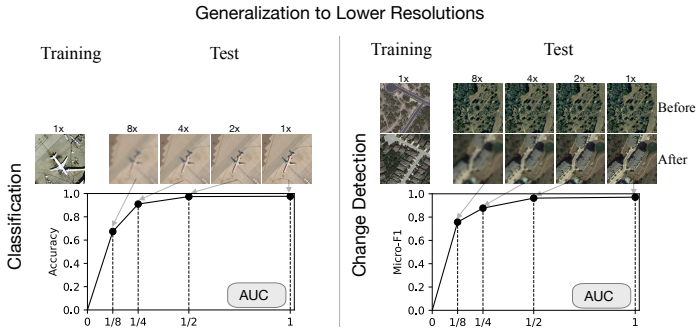
DO SATELLITE TASKS NEED SPECIAL PRETRAINING?

Ani Vanyan*[1,2], Alvard Barseghyan*[1,2], Hakob Tamazyan[1,2], Tigran Galstyan[1,2], Vahan Huroyan[3], Hrant Khachatrian[1,2]
[1] YerevaNN
[2] Yerevan State University
[3] Saint Louis University
{ani,alla,hakob,tigrann,hrant}@yerevann.com



ABSTRACT

Foundation models have advanced machine learning across various modalities, including remote sensing. Recent efforts focus on developing specialized models using masked image modeling techniques. This work explores whether specialized pretraining, such as training on large-scale remote sensing datasets or domain-specific techniques, is necessary for remote sensing foundation models. We introduce a benchmark evaluating model performance on tasks like change detection and scene classification using RESISC45, UC Merced, LEVIR-CD, and CDD datasets. We assess recent foundation models and analyze the impact of various pretraining and fine-tuning choices. Specifically, we pretrain self-distillation-based self-supervised models on aerial imagery, including variations without scale augmentations and with a pretrained mask decoder module.

1 INTRODUCTION

The rapid advancements in remote sensing technologies have led to an increased reliance on foundation models for interpreting vast amounts of imagery data captured by satellites (e.g., Sentinel-1, Sentinel-2) (Akiva et al., 2022; Mall et al., 2023; Mañas et al., 2021; Wanyan et al., 2023; Cong et al., 2022; Reed et al., 2023; Sun et al., 2023; Hong et al., 2024; Muhtar et al., 2023; Mendieta et al., 2023; Tang et al., 2023; Fuller et al., 2023; Bao et al., 2023; Guo et al., 2023; Wang et al., 2023b; Bastani et al., 2023). Usually, this data is raw and unlabeled, whereas creating labels is time-consuming and expensive. Many critical tasks, like change detection, image classification, and semantic segmentation (Used for land cover mapping, disaster monitoring, urban growth, vegetation health, and terrain analysis), require abundant labeled data for effective model training. In line with recent advancements in self-supervised and semi-supervised learning for vision tasks, the current trend is to train a self-supervised model (either contrastive or based on masked image modeling) which later serves as a backbone for subsequent downstream tasks. Subsequently, fine-tuning these backbones with a small labeled data produces a strong model for downstream tasks.

We evaluate the performance of foundation models for remote sensing imagery on scene classification (Cheng et al., 2017; Yang & Newsam, 2010; Sumbul et al., 2019) and change detection (Chen & Shi, 2020; Lebedev et al., 2018; Caye Daudt et al., 2018), assessing their generalization across image resolutions. To analyze the impact of design choices, we pretrain a self-distillation-based model

with variations, including one without scale augmentations and another with a pretrained mask decoder module. Our contributions include (a) developing a benchmark to evaluate remote sensing foundation models' generalization across scales, (b) pretraining multiple iBOT-based (Zhou et al., 2022) models, a self-distillation based ViT (Dosovitskiy et al., 2021), on MillionAID (Long et al., 2021), including one with a pretrained UperNet-like (Xiao et al., 2018) head for segmentation and change detection, and (c) demonstrating that existing foundation models (Bao et al., 2023; Bastani et al., 2023; Mendieta et al., 2023; Zhou et al., 2022) still have significant room for improvement in generalization and transferability to downstream tasks. The related work is discussed in Appendix A

2 EVALUATION

Generalization can be evaluated across various aspects, including adaptation to different spatial resolutions, spectral bands, seasonal variations, times of day, and diverse geographical locations. However, many of these assessments are constrained by dataset availability. In this work, we focus on evaluating the foundation model's ability to generalize to unseen resolutions across two key tasks: scene classification and change detection. We emphasize that our evaluation focuses solely on **generalization to lower spatial resolutions**. Low-resolution satellites, such as Landsat and Sentinel, provide publicly available imagery, whereas higher-resolution imagery is often more difficult to obtain. In many scenarios, image labeling is performed on high-quality imagery, but at test time, the images may come from satellites with lower resolution. Therefore, we expect models to perform robustly under such distribution shifts. While generalization to higher spatial resolutions can also occur in practical applications, retaining performance at higher resolutions is trivial by simply downsampling images to the original resolution.

