FoMu-SSL: Foundation Model-Guided Multi-Sensor Self-Supervised Learning for Remote Sensing

Dabin Seo [1] **Haeji Jung** [1] **Jinkyu Kim** [1]

Abstract

The application of remote sensing in computer vision struggles with domain shifts among datasets, where models trained on one satellite dataset may not generalize well to others due to diverse geographic and environmental conditions. These differences hinder the self-supervised representation learning, hence this paper introduces an innovative strategy that employs the ImageNet-pretrained foundation model as a guide to enhance the semantic feature extraction process. We also incorporate radar sensor to complement optical sensor inputs, without additional training. Our approach significantly improves performances in segmentation, detection, and classification tasks, offering a robust and efficient method for self-supervised learning in remote sensing.

1. Introduction

With recent advancements in deep learning, its application to various fields including remote sensing has become increasingly prevalent (Mendieta et al., 2023; Muhtar et al., 2023; Cong et al., 2022; Mañas et al., 2021). However, various challenges persist in remote sensing that hurdle the application of computer vision models. Labeling data remains labor-intensive and demands many resources, especially with satellite images, making self-supervised learning increasingly relevant in this field.

Remote sensing datasets often exhibit unique characteristics, differing in many geographical features (*e.g.*, vegetation or color of forests due to weather) due to specific urban planning and design across different locations and time frames. Due to such differences, a deep learning model trained with one particular remote sensing dataset degrades in their performances when finetuned on another dataset, suffering from domain shift. Our approach utilize ImageNet pretrained foundation model (FM) constantly throughout the entire process, inspired by a domain generalized approach with oracle model suggested by MIRO (Cha et al., 2022) and geospatial foundation model (GFM) (Mendieta et al., 2023). This helps to mitigate domain shifts between pretraining and downstream datasets, ensuring more robust pretraining method across diverse remote sensing tasks.

Additionally, satellites also use other sensors such as radar, which operates with radio waves outside the visible spectrum. Radar can capture unique characteristics not detected by optical sensors (*i.e.* penetrating through obstacles like clouds). Numerous datasets and methods integrate paired optical and radar sensor data, utilizing the combination to enrich representation during the training process (Wang et al., 2022a; 2023; Fuller et al., 2024; Scheibenreif et al., 2022; Schmitt et al., 2019; Xu et al., 2023; Chen et al., 2022). By integrating paired optical and radar data with ImageNet pretrained FM, we enhance representation learning with information more nuanced and specialized to satellite imagery tasks. We offer an effective solution, combining FM and multi-sensor data, to enrich the representation learning process and overcome domain shifts in remote sensing.

We present **FoMu-SSL**, **Fo**undation Model-Guided **Mu**lti-Sensor **S**elf-**S**upervised **L**earning for Remote Sensing. Our contributions are as follows:

- We expand contrastive learning based self-supervised learning method by adding simple loss terms to pull the output representations towards the distribution of oracle (*i.e.* ImageNet pretrained FM), reducing the domain gap between different datasets in remote sensing.
- To enrich input semantics, we employ paired radar and optical inputs, bringing enhanced performances on multiple downstream tasks.
- We evaluate our proposed method on various remote sensing tasks–semantic segmentation, rotated object detection, and classification–and demonstrate the effectiveness of each FM application usage step.

2. Method

In this section, we discuss our proposed method to effectively apply ImageNet pretrained foundation model (FM) for satellite imagery pretraining. While adopting self-

[1]Department of Computer Science and Engineering, Korea University, Seoul, South Korea. Correspondence to: Jinkyu Kim <jinkyukim@korea.ac.kr>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

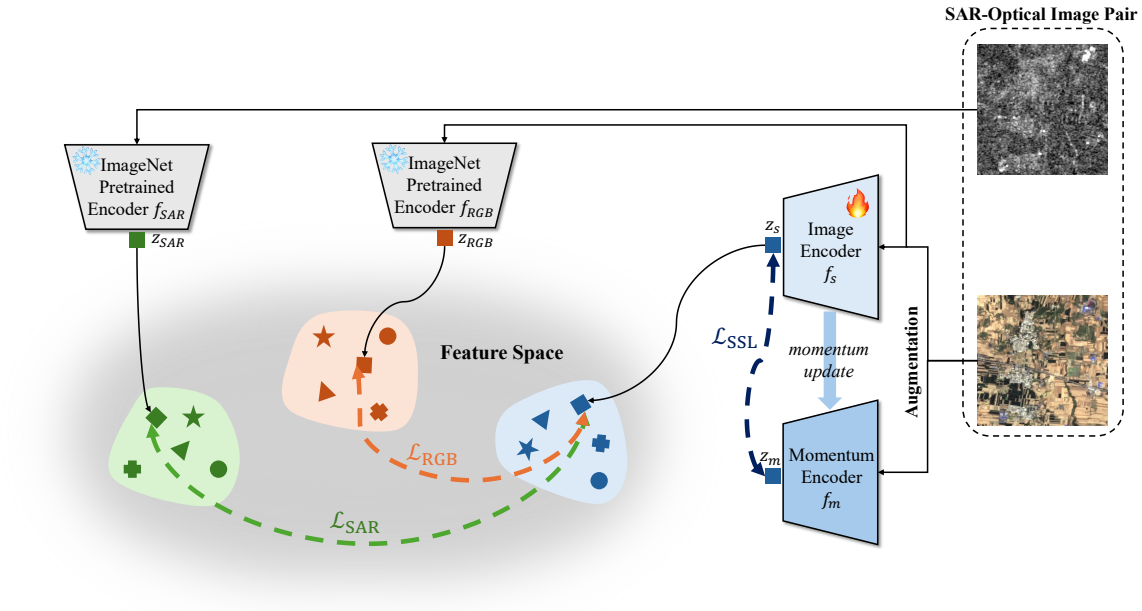


Figure 1: Based on existing MoCo-v2 architecture, we propose a powerful method utilizing ImageNet pretrained foundation model. Since directly finetuning the ImageNet FM performs well on remote sensing downstream tasks, we opt to take advantage of ImageNet, using FM as an oracle model. We minimize the cosine distance between (i) RGB input embedding from updating encoder and RGB input embedding from ImageNet pretrained encoder, and (ii) RGB input embedding from updating encoder and SAR input embedding from ImageNet pretrained encoder.

supervised learning (SSL) based approach to leverage large amounts of unannotated remote sensing data, we also incorporate the FM as a teacher oracle model to tackle the discrepancies (*i.e.* domain gaps) between datasets. We utilize the FM in three ways:

- We initialize the weight of encoder updated during self-supervised pretraining with ImageNet pretrained weight.
- We measure similarity between RGB feature extracted from the updating target model and RGB feature extracted from FM.
- We incorporate an additional sensor, SAR, by extracting its features with FM.

