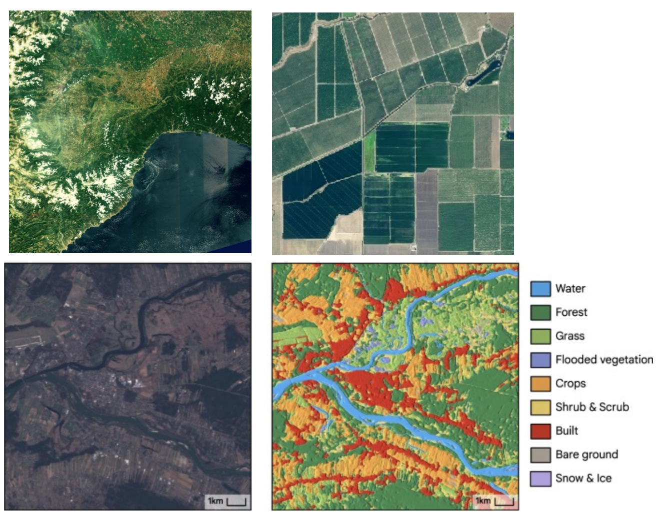
GFM: Building Geospatial Foundation Models via Continual Pretraining

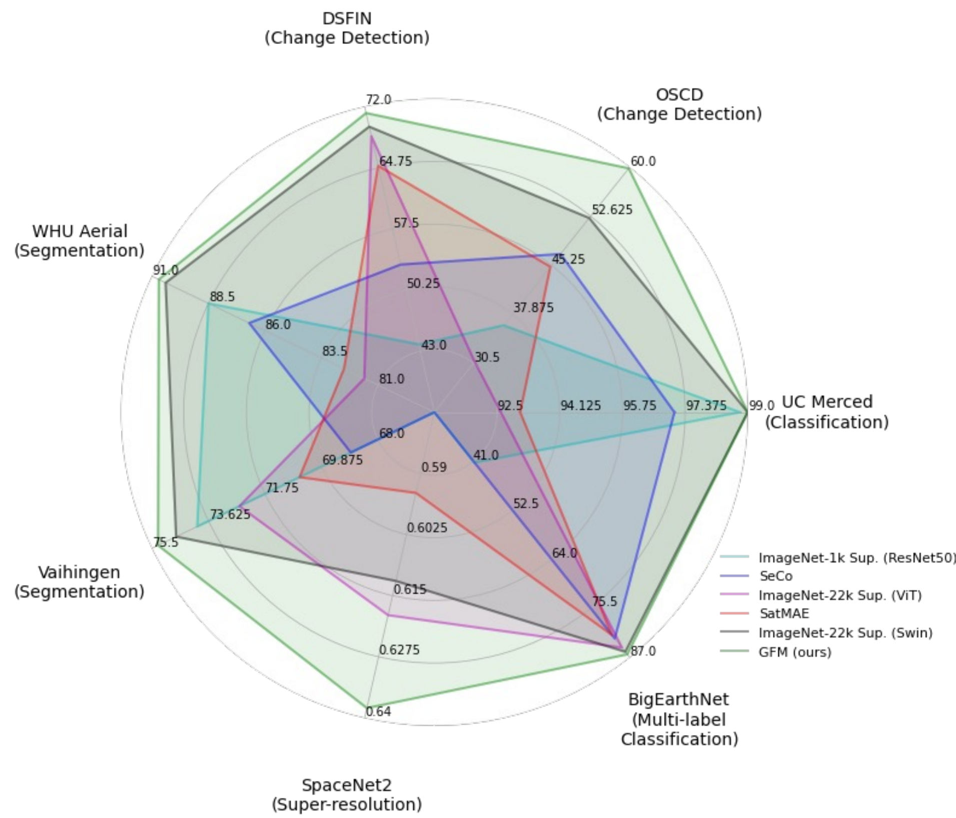
Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, Chen Chen, Mu Li
AWS AI Research and Education Labs
Deepearth team

