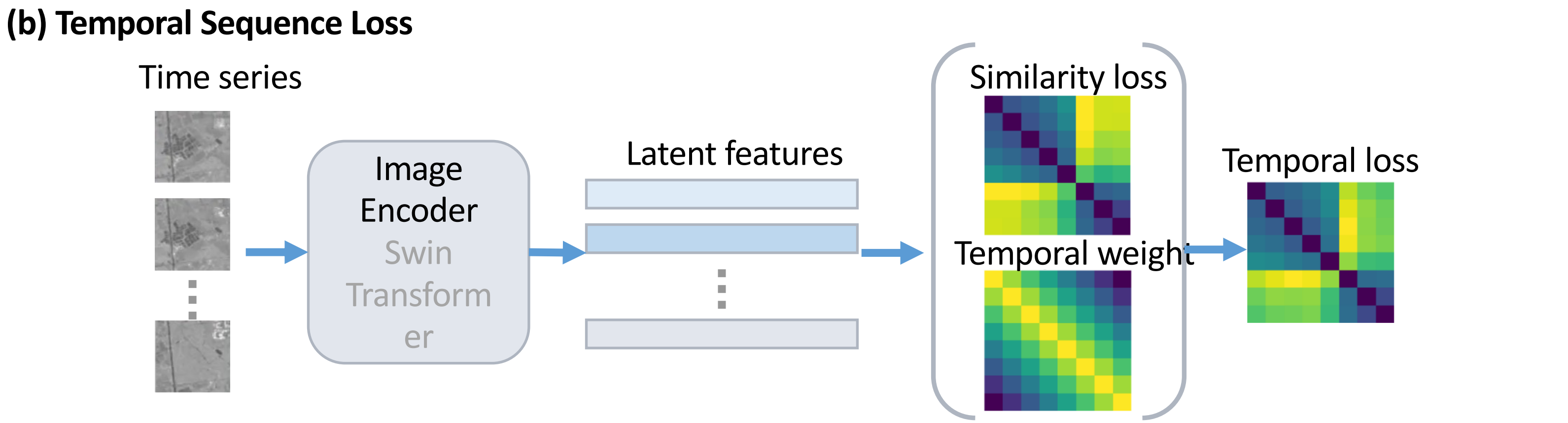
Categories

Build-up　Cropland

Water　Others

Spatial-Temporal Masked Image Modeling





- **Masked Image Modeling**

Spatial feature

① Split: 4x4 non-overlapping patches

② Random Mask: 80%

③ Feature Extraction: Swin Transformer[8]

④ Decoder: Conv Layer

- **Sequence Contrastive Loss**

Temporal feature

$$l = l_{\text{rec}} + \alpha \cdot l_{\text{sim}}$$

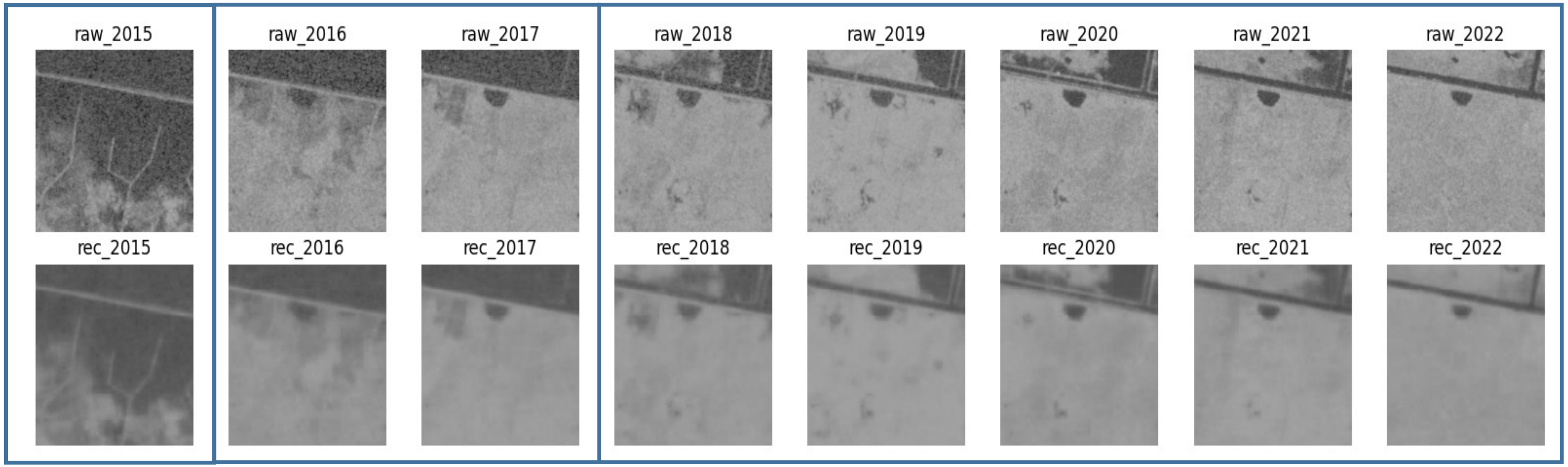
$$l_{\text{sim}} = w_{ij}\left(1 - \frac{\mathbf{uv}}{|\mathbf{u}||\mathbf{v}|}\right)$$

$$l_{\text{rec}} = \text{smooth}\, l_1 \,\text{loss}(\mathbf{x}, \hat{\mathbf{x}})$$

