Joint Multi-Modality SAR and Optical Representation Learning

Limeng Zhang[1], Zenghui Zhang[1], Weiwei Guo[2], Tao Zhang[1], and Wenxian Yu[1]
[1]Shanghai Jiao Tong University, [2]Tongji University.

1. Abtract

The remote sensing community has shown increasingly interest in self-supervised learning for its ability to learn representations without labeled data. These representations can be easily adapted to downstream tasks through pre-training and fine-tuning. Recently, Masked Autoencoders (MAE) achieve better semantic representation by masking out a significant portion of the input image. However, the original design of MAE for RGB natural images may not be optimal for remote sensing (RS) images, which exhibit considerable variation between modalities like SAR and optical. To address this, we propose a masking methods that enhances feature extraction. After fine-tuning, proposed model outperforms state-of-the-art contrastive and MAE-based models on BigEarthNet-MM classification and significantly reduces input data volume by at least 50%, resulting in a more efficient model. Generalization experiments show a significant F1-score improvement when applied to the SEN12MS dataset, which has diverse data distributions.

2. Introduction

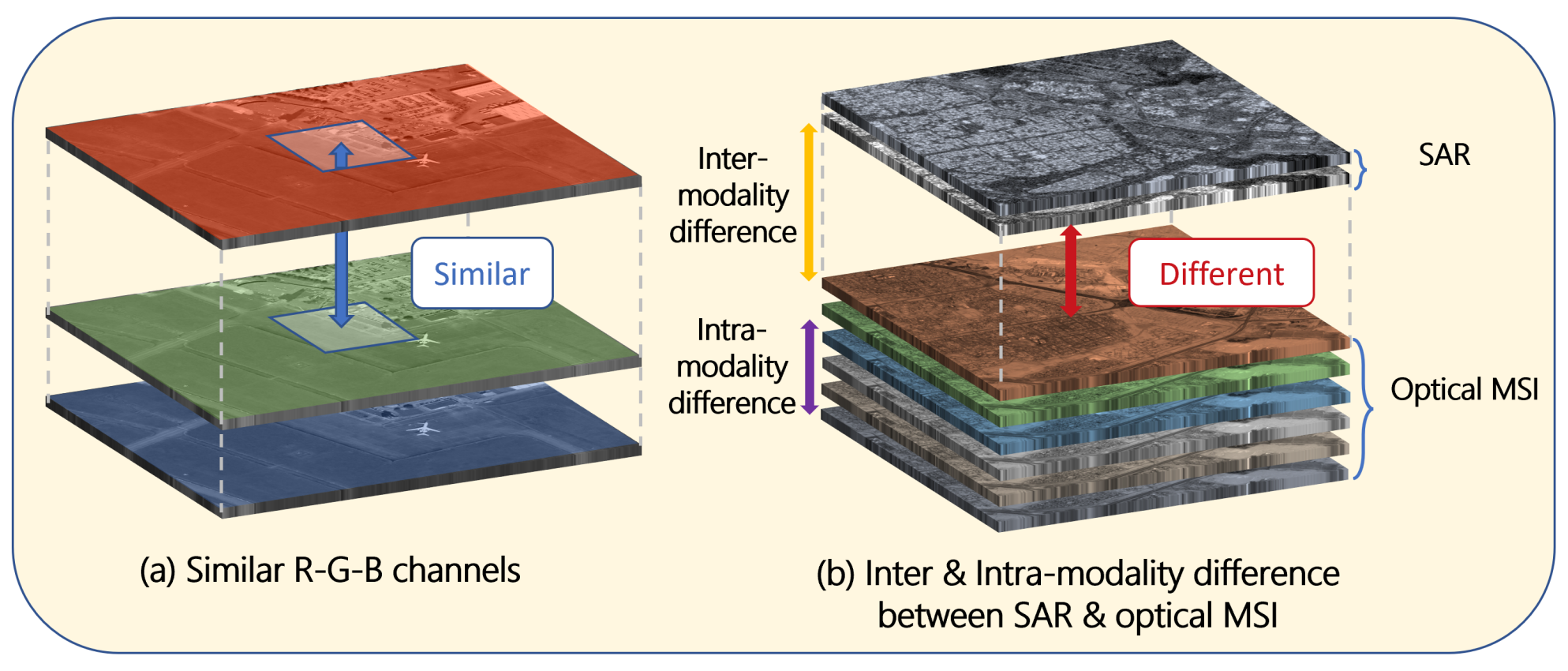
Similar
Inter-modality difference
Intra-modality difference
SAR
Different
Optical MSI
(a) Similar R-G-B channels
(b) Inter & Intra-modality difference between SAR & optical MSI