Introduction

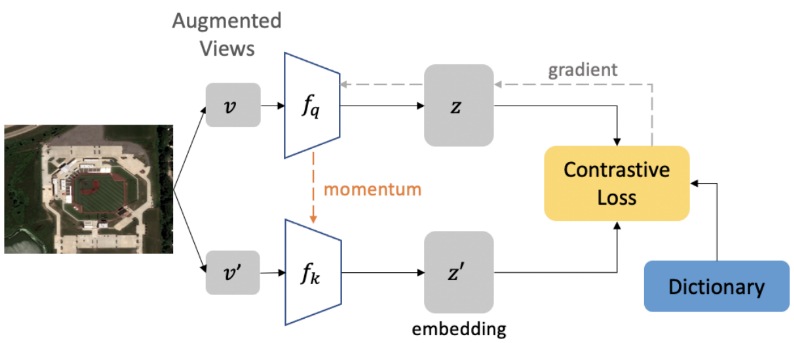
- Geospatial technologies
  - Understand the earth
  - How we interact with it
  - Various features

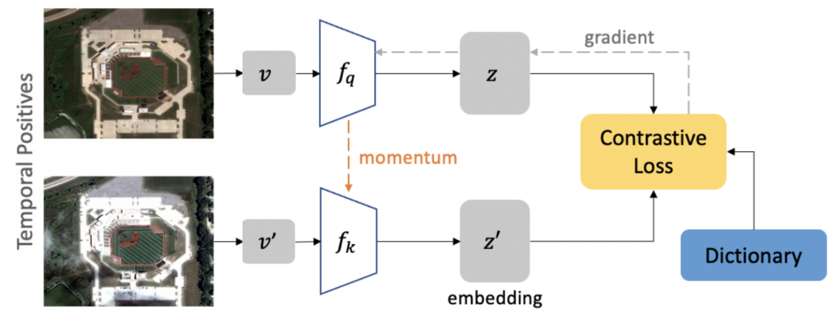


GFM (our model)



Background and Related Work: Contrastive Learning





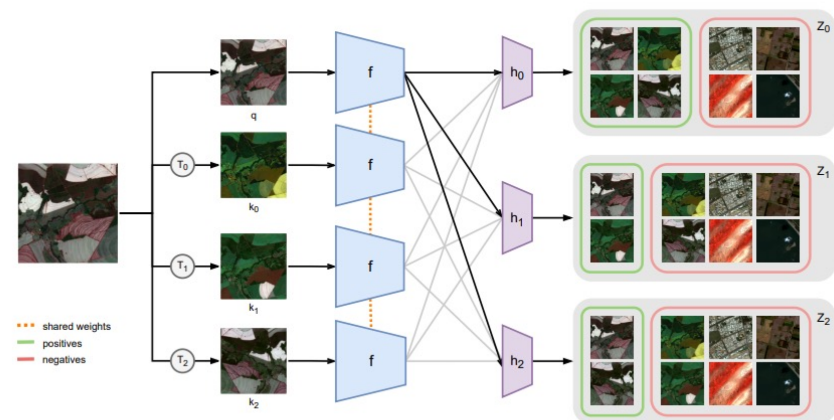
Ayush et, al, Geography-aware self-supervised learning.

- Limitation in augmentations:

  Data augmentation that affects the intensity of the values should be discarded.

  Neumann, et. al, In-domain representation learning for remote sensing.

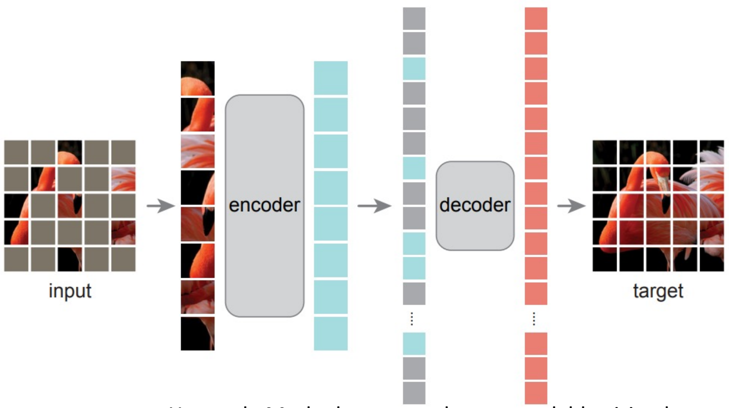
- Inconsistent performance on downstream tasks