$$w_{ij} = 1 - \beta|i - j|$$

Visualization of STMIM Representations

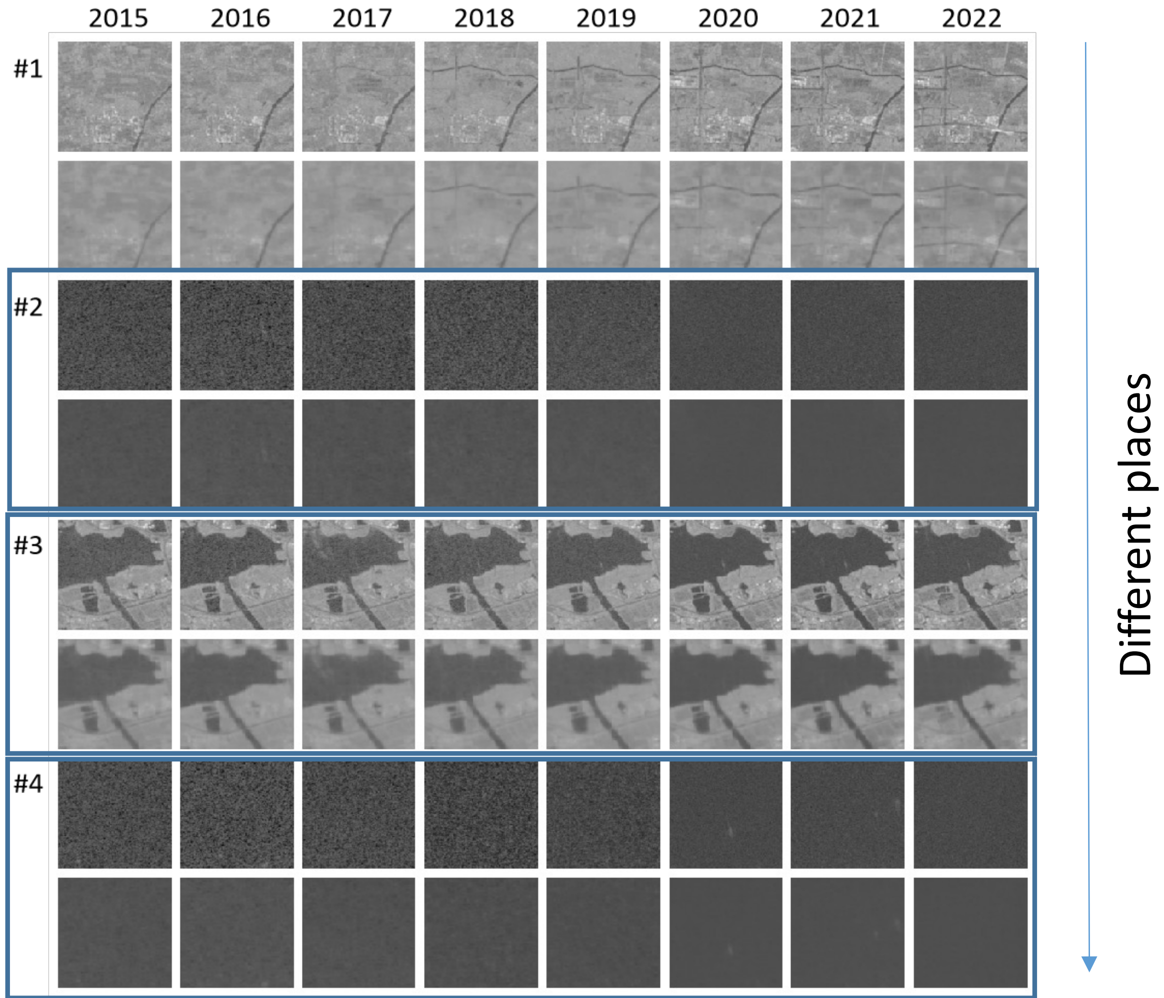
□ Reconstruction Results of sampled image patches



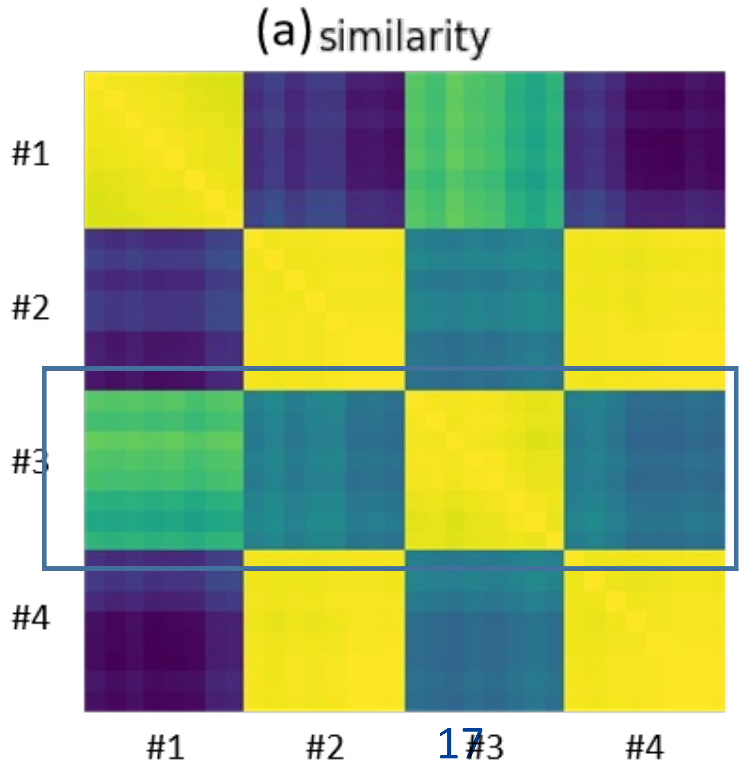
- ✓ STMIM can restore the outline and detail information of the image, indicating that the model has learned certain image features from a large number of unlabeled images.
- ✓ STMIM can effectively extract the consistent features of similar images for the same geographical location at different times. Moreover, it preserves the differentiated features of images with semantic differences

Visualization of STMIM Representations

- Reconstruction Results of a Multiple Time Series Image

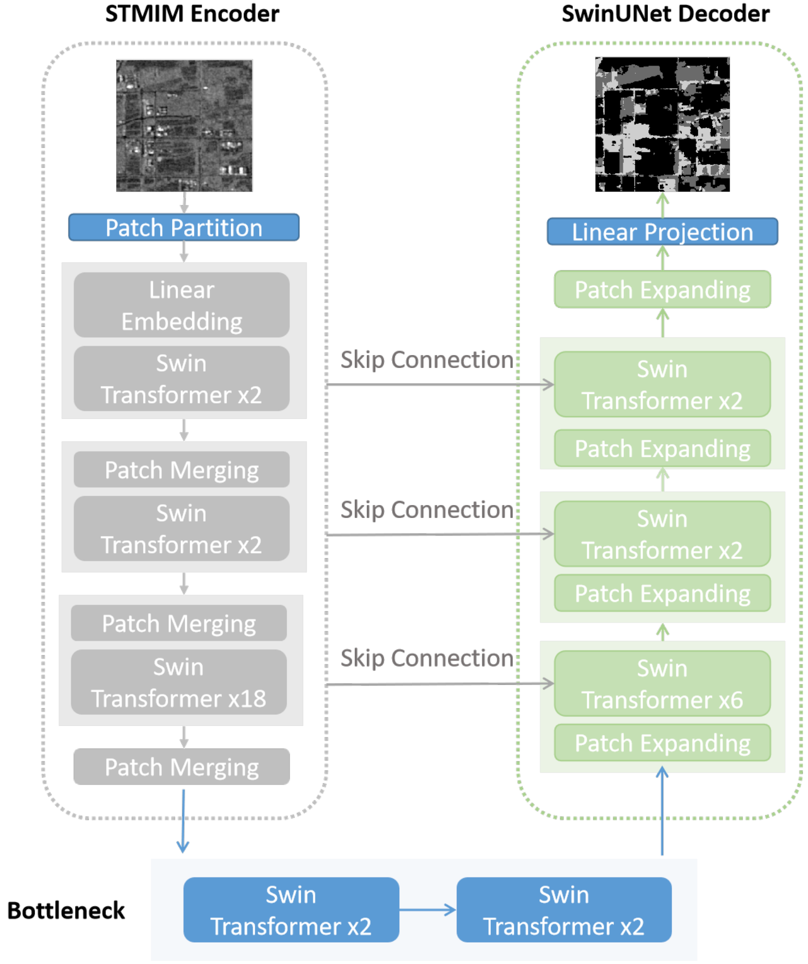


- Sequence Contrastive Loss



Recent Work 2：Spatial-Temporal Masked Image Modeling for Satellite Image Time Series