Scene Classification. We use two commonly used benchmark datasets in the literature: RESISC45 (Cheng et al., 2017) and UC Merced (Yang & Newsam, 2010) (Appendix B). Performance is measured at the original resolution and at reduced resolutions (1/2, 1/4 and 1/8). Images are downscaled by a factor of $1/x$ and then upscaled back by x, preserving pixel count but reducing quality. This simulates lower-resolution satellite imagery. As an evaluation metric, we plot a curve with the scaling factor ($1/8$, $1/4$, $1/2$, 1) on the x-axis and accuracy on the y-axis. The area under this curve (**AUC-Acc**) serves as our final metric. We restrict the models to use 50 GFLOPs on a single image. This threshold is independent from the neural architecture, and ViT-B/16 on an image of size 256x256px is within the limits.

Change Detection. We use another two commonly used datasets: CDD (Lebedev et al., 2018) and LEVIR-CD (Chen & Shi, 2020); see Appendix B. We create partially scaled versions of the test sets of these datasets. We maintain the scale of the first image unchanged, while for the second image, we distort it by reducing its quality by a factor of 2, 4, and 8. Note that a similar setup has been first proposed in (Liu et al., 2022). We evaluate on the original resolution, as well as on the scaled versions. We compute micro-averaged F1 score for each of the versions. Finally we draw a curve where x-axis is the scaling parameter and y-axis is the micro-averaged F1 score for each version. We report the area under this curve as our final metric, and call it **AUC-F1**. For this benchmark, we restrict the models to use 100 GFLOPs on a pair of images.

3 FACTORS CONTRIBUTING TO THE PERFORMANCE

iBOT pretraining. We analyzed various factors of generalization of fine-tuned models by pretraining several iBOT models on satellite imagery. As shown by Vanyan et al. (2023a), self-

LEVIR-CD	1:1	1:2	1:4	1:8	AUC-F1
Without Mask Decoder	90.6 ± 0.2	87.6 ± 0.9	50.4 ± 15.1	2.0 ± 1.0	65.2 ± 3.2
With Mask Decoder	90.6 ± 0.1	89.2 ± 0.1	66.6 ± 5.0	4.3 ± 1.1	69.1 ± 1.0
CDD					
Without Mask Decoder	97.4 ± 0.0	96.8 ± 0.0	91.4 ± 0.6	79.2 ± 0.9	87.7 ± 0.2
With Mask Decoder	97.1 ± 0.0	96.7 ± 0.0	91.5 ± 0.5	80.1 ± 0.9	87.7 ± 0.2

Table 1: The effect of a pretrained mask decoder on change detection tasks. All models are iBOTs pretrained on MillionAID with scale augmentation.

distillation models like iBOT outperform MIM-based models in learning robust image representations. We pre-trained iBOT with the MillionAID dataset (Long et al., 2021), dividing images into a maximum of 550-pixel square tiles, yielding 2106700 images. Since the original iBOT pre-trained on ImageNet is already strong, we included it in our comparisons. We trained iBOT for 200 epochs with peak learning rate 5×10^{-4} that linearly decreases to 2×10^{-6} over 5 warmup epochs. All RandomResizeCrops were converted to RandomCrops in the transforms. The training was conducted using PyTorch Distributed Data Parallel to utilize multiple GPUs and used 100 batch size per GPU. The experiments were performed on NVIDIA DGX A100 at the local university and an instance with 8 NVIDIA H100s kindly provided by Nebius.ai. The loss curve followed the typical pattern of similar networks (Figure 3 in Appendix). The resulting model is labeled as **iBOT-MillionAID**.

Augmentation. We analyze scale augmentation's impact on robustness to scale changes. iBOT's augmentation module resizes and crops images. We pre-trained two iBOTs: with and without resizing. The hypothesis is that scale augmentation improves robustness, transferring to fine-tuned models and increasing AUC scores. We also test scale augmentation during fine-tuning by shrinking images (or the second image in change detection) by 2, 4, and 8 times, then resizing them back These serve as an upper bound for scale robustness (see Table 5 of Appendix).

When augmentations are not applied during fine-tuning, augmentations during pretraining at 1:1 and 1:2 resolutions consistently give better results across all datasets. However, this trend does not hold for smaller resolutions. Augmentations during fine-tuning have significantly higher impact on the generalization. In case of classification, we leverage $2\times$, $4\times$, and $8\times$ versions of the original dataset. Although we obtain $4\times$ more data, this does not add new information, and we keep the total number of optimization steps constant by decreasing the number of epochs by $4\times$. In case of change detection, we randomly choose one of the augmented versions of the second image at each epoch, and train for the same number of epochs as in the experiment without augmentations. These experiments indicate that scale augmentation during pretraining still does not produce generalization capabilities at a level comparable to what one can obtain by augmenting during fine-tuning.

RESISC45					AUC-ACC
Full fine-tuning	93.4 ± 0.2	84.3 ± 1.2	47.4 ± 5.6	18.7 ± 2.0	66.2 ± 1.8
Frozen backbone	$\mathbf{94.6 \pm 0.1}$	$\mathbf{92.2 \pm 0.2}$	$\mathbf{66.5 \pm 1.5}$	$\mathbf{25.1 \pm 1.3}$	$\mathbf{73.8 \pm 0.5}$
LEVIR-CD	1:1	1:2	1:4	1:8	AUC-F1
Full fine-tuning	90.6 ± 0.2	87.6 ± 0.9	50.4 ± 15.1	2.0 ± 1.0	65.2 ± 3.2
Frozen backbone	84.4 ± 0.0	84.4 ± 0.2	61.6 ± 7.8	3.4 ± 4.0	64.7 ± 2.0
UC Merced					AUC-ACC
Full fine-tuning	98.7 ± 0.8	97.9 ± 1.3	84.3 ± 4.3	46.0 ± 8.3	82.9 ± 1.0
Frozen backbone	99.5 ± 0.1	99.2 ± 0.3	75.7 ± 2.9	31.3 ± 3.9	80.2 ± 0.7

Table 2: The impact of full fine-tuning on loss. All models are iBOTs pretrained on MillionAID with scale augmentation. No scale-augmentation was performed during fine-tuning (or linear probing).