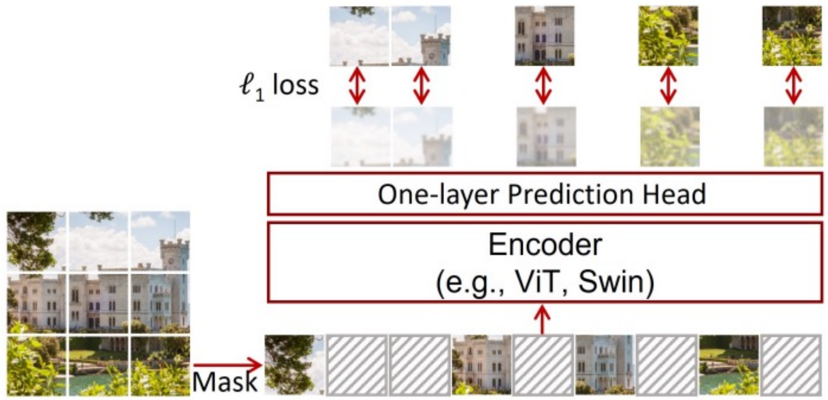
Oscar Ma˜nas, et, al.. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data.

Background and Related Work: Masked Image Modeling

- Masked Image Modeling
  - Strong downstream transfer
  - Simple spatial augmentations
  - Underexplored in geospatial applications



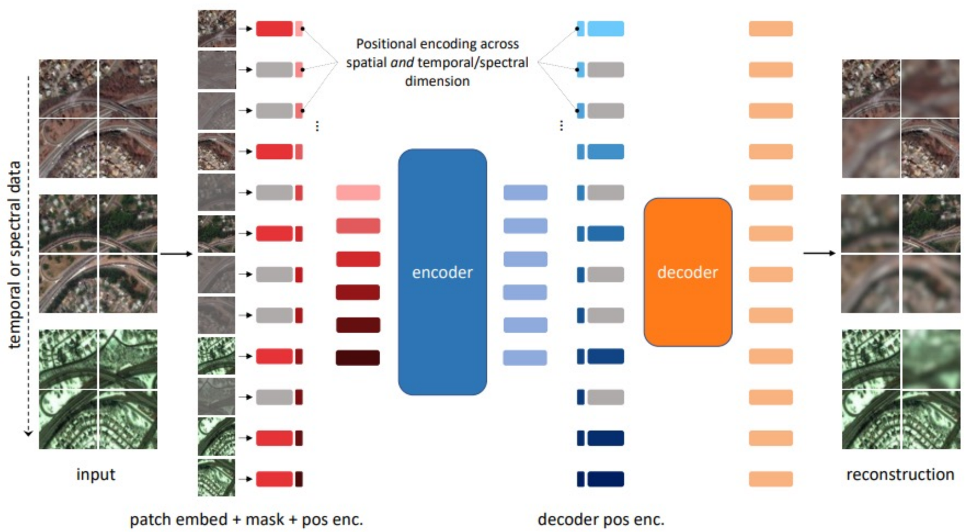
He et. al., Masked autoencoders are scalable vision learners



Xie et. al., Simmim: A simple framework for masked image modeling.

Background and Related Work: Masked Image Modeling

- SatMAE
  - Multiple temporal inputs
  - Temporal positional encoding
  - Independent masking