Figure 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.

As shown in Figure 1(a), for a 3-channel natural image, there is usually little difference among the information in the channels. However, as shown in Figure 1(b), RS images often exhibit substantial variation across different modalities. Optical images offer detailed information but degrade in challenging weather and lighting conditions, while SAR provides complementary information but faces noise interference. While research on single-modality data is well-established, it is challenging but meaningful to study the correlation along the modalities to leverage complementary information between modalities.

3. Methods

Figure 2 illustrates the overall architecture, which introduces a novel masking approach distinct from MAE. The upper section represents pre-training, while the lower section depicts fine-tuning. In pre-training, a pair of SAR I_1 and optical image I_2 of the same location are first concatenated along channel dimension as input. Subsequently, a large portion of the patches will be masked out by our proposed Module.

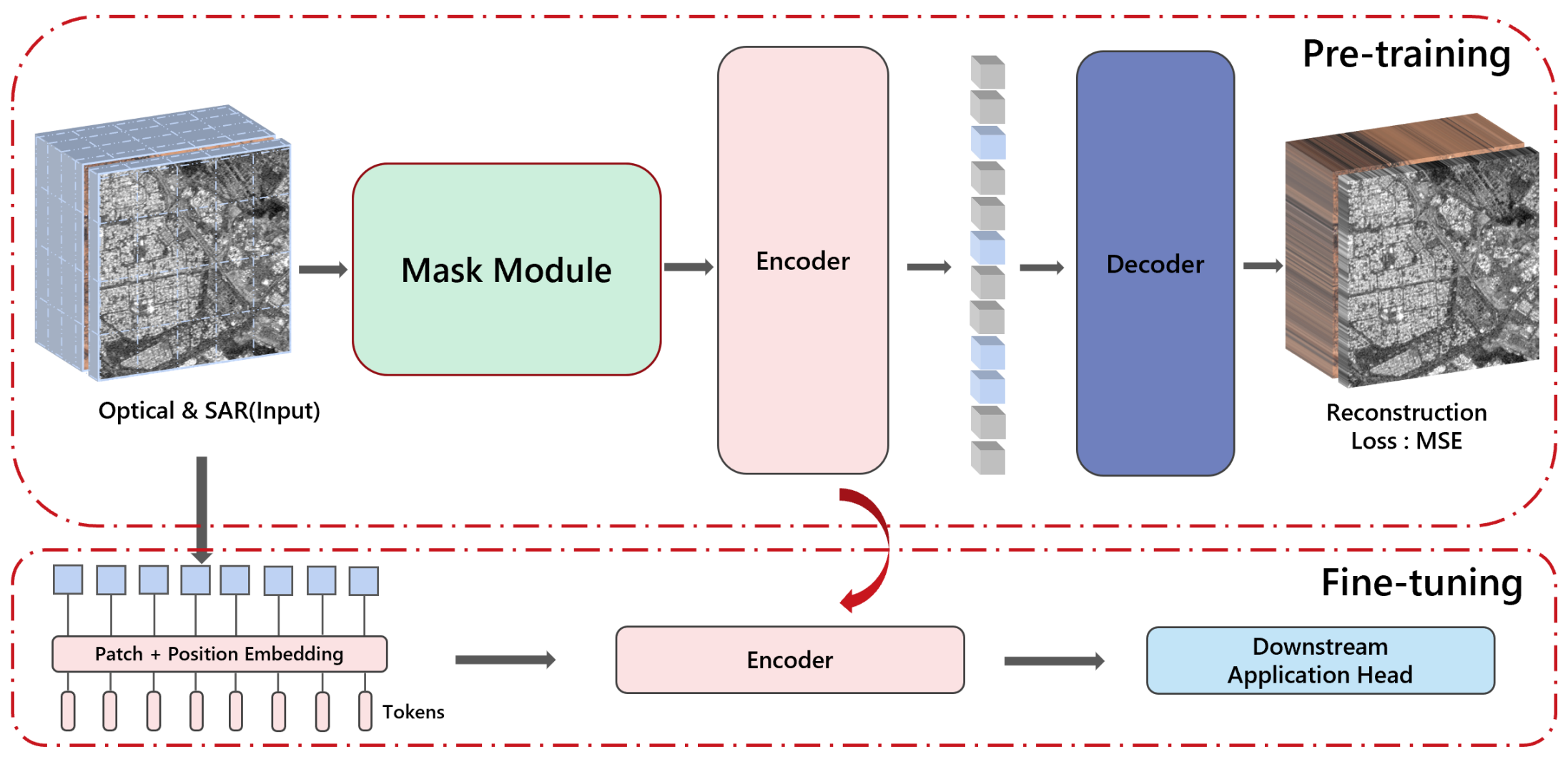

Figure 2: Overall architecture of proposed method. When fine-tuning, the pre-trained encoder is utilized to encode the entire image, extracting features that can be applied to various downstream applications.

The Mask Module is designed to enhance the ability to extract complementary modality information and increase the scenarios encountered by the mask. Careful design is needed to prevent pre-training from becoming a simple interpolation reconstruction job.