Pretrained Mask Decoder. We extend iBOT-MillionAID with a pretrained mask decoder for segmentation and change detection tasks, requiring a binary mask, and leverage a module pretrained on large datasets. Since MillionAID lacks segmentation or change masks, we use iBOT's teacher-student framework to generate them. The teacher processes two global crops, while the student handles those plus ten local crops. We map the second global crop's mask to the first crop's coordinate space as the target mask. Patch representations from both crops are concatenated and fed into an UperNet (Xiao et al., 2018) decoder to generate the binary mask with a pixel-wise cross-entropy loss. Note that UperNet uses features from ViT layers 3, 5, 8, and 12. We explored two methods to integrate mask loss into iBOT training: using only the student for patch representations or incorporating the teacher for one. The first approach led to unstable training with spiking activations, while the teacher-student method ensured stable joint training. The final architecture is shown in Fig. 1.

We used 2.5×10^{-4} peak learning rate and cosine decay with 5 warmup epochs. The model is trained for $\approx$ 800 H100 GPU hours on an instance with 8 NVIDIA H100s provided by Nebius.ai.

As shown in Table 1, there is a slight improvement in performance and significantly lower variance across all scales with the pretrained mask decoder on LEVIR-CD. There is no visible change on CDD. This can be explained by the large size of the CDD dataset. It is likely that the additional power of the pretrained models is not critical when the fine-tuning dataset is large enough. Another

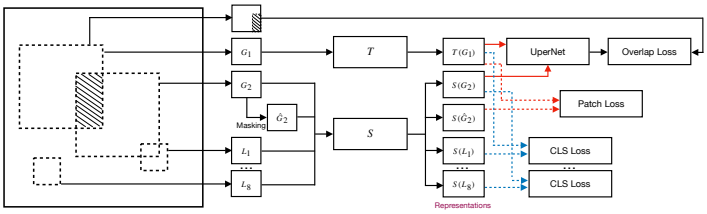


Figure 1: iBOT pretraining architecture with an additional UperNet mask decoder that is trained using the "overlap loss". There are two global and eight local crops of the original image that pass through Teacher (T) and Student (S) networks. Dotted lines imply that only the representations of the last layers are used. Solid lines imply that representations of four layers are used (as an input to UperNet). Red lines correspond to patch representations, the blue lines correspond to CLS vectors.

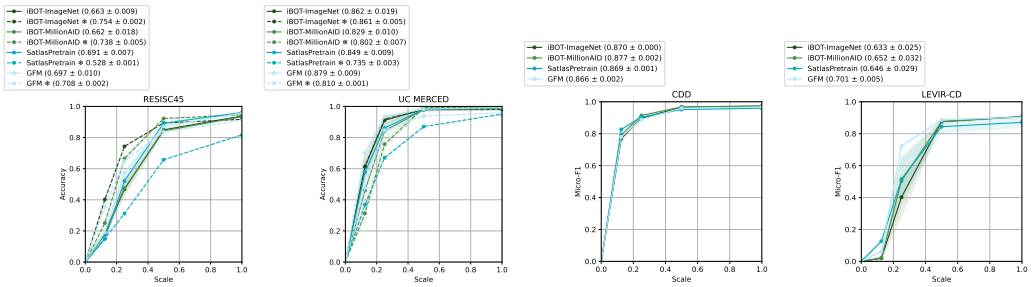


Figure 2: The results of the baselines on our benchmark tasks for generalization across image resolution. The top row shows classification on RESISC and UC Merced, while the bottom row shows change detection on CDD and LEVIR-CD. X-axis: Scale of Distortions, Y-axis: Micro-F1 Scores.

potential way to enhance the impact of pretrained decoders is to pretrain it with denser supervision signal. While we used a binary mask calculated during pretraining, Wang et al. (2024) uses segmentation pseudo-labels generated by a strong domain-agnostic segmentation model.

Catastrophic Forgetting During Fine-Tuning. Pretrained models may lose generalization during fine-tuning. To assess this, we repeat fine-tuning with frozen backbones, ensuring the final linear layer or decoder lacks exposure to diverse scales. Table 2 shows that the effect varies by dataset. For RESISC45, freezing the backbone improves robustness to lower resolutions. LEVIR-CD follows this trend at 1:4 and 1:8 resolutions, though full fine-tuning performs better at 1:1 and 1:2. In contrast, UC Merced benefits from a frozen backbone at higher resolutions, while full fine-tuning excels at lower resolutions.

4 BASELINES

We used SatlasPretrain (Bastani et al., 2023) trained on high-resolution imagery (Aerial) and on the RGB subset of Sentinel-2 imagery (Sentinel2), GFM (Mendieta et al., 2023), and general-purpose iBOT pretrained on ImageNet as baseline. Each of these models have a different training paradigm and pretraining dataset. iBot is a self-supervised method pretrained on ImageNet. GFM combines two concepts: self-supervised pretraining on a custom-collected dataset, GeoPile, and continual pretraining to retain knowledge obtained from pretraining on ImageNet. SatlasPretrain is pretrained on a custom-collected dataset, Satlas, in a supervised manner. Prithvi (Jakubik et al., 2023) is a modification of a MAE model to support 3D inputs with 6 channels.