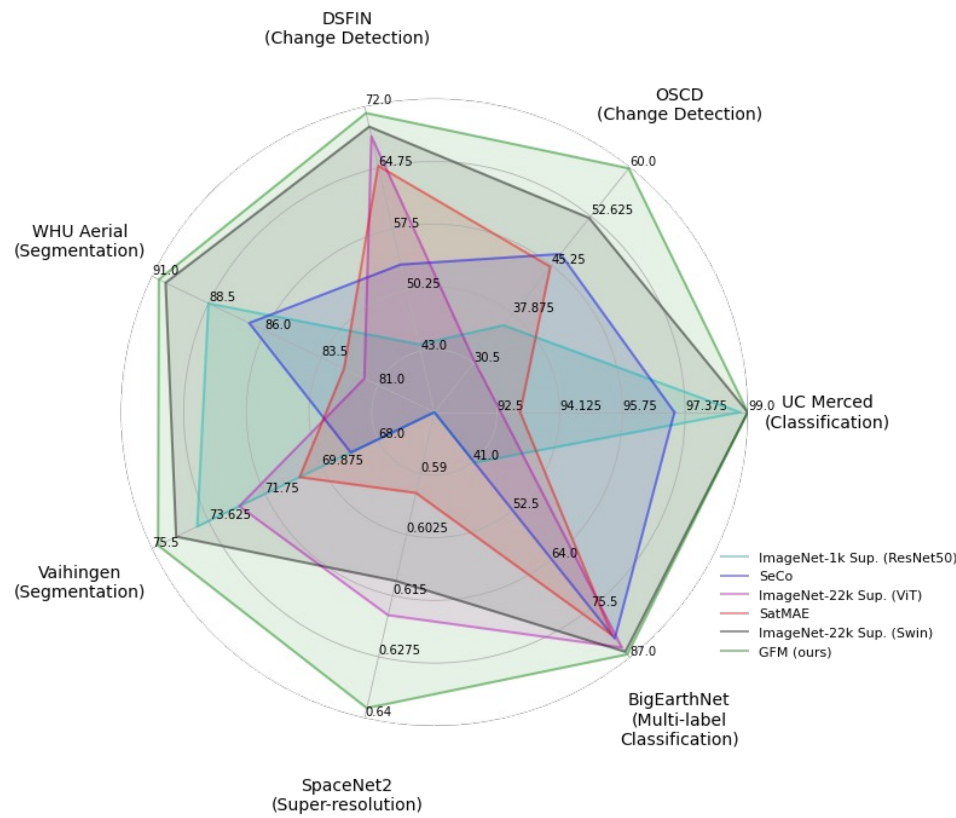


- Downsides
  - Carbon footprint
    - 109.44 kg. CO2 eq.
    - More than 12x our GFM
  - Often falls behind ImageNet-22k counterpart

Cong et. al., Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery

GFM – A sustainable approach

- Investigation
  - Pretraining data matters
    - GeoPile
    - Effective and sample efficient
  - Multi-objective continual pretraining
    - Leverage diverse representations
    - Adapt in-domain knowledge
- GFM
  - Strong performance
  - Broad set of tasks



Methodology: Pretraining data matters

- Sentinel-2 imagery
  - Common choice
  - Gather 1.3M images
  - Train Swin-B with MIM
- 7 downstream datasets
  - Change detection
  - Single and multi-label classification
  - Segmentation
  - Super-resolution

Method	# Images	Epochs	ARP ↑	CO2 ↓
ImageNet-22k Sup.	14M	-	0.0	-
Sentinel-2 [29]	1.3M	100	-5.53	17.76

$$\mathrm{ARP}(M) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{score}(M, \mathrm{task}_i) - \mathrm{score}(\mathrm{baseline}, \mathrm{task}_i)}{\mathrm{score}(\mathrm{baseline}, \mathrm{task}_i)}$$

Methodology: Pretraining data matters

- Sentinel-2 imagery
  - Common choice
  - Gather 1.3M images
  - Train Swin-B with MIM
- ImageNet-1k
  - Domain gap
  - Higher entropy (5.1 vs 3.9)
  - Diverse features
    - Within image
    - Across images

Method	# Images	Epochs	ARP ↑	CO2 ↓
ImageNet-22k Sup.	14M	-	0.0	-
ImageNet-1k	1.3M	100	1.82	17.76
Sentinel-2 [29]	1.3M	100	-5.53	17.76

$$\mathrm{ARP}(M) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{score}(M, \mathrm{task}_i) - \mathrm{score}(\mathrm{baseline}, \mathrm{task}_i)}{\mathrm{score}(\mathrm{baseline}, \mathrm{task}_i)}$$

Methodology: Pretraining data matters

- GeoPile
  - Labeled and unlabeled sources
  - Informative samples
    - Ground sample distance (GSD) variations
    - Wide variety of classes and scenes

Dataset	# Images	GSD	# Classes
NAIP [31]	300,000	1m	n/a
RSD46-WHU [28]	116,893	0.5m - 2m	46
MLRSNet [33]	109,161	0.1m - 10m	60
RESISC45 [8]	31,500	0.2m - 30m	45
PatternNet [47]	30,400	0.1m - 0.8m	38

Methodology: Pretraining data matters

- GeoPile
  - Labeled and unlabeled sources
  - Informative samples
    - Ground sample distance (GSD) variations
    - Wide variety of classes and scenes
- Further improvement
  - Train longer, scale up training data.
  - Vast datasets with poor quality
  - More cost, marginal benefit.

