Multispectral Masked Autoencoder for Remote Sensing Representation Learning

Yibing Wei[1,2], Zhicheng Yang[2], Hang Zhou[2], Mei Han[2], Pedro Morgado[1], Jui-Hsin Lai[2]
[1]University of Wisconsin, Madison, WI, USA
[2]PAII Inc., CA, USA

Introduction

➢ Automated analysis of remote sensing (RS) imagery is the key to monitoring global issues. Hundreds of satellites collect plentiful RS data on a daily basis. However, most images remain unlabeled, thus supervised learning algorithms are unable to make full use of the massive amounts of RS data.

➢ Self-supervised learning (SSL) aims to obtain image representations without explicit annotation efforts and has two categories: **contrastive methods** and **generative methods**. However, most of these methods are pre-trained on ImageNet.

➢ Their generalization to multispectral RS tasks is not guaranteed due to the domain gap. Some data augmentation in contrastive methods such as color jittering and center-focused features might lose the spectrum information and the spatial details.

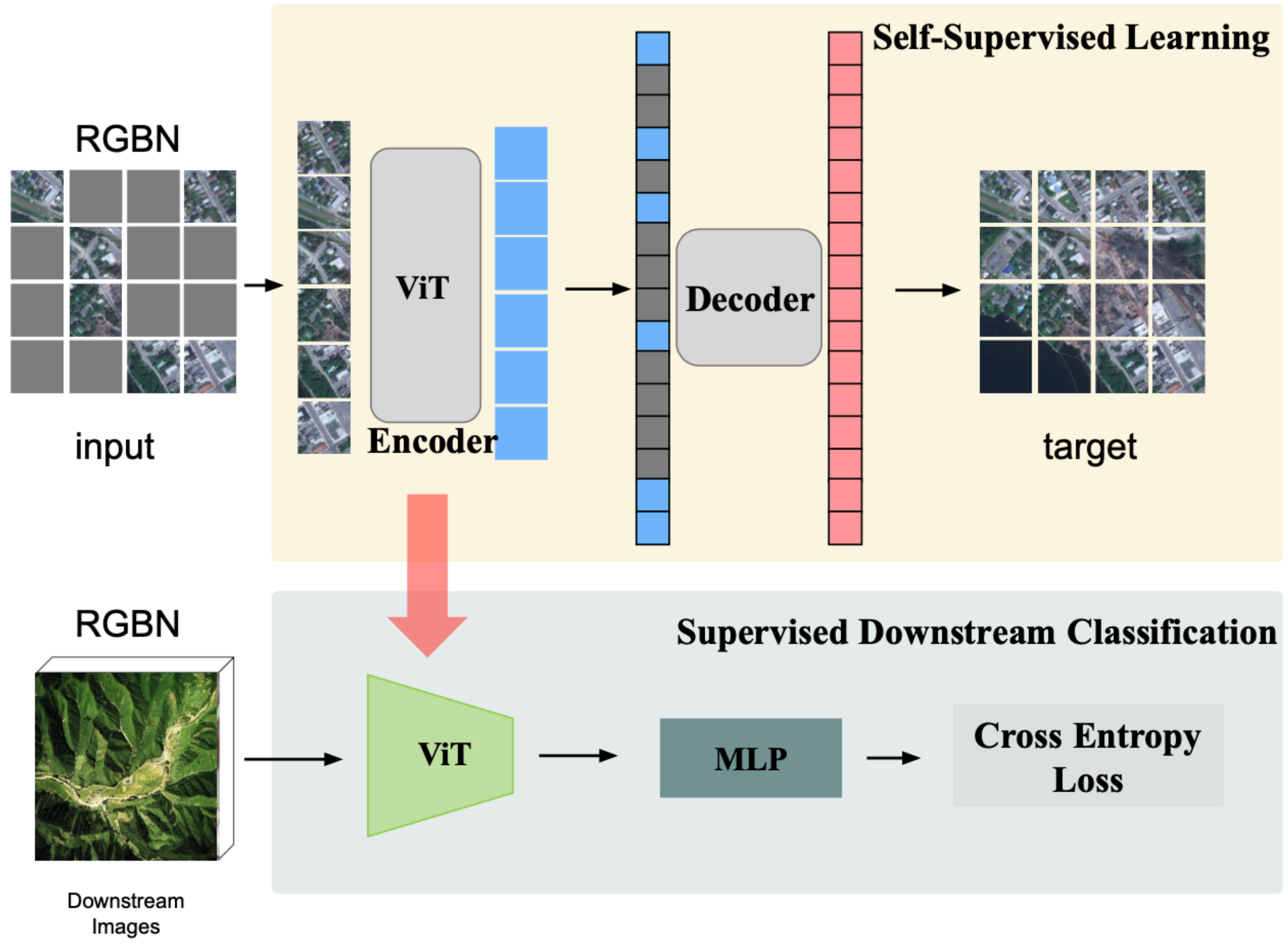
➢ To address this issue, we leverage the benefits of generative methods and build a multispectral masked autoencoder (MAE) to learn RS representation from RGB and Near-infrared (RGBN) data.

➢ Following [2], it has a Vision Transformer (ViT) encoder that maps the unmasked data to a latent representation and a lightweight decoder that reconstructs the original image from the latent representation.

Methods

➢ We perform self-supervised pre-training on the BigEarthNet [3] dataset. The dataset provides **590,326** multispectral Sentinel-2 satellite images of a resolution of **10m**.

➢ To evaluate the representation, we conduct supervised training on NaSC-TG2 [4] dataset with end-to-end fine-tuning, which contains **20,000** multispectral images at 100/200/300m resolution with 10 scene classification labels.



Results

➢ The top-1 validation accuracies are 94.05%, 94.45% and **98.3%** for the ImageNet-1K (IN-1K) pretrain model, BigEarthNet RGB SSL model and RGBN SSL model, respectively. When trained on RGB data only, the RS SSL model still outperformed the IN-1K baseline on the RS classification task, even though IN-1K has more than twice as much data as the BigEarthNet.

➢ The results indicate that the features extracted from RS images are more effective than those from the natural images for the RS task. This domain gap makes RS self-supervised pre-training generalized better in RS tasks. Moreover, the multispectral feature learned with a near-infrared signal increases the top-1 validation accuracy by **3.8%**, showing that the multispectral feature is crucial in RS representation learning.

Method	Train/Validation Ratio 2:8	5:5	8:2
ResNet50 w/ ImageNet(RGB)	92.85	93.64	94.05
SimCLR+ResNet50 w/ BEN3	92.25	93.37	94.45
SimCLR+ResNet50 w/ BEN4	92.90	95.59	97.13
MAE+ViT-B w/ BEN4	**93.28**	**96.11**	**98.30**

Top-1 accuracies of the downstream scene classification task on NaSC-TG2 dataset with CNN or ViT backbones under different SSL settings. (BEN3 or 4: BigEarthNet(RGB) or (RGBN))

Future Work

➢ Our future work includes utilizing spatial-temporal information to introduce strong inductive bias to downstream tasks for more effective transfer learning.

[1] Khan, Adnan, et al. "Contrastive Self-Supervised Learning: A Survey on Different Architectures." *2022 2nd International Conference on Artificial Intelligence (ICAI)*. IEEE, 2022.
[2] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE CVPR*. 2022.
[3] Sumbul, Gencer, et al. "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding." *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019.
[4] Zhou, Zhuang, et al. "NaSC-TG2: Natural scene classification with Tiangong-2 remotely sensed imagery." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021): 3228-3242.