2.1. Self-supervised Pretraining

MoCo-v2 (Chen et al., 2020b) is a powerful SSL method, integrating MoCo (He et al., 2020) and SimCLR (Chen et al., 2020a), and it can effectively extract meaningful representations from remote sensing data. We adopt the training mechanism of MoCo-v2 for the baseline self-supervised contrastive learning. It consists of two encoders, where one is the target encoder $f_s(\cdot)$ updated during training and the other is a momentum encoder $f_m(\cdot)$ updated using exponential moving average of the target encoder's parameters. Each query image is passed through $f_s(\cdot)$, and from the queue in $f_m(\cdot)$, each query finds one positive pair (*i.e.* augmented version of the query), and the rest of the keys in the queue are negative pairs, which is progressively updated with new batch of images during training. More details on the training mechanism can be found in (He et al., 2020) and (Chen et al., 2020b). As in MoCo-v2, we utilize a well-known objective function for contrastive learning, InfoNCE loss (van den Oord et al., 2018), which is formulated as:

$$\mathcal{L}_{\text{SSL}} = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)}, \tag{1}$$

where q is the feature of query image extracted by f_s (*i.e.* $q = z_s$), and k_+ is the positive pair from the queue of keys, and k_i is the entire queue of keys including both positive and negative pairs.

Though target encoder f_s can be trained from scratch, we initialize the encoder with ImageNet pretrained weights to utilize the FM. FM is trained with many RGB images, hence it is already a good image feature extractor, serving as a good starting point for encoder training.

2.2. ImageNet FM on Satellite Images

2.2.1. Weight Initialization with FM

ImageNet (Deng et al., 2009) is a large-scale dataset containing many common objects from the real world. On the other hand, remote sensing images capture geographical characteristics that vary by location on Earth. The unique aerial perspective of satellite images makes it difficult to identify objects from an unfamiliar view. For instance, buildings under construction or natural disasters may complicate earth observation. Despite the differences in object classes and

distinctive features between ImageNet and remote sensing datasets, using ImageNet pretrained FMs as backbones in tasks like semantic segmentation and rotated object detection on remote sensing datasets has shown promising results. Hence, initializing models with ImageNet pretrained weights provides a better starting point for remote sensing representation learning. By continually pretraining the ImageNet pretrained FM with remote sensing datasets, we can integrate the rich image features from ImageNet with the domain-specific features from remote sensing data.

2.2.2. ImageNet FM as Teacher Model

The ImageNet pretrained model can serve as a teacher model, guiding a general distribution of input images in the feature space. This allows resolving domain gaps between multiple remote sensing dataset during pretraining and finetuning processes. To leverage the teacher FM for guidance, we measure the cosine similarity between the features extracted from the encoders.

Cosine distance is a metric used to measure the cosine similarity between two vectors in feature space, based on the cosine of the angle between them. It is defined as:

$$\mathcal{D}_{\cos}(a, b) = 1 - \frac{a \cdot b}{\|a\| \cdot \|b\|}, \tag{2}$$

where a and b are vectors. The value of the cosine distance ranges from 0 to 2, where 2 indicates that the vectors point in opposite directions, 1 means they are perpendicular, and 0 means the vectors are pointing at the same direction. The loss term that enables the teacher model f_{RGB} to guide the target image encoder f_s is defined using cosine distance as:

$$\mathcal{L}_{\text{RGB}} = \mathcal{D}_{\cos}(z_{\text{RGB}}, z_s) \tag{3}$$

The latent feature vector z_{RGB} is the output of input query image passed through the ImageNet pretrained FM, and z_s is the output of same image passed through the the student target encoder f_s. The cosine distance between the vectors are minimized to make the target feature vector to have similar representations as that of the teacher encoder.

2.3. FM with Additional Sensor in Remote Sensing

While most image data, including ImageNet, consist of RGB images with three channels, satellite data often include additional channels collected by specialized sensors. Other than optical sensor that collects information including RGB values, synthetic aperture radar (SAR) is a widely-used sensor on satellites orbiting around Earth. Unlike optical sensor which often suffer from limited views due to instances like bad weather, radar uses longer wavelength compared to optical, hence it can penetrate through obstacles like clouds.

We utilize both *RGB image* and *SAR input* to obtain representations from the FM. While RGB images are directly compatible with ImageNet pretrained models, SAR input consists of 2 channels (pre-processed to 1 channel), requiring the original model to be adjusted. To extract SAR features, we employ the same network architecture (*i.e.* ResNet-50) and the ImageNet pretrained foundation model weights, but use only a subset of filters of the first layer, enabling the model $f_{\text{SAR}}(\cdot)$ to process inputs with less channels. Although ImageNet consists only of RGB images, FM can still be a powerful feature extractor for other modalities like SAR input. The loss function to pull together SAR features z_{SAR} with the target encoder's features z_s is:

$$\mathcal{L}_{\text{SAR}} = \mathcal{D}_{\cos}(z_{\text{SAR}}, z_s). \tag{4}$$

2.4. Loss Function

Ultimately, we minimize the following loss $\mathcal{L}$ for training:

$$\mathcal{L} = \mathcal{L}_{\text{SSL}} + \lambda_{\text{RGB}}\mathcal{L}_{\text{RGB}} + \lambda_{\text{SAR}}\mathcal{L}_{\text{SAR}}, \tag{5}$$

where λ_{RGB} and λ_{SAR} control the strength of each term.

3. Experiments

To validate our proposed self-supervised pretraining approach, we experiment with three downstream tasks: semantic segmentation, rotated object detection, and scene classification. See appendix for details on implementations.

3.1. Self-supervised Pretraining

We employ existing self-supervised learning method, MoCo-v2 (Chen et al., 2020b) as a base method, and add additional loss terms to utilize the ImageNet foundation model (FM). We implement MoCo-v2 (Chen et al., 2020b) pretraining with ResNet-50 (He et al., 2016) backbone from MMSelfSup toolbox (Contributors, 2021) and apply our losses into the pipeline. We initialize the model weights with FM weights, then we further conduct pretraining for 200 epochs on fMoW dataset (Christie et al., 2018), and continually train on SEN1-2 dataset (Schmitt, 2018) for 50 epochs.

For the first 200 epochs of pretraining, we use Functional Map of the World (fMoW) (Christie et al., 2018) dataset during pretraining. For the last 50 epochs, we use Sentinel 1-2 Pair (SEN1-2) dataset (Schmitt, 2018) along with SAR loss $\mathcal{L}_{SAR}$ to further train the model with an additional sensor with copious features unavailable with RGB dataset. We set the value of λ_{RGB} and λ_{SAR} as 0.25.