4. Results

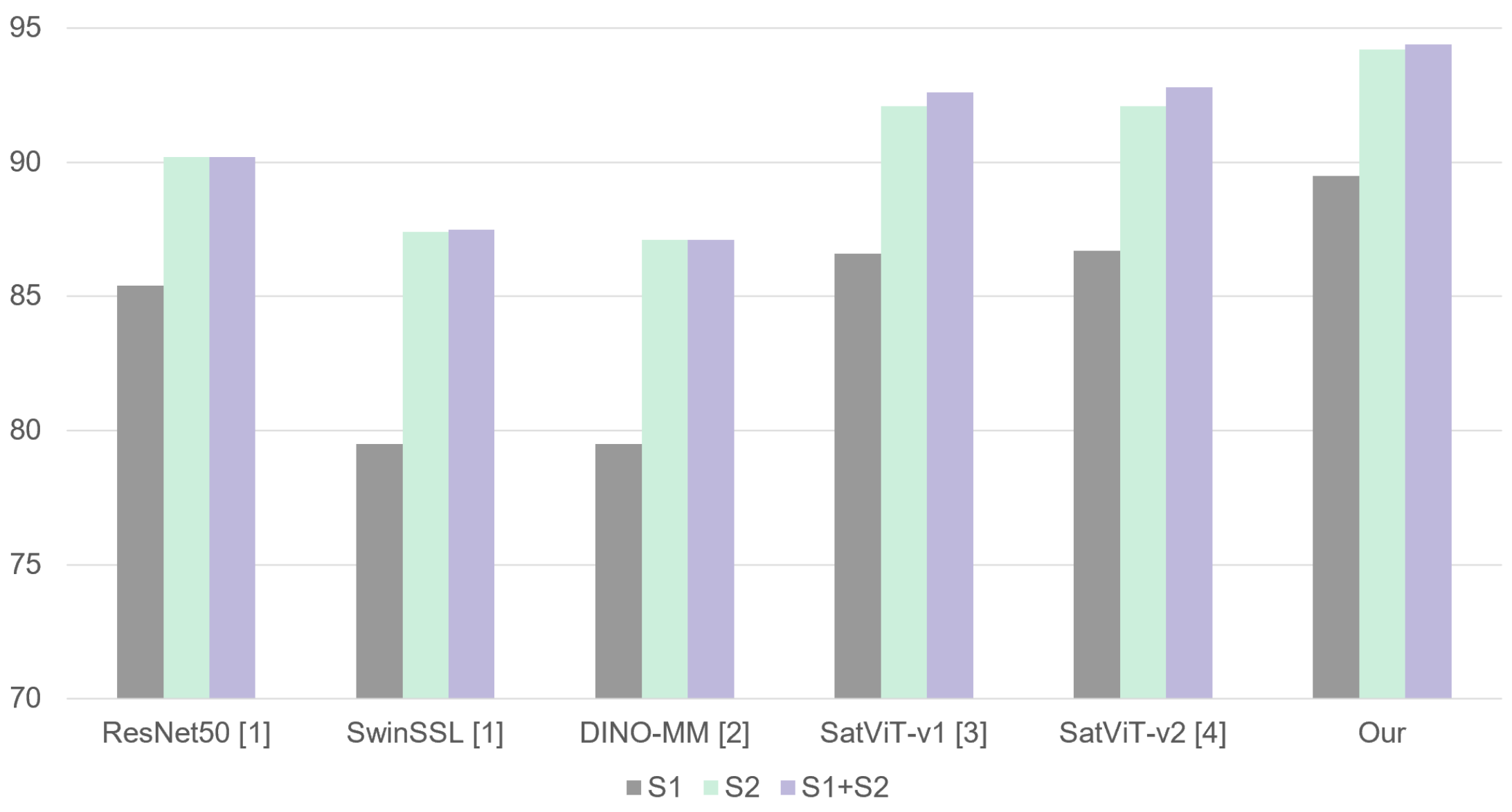

Figure 3: Fine-tuned results on BigearthNet-MM validation set. The evaluation metric is mAP. S1 represents SAR, S2 represents optical image, and + represent the concatenation.

According to Figure 3, proposed method demonstrates strong multi-modality learning, yielding improved S1+S2 performance over S2 alone and surpassing SatViT overall. The relatively slight improvement of S1+S2 compared to S2, is due to the model's already high performance level, making further enhancements challenging.