We apply the STMIM pre-trained model for land cover classification and construce a U-like network and take STMIM as the encoder



STMIM-UNet

☐ STMIM Encoder

- ✓ initialized with pre-trained weights from STMIM
- ✓ to extract features from remote sensing images
- ✓ significantly reduces the amount of labeled training samples required

☐ SwinUNetDecoder

- ✓ for fine-grained land cover classification

■ Labelled Data Ratio Experiments

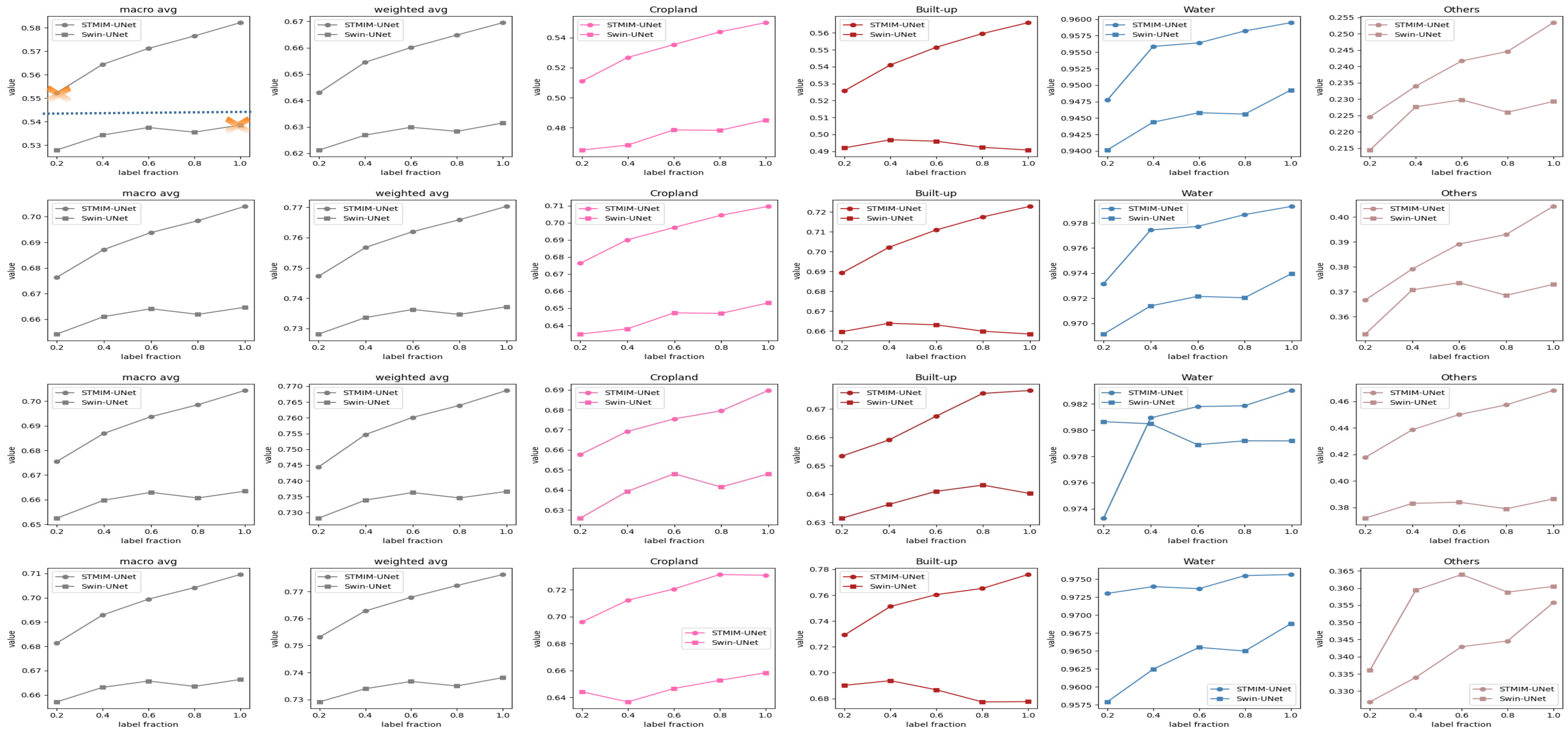
- our model demonstrates remarkable performance while requiring **only less labeled data**, outperforming the full supervised baseline model that necessitates 100% labeled data.