Experimental Setup. To adapt the models for classification, we add a linear layer on top of the [CLS] token representation, if available, or on top of the global average pooled vector of all patch representations. To test the models for change detection, we take the backbone, which is either a Swin Transformer, or a ViT, and integrate the UperNet head (Xiao et al., 2018). The two source images go through identical backbones, and the resulting representations are substracted from each other and passed to the head. In the case of ViTs, we use an additional *neck* module between the

backbone and UperNet. The backbone is initialized with the pre-trained weights and further fine-tuned using the change detection datasets. In case of our iBOT trained on MillionAID, the neck and the head modules are also initialized, and we take the concatenation of features instead of the difference. For more details see Appendix C.

Results and Conclusions. The results are shown in Figure 2, (more detailed results are in Table 3 in Appendix). The general conclusion is that all tested models struggle with generalizability across scales. There are cases where the same model with a frozen backbone performs slightly better than its fine-tuned counterpart (e.g. SatlasPretrain and DINOv2). There is a noticable performance gap for classification, with the SatlasPretrain model, which could be due to its supervised pretraining. However, we can observe that, when fine-tuned on larger datasets, the weakness of supervised pretraining becomes less significant, as seen with SatlasPretrain on BigEarthNet. As mentioned in some studies (He et al., 2022; Vanyan et al., 2023b), models trained with masked image modeling exhibit their advantages when fully fine-tuned; their representations are not designed for linear probing.

Compute constraints. Foundation models should target specific compute requirements. Many downstream applications require the models to run on low power devices or need to support large volumes of data in deployment, and hence require limited number of FLOPs per image. It is important to note that this requirement refers to the fine-tuning stage and the deployment of the final model, and not to the pretraining process. For example, DINOv2 (Oquab et al., 2023) has a ViT-B version which is distilled from a larger ViT-g model. While the large model was trained using hundreds of GPUs, the distilled version can be easily fine-tuned on a single consumer-grade GPU. To keep the comparisons fair, for all models, we relatively small versions with around 100M parameters and trained on fewer than 3M images.

ACKNOWLEDGEMENTS

This work was supported by the Higher Education and Science Committee of RA (Research project No 24RL-1B049). Hakob Tamazyan's work was supported by Yandex Armenia fellowship. Ani Vanyan's work was supported by Fimetech fellowship. Alvard Barseghyan's work was supported by Layerswap scholarship. Vahan Huroyan's work was supported by Higher Education and Science Committee of RA (Research project No 23PostDoc-1B009).

REFERENCES

Peri Akiva, Matthew Purri, and Matthew J. Leotta. Self-supervised material and texture representation learning for remote sensing tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8193–8205. IEEE, 2022.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth C x 16 x 16 words. *CoRR*, abs/2309.16108, 2023.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16772–16782, 2023.

R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, July 2018.

Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote. Sens.*, 12(10):1662, 2020.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE*, 105(10):1865–1883, 2017.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Anthony Fuller, Koreen Millard, and James R. Green. CROMA: remote sensing representations with contrastive radar-optical masked autoencoders. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *CoRR*, abs/2312.10115, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 15979–15988. IEEE, 2022.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi (Steve) Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation Models for Generalist Geospatial Artificial Intelligence. *Preprint Available on arxiv:2310.18660*, October 2023.

MA Lebedev, Yu V Vizilter, OV Vygolov, Vladimir A Knyaz, and A Yu Rubis. Change detection in remote sensing images using conditional adversarial networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:565–571, 2018.

Mengxi Liu, Qian Shi, Andrea Marinoni, Da He, Xiaoping Liu, and Liangpei Zhang. Super-resolution-based change detection network with stacked attention module for images with different resolutions. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–18, 2022.

Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 14:4205–4230, 2021.

Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 5261–5270. IEEE, 2023.

Oscar Mañas, Alexandre Lacoste, Xavier Giró-i-Nieto, David Vázquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9394–9403. IEEE, 2021.

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 16760–16770. IEEE, 2023.

Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Zhenshi Li, and Feng Gu. CMID: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–17, 2023.

Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4065–4076. IEEE, 2023.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *2019 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2019, Yokohama, Japan, July 28 - August 2, 2019*, pp. 5901–5904. IEEE, 2019.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–22, 2023. doi: 10.1109/TGRS.2022.3194732. URL https://doi.org/10.1109/TGRS.2022.3194732.

Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale MAE: A tale of multiscale exploitation in remote sensing. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V. Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, Theo Moutakanni, Piotr Bojanowski, Tracy Johns, Brian White, Tobias Tiecke, and Camille Couprie. Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment*, 300:113888, 2024. ISSN 0034-4257.

Ani Vanyan, Alvard Barseghyan, Hakob Tamazyan, Vahan Huroyan, Hrant Khachatrian, and Martin Danelljan. Analyzing local representations of self-supervised vision transformers. *arXiv preprint arXiv:2401.00463*, 2023a.

Ani Vanyan, Alvard Barseghyan, Hakob Tamazyan, Vahan Huroyan, Hrant Khachatrian, and Martin Danelljan. Analyzing local representations of self-supervised vision transformers. *arXiv preprint arXiv:2401.00463*, 2023b.

Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–20, 2023a.

Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang. MTP: advancing remote sensing foundation model via multi-task pretraining. *CoRR*, abs/2403.13430, 2024.

Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decur: decoupling common & unique representations for multimodal self-supervision. *CoRR*, abs/2309.05300, 2023b.

Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. DINO-MC: self-supervised contrastive learning for remote sensing imagery with multi-sized local crops. *CoRR*, abs/2303.06670, 2023.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. *CoRR*, abs/1807.10221, 2018.

Yi Yang and Shawn D. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In Divyakant Agrawal, Pusheng Zhang, Amr El Abbadi, and Mohamed F. Mokbel (eds.), *18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings*, pp. 270–279. ACM, 2010.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

A Related Work

Some recent developments in the field include various approaches using either supervised or self-supervised learning algorithms. Surprisingly, for some transformer-based models, performance on ImageNet (Deng et al., 2009) in certain instances outperforms those pre-trained on remote sensing imagery (Vanyan et al., 2023a). The effect of pre-training on ImageNet vs a large remote sensing scene recognition dataset is studied in Remote Sensing Pretraining (RSP) (Wang et al., 2023a). To serve as a pre-training dataset, some existing techniques involve gathering data from available open-source large remote sensing datasets and employing it to train the self-supervised algorithm. The two main methods to train self-supervised foundation models are contrastive learning-based methods and generative-based methods (masked image modeling).

Similar to classical contrastive learning-based methods, recent advancements include SECO (Mañas et al., 2021), CACo (Mall et al., 2023), MATTER (Akiva et al., 2022), Dino-MC (Wanyan et al., 2023), and (Tolan et al., 2024), among others. Another line of research builds on Masked Autoencoders (MAE) (He et al., 2022), a successful foundation model utilizing masked image modeling, where the pretext task is to reconstruct an image from its masked version. Notable extensions include SatMAE (Cong et al., 2022), Scale-MAE (Reed et al., 2023), RingMO (Sun et al., 2023), and SpectralGPT (Hong et al., 2024). A more recent direction aims to integrate reconstruction-based and contrastive learning-based approaches. Notable examples include CMID (Muhtar et al., 2023), GFM (Mendieta et al., 2023), SECO (Mañas et al., 2021), CROMA (Fuller et al., 2023), and Cross-Scale-MAE (Tang et al., 2023). Mendieta et al. (2023) observed that some state-of-the-art methods for aerial imagery often do not outperform ImageNet-22k pretrained Vision Transformers (ViTs). Another research focus is multi-task pretraining, with works such as Satlas (Bastani et al., 2023) and Multi-Task Pretraining (MTP) (Wang et al., 2024). Recently, for change detection, an end-to-end super-resolution-based network, SRCDNet (Liu et al., 2022), was introduced to address change detection across varying image resolutions. We extend this idea to additional classification and change detection datasets.

B Datasets

RESISC45 (Cheng et al., 2017) and UC Merced (Yang & Newsam, 2010) datasets contain 256x256px images. Image resolution is 30cm/px for UC Merced and varies 20-600cm/px for RESISC45. Both datasets use RGB bands only. We take the splits defined in (Neumann et al., 2019).
The LEVIR-CD dataset (Chen & Shi, 2020) comprises a substantial collection of bitemporal Google Earth images. It includes 637 image pairs, each sized 1024×1024px, with 400 images designated for training. The images in the training set have a resolution of 50cm/px. Originating from 20 distinct regions within cities in Texas, USA, these images showcase the construction-induced changes. The fully annotated LEVIR-CD dataset encompasses a total of $31,333$ individual changed buildings. The changes in the LEVIR-CD dataset primarily come from the construction of new buildings. The average size of each changed area is approximately 987 pixels.
The CDD (Lebedev et al., 2018) dataset contains season-varying remote sensing images of the same region, obtained from Google Earth (DigitalGlobe). The dataset comprises $16,000$ image sets (two images of the same location and the annotated change), each with an image size of 256×256 pixels and 0.03-1m/px ground sample distance.
Onera Satellite Change Detection (OSCD) dataset contains pairs of aerial images of the same location captured at different times, with changes manually annotated at the pixel level (Caye Daudt et al., 2018). The dataset contains images from a total of 24 cities, divided into smaller chunks (192×192) of images. Similar to the classification benchmark, we train on the RGB channels and evaluate on four tri-channel triplets and one bi-channel pair: RGB, RGE1, RE1E2, N'S1S2, and VV VH (bi-channel). We note that for the evaluation, we always keep the first picture as RGB and the second figure with the corresponding band channels. We compute the micro F1 score for each experiment and report the average over these five values.

C Implementation Details

All the codes for pretraining, as well as the benchmarks proposed by us with all the hyperparameters, can be found at: `https://anonymous.4open.science/r/rs_foundation_models-42DC/README.md`.

C.1 Classification

We perform two kinds of fine-tuning: full fine-tuning and linear probing. For both setups, we train for 100 epochs. For all experiments in the full fine-tuning setup or linear probing, we evaluate using the last checkpoint. However, for full fine-tuning on the BigEarthNet dataset, we select the best checkpoint based on performance on the validation set. In all experiments within the full fine-tuning setup, we use the $AdamW$ optimizer with a learning rate of 10^{-4} employing $WarmupCosineAnnealing$ scheduling and an estimated minimum value of 10^{-5}. In experiments within the linear probing setup, we use the $AdamW$ optimizer with a learning rate of 10^{-3} employing $MultiStep$ scheduling and an estimated minimum value of 10^{-5}.