3.2. Semantic Segmentation

We use UperNet (Xiao et al., 2018) for ISPRS Potsdam (for Photogrammetry & , ISPRS) and ISPRS Vaihingen (for Photogrammetry & , ISPRS) segmentation datasets, for evaluation on this task, experimenting with different self-supervised pretraining models.

Table 1 demonstrates semantic segmentation results of our proposed method, adding each usage of FM. Based on the fact that directly finetuning with the supervised ImageNet model shows better segmentation performance, we initialize target encoder with ImageNet pretrained weights. To

Pretrained Method			Seg. (mIoU)↑		Det. (mAP)↑	Cls. (Acc.)↑	
Dataset	Backbone	SSL Method	Potsdam	Vaihingen	DOTA	UCM	EuroSAT
ImageNet	ResNet50	(Supervised)	85.72	73.39	**75.47**	90.06	99.87
GeoPile (Mendieta et al., 2023)	Swin-B	GFM (Mendieta et al., 2023)	80.56	65.32	63.52	96.55	98.26
fMoW	ResNet50	SeCo-1M (Mañas et al., 2021)	79.50	70.02	68.11	96.31	99.57
fMoW	ResNet50	CMID (Muhtar et al., 2023)	83.89	69.86	70.54	94.17	99.75
Million-AID (Long et al., 2021)	ResNet50	CMID (Muhtar et al., 2023)	84.66	69.91	74.59	95.83	99.81
SSL4EO (Wang et al., 2022b)	ViT-B	CROMA (Fuller et al., 2024)	74.38	59.68	–	96.19	91.32
SEN1-2	ResNet50	SSLTransformerRS (Scheibenreif et al., 2022)	81.96	70.18	61.62	96.07	99.79
fMoW	ResNet50	MoCo-v2 IN init.	86.82	71.83	73.29	95.42	99.94
fMoW	ResNet50	Ours (RGB Only)	87.04	72.40	74.11	96.43	98.09
fMoW+SEN1-2	ResNet50	Ours (RGB+SAR)	**87.36**	**73.55**	74.33	**96.96**	**99.99**

Table 1: Performance comparison with existing state-of-the-art SSL approaches in remote sensing. We report scores in three downstream tasks: semantic segmentation, rotated object detection, and scene classification. For each task, we measure mean intersection over union (mIoU) among all classes, mean average precision (mAP) among all classes, and accuracy (Acc.) of class prediction, respectively.

further maximize the usage of FM, adding $\mathcal{L}_{\text{RGB}}$ and $\mathcal{L}_{\text{SAR}}$ enhances mIoU score by following the ImageNet feature distribution in the feature space. Especially, exploiting SAR sensor outperforms the outstanding ImageNet FM, since SAR inputs contain features cannot be detected with optical sensor. On both Potsdam and Vaihingen dataset, our proposed method outperforms other self-supervised remote sensing pretraining methods trained with only optical sensor inputs (Mendieta et al., 2023; Mañas et al., 2021; Muhtar et al., 2023). In addition, ours also outperforms multi-sensor (*i.e*. SAR-optical data pairs) remote sensing SSL methods (Fuller et al., 2024; Scheibenreif et al., 2022).

3.3. Rotated Object Detection

Implemented from MMRotate (Zhou et al., 2022b), we utilize Oriented R-CNN (Xie et al., 2021) for detection, on DOTA (Xia et al., 2018) dataset, which contains 15 classes.

Table 1 depicts the effectiveness of incorporating ImageNet pretrained model within self-supervised learning. Typically, difference in backbone architecture greatly affects the detection performance, *e.g*. Swin-B better than ResNet50. However, comparing the SSL methods specialized for satellite images, there is no significant correlation between performance and backbone architecture. Though finetuning with ImageNet model for detection yield better results, adding $\mathcal{L}_{\text{RGB}}$ and $\mathcal{L}_{\text{SAR}}$ to baseline MoCo-v2 pipeline improves mAP score. Our method suggest a powerful approach to improve detection performance with satellite images.

3.4. Scene Classification

For classification on UC Merced Land Use (UCM) (Yang & Newsam, 2010) and EuroSAT (Helber et al., 2019) datasets, we validate the feature extracting ability of each pretrained method by using the frozen pretrained model and simply training an additional linear layer.

Table 1 depicts the class accuracy of different pretrained models. The results based on our proposed methodologies, with and without cosine distance losses, demonstrate exceeding performances in both UCM and EuroSAT datasets. Our multi-sensor approach outperforms other SSL methods for remote sensing in both datasets. With comparison to other remote sensing-specific methods, our approach demonstrates outstanding performances on both UCM and EuroSAT datasets. Because this task does not distort pretrained backbone, *i.e*. only training an extra projection layer, the results illustrates the exceptional feature extracting capability of our pretraining approach, benefitting from FM.

4. Conclusion

This study explores novel self-supervised pretraining for remote sensing, mitigating the domain gaps among different datasets using ImageNet pretrained model (*i.e*. foundation model). We observe that the ImageNet pretrained backbone performs well when finetuned on downstream tasks, so we propose to use representations from the foundation model (FM) as a guidance in the feature space, and also incorporating paired multi-sensor datasets with FM. We demonstrate substantial improvements in downstream tasks, approving the capability of utilizing FM for remote sensing. Our study paves the way to enhance the performance of segmentation, detection, and classification in application of computer vision with satellite images.

Acknowledgements

This work was partly supported by IITP under the Leading Generative AI Human Resources Development(IITP-2024-RS-2024-00397085, 30%) grant, IITP grant (RS-2022-II220043, Adaptive Personality for Intelligent Agents, 30% and IITP-2024-2020-0-01819, ICT Creative Consilience program, 10%). This work was also partly supported by Basic Science Research Program through the NRF funded by the Ministry of Education(NRF-2021R1A6A1A13044830, 30%).

References

Akiva, P., Purri, M., and Leotta, M. Self-supervised material and texture representation learning for remote sensing tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8203–8215, June 2022.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., and Ermon, S. Geography-aware self-supervised learning. *ICCV*, 2021.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Blanchard, G., Deshmukh, A. A., Dogan, Ü., Lee, G., and Scott, C. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22 (1):46–100, 2021.

Cha, J., Lee, K., Park, S., and Chun, S. Domain generalization by mutual-information regularization with pre-trained models. In *European Conference on Computer Vision*, pp. 440–457. Springer, 2022.

Chen, S., Zhang, W., Li, Z., Wang, Y., and Zhang, B. Cloud removal with sar-optical data fusion and graph-based feature aggregation network. *Remote Sensing*, 14(14): 3374, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *CVPR*, 2018.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Contributors, M. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Contributors, M. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. https://github.com/open-mmlab/mmselfsup, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

for Photogrammetry, I. S. and (ISPRS), R. S. 2d semantic labeling contest - potsdam. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx, 2022.

for Photogrammetry, I. S. and (ISPRS), R. S. 2d semantic labeling - vaihingen data. https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx, 2024.