IOU

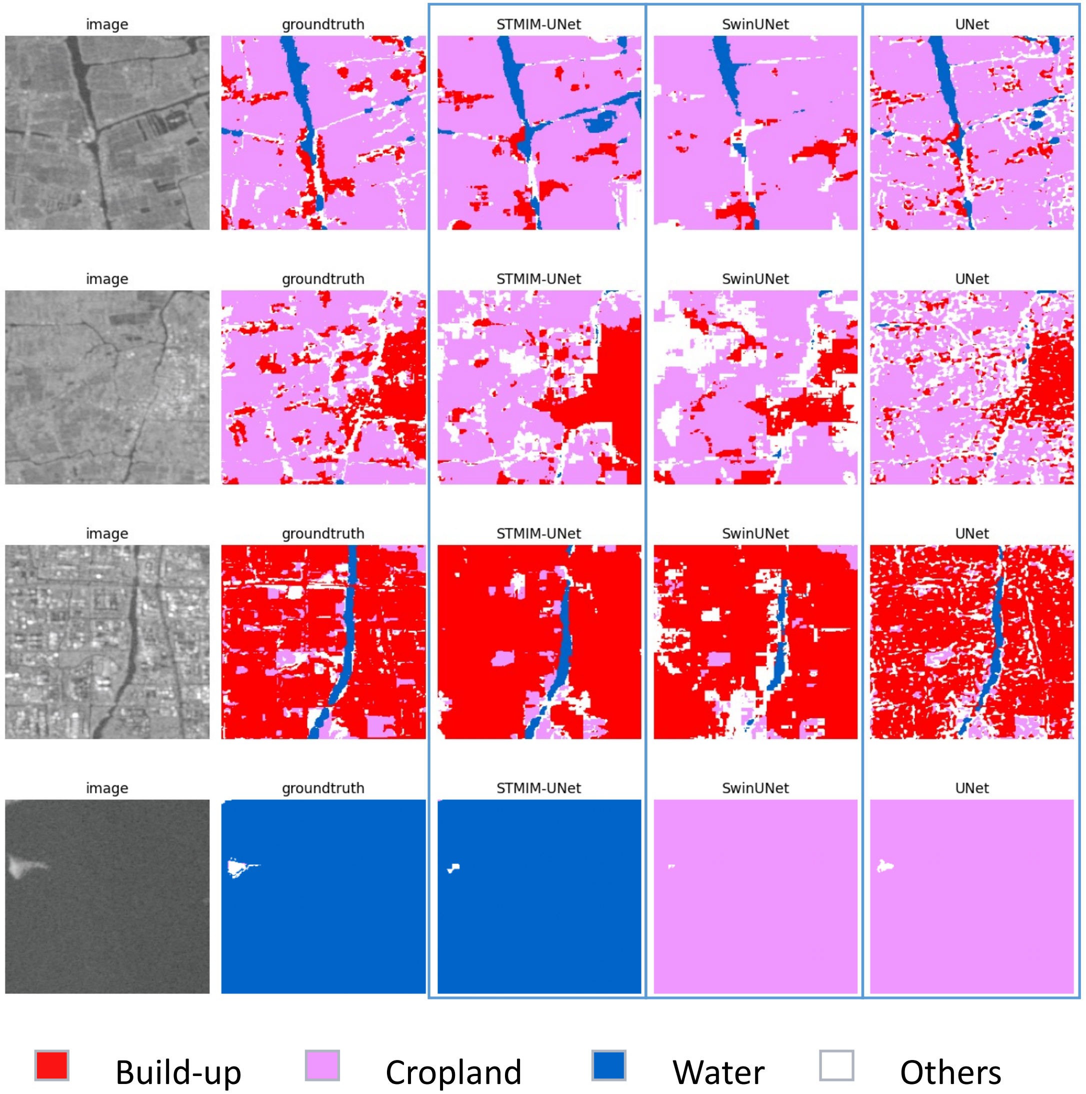
f1-score

precision

recall



Comparison with Baseline Methods



Main Results

- **UNet**
  poor at noise resistance, with the issue of over-segmentation
- **SwinUNet**
  the accuracy still needs to be improved
- **STMIM-UNet**
  shows superior performance with better classification results

Recent Work 2: Spatial-Temporal Masked Image Modeling for Satellite Image Time Series

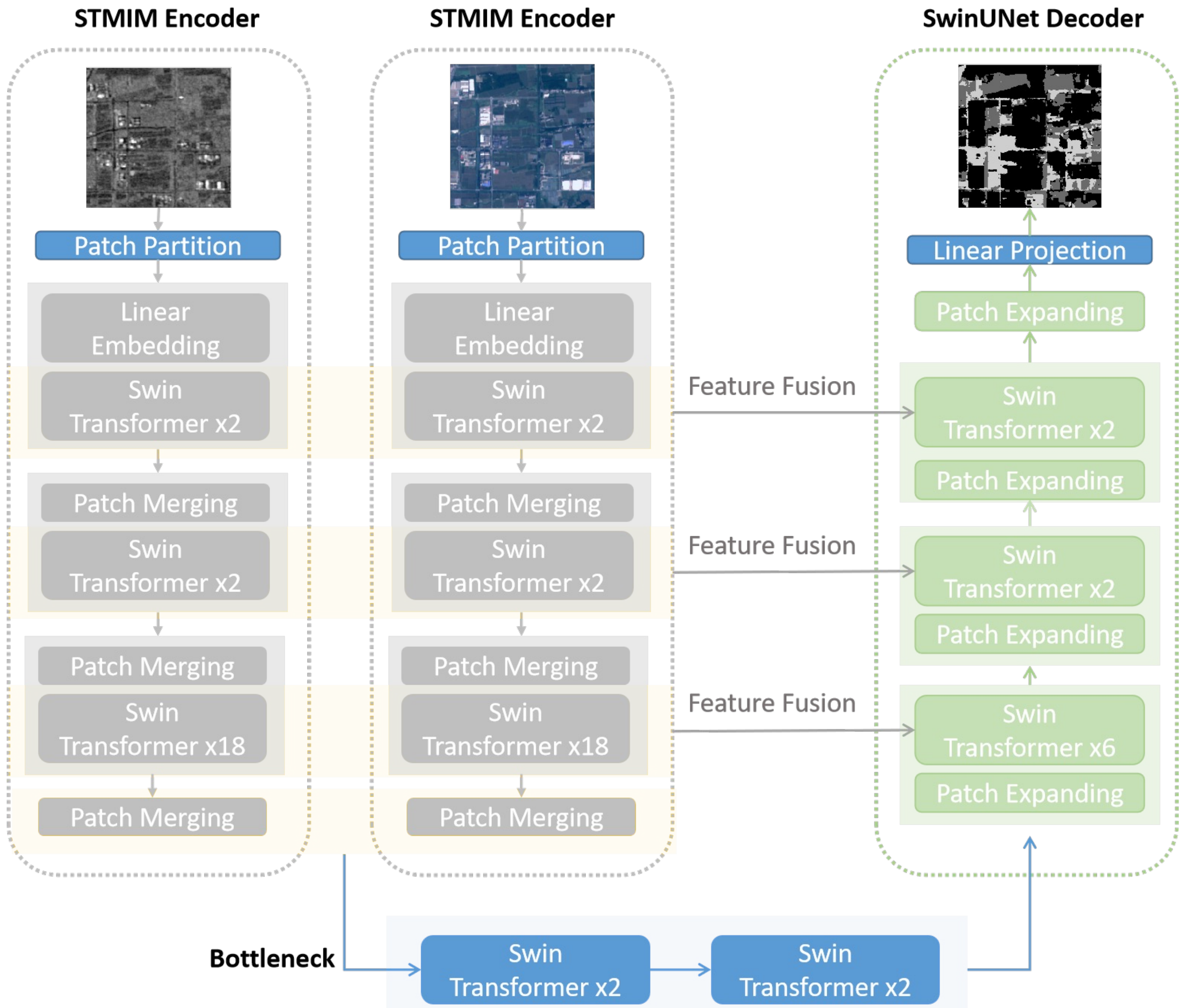
■ Comparison with Baseline Methods

Metric	Model	macro avg	weighted avg	Class			
				Cropland	Build-up	Water	Others
IOU	UNet	0.55	0.64	0.50	0.50	0.96	0.24
	SwinUNet	0.54	0.63	0.48	0.49	0.95	0.23
	STMIM-UNet	**0.58**	**0.67**	**0.55**	**0.57**	**0.96**	**0.25**
precision	UNet	0.67	0.74	0.65	0.66	0.98	0.40
	SwinUNet	0.66	0.74	0.65	0.64	0.98	0.39
	STMIM-UNet	**0.70**	**0.77**	**0.69**	**0.68**	**0.98**	**0.47**
recall	UNet	0.68	0.75	0.68	0.67	0.97	**0.38**
	SwinUNet	0.67	0.74	0.66	0.68	0.97	0.36
	STMIM-UNet	**0.71**	**0.77**	**0.73**	**0.78**	**0.98**	0.36
f1-score	UNet	0.67	0.75	0.67	0.67	0.98	0.39
	SwinUNet	0.66	0.74	0.65	0.66	0.97	0.37
	STMIM-UNet	**0.70**	**0.77**	**0.71**	**0.72**	**0.98**	**0.40**