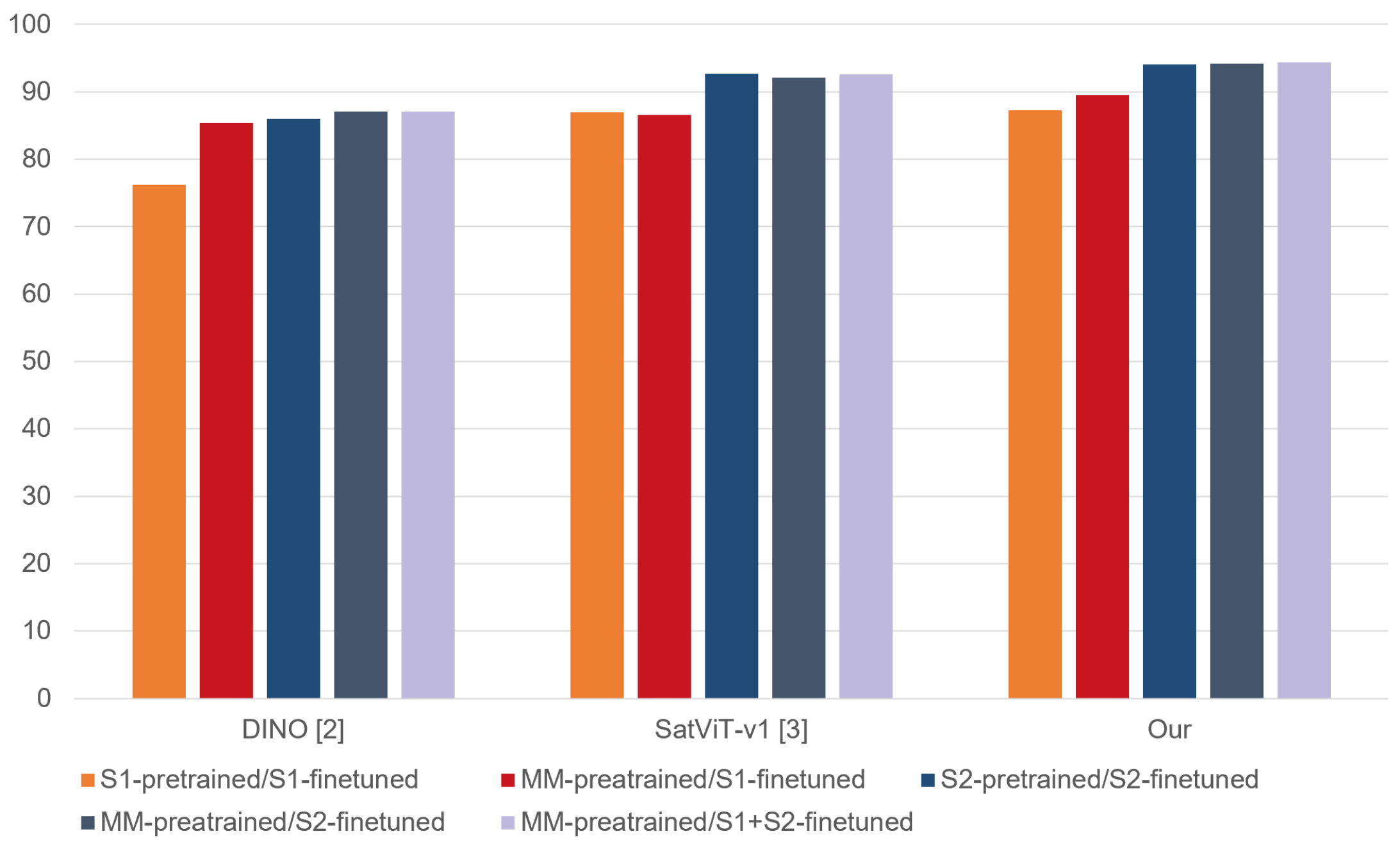

Figure 4: Fine-tuned results on the single-modality BigearthNet-MM validation set using mAP as the evaluation metric. S1 represents SAR, S2 represents the optical image, MM represents multi-modality. The evaluation metric is mAP.

Moreover, proposed method demonstrated superior performance in single-modality settings in Figure 4.