Fuller, A., Millard, K., and Green, J. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Gao, P., Ma, T., Li, H., Lin, Z., Dai, J., and Qiao, Y. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022.

Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pp. 2551–2559, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Huang, Z., Wang, H., Xing, E. P., and Huang, D. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 124–140. Springer, 2020.

Kim, D., Yoo, Y., Park, S., Kim, J., and Lee, J. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9619–9628, 2021.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Larsson, G., Maire, M., and Shakhnarovich, G. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 577–593. Springer, 2016.

Lee, G., Jang, W., Kim, J. H., Jung, J., and Kim, S. Domain generalization using large pretrained models with mixture-of-adapters. *arXiv preprint arXiv:2310.11031*, 2023.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Li, Z., Ren, K., Jiang, X., Li, B., Zhang, H., and Li, D. Domain generalization using pretrained models without fine-tuning. *arXiv preprint arXiv:2203.04600*, 2022a.

Li, Z., Ren, K., Jiang, X., Shen, Y., Zhang, H., and Li, D. Simple: Specialized model-sample matching for domain generalization. In *The Eleventh International Conference on Learning Representations*, 2022b.

Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M. Y., Zhu, X. X., Zhang, L., and Li, D. On creating benchmark dataset for aerial image interpretation: Reviews, guidances and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 4205–4230, 2021.

Mañas, O., Lacoste, A., Giro-i Nieto, X., Vazquez, D., and Rodriguez, P. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *arXiv preprint arXiv:2103.16607*, 2021.

Mendieta, M., Han, B., Shi, X., Zhu, Y., Chen, C., and Li, M. Gfm: Building geospatial foundation models via continual pretraining. *arXiv preprint arXiv:2302.04476*, 2023.

Meraner, A., Ebel, P., Zhu, X. X., and Schmitt, M. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020.

Muhtar, D., Zhang, X., Xiao, P., Li, Z., and Gu, F. Cmid: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Neumann, M., Pinto, A. S., Zhai, X., and Houlsby, N. In-domain representation learning for remote sensing. In *AI for Earth Sciences Workshop at International Conference on Learning Representations (ICLR)*, pp. 1–20, April 2020. URL https://arxiv.org/abs/1911.06721.

Noroozi, M. and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pp. 69–84. Springer, 2016.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., Shapira, O., and Birchfield, S. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7249–7255. IEEE, 2019.

Risojević, V. and Stojnić, V. Do we still need imagenet pre-training in remote sensing scene classification? *arXiv preprint arXiv:2111.03690*, 2021.

Scheibenreif, L., Mommert, M., and Borth, D. Contrastive self-supervised data fusion for satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:705–711, 2022.

Schmitt, M. Sen1-2, 2018.

Schmitt, M., Hughes, L. H., Qiu, C., and Zhu, X. X. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019.

Sebastianelli, A., Nowakowski, A., Puglisi, E., Del Rosso, M. P., Mifdal, J., Pirri, F., Mathieu, P. P., and Ullo, S. L. Sentinel-1 and sentinel-2 spatio-temporal data fusion for clouds removal. *arXiv e-print*, 2021.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. URL https://api.semanticscholar.org/CorpusID:49670925.

Wang, Y., Albrecht, C. M., and Zhu, X. X. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 139–142. IEEE, 2022a.

Wang, Y., Braham, N. A. A., Xiong, Z., Liu, C., Albrecht, C. M., and Zhu, X. X. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation. *arXiv preprint arXiv:2211.07044*, 2022b.

Wang, Y., Albrecht, C. M., Braham, N. A. A., Liu, C., Xiong, Z., and Zhu, X. X. Decur: decoupling common unique representations for multimodal self-supervision. *arXiv preprint arXiv:2309.05300*, 2023.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Xie, X., Cheng, G., Wang, J., Yao, X., and Han, J. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3520–3529, 2021.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Xu, G., Jiang, X., Li, X., Zhang, Z., and Liu, X. Exploring self-supervised learning for multi-modal remote sensing pre-training via asymmetric attention fusion. *Remote Sensing*, 15(24):5682, 2023.

Yang, Y. and Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., and Gong, B. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2100–2110, 2019.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 649–666. Springer, 2016.

Zhang, R., Isola, P., and Efros, A. A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1058–1067, 2017.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022a.

Zhou, Y., Yang, X., Zhang, G., Wang, J., Liu, Y., Hou, L., Jiang, X., Liu, X., Yan, J., Lyu, C., Zhang, W., and Chen, K. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022b.

Appendix

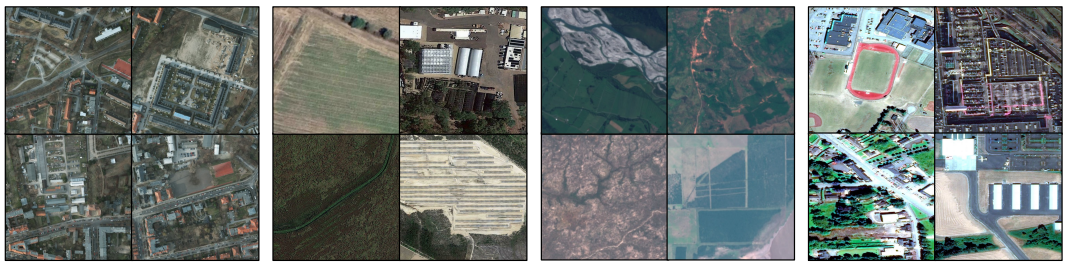
A. Impact Statement

We believe there is no potential of critical harm or misuse of our work that will cause negative societal issues. Foundation model used in this paper is widely used ImageNet pretrained model. It can be employed for various purposes, but are not limited to misuses only. Remote sensing dataset may have ethical or privacy issues, especially with high resolution images that may contain private sensitive information. However, this research uses publicly opened dataset that are, to the extent of our understanding, collected with legal and institutional protocols.

B. Limitations

While we demonstrate the advantage of utilizing ImageNet-pretrained foundation model as an oracle model for remote sensing tasks in terms of domain generalization, there remains several limitations we might need to explore further. It is important to try applying our proposed losses, to other SSL frameworks, testing the compatibility of our method. While the losses are compatible with the methods incorporated in this work, it may not be as effective with other baselines. Moreover, it is worth running in-depth experiments to search for the optimal weight of each additional loss term, either scheduling to adjust its strength over each time step or simply fixing it to a different numerical value. Lastly, despite the fact that we evaluate on 5 datasets of 3 essential downstream tasks, exploration on other tasks such as super-resolution and change detection will be useful to further expand the application of our pretraining method.