mIOU: 0.58

acc: 0.77

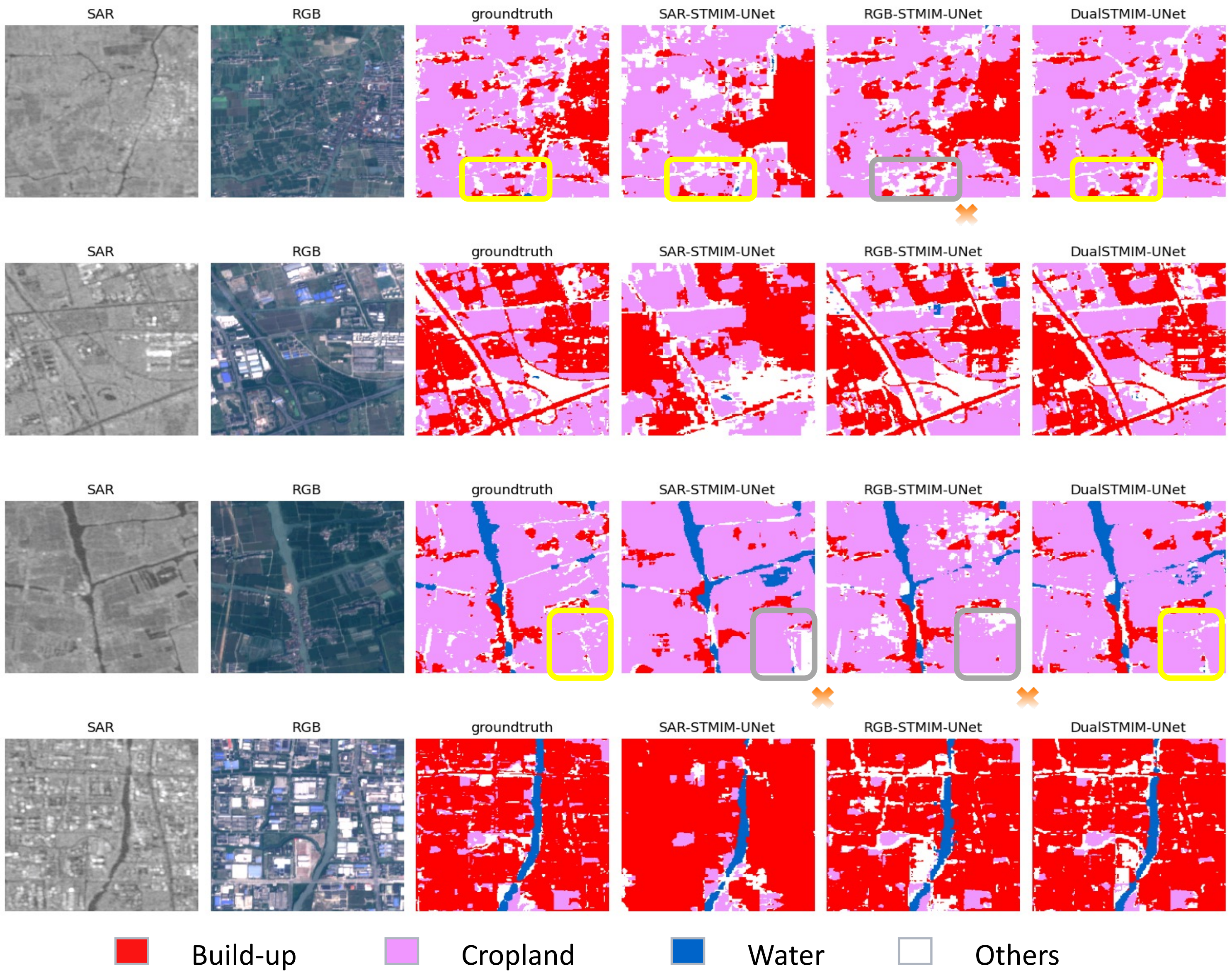
Recent Work 2： Spatial-Temporal Masked Image Modeling for Satellite Image Time Series



- Optical images can provide high spatial resolution, rich spectral and texture information, but optical sensors are sensitive to weather
- SAR sensors adapt to different weather conditions and can penetrate the surface to provide rich spatial information
- We proposed **dual-STMIM-Unet** to fuse SAR and optical image time series.

Fusion SAR and RGB image features layer by layer

Recent Work 2： Spatial-Temporal Masked Image Modeling for Satellite Image Time Series



Main Results

- DualSTMIM-UNet can effectively correct the misclassified pixel blocks in both SAR and RGB images
- Its classification performance is superior to that of single-modal models.

Recent Work 2： Spatial-Temporal Masked Image Modeling for Satellite Image Time Series

Metric	Data	Model	macro avg	weighted avg	Class			
					Cropland	Build-up	Water	Others
IOU	SAR	SwinUNet	0.54	0.63	0.48	0.49	0.95	0.23
		STMIM-UNet	0.58	0.67	0.55	0.57	0.96	0.25
	RGB	SwinUNet	0.66	0.73	0.62	0.69	0.95	0.37
		STMIM-UNet	0.69	0.75	0.66	0.71	0.96	0.42
	SAR+RGB	DualSTMIM-UNet	**0.71**	**0.77**	**0.70**	**0.72**	**0.97**	**0.44**
precision	SAR	SwinUNet	0.66	0.74	0.65	0.64	0.98	0.39
		STMIM-UNet	0.70	0.77	0.69	0.68	0.98	0.47
	RGB	SwinUNet	0.77	0.82	0.75	0.79	0.98	0.57
		STMIM-UNet	0.80	0.84	0.79	0.80	0.98	0.63
	SAR+RGB	DualSTMIM-UNet	**0.81**	**0.85**	**0.80**	**0.80**	**0.99**	**0.66**
recall	SAR	SwinUNet	0.67	0.74	0.66	0.68	0.97	0.36
		STMIM-UNet	0.71	0.77	0.73	0.78	0.98	0.36
	RGB	SwinUNet	0.78	0.82	0.77	0.85	0.97	0.51
		STMIM-UNet	0.80	0.84	0.81	0.86	0.97	0.55
	SAR+RGB	DualSTMIM-UNet	**0.82**	**0.86**	**0.84**	**0.87**	**0.98**	**0.57**
f1-score	SAR	SwinUNet	0.66	0.74	0.65	0.66	0.97	0.37
		STMIM-UNet	0.70	0.77	0.71	0.72	0.98	0.40
	RGB	SwinUNet	0.77	0.82	0.76	0.82	0.97	0.54
		STMIM-UNet	0.80	0.84	0.80	0.83	0.98	0.59
	SAR+RGB	DualSTMIM-UNet	**0.81**	**0.85**	**0.82**	**0.84**	**0.98**	**0.61**