In the linear probing setup for the Prithvi model, we conducted a grid search to optimize the hyperparameters. The optimization process involved testing three different optimizers: $\{Adam, AdamW, SGD\}$. For the learning rate, we evaluated three values: $\{10^{-3}, 10^{-4}, 10^{-6}\}$ setting one of the following schedulers: $\{MultiStep, WarmupCosineAnnealing\}$. We selected the $AdamW$ optimizer with a learning rate of 10^{-3} and the $WarmupCosineAnnealing$ scheduler for our final configuration based on the performance on the validation set. For linear probing with the ChannelVit model, we use the initial hyperparameters for linear probing provided by the authors and perform the same grid search. Ultimately, we choose the $Adam$ optimizer with an initial learning rate of 10^{-3} and $MultiStep$ scheduling.

C.2 Change Detection

For change detection experiments, we train our models for 200 epochs. We use the $AdamW$ optimizer with a learning rate of 6×10^{-5} along with $WarmupCosineAnnealing$ which includes warmup steps of 10 and batch size of 32. For experiments on OSCD dataset we choose learning rate 3×10^{-5} decrease the training epochs to 100 and use warmup steps of 5 with a batch size of 4.

D Detailed Results

In Table 3 we present the benchmark results for proposed and existing models in change detection (LEVIR-CD and CDD) and classification (RESISC45 and UC Merced). For classification, we demonstrate results for both full fine-tuning and linear probing. All experiments are conducted with scale distortions of 1:1, 1:2, 1:4, and 1:8. The AUC-F1 score is reported for change detection, and the AUC-ACC score is reported for classification. For change detection, we compare iBOT trained on ImageNet, our trained iBOT for MillionAID, Satlas, and GFM. For the LEVIR-CD dataset, the results are generally comparable across methods. However, GFM shows a clear advantage over the other methods for the 1:2 and 1:4 scale distortions. Specifically, while all four methods produce comparable results at 1:2, GFM demonstrates a clear advantage at 1:4. However, we remark that the pretraining dataset for GFM GeoPile contains RESISC45, which could possibly cause its superior performance over the other methods. For CDD dataset, we observe that all the results are comparable, however, we observe that GFM does not have superior performance over the other methods. The little AUC-F1 score difference between various scale distortions could be explained by the fact that the CDD dataset contains samples from different GSD (0.03m-1m). For classification, we compare iBOT trained on ImageNet, our trained iBOT for MillionAID, the two versions of Satlas and GFM. We observe that for iBOT (both trained on ImageNET and MillionAID) linear probing has a clear advantage over full-finetuning for lower resolutions.

In Table 4, we report the performance of our trained iBOT on the MillionAID dataset, comparing results with and without augmentations, as well as between a frozen backbone or linear probing and full fine-tuning. For change detection on the LEVIR-CD dataset, we observe that full fine-tuning has a clear advantage over a frozen backbone. Additionally, we note that augmentations do not improve performance for this task. For the classification task (RESISC45 and UC Merced), we

Table 3: Benchmark Results for Change Detection (LEVIR-CD, CDD) and Classification (RESISC45, UC Merced) tasks with Different Scale Distortions.