C. Domain Gap in Remote Sensing



(a) ISPRS Potsdam (b) Million-AID (c) SEN12MS (d) fMoW

Figure A1: Typical examples of four public remote sensing datasets (ISPRS Potsdam (for Photogrammetry & , ISPRS), Million-AID (Long et al., 2021), SEN12MS (Schmitt et al., 2019), and fMoW (Christie et al., 2018)), each of which was collected under different conditions, such as locations.

Dataset	Sensor	Patch Size	Cropped Size	GSD	Time	Location	# Categories	# Images
fMoW (Christie et al., 2018)	Optical	700-9000	256	-	2002-2017	Worldwide	62	523,842
SEN1-2 (Schmitt, 2018)	SAR-Optical	(256, 256)	256	-	2016-2017	Worldwide	-	282,384
Potsdam (for Photogrammetry & , ISPRS)	Optical	(6000, 6000)	512	0.5 m	-	Potsdam, Germany	6	5,472
Vaihingen (for Photogrammetry & , ISPRS)	Optical	(2494, 2064)	512	0.9 m	-	Vaihingen, Germany	6	742
DOTA (Xia et al., 2018)	Optical	(1024, 1024)	1024	-	-	Worldwide	15	2,806
UCMerced (Yang & Newsam, 2010)	Optical	(256, 256)	256	0.3 m	-	USA	21	2,100
EuroSAT (Helber et al., 2019)	Optical	(64, 64)	64	$\sim$ 0.1m	-	34 countries in Europe	10	27,000

Table A1: Metadata comparison of various public remote sensing datasets, which were collected under different sensing configurations, such as sensors, time, locations, etc. *Abbr.* Ground Sample Distance (GSD).

Different datasets in the remote sensing field each have specific characteristics varying from each other, demonstrated in Figure A1. As summarized in Table A1, satellite images are collected at different times, and many geographical characteristics (*e.g.*, vegetation or color of forests due to weather) differ significantly in a short time frame. Additionally, distinct locations each have their unique urban planning and design, resulting in substantial variances between datasets. Due

to these differences, a deep learning model trained with one particular remote sensing dataset degrades in their performances when finetuned on another dataset, suffering from domain shift.

D. Experimental Details

In this section, we discuss the detailed hyperparameter settings and datasets of our experiments, including self-supervised pretraining, segmentation finetuning, detection finetuning, and classification linear probing.

D.1. Self-supervised Pretraining

Pretrain Dataset. For the first 200 epochs of pretraining, we use Functional Map of the World (fMoW) (Christie et al., 2018) dataset during pretraining. The fMoW dataset contains bounding box annotations for 63 categories, where one is a 'false detection' class. This dataset covers local and temporal features of satellite images, also with other features such as physical sizes or position of the Sun. We use 363,568 images for training, following the pre-defined train split.

For the last 50 epochs, we use Sentinel 1-2 Pair (SEN1-2) dataset (Schmitt, 2018) along with SAR loss $\mathcal{L}_{SAR}$ to further train the model with an additional sensor with copious features unavailable with RGB dataset. SEN1-2 consists of 282,384 pairs of synthetic aperture radar (SAR) and multispectral (including RGB) image patches obtained from Sentinel-1 and Sentinel-2 satellites, respectively. Out of the 14 channels of Sentinel-2 optical sensor, the dataset extracts 3 channels (RGB) only; out of the 2 channels (VV and VH) of Sentinel-1 SAR sensor, the dataset only considers 1 channel, with only vertically polarized (VV) data for simplicity.

Implementation Details. Pretraining experiments are based off the MoCo-v2 implementation from MMSelfSup (Contributors, 2021) with fMoW (Christie et al., 2018) and SEN1-2 (Schmitt, 2018) datasets. We adopt the provided configuration file for the hyperparameter settings. We pretrain the model with batch size of 256 on 2 NVIDIA A100 80GB GPUs. The applied data augmentations for multi-view contrastive learning includes Random Resized Crop, Color Jitter, Random Grayscale, Random Gaussian Blur, and Random Flip. All images are cropped into size (224, 224). The queue size of the momentum encoder is 65,536, and the momentum is 0.9999. We employ the Cosine Annealing learning rate scheduler. SGD optimizer is used with learning rate 0.03, momentum 0.9, and weight decay 0.0001. The contrastive learning head is updated by cross entropy loss with temperature of 0.2.

D.2. Semantic Segmentation

Dataset. ISPRS Potsdam (for Photogrammetry & , ISPRS) contains high-resolution images collection from Potsdam, Germany, which composed of 6 classes: impervious surface, building, low vegetation, tree, car, and clutter. We implement the provided code from (Muhtar et al., 2023), which ignores the 'clutter' class, and only consider the other 5 classes. Moreover, ISPRS Vaihingen (for Photogrammetry & , ISPRS) is another dataset for 2D semantic segmentation collected over Vaihingen, Germany. Similar to Potsdam (for Photogrammetry & , ISPRS), Vaihingen also contains 6 classes: impervious surface, building, low vegetation, tree, car, and clutter. We use the default dataset setup from MMSegmentation (Contributors, 2020) for both datasets.

Implementation Details. For finetuning the pretrained model for segmentation tasks, we utilize UperNet(Xiao et al., 2018) architecture with the code provided by MMSegmentation(Contributors, 2020) on both Potsdam(for Photogrammetry & , ISPRS) and Vaihingen(for Photogrammetry & , ISPRS) datasets. The Potsdam experiments are based on the configuration files provided by (Muhtar et al., 2023), and Vaihingen experiments are based on default settings in (Contributors, 2020). To finetune on Potsdam dataset, we only load 5 classes, disregarding the clutter class. The input images are resized into (512, 512), and Random Crop and Random Flip is applied for augmentation. The UperNet model takes batch size of 8 and trains for 50 epochs. Cosine annealing learning rate scheduler is implemented with linear warmup for 100 iterations and minimum learning rate of 0.000001. For optimizer, SGD with learning rate 0.01, momentum 0.9, and weight decay 0.0005 is applied. For Vaihingen dataset, we load all 6 classes, resize into (512, 512), and apply Random Crop and Random Flip for augmentation. The batch size is 4, and the total training step is 40,000 iterations. For all 40,000 iterations, Polynomial learning rate scheduler is employed with power 0.9, and the minimum learning rate at the end is set to 0.0001. We implement the SGD optimizer, with the same parameters as Potsdam experiments.

D.3. Rotated Object Detection

Dataset. DOTA (Xia et al., 2018) contains 15 classes: plane, baseball diamond, bridge, ground track field, small vehicle, large vehicle, ship tennis court, baseball court, storage tank, soccer ball field, roundabout, harbor, swimming pool, and helicopter. It is a large-scale dataset for detecting oriented bounding boxes. Each original image is cropped into patches with size (1024, 1024) with an overlap of 200 pixels, provided in MMRotate (Zhou et al., 2022b).