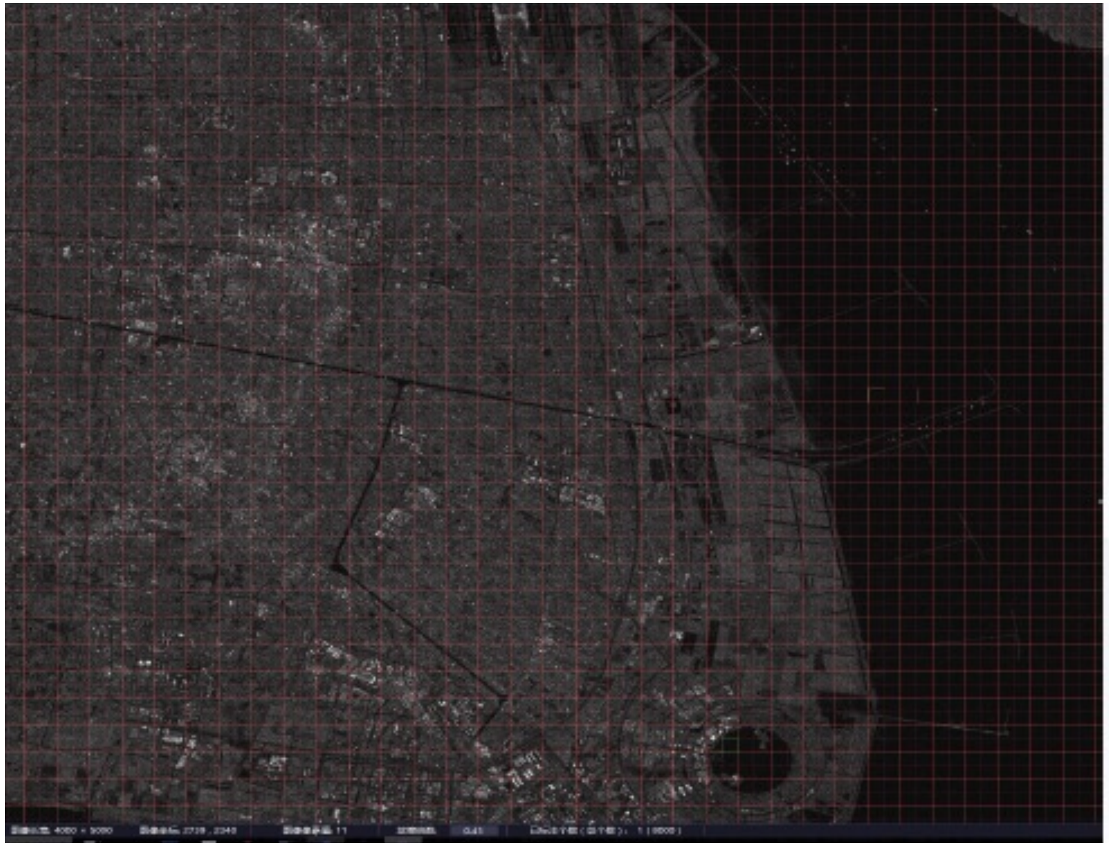
mIOU：0.71

acc：0.86

Recent Work 3：Optical-Aided SAR Image Classification

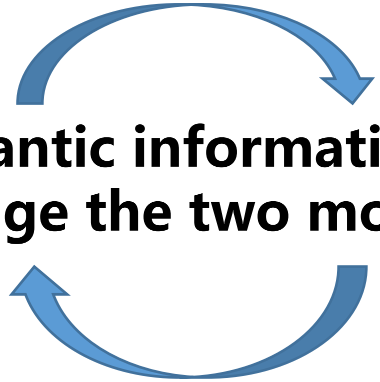
- Land cover and land use classification and change detection requires pixel-wise annotation

  - Low resolution
  - Texture and boundaries are vague
  - Some structures are lost



Sentinel-1 SAR images @Shanghai

- Domain experts interpret SAR image aided by the optical images

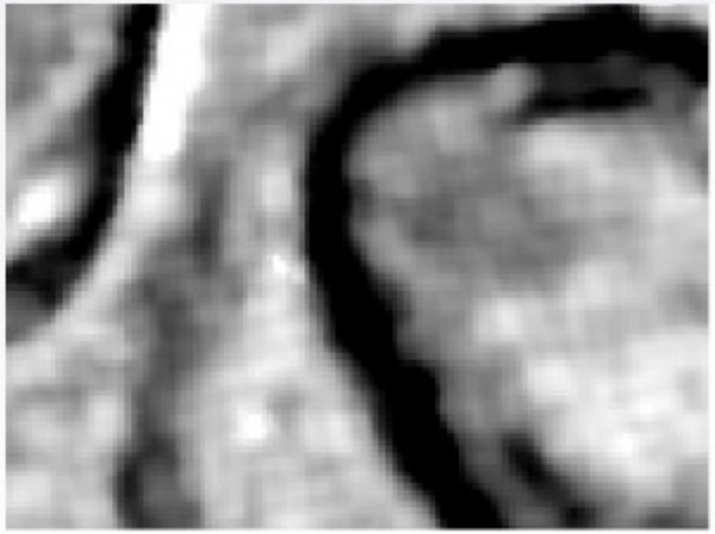


Semantic information to bridge the two modals

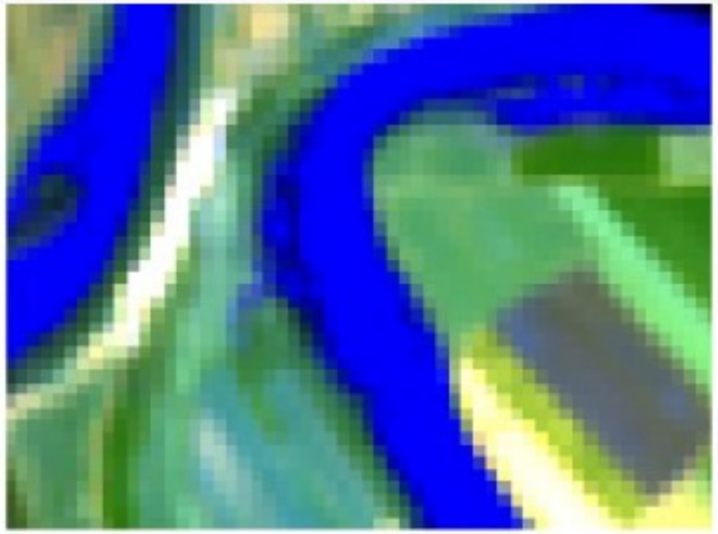
Recent Work 3： Optical-Aided SAR Image Classification

SAR and Optical image fusion?

