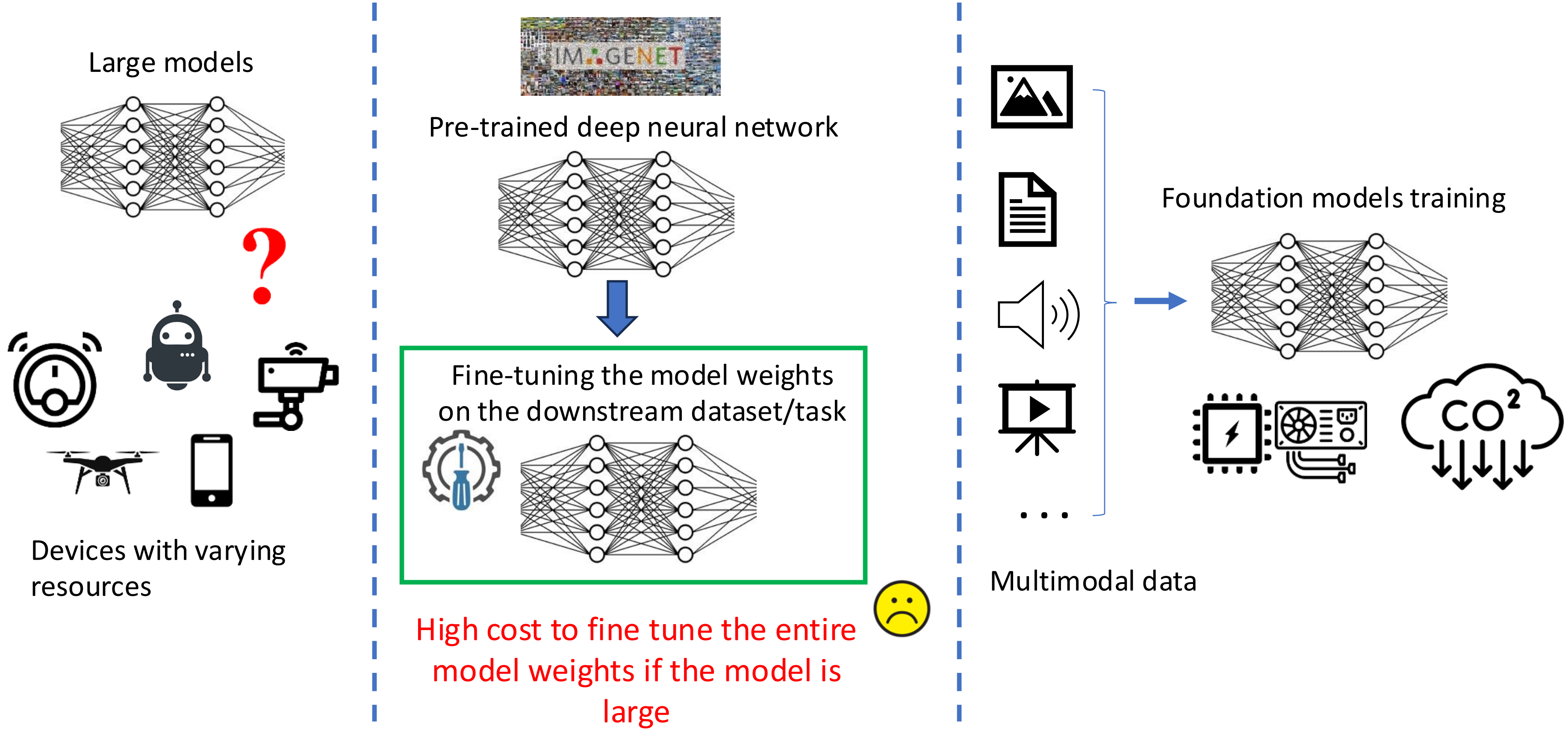
CAP 5516
Medical Image Computing
(Spring 2025)

Dr. Chen Chen
Associate Professor
Center for Research in Computer Vision (CRCV)
University of Central Florida
Office: HEC 221
Email: chen.chen@crcv.ucf.edu
Web: https://www.crcv.ucf.edu/chenchen/

Lecture 14

Efficient Deep Learning (4)


Large models
Devices with varying resources
IM GENET
Pre-trained deep neural network
Fine-tuning the model weights on the downstream dataset/task
High cost to fine tune the entire model weights if the model is large
Multimodal data
Foundation models training
CO2

Building Geospatial Foundation Models with Minimal Resource Costs

Mendieta, Matías, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. "Towards geospatial foundation models via continual pretraining." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16806-16816. 2023.