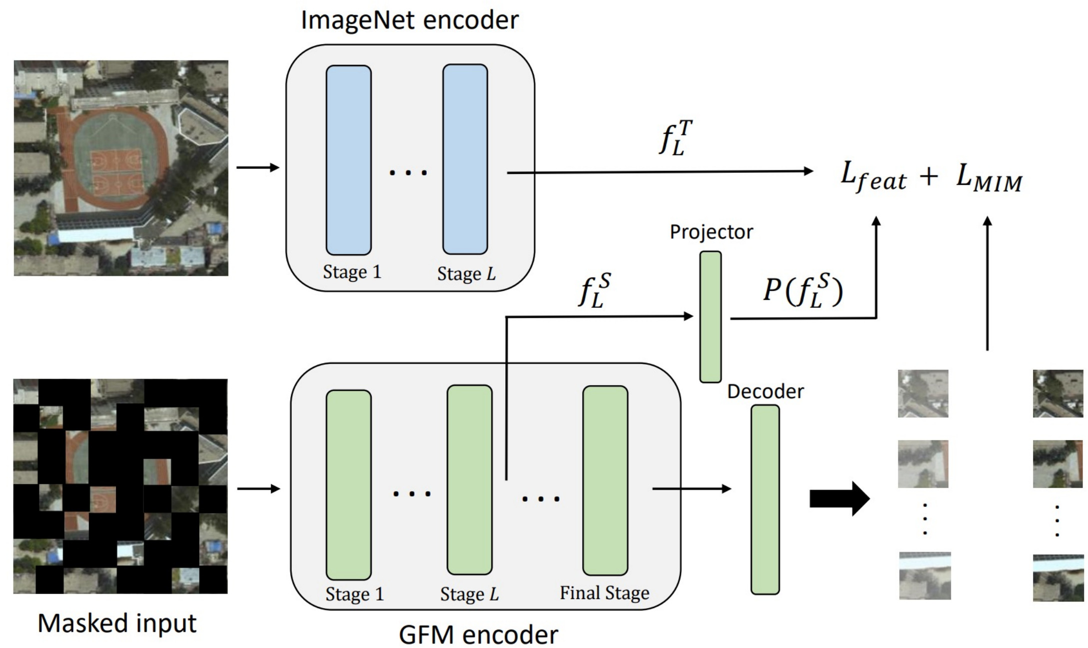
Method	# Images	Epochs	ARP $\uparrow$	CO2 $\downarrow$
ImageNet-22k Sup.	14M	-	0.0	-
ImageNet-1k	1.3M	100	1.82	17.76
Sentinel-2 [29]	1.3M	100	-5.53	17.76
GeoPile	600k	200	2.02	12.64
GeoPile	600k	800	2.44	50.56

Can we significantly improve performance with minimal compute and carbon footprint overhead?

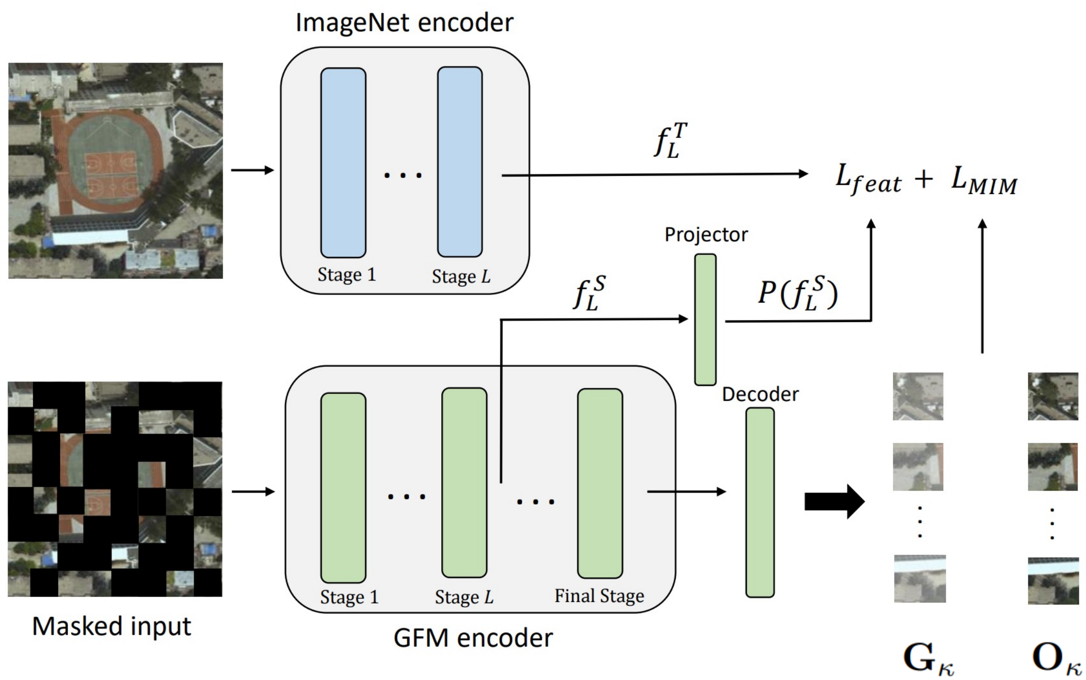
Methodology: GFM via continual pretraining