Implementation Details. We employ Oriented R-CNN(Xie et al., 2021) model from MMRotate(Zhou et al., 2022b) with DOTA(Xia et al., 2018) dataset, following the configuration setting from (Muhtar et al., 2023). The detection model is finetuned for 12 epochs, with batch size of 2. Random resize is applied with size (1024, 1024), and followed by Random Flip for data augmentation stage. For scheduling the learning rate, we implement the step learning rate scheduler incremented at epochs 8 and 11 , with linear warmup for 500 iterations and warmup ratio of $\frac{1}{3}$. We use SGD optimizer with learning rate 0.03, momentum 0.9, and weight decay 0.0001.

D.4. Classification (Linear Probing)

Dataset. UC Merced Land Use (UCM) (Yang & Newsam, 2010) dataset consists of 21 classes with 100 images for each class. 2,100 images are randomly split (not class-balanced) into train and test set with ratio of 8:2. EuroSAT (Helber et al., 2019) dataset for land use and land cover classification consists of 27,000 images total with 10 classes of land usage in 34 European countries. We follow the data split from (Neumann et al., 2020) and train on the train set, then evaluate on both 5,400 validation and 5,400 test sets.

Implementation Details. To evaluate the feature representation ability of pretrained models, we implement linear probing with a single linear layer that classifies the input data. The linear probing layer is same for both UC Merced Land Use (UCM) (Yang & Newsam, 2010) and EuroSAT (Helber et al., 2019) datasets. For UCM dataset, we update the linear layer for 200 epochs, with crop size of (224, 224) and batch size of 32. For EuroSAT, the epoch, crop size, and batch size is 50, (224, 224), and 256 respectively.

E. Related Work

E.1. Self-supervised Learning (SSL)

In the era of recent deep learning, where an abundant amount of data is available, the cost of annotating data is a big obstacle for practical use of the data. Self-supervised learning (SSL) has gained attention within this context, enabling training the model without the need for extensive data annotation. In SSL, the model learns representations of training data through techniques such as pretext tasks, contrastive learning (He et al., 2020; Chen et al., 2020a;b) or masked image modeling (Bao et al., 2021; Xie et al., 2022; Zhou et al., 2022a; Gao et al., 2022).

Initial approaches of SSL exploit pretext tasks that enable the model understand inputs by learning through pseudo-labels. Examples of pretext tasks include patch position prediction (Doersch et al., 2015), solving jigsaw puzzle (Noroozi & Favaro, 2016), reconstructing part of the image (Pathak et al., 2016), colorization (Larsson et al., 2016; Zhang et al., 2016), and channel prediction (Zhang et al., 2017).

Contrastive learning based methods try to pull together the positive pairs while pushing away the negative pairs in the latent space, enabling the model to create a robust representation space. One of the pioneering works in SSL, MoCo (He et al., 2020), suggests utilizing a large number of negative samples with momentum update of the encoder. On the other hand, SimCLR (Chen et al., 2020a) employs various image augmentation methods to construct well-functioning positive pairs, and projects the feature vectors into a high-dimensional space to compute their contrastive loss. MoCo-v2 (Chen et al., 2020b) integrates the two methodologies to maximize the performance and efficiency, and demonstrates its outstanding performance. Another popular approach is utilizing masked image modeling techniques to learn to reconstruct the image (He et al., 2022; Bao et al., 2021; Zhou et al., 2022a; Gao et al., 2022). Inspired by the learning mechanism of language models, *i.e.*, masked language modeling, this enables the model to learn proper latent representations by taking the surrounded context into account.

E.2. SSL in Remote Sensing

The field of remote sensing is particularly suitable for self-supervised learning. While there is a vast amount of data available from various types of satellites, it is impractical to annotate all accessible data since it requires detailed geographical information of the region. To this end, self-supervised learning techniques are increasingly adopted to the field of remote sensing tasks in order to leverage large amount of data (Ayush et al., 2021; Mañas et al., 2021; Muhtar et al., 2023; Cong et al., 2022; Akiva et al., 2022; Sun et al., 2022). Self-supervised learning for remote sensing share the two main approaches with general SSL. Methods that utilize contrastive learning include (Ayush et al., 2021; Mañas et al., 2021; Akiva et al., 2022). These methods take advantage of the characteristics of remote sensing data to assign positive and negative pairs. For example, in (Mañas et al., 2021), they use temporal information of the images and assign the images from the same region with different season as positive pairs. On the other hand, (Cong et al., 2022; Sun et al., 2022) adopt generative approach, *i.e*., masked image modeling, to learn representations. (Muhtar et al., 2023) brought the two approaches together to maximize the advantage of self-supervised learning. With the proposed framework, they achieve competitive or better performance in remote sensing downstream tasks compared to supervised or other SSL methods.

Data collected from satellites are not limited to the visible spectrum. It encompasses information gathered from multiple sensors, *e.g*. Synthetic Aperture Radar (SAR) using radio waves, and Light Detection and Ranging (LiDAR) employing laser beams to detect the distance to objects. These sensors have the capability to penetrate through atmospheric obstacles (*e.g*. cloud), enabling us to comprehensively observe the top-view information. There are methods (Sebastianelli et al., 2021; Chen et al., 2022; Meraner et al., 2020) that embody sensor fusion, aggregating multiple sensors, to execute cloud removal tasks. Several studies integrate additional sensor to enrich the self-supervised representation learning. For instance, (Wang et al., 2022a) introduces a DINO-based network that utilizes concatenated SAR-optical image as input, randomly masking out one or none of the modalities. Besides, other works (Wang et al., 2023; Fuller et al., 2024; Scheibenreif et al., 2022) also jointly train paired multi-sensor data using self-supervised techniques, leveraging sensor fusion to learn representations from multiple modalities.

E.3. Domain Generalization with Pretrained Models

Domain gap between training and test data is a well-known problem that deep learning models suffer from. Although the model is optimized enough to sufficiently express the training dataset, the model fails to deal with out-of-distribution data. Being aware of this limitation in the model's generalizability, there has been various works that tackles this problem (Yue et al., 2019; Prakash et al., 2019; Blanchard et al., 2021; Ghifary et al., 2015; Li et al., 2018; Ganin et al., 2016; Arjovsky et al., 2019; Huang et al., 2020; Krueger et al., 2021; Kim et al., 2021).

While these algorithms do handle domain gap, it still suffers from bias of source domain and distribution shift for out-of-distribution data as (Cha et al., 2022) pointed out. As the generalizability of pretrained models are known to be powerful, recent works try to tackle domain generalization problem with large pretrained models (Li et al., 2022a; Lee et al., 2023; Li et al., 2022b; Cha et al., 2022).