LEVIR-CD	**1:1**	**1:2**	**1:4**	**1:8**	**AUC-F1**
iBOT-ImageNet	90.7 ± 0.1	87.6 ± 0.5	40.2 ± 12.0	2.0 ± 1.4	63.3 ± 2.5
iBOT-MillionAID	90.6 ± 0.2	87.6 ± 0.9	50.4 ± 15.1	2.0 ± 1.0	65.2 ± 3.2
SatlasPretrain (S2_SwinB_SI_RGB)	87.1 ± 3.2	84.4 ± 3.5	51.5 ± 12.4	12.6 ± 1.8	64.6 ± 2.9
GFM	90.3 ± 1.1	88.6 ± 1.0	72.3 ± 1.5	6.2 ± 1.1	70.1 ± 0.5
Prithvi	85.2 ± 0.1	84.4 ± 0.1	76.4 ± 1.1	14.5 ± 1.2	69.1 ± 0.4
DINOv2	88.0 ± 0.1	86.5 ± 0.2	70.4 ± 1.5	12.2 ± 2.5	69.1 ± 0.6
CDD					**AUC-F1**
iBOT-ImageNet	97.3 ± 0.0	96.6 ± 0.0	89.7 ± 0.2	76.9 ± 0.4	87.0 ± 0.0
iBOT-MillionAID	97.4 ± 0.0	96.8 ± 0.0	91.4 ± 0.6	79.2 ± 0.9	87.7 ± 0.2
SatlasPretrain (S2_SwinB_SI_RGB)	96.0 ± 0.0	95.1 ± 0.0	90.4 ± 0.3	82.7 ± 0.4	86.9 ± 0.1
GFM	96.8 ± 0.0	96.0 ± 0.1	88.9 ± 0.3	78.0 ± 0.6	86.6 ± 0.2
Prithvi	90.9 ± 0.2	90.5 ± 0.2	88.5 ± 0.3	82.9 ± 0.8	83.6 ± 0.3
DINOv2	92.4 ± 0.0	91.3 ± 0.1	87.5 ± 0.1	78.2 ± 0.1	83.5 ± 0.0
RESISC45: full fine-tuning					**AUC-ACC**
iBOT-ImageNet	93.8 ± 0.2	84.9 ± 0.8	46.8 ± 3.3	18.1 ± 0.7	66.3 ± 0.9
iBOT-MillionAID	93.4 ± 0.2	84.3 ± 1.2	47.4 ± 5.6	18.7 ± 2.0	66.2 ± 1.8
DINOv2	94.1 ± 0.4	84.3 ± 1.7	46.7 ± 5.2	19.3 ± 2.6	66.3 ± 1.6
SatlasPretrain (S2_SwinB_SI_RGB)	96.1 ± 0.1	89.2 ± 1.2	61.4 ± 3.3	23.7 ± 2.6	71.9 ± 1.4
SatlasPretrain (Aerial_SwinB_SI)	96.1 ± 0.1	89.2 ± 0.6	52.1 ± 2.3	14.9 ± 1.5	69.1 ± 0.7
GFM	95.7 ± 0.1	87.1 ± 0.9	57.4 ± 3.4	19.1 ± 3.0	69.7 ± 1.0
RESISC45: linear probing					**AUC-ACC**
iBOT-ImageNet	91.7 ± 0.1	89.3 ± 0.2	74.3 ± 0.6	40.2 ± 0.9	75.4 ± 0.2
iBOT-MillionAID	94.6 ± 0.1	92.2 ± 0.2	66.5 ± 1.5	25.1 ± 1.3	73.8 ± 0.5
DINOv2	91.1 ± 0.7	87.2 ± 1.0	72.9 ± 1.4	40.3 ± 1.0	74.2 ± 0.9
SatlasPretrain (S2_SwinB_SI_RGB)	72.8 ± 0.1	58.0 ± 0.2	25.4 ± 0.4	15.0 ± 0.3	46.6 ± 0.1
SatlasPretrain (Aerial_SwinB_SI)	81.7 ± 0.1	65.7 ± 0.1	31.1 ± 0.3	15.1 ± 0.1	52.8 ± 0.1
GFM	91.1 ± 0.0	83.6 ± 0.1	64.9 ± 0.4	35.6 ± 0.6	70.8 ± 0.2
UC Merced: full fine-tuning					**AUC-ACC**
iBOT-ImageNet	98.6 ± 0.7	98.2 ± 1.0	91.0 ± 2.7	61.3 ± 7.7	86.2 ± 1.9
iBOT-MillionAID	98.7 ± 0.8	97.9 ± 1.3	84.3 ± 4.3	46.0 ± 8.3	82.9 ± 1.0
DINOv2	98.1 ± 0.5	97.9 ± 0.3	98.1 ± 0.4	97.3 ± 0.3	91.8 ± 0.1
SatlasPretrain (S2_SwinB_SI_RGB)	98.7 ± 0.2	98.0 ± 0.3	87.3 ± 2.6	61.9 ± 5.9	85.5 ± 1.3
SatlasPretrain (Aerial_SwinB_SI)	99.1 ± 0.2	98.1 ± 0.3	86.1 ± 3.1	57.7 ± 3.9	84.9 ± 0.9
GFM	99.2 ± 0.2	98.3 ± 0.6	93.3 ± 1.6	69.9 ± 3.8	87.9 ± 0.9
UC Merced: linear probing					**AUC-ACC**
iBOT-ImageNet	98.0 ± 0.3	97.9 ± 0.3	91.8 ± 0.7	61.4 ± 3.6	86.1 ± 0.5
iBOT-MillionAID	99.5 ± 0.1	99.2 ± 0.32	75.7 ± 2.9	31.3 ± 3.9	80.2 ± 0.7
DINOv2	97.4 ± 0.2	97.0 ± 0.1	96.8 ± 0.1	91.8 ± 0.4	90.3 ± 0.1
SatlasPretrain (S2_SwinB_SI_RGB)	85.7 ± 0.8	79.6 ± 0.4	55.6 ± 1.6	27.2 ± 0.5	65.1 ± 0.3
SatlasPretrain (Aerial_SwinB_SI)	95.0 ± 0.3	87.0 ± 0.4	67.0 ± 0.8	36.8 ± 0.3	73.5 ± 0.3
GFM	95.8 ± 0.1	93.9 ± 0.2	84.7 ± 0.4	47.7 ± 0.4	81.0 ± 0.1

observe that for both full fine-tuning and linear probing the model trained with augmentations has a clear advantage over the one trained without augmentation.

Experiments with augmentations and the results of the default setup for RESISC45 and CDD datasets show that the diversity of the dataset in terms of real resolutions (GSD) improves the generalization capabilities of the finetuned model, even if the backbone weights are frozen.

Table 4: The impact of full fine-tuning on the loss of generalization capabilities. All models are iBOTs pretrained on MillionAID.