- ImageNet-22k models
  - Not perfect for geospatial
  - Still valuable features
  - Ignoring not ideal
    - Especially for data hungry transformers

- Multi-objective continual training paradigm



Methodology: GFM via continual pretraining



Loss functions:

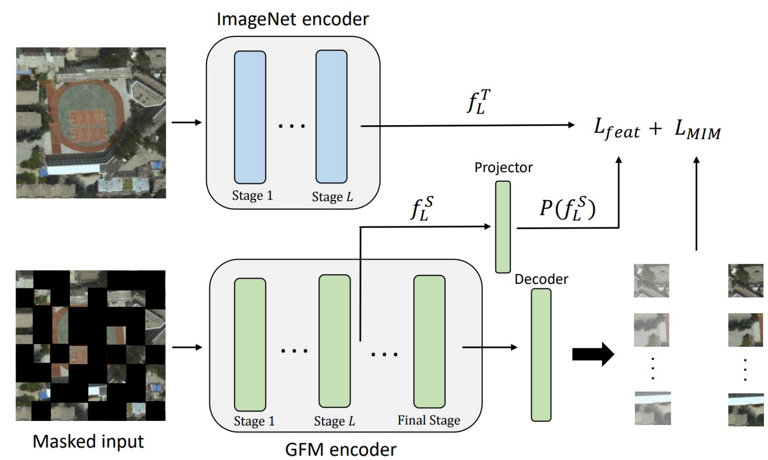
$$\mathcal{L}_{feat} = -\frac{P(f_L^S)}{\|P(f_L^S)\|_2} \cdot \frac{f_L^T}{\|f_L^T\|_2}$$

$$\mathcal{L}_{MIM} = \frac{\|\mathbf{O}_\kappa - \mathbf{G}_\kappa\|_1}{N}$$

$$\mathcal{L} = \mathcal{L}_{MIM} + \alpha \mathcal{L}_{feat}$$

Methodology: GFM via continual pretraining

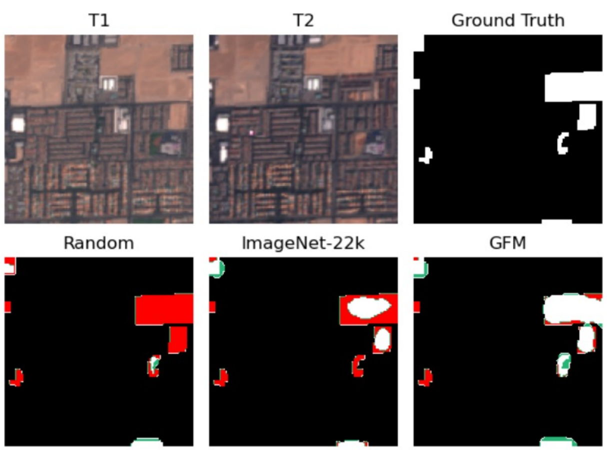
- Distillation
  - Benefit from diverse knowledge
  - Learn more in less time
- Masked Image Modeling
  - Freedom for in-domain adaptation
  - Geospatial features
    - Improved performance



Method	# Images	Epochs	ARP ↑	CO2 ↓
ImageNet-22k Sup.	14M	-	0.0	-
ImageNet-1k	1.3M	100	1.82	17.76
Sentinel-2 [29]	1.3M	100	-5.53	17.76
GeoPile	600k	200	2.02	12.64
GeoPile	600k	800	2.44	50.56
GFM	600k	100	4.47	8.56