On the other hand, (Risojević & Stojnić, 2021) demonstrates the effectiveness of domain-adaptive pretraining. While the self-supervised ImageNet-pretrained model has its power to generalize over various visual inputs, additional pretraining on in-domain datasets that are different from the target dataset, helps the model to perform even better with the target dataset. Due to the unique characteristics of remote sensing datasets, adaptive pretraining yields better results, mitigating the domain gap that happens within the domain. To effectively incorporate remote sensing generalizability with ImageNet pretrained model, (Risojević & Stojnić, 2021) train the model with remote sensing data with its weights initialized with ImageNet-pretrained weights. In this work, we also utilize ImageNet-pretrained model to leverage general knowledge of visual data. Rather than merely initializing the model with pretrained weights, we treat ImageNet-pretrained model as an oracle model that guides the newly training model, preventing the model from overfitting on the specific domain it is newly trained on and effectively enhances generalizability.

F. Ablations

Here, we perform ablations studies on various factors of our work.

Pretraining				Seg. (mIoU)		Det. (mAP)
Dataset	Method	$\mathcal{L}_{RGB}$	$\mathcal{L}_{SAR}$	Potsdam	Vaihingen	DOTA
fMoW	MoCo-v2	-	-	83.47	70.92	68.65
fMoW	MoCo-v2	MIRO	-	83.60 (+0.13)	71.34 (+0.42)	64.98 (-3.67)
fMoW+SEN1-2	MoCo-v2	MIRO	MIRO	83.18 (-0.29)	71.14 (+0.22)	63.63 (-5.02)
fMoW	MoCo-v2	MSE	-	85.16 (+1.69)	71.21 (+0.33)	70.88 (+2.23)
fMoW+SEN1-2	MoCo-v2	MSE	MSE	82.98 (-0.49)	70.32 (-0.60)	64.03 (-4.62)
fMoW	MoCo-v2	CosD	-	87.04 (+3.57)	72.40 (+1.48)	74.11 (+5.46)
fMoW+SEN1-2	MoCo-v2	CosD	CosD	**87.36** (+3.89)	**73.55** (+2.63)	**74.33** (+5.68)

Table A2: Comparison of different loss functions for pretraining. FMoW (Christie et al., 2018) is used for RGB inputs, and SEN1-2 (Schmitt, 2018) for SAR inputs. While adopting training mechanism of MoCo-v2 (Chen et al., 2020b), we employ additional loss terms to leverage the oracle model(*i.e.* ImageNet-pretrained model). For each loss design, we evaluate the impact of the loss by the performance of downstream tasks; semantic segmentation on ISPRS Potsdam/Vaihingen (for Photogrammetry & , ISPRS;I) and object detection on DOTA (Xia et al., 2018).

F.1. Comparative Analysis of Loss Functions.

The empirical evidence demonstrated in Table A2 underscores the efficacy of distilling the network with ImageNet pretrained model. For all experiments with additional branches, the loss weight λ_{RGB} and λ_{SAR} are set to 0.25 to ensure a fair and uniform comparison of the loss functions. The baseline model is trained on fMoW dataset with MoCo-v2 SSL method.

The rows highlighted in green represent the integration of solely the RGB branch into the original SSL framework. Applying any of the three losses enhances segmentation mIoU performances across both Potsdam and Vaihingen datasets. In particular, CosD loss shows the largest progress of 3.57 mIoU in Potsdam and 1.48 in Vaihingen. Observing the detection downstream task, MIRO and CosD shows improvement of 2.23 and 5.46 point increase in mAP respectively. Contrastively, selecting MIRO loss to minimize the difference between feature distributions results in decrease of 3.67 on detection task.

The rows highlighted in orange represent additional implementation of SAR branch, leveraging both RGB and SAR features. To further advance the model's ability to analyze remote sensing images, SAR features are extracted from the ImageNet pretrained encoder and are compared with features from updating encoder. Unlike applying losses with fMoW RGB data only, introducing SAR-RGB pair data with ImageNet encoder is only effective with CosD function. Exploiting MIRO and MSE with SAR branch rather worsen segmentation and detection performances, while CosD improves by 3.89, 2.63, and 5.68 on Potsdam, Vaihingen, and DOTA respectively.

The results suggest to use an additional loss term to follow the general ImageNet feature distribution while further training with satellite images. Though implementing RGB image features enhance performance with any of the three suggested losses, Cosine Distance (CosD) is the most effective one, and even applicable with SAR features as well. Hence, implementing CosD loss function is the best way to utilize the ImageNet pretrained model as a teacher, and further adapt to remote sensing datasets specifically during training.

F.2. Application on Other SSL Frameworks.

We verify the necessity of following the ImageNet distribution in feature space. The study involves Contrastive Mask Image Distillation (CMID) (Muhtar et al., 2023), a SSL framework dedicated to remote sensing, combining both masked image modeling and contrastive objective. To further improve CMID, we adhere our RGB branch to the provided CMID code. Figure A2 illustrates incorporating additional guidance from the pretrained model is effective across other remote sensing SSL frameworks as well, thereby supporting our proposed method.

Pretraining			Seg. (mIoU)	Det. (mAP)
Dataset	Method	$\mathcal{L}_{RGB}$	Potsdam	DOTA
fMoW	CMID	-	83.99	70.54
fMoW	CMID	MIRO	85.88	70.58
fMoW	CMID	MSE	84.09	69.28
fMoW	CMID	CosD	86.15	71.55

Figure A2: Performances report on CMID (Muhtar et al., 2023), a pretraining method specialized in remote sensing domain. We attach the proposed RGB branch to the existing framework to examine the applicability of our proposed technique to methods other than (Chen et al., 2020b).

F.3. Finetuning with Fewer Data.

Method	mIoU with Reduced Train Set				
	1%	10%	20%	50%	100%
MoCo-v2	47.08	72.97	75.67	79.49	83.47
MoCo-v2+RGB	54.71	77.02	79.46	82.26	87.04
MoCo-v2+RGB+SAR	59.48	80.91	83.39	85.64	87.36

Method	mAP with Reduced Train Set			
	10%	20%	50%	100%
MoCo-v2	44.10	51.47	61.25	68.65
MoCo-v2+RGB	49.85	59.47	68.60	74.11
MoCo-v2+RGB+SAR	48.85	60.14	69.07	74.33

Table A3: Segmentation *(left)* and detection *(right)* results with downsampled train set. Finetuning our method with a smaller sampled training set achieves high performances similar to the baseline. Looking at the highlighted cells, our approach maintains high performances even with a smaller training set. When training sample size is decreased, our methods exhibits minimal performance decline compared to baseline SSL model.