LEVIR-CD: full fine-tuning	1:1	1:2	1:4	1:8	AUC-F1
iBOT-MillionAID	88.7 ± 0.1	86.5 ± 0.2	63.6 ± 3.3	7.5 ± 0.5	67.5 ± 0.7
iBOT-MillionAID-augm	90.6 ± 0.2	87.6 ± 0.9	50.4 ± 15.1	2.0 ± 1.0	65.2 ± 3.2
LEVIR-CD: frozen backbone					
iBOT-MillionAID	81.5 ± 0.1	81.0 ± 0.4	69.3 ± 3.1	17.0 ± 7.9	65.9 ± 1.6
iBOT-MillionAID-augm	84.4 ± 0.0	84.4 ± 0.2	61.6 ± 7.8	3.4 ± 4.0	64.7 ± 2.0
RESISC45: full fine-tuning					AUC-ACC
iBOT-MillionAID	94.6 ± 0.2	92.8 ± 0.3	70.4 ± 4.0	16.6 ± 4.0	73.7 ± 1.3
iBOT-MillionAID-augm	93.4 ± 0.2	84.3 ± 1.2	47.4 ± 5.6	18.7 ± 2.0	66.2 ± 1.8
RESISC45: linear probing					
iBOT-MillionAID	91.0 ± 0.1	87.5 ± 0.1	60.8 ± 0.2	9.3 ± 0.2	68.1 ± 0.1
iBOT-MillionAID-augm	94.6 ± 0.1	92.2 ± 0.2	66.5 ± 1.5	25.1 ± 1.3	73.8 ± 0.5
UC Merced: full fine-tuning					
iBOT-MillionAID	98.0 ± 0.3	97.2 ± 0.6	87.2 ± 1.9	38.7 ± 3.0	82.2 ± 0.7
iBOT-MillionAID-augm	98.7 ± 0.8	97.9 ± 1.3	84.3 ± 4.3	46.0 ± 8.3	82.9 ± 1.0
UC Merced: linear probing					
iBOT-MillionAID	96.9 ± 0.0	97.1 ± 0.2	93.6 ± 0.2	34.0 ± 1.3	82.5 ± 0.2
iBOT-MillionAID-augm	99.5 ± 0.1	99.2 ± 0.32	75.7 ± 2.9	31.3 ± 3.9	80.2 ± 0.7

Table 5: Dependence of the performance of fine-tuned models on scale augmentation performed during pretraining and fine-tuning. All models are iBOTs trained on MillionAID.

Augmentation Phase	1:1	1:2	1:4	1:8	
LEVIR-CD					AUC-F1
Pretraining / Fine-tuning	88.7 ± 0.1	86.5 ± 0.2	63.6 ± 3.3	7.5 ± 0.5	67.5 ± 0.7
Pretraining / Fine-tuning	90.6 ± 0.2	87.6 ± 0.9	50.4 ± 15.1	2.0 ± 1.0	65.2 ± 3.2
Pretraining / **Fine-tuning**	88.2 ± 0.1	88.4 ± 0.1	87.9 ± 0.1	86.1 ± 0.1	82.4 ± 0.1
Pretraining / **Fine-tuning**	89.9 ± 0.1	89.9 ± 0.1	89.4 ± 0.1	87.7 ± 0.1	83.9 ± 0.1
CDD					AUC-F1
Pretraining / Fine-tuning	95.8 ± 0.0	95.3 ± 0.0	92.3 ± 0.1	80.1 ± 0.5	87.0 ± 0.1
Pretraining / Fine-tuning	97.4 ± 0.0	96.8 ± 0.0	91.4 ± 0.6	79.2 ± 0.9	87.7 ± 0.2
UC Merced					AUC-ACC
Pretraining / Fine-tuning	98.0 ± 0.3	97.2 ± 0.6	87.2 ± 1.9	38.7 ± 3.0	82.2 ± 0.7
Pretraining / Fine-tuning	98.7 ± 0.8	97.9 ± 1.3	84.3 ± 4.3	46.0 ± 8.3	82.9 ± 1.0
Pretraining / **Fine-tuning**	98.2 ± 0.6	98.3 ± 0.6	98.0 ± 0.6	95.7 ± 1.2	91.8 ± 0.6
Pretraining / **Fine-tuning**	95.3 ± 1.8	94.7 ± 2.0	94.0 ± 2.4	91.8 ± 3.6	88.4 ± 2.1

In Figure 3 the left subfigure shows the iBOT loss (total training loss and its components) trained on the MillionAID dataset. The right subfigure displays the iBOT loss (total training loss and its components: train cls, train patch, and train overlap) for the model trained on the MillionAID dataset with the additional mask decoder proposed by us.

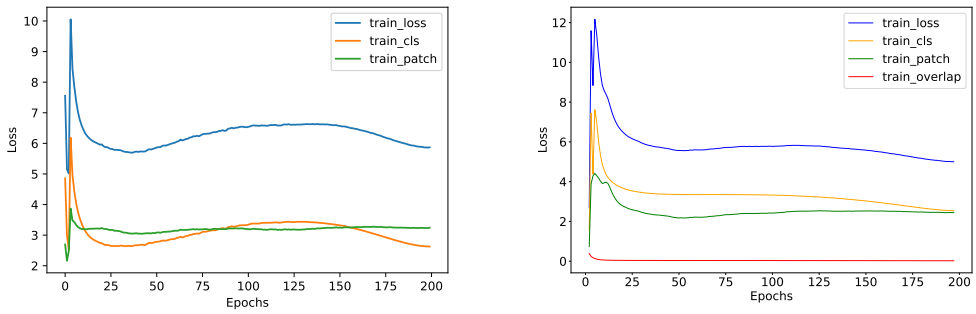


Figure 3: Overall loss and loss components of the iBOT trained on MillionAID dataset for 200 epochs with scale augmentation and without a mask decoder on the left and with mask decoder on the right.