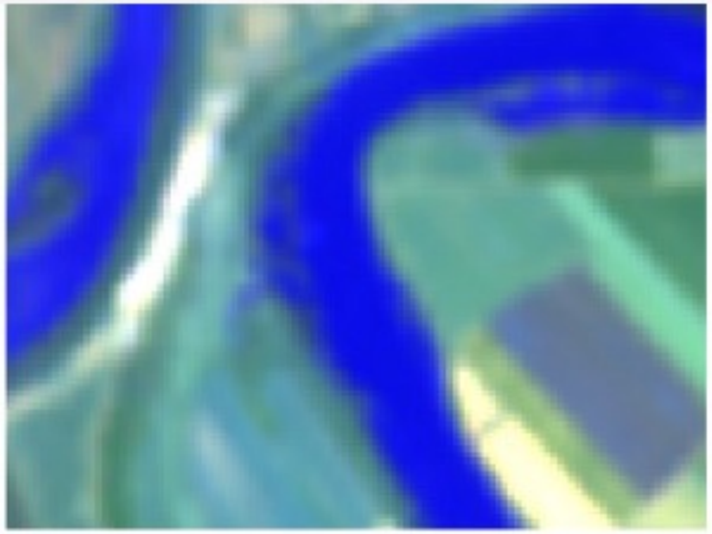
It is very difficult to be registered and can not capture the semantic information



Original Sentinel-1 image
(REFERENCE IMAGE)



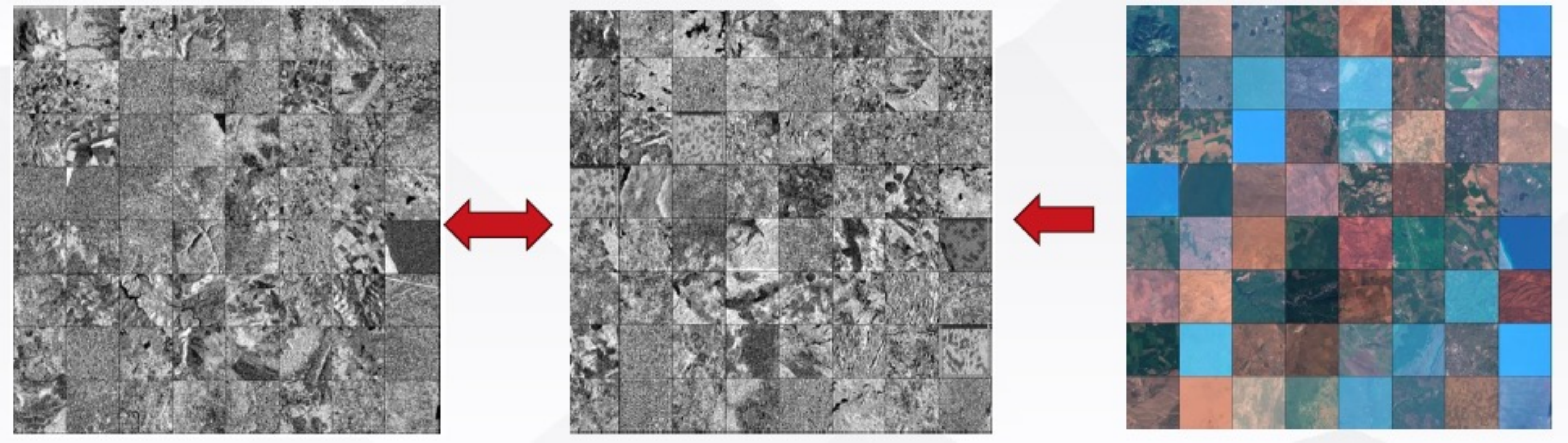
Original Sentinel-2 image
(TARGET IMAGE)



Co-registred Sentinel-2
image

Recent Work 3： Optical-Aided SAR Image Classification

We bridge two modalities through high-level semantic but loose the low-level features.



Original SAR

Generated SAR

Geo-matched
Original optical

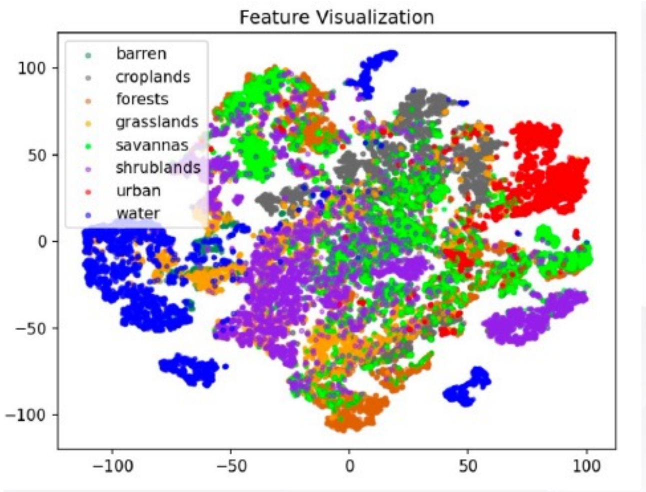
Based adversarial Leaning to generate SAR images at middle
domain that are semantically similar but loose details

The method are based on contrastive learning with three strategies to construct the samples

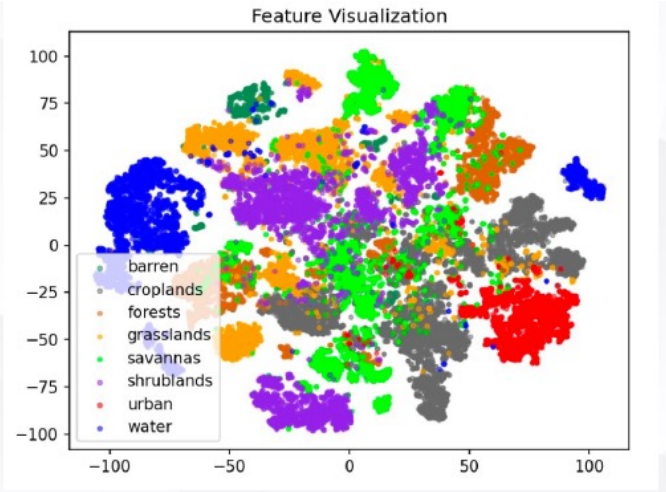
- Instance-level: image augmentation
- Optical-aid: generate SAR image that are semantically similar from the geo-matched optical images
- Cluster-level: cluster the samples

Recent Work 3： Optical-Aided SAR Feature Learning

- The experiments are conduceted SEN12MS dataset
- Achieving SOTA with different backbone
- Our method can improve the performance by 20%
- It is comparable or outperform the full-supervised methods