To test the robustness of our method, we explore experiments training with a smaller samples for segmentation and detection. We train with randomly sampled sectors of the Potsdam (for Photogrammetry & , ISPRS) and DOTA (Xia et al., 2018) train set, and test with the same test set. The training data of Potsdam is sampled into 34, 345, 691, and 1728 images (1%, 10%, 20%, and 50% of original train split). DOTA train set is split into 2104, 4209, and 10523 images (10%, 20%, and 50%). Table A3 depicts comparative results of 3 settings, fMoW dataset pretrained with (i) MoCo-v2 (Chen et al., 2020b) only, (ii) with MoCo-v2 and RGB branch, and (iii) MoCo-v2 with both RGB and SAR branches. We report the segmentation and detection downstream tasks while progressively add our proposed branches.

As shown in Table A3, training with a smaller portion of the train set with our proposed method (*i.e*. SSL+RGB+SAR) outperforms baseline SSL model (*i.e*. MoCo-v2) trained with 100% train set of the downstream task. This observation signifies effectiveness of our methodologies, showcasing the ability to obtain competitive performances in downstream tasks with half the amount of training data. The results are promising even with a reduced number of images in the training subset, validating the capabilities of our methods.

F.4. Further Analysis on Multi-Sensor Approach

Pretraining					Seg. (mIoU)↑		Det. (mAP)↑
Dataset	Init.	Method	λ_{RGB}	λ_{SAR}	Potsdam	Vaihingen	DOTA
fMoW	✓	MoCo-v2	0.25	-	87.04	72.40	74.11
fMoW+SEN2	✓	MoCo-v2	0.25	-	86.92	73.25	73.35
fMoW+SEN1-2	✓	MoCo-v2	0.25	0.25	**87.36**	**73.55**	**74.33**

Table A4: Performance comparison of our proposed methods and Sentinel 2 RGB only on segmentation and detection downstream tasks.

To validate our proposed SAR branch that continually pretraining for additional 50 epochs, we conduct another experiment. Our suggested approach implements both RGB and SAR branch for 50 more epochs with Sentinel 1-2 Pair (SEN1-2) (Schmitt, 2018) dataset. To confirm the effectiveness multi-sensor approach, we pretrain with only the RGB Sentinel 2 images from the paired SEN1-2 dataset. Solely training for a longer time with the new RGB dataset hinders finetuning performance on Potsdam and DOTA. Implementing both sensors, RGB and SAR outperforms single-sensor pretraining on all finetuning tasks.

G. Qualitative Results of Segmentation Finetuning

In Figure A3, we visualize the segmentation inference results on the test set of Vaihingen dataset. Our proposed methods result in better segmentation maps similar to ground truth labels, compared to other existing approaches.

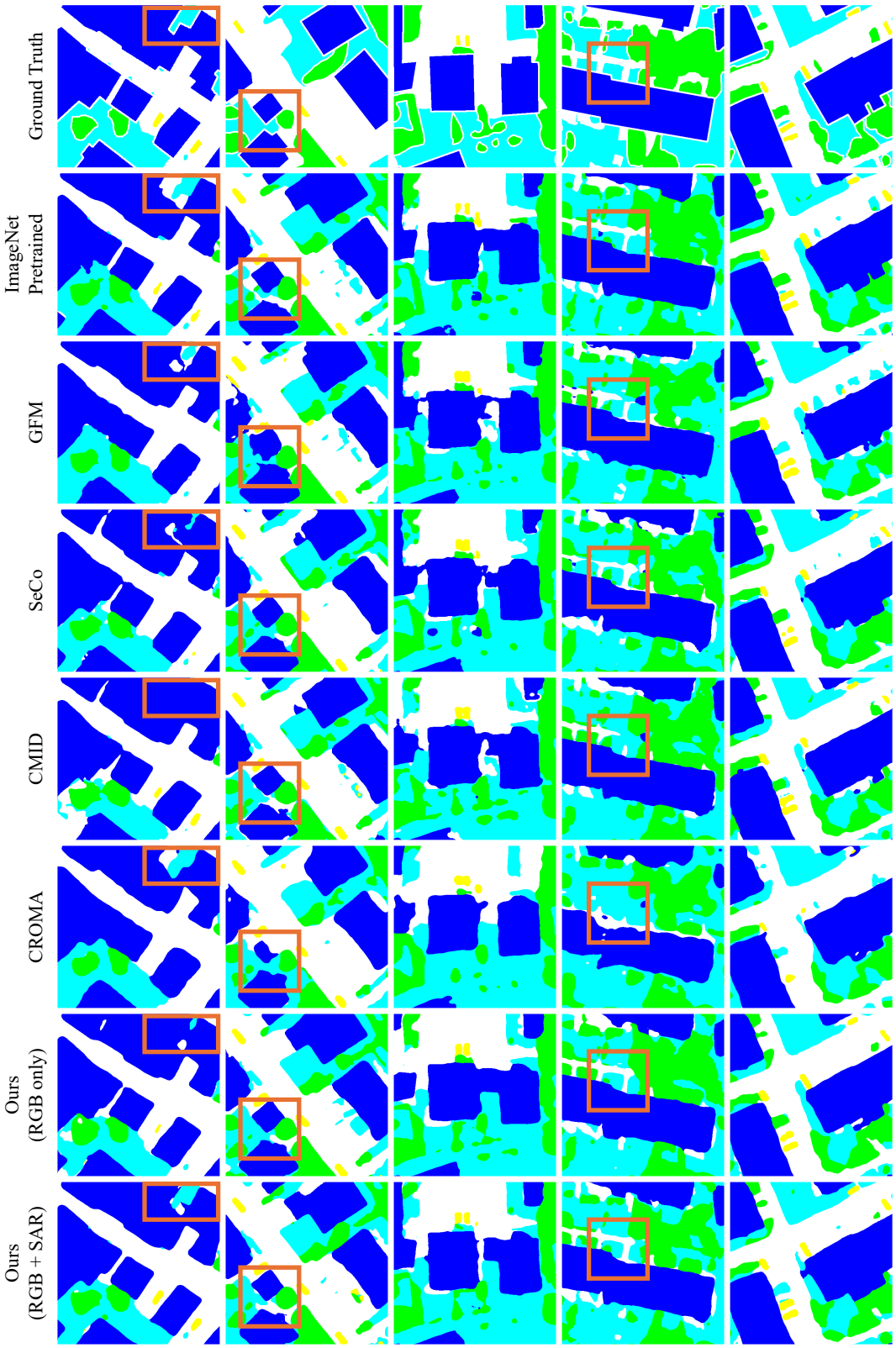


Figure A3: Visualization of different pretraining methods finetuned on Vaihingen dataset. The differences are emphasized with orange boxes. (Blue: Building, Cyan: Low Vegetation, Green: Tree, Yellow: Car, White: Impervious Surface).