Experiments: Change Detection

- Onera Satellite Change Detection
  - Sentinel-2 imagery
    - 10m GSD
- DSFIN
  - WorldView-3 and GeoEys-1
    - 1m GSD



White – True Positive
Green – False Positive
Red – False Negative

Onera Satellite Change Detection

Method	Precision ↑	Recall ↑	F1 ↑
ResNet50 (ImageNet-1k) [20]	**70.42**	25.12	36.20
SeCo [29]	65.47	38.06	46.94
MATTER [1]	61.80	57.13	59.37
ViT (ImageNet-22k) [14]	48.34	22.52	30.73
SatMAE [9]	48.19	42.24	45.02
Swin (random) [26]	51.80	47.69	49.66
Swin (ImageNet-22k) [26]	46.88	59.28	52.35
GFM	58.07	**61.67**	**59.82**

DSFIN

Method	Precision ↑	Recall ↑	F1 ↑
ResNet50 (ImageNet-1k) [20]	28.74	**92.07**	43.80
SeCo [29]	39.68	81.02	53.27
ViT (ImageNet-22k) [14]	70.77	66.34	68.49
SatMAE [9]	70.45	60.29	64.98
Swin (random) [26]	57.97	62.06	59.94
Swin (ImageNet-22k) [26]	67.11	72.33	69.62
GFM	**74.83**	67.98	**71.24**

Experiments: Classification

- UC Merced
  - 21 classes
  - 1 foot GSD
- BigEarthNet
  - 19 classes
  - 10m GSD
- Baseline comparisons
  - SeCo lower in UCM
  - SatMAE lower in BEN

Method	UCM	BEN 10%	BEN 1%
ResNet50 (ImageNet-1k) [20]	98.8	80.0	41.3
SeCo [29]	97.1	82.6	63.6
ViT (ImageNet-22k) [14]	93.1	84.7	73.6
SatMAE [9]	92.6	81.8	68.9
Swin (random) [26]	66.9	80.6	65.7
Swin (ImageNet-22k) [26]	**99.0**	85.7	79.5
GFM	**99.0**	**86.3**	**80.7**

- Sample efficiency
  - BigEarthNet 10% and 1%
  - Maintain strong performance

Experiments: Segmentation and Super resolution

- WHU Aerial
  - Building segmentation
  - GSD 0.3m
- Vaihingen
  - 6 class
  - GSD 0.9m

Method	WHU Aerial	Vaihingen
ResNet50 (ImageNet-1k) [20]	88.5	74.0
SeCo [29]	86.7	68.9
ViT (ImageNet-22k) [14]	81.6	72.6
SatMAE [9]	82.5	70.6
Swin (random) [26]	88.2	67.0
Swin (ImageNet-22k) [26]	90.4	74.7
GFM	**90.7**	**75.3**

SpaceNet2

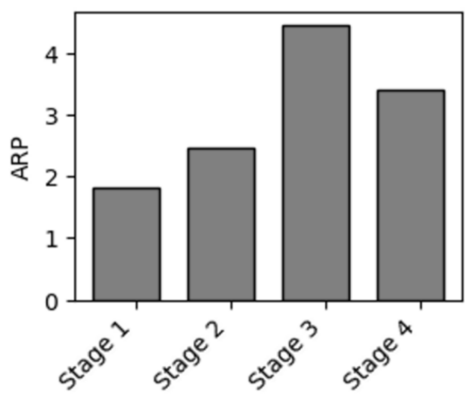
1.24m 8-band input

Generate 0.3m pan-sharpened equivalent

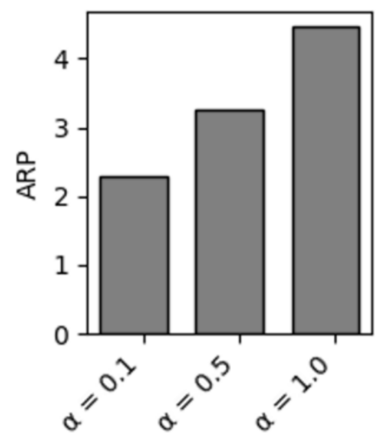
Method	PSNR ↑	SSIM ↑
ViT (ImageNet-22k) [14]	**23.279**	0.619
SatMAE [9]	22.742	0.621
Swin (random) [26]	21.825	0.594
Swin (ImageNet-22k) [26]	21.655	0.612
GFM	22.599	**0.638**