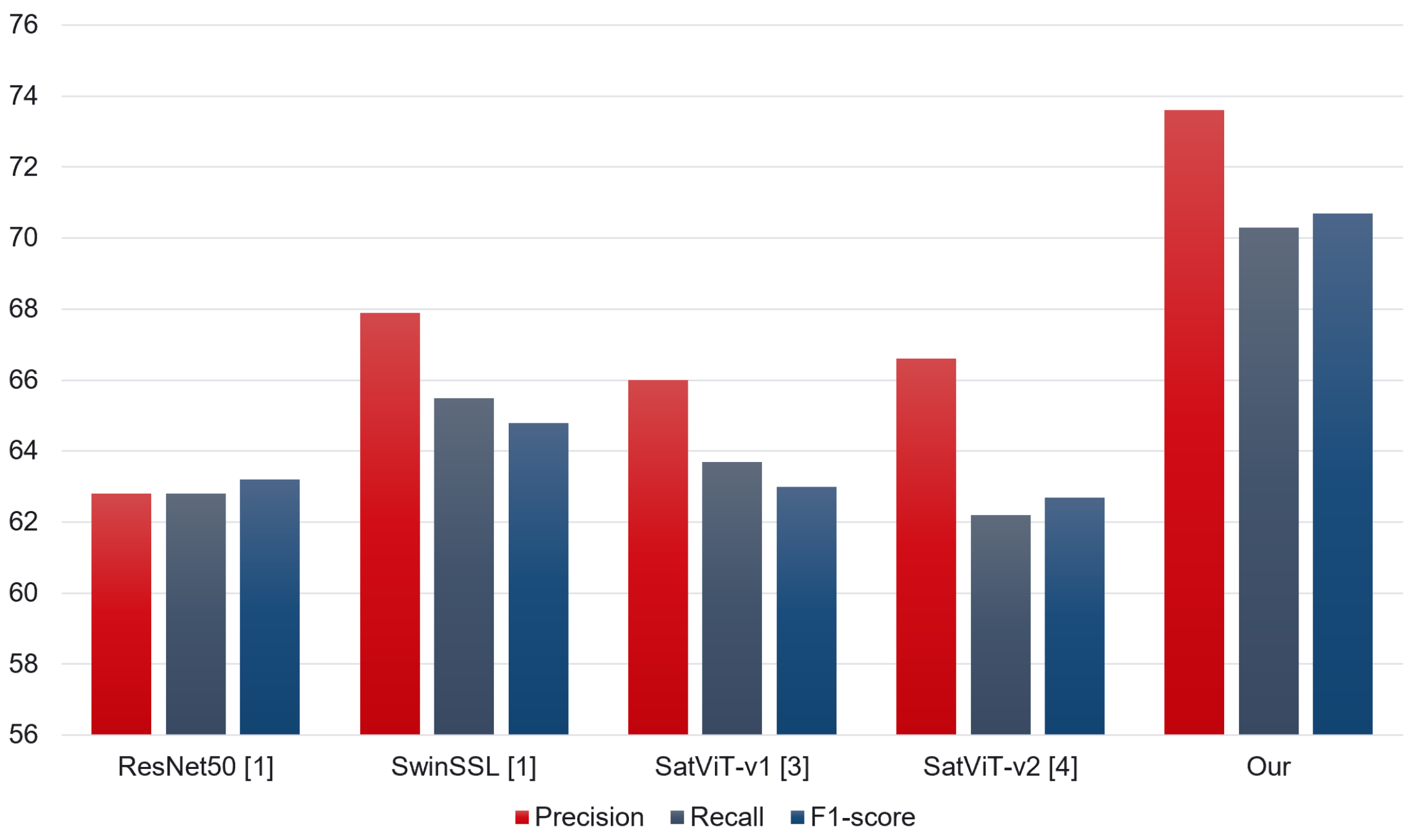

Figure 5: Fine-tuned results on the SEN12MS validation set. The evaluation metrics are the weighted Precision, Recall, and F1-score.

As indicated in Figure 5, the proposed model excels in classification, even with varied data distribution, despite ResNet50 & SwinSSL's direct pre-training on SEN12MS, which should theoretically benefit them.