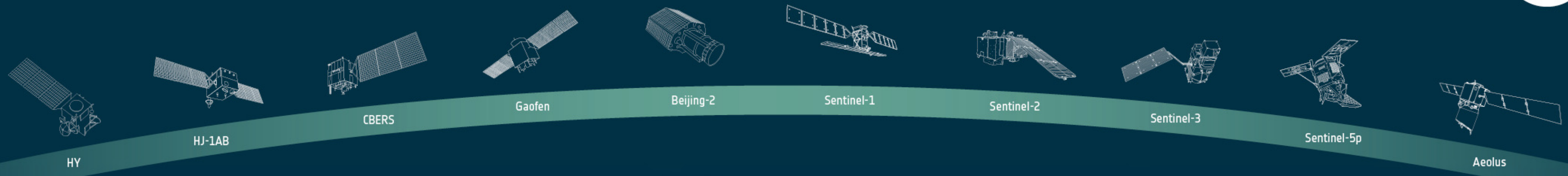
Only SAR



SAR with optical assistance

BackBone	Method	Setting	OA	AA	AP	F1	Kappa	NMI	ARI
ResNet18	DANN [23]	DA	0.2527	0.2706	0.2731	0.1944	0.1329	-	-
	CDAN [24]		0.2377	0.2639	0.2527	0.1688	0.1163	-	-
	Deepcluster [5]	SSL	0.3290	0.3191	0.3233	0.3080	0.2316	0.2070	0.1335
	IIC [20]		0.3406	0.3541	0.3561	0.3358	0.2441	0.1994	0.1211
	MoCo [36]		0.3253	0.3527	0.3282	0.3156	0.2272	0.1780	0.1011
	CC [6]		0.3748	0.3917	0.3766	0.3738	0.2794	0.2513	0.1575
	GCC [7]		0.3780	0.4060	0.3752	0.3763	0.2867	0.2656	0.1606
	GCC-RS (our)		**0.5608**	**0.6320**	**0.5610**	**0.5500**	**0.4983**	**0.5155**	**0.3793**
	Train from Scratch	FT	0.8886	0.8815	0.8915	0.8844	0.8686	-	-
	GCC-RS-FT (our)		0.9167	0.9159	0.9165	0.9159	0.9019	-	-
	Pretrained with ImageNet		0.9223	0.9206	0.9082	0.9093	0.9088	-	-
	Concat-4 channel	DF	0.9350	0.9317	0.9470	0.9383	0.9234	-	-
ResNet50	DANN [23]	DA	0.2499	0.2761	0.3100	0.1856	0.1350	-	-
	CDAN [24]		0.2218	0.2469	0.2352	0.1583	0.1031	-	-
	Deepcluster [5]	SSL	0.3503	0.3537	0.3170	0.3257	0.2465	0.2233	0.1438
	IIC [20]		0.3536	0.3748	0.3458	0.3459	0.2580	0.2217	0.1456
	MoCo [36]		0.3633	0.3194	0.3144	0.2867	0.2392	0.2235	0.1193
	CC [6]		0.3859	0.3959	0.3780	0.3790	0.2920	0.2694	0.1683
	GCC [7]		0.3755	0.4001	0.3722	0.3721	0.2845	0.2670	0.1602
	GCC-RS (our)		**0.5379**	**0.6109**	**0.5345**	**0.5307**	**0.4701**	**0.4923**	**0.3667**
	Train from Scratch	FT	0.8960	0.8802	0.9003	0.8865	0.8776	-	-
	GCC-RS-FT (our)		0.9202	0.9077	0.9213	0.9135	0.9060	-	-
	Pretrained with ImageNet		0.9192	0.8300	0.8153	0.8204	0.9047	-	-
	Concat-4 channel	DF	0.9160	0.9187	0.9210	0.9176	0.9012	-	-
VGG16	DANN [23]	DA	0.3032	0.2583	0.2159	0.1775	0.1287	-	-
	CDAN [24]		0.2713	0.2040	0.1994	0.1362	0.0666	-	-
	Deepcluster [5]	SSL	0.3693	0.3836	0.3893	0.3779	0.2697	0.2534	0.1471
	IIC [20]		0.3650	0.3780	0.3530	0.3544	0.2682	0.2352	0.1584
	MoCo [36]		0.3444	0.3929	0.3452	0.3488	0.2464	0.2003	0.1274
	CC [6]		0.4221	0.4389	0.4166	0.4099	0.3374	0.3094	0.2101
	GCC [7]		0.4246	0.4400	0.4247	0.4145	0.3418	0.3230	0.2185
	GCC-RS (our)		**0.5442**	**0.6236**	**0.5523**	**0.5385**	**0.4791**	**0.5315**	**0.3760**
	Train from Scratch	FT	0.9028	0.9038	0.9153	0.9091	0.8854	-	-
	GCC-RS-FT (our)		0.9424	0.9407	0.9498	0.9449	0.9320	-	-
	Pretrained with ImageNet		0.9328	0.9383	0.9207	0.9292	0.9207	-	-
	Concat-4 channel	DF	0.9213	0.9227	0.9303	0.9263	0.9074	-	-
Inception_v3	DANN [23]	DA	0.2196	0.2170	0.3148	0.1585	0.0941	-	-
	CDAN [24]		0.1973	0.2014	0.3048	0.1553	0.070	-	-
	Deepcluster [5]	SSL	0.3538	0.3774	0.3779	0.3624	0.2592	0.2201	0.1309
	IIC [20]		0.3507	0.3734	0.3650	0.3540	0.2558	0.2328	0.1417
	MoCo [36]		0.3332	0.2336	0.1663	0.1564	0.1637	0.1785	0.0827
	CC [6]		0.4073	0.4114	0.4422	0.3739	0.3239	0.3078	0.2013
	GCC [7]		0.4552	0.5285	0.4535	0.4526	0.3758	0.3972	0.2639
	GCC-RS (our)		**0.6027**	**0.5690**	**0.6015**	**0.5697**	**0.5458**	**0.5781**	**0.4372**
	Train from Scratch	FT	0.9214	0.8676	0.9280	0.8836	0.9072	-	-
	GCC-RS-FT (our)		0.9804	0.9811	0.9817	0.9813	0.9770	-	-
	Pretrained with ImageNet		0.9521	0.9465	0.9524	0.9486	0.9437	-	-
	Concat-4 channel	DF	0.9489	0.9558	0.9529	0.9540	0.9399	-	-
Swin-Transformer Tiny	MoCo [36]	SSL	0.3491	0.3548	0.3548	0.3358	0.2578	0.1810	0.1221
	CC [6]		0.4055	0.4291	0.4003	0.3971	0.3188	0.2935	0.1947
	GCC [7]		0.4356	0.4336	0.4369	0.4163	0.3560	0.3328	0.2275
	GCC-RS (our)		**0.5990**	**0.5677**	**0.6002**	**0.5693**	**0.5403**	**0.5395**	**0.4162**
	GCC-RS-FT (our)	FT	0.9572	0.9573	0.9619	0.9595	0.9497	-	-
	Pretrained with ImageNet		0.9184	0.8256	0.8073	0.8158	0.9035	-	-
ViT	MAE [37]	SSL	0.3913	0.3951	0.4189	0.3843	0.2944	0.2590	0.1557
	SaViT-4channel [2]	DF	0.9678	0.9705	0.9699	0.9702	0.9621	-	-

Bold: Optimal performance of self-supervised trained models　　Underline: Optimal performance of supervised trained models


HY
HJ-1AB
CBERS
Gaofen
Beijing-2
Sentinel-1
Sentinel-2
Sentinel-3
Sentinel-5p
Aeolus

THANK YOU !