Ablation Studies

- Distillation Stage
  - Stage 3 is best
    - Appropriate distillation supervision
    - Purely in-domain features for final layers



- Balancing Term α
  - $\mathcal{L} = \mathcal{L}_{MIM} + \alpha \mathcal{L}_{feat}$
  - Simply α = 1.0 is best



Ablation Studies

- GeoPile pretraining dataset
  - Labeled datasets
    - Better performance, less images
  - Unlabeled data
    - Easily sourced and scaled
  - Further scaling
    - Increased training and CO2 impact

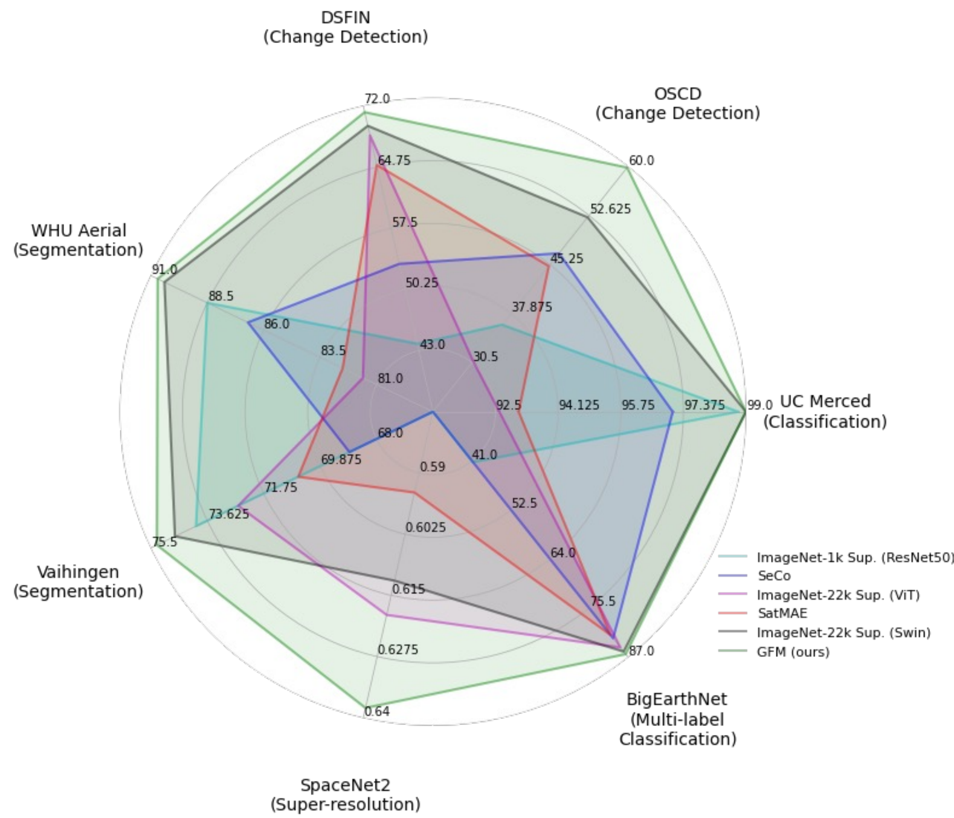
Data	# Images	ARP ↑
w/o WHU-RSD46	444,061	2.87
w/o MLRSNet	451,793	3.30
w/o Resisc45	529,454	2.72
w/o PatternNet	557,554	2.98
w/o curated datasets	300,000	1.62
w/o NAIP	260,954	2.65

- Continual pretraining comparison
  - Basic approach
    - Initialize with ImagNet-22k
    - MIM training on GeoPile
  - GFM
    - More effective and efficient

Method	Epochs	ARP ↑	CO2 ↓
ImageNet-22k Init.	200	2.66	12.64
ImageNet-22k Init.	800	2.98	50.56
GFM	100	4.47	8.56

Conclusion

- GFM
  - Sustainable approach
  - Strong performance
  - Broad set of tasks
- GeoPile
  - Diverse and effective
- Multi-objective continual pretraining
  - Leverage diverse representations
  - Adapt in-domain knowledge



Thank you