5. Discussion

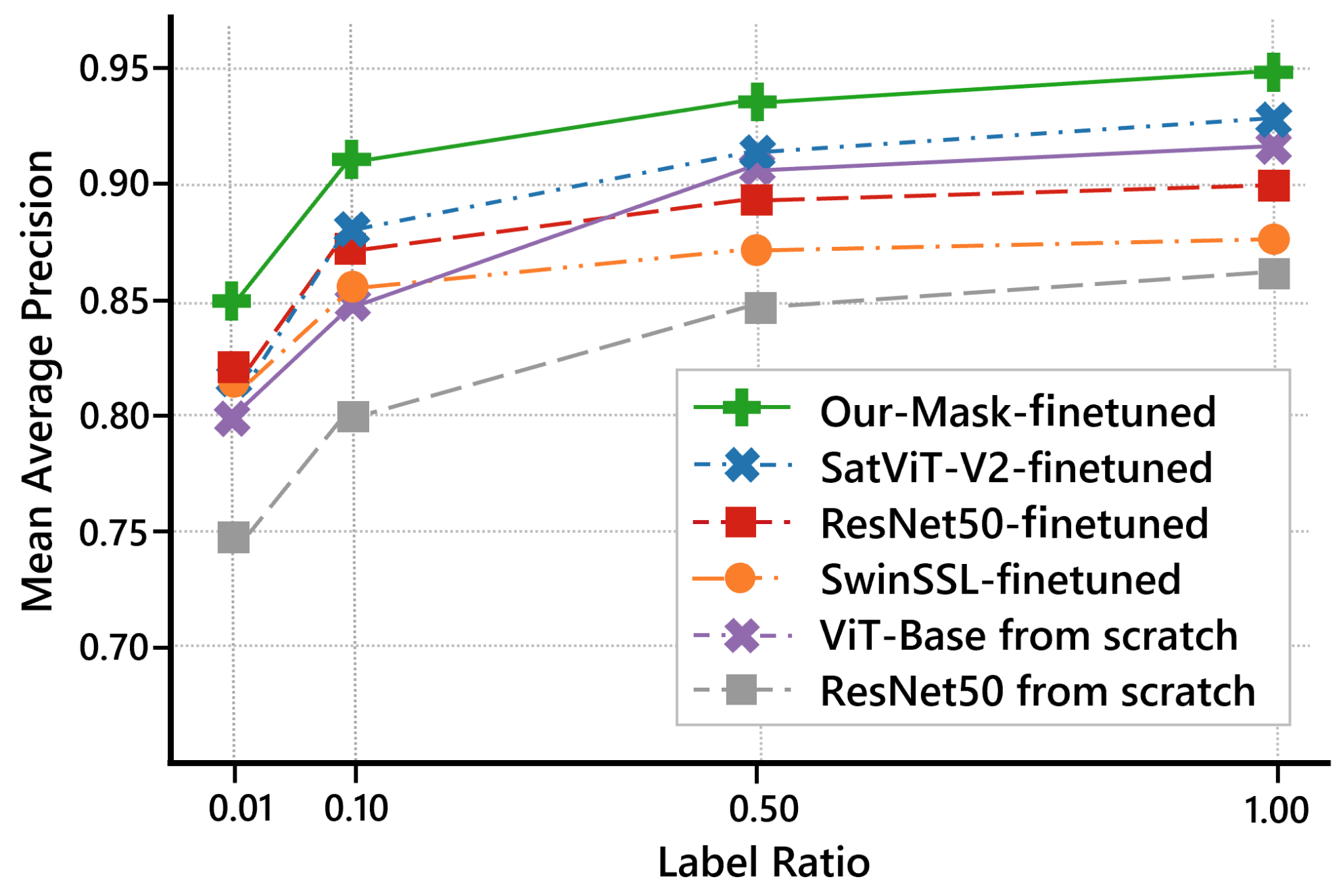

Figure 6: Proposed method outperforms other self-supervised pre-trained models when applied to small-scale data. Furthermore, it shows significantly better performance compared to models that underwent supervised training from scratch on small datasets of BigEarthNet-MM.

We also assessed self-supervised pre-trained and supervised methods on limited datasets with label ratios of 0.01, 0.1, 0.5, and 1. These ratios reflect the labeled data percentage used during fine-tuning.

6. Conclusions

We present a new self-supervised model that employs to enhance the extraction of improved correlations between SAR and optical images. Our model reduces input data by 50%, achieving top performance and semantically rich representations. It generalizes well, even with limited data. Furthermore, it's significant that the model enhances performance by leveraging diverse modalities in scenarios like emergencies with only SAR data and it's versatile and applicable in both single and multi-modality settings.

7. Forthcoming Research

Due to computational resource constraints, we were unable to explore the optimal mask ratio, indicating potential for further improvement in the future.

References

[1] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised vision transformers for land-cover segmentation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1422–1431, 2022.

[2] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 139–142. IEEE, 2022.

[3] Anthony Fuller, Koreen Millard, and James R Green. Satvit: Pretraining transformers for earth observation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[4] Anthony Fuller, Koreen Millard, and James R Green. Transfer learning with pretrained remote sensing transformers. *arXiv preprint arXiv:2209.14969*, 2022.

[5] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *arXiv preprint arXiv:2206.13188*, 2022.