Introduction

- Geospatial technologies
  - Understand the earth
  - How we interact with it
- Applications
  - Agriculture
  - Urban planning
  - Disaster response

- Geospatial foundation models
  - Enable strong performance
  - Various downstream tasks

Introduction

- Publicly available ImageNet
  - Directly fine-tune
  - Useful, but suboptimal
  - Domain gap

- Geospatial models from scratch
  - Immense data, time, resources
  - Not consistently better than ImageNet

SatMAE, requires 768 hours on a V100 GPU for training a vision transformer

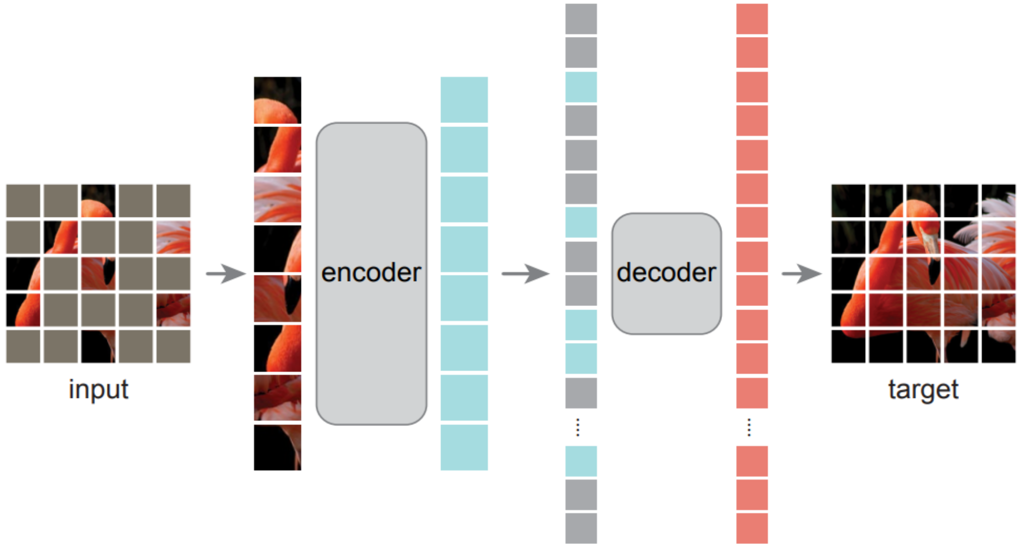
Cong, Yezhen, et al. "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery." Advances in Neural Information Processing Systems 35 (2022): 197-211.

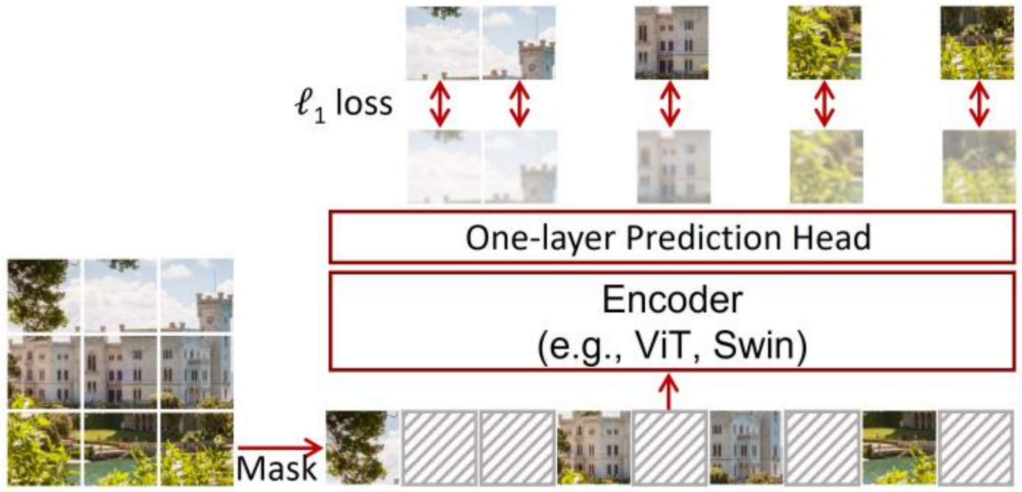
Contributions

- GFM paradigm investigation
  - Highly effective and sustainable
- Pretraining data
  - GeoPile
  - Various remote sensing sources
- Multi-objective continual pretraining
  - Leverage ImageNet representations
  - Freedom to learn in-domain features

Background

- Masked Image Modeling
  - Pre-training strategy using self-supervised learning
  - Strong downstream transfer





Pretraining Data Selection

Time: training time in hours on a V100 GPU
CO2: carbon impact estimations in kg CO2 equivalent

- Sentinel-2 imagery
  - Common choice
  - Gather 1.3M images
- Experiment Setup
  - Train Swin-B with MIM
  - Fine-tune the pretrained model on downstream datasets
- 7 downstream datasets
  - Change detection
  - Single and multi-label classification
  - Segmentation
  - Super-resolution

Method	# Images	Epochs	ARP ↑	Time ↓	CO_2 ↓
ImageNet-22k Sup.	14M	-	0.0	-	-
Sentinel-2 [30]	1.3M	100	-5.83	155.6	22.2

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

ARP: average relative performance

Pretraining Data Selection

- Sentinel-2 imagery
  - Perceivably low feature diversity
  - Low entropy

- MIM objective
  - Easier reconstruction with Sentinel
  - Similar image regions for masking

Method	# Images	Epochs	ARP ↑	Time ↓	CO_2 ↓
ImageNet-22k Sup.	14M	-	0.0	-	-
Sentinel-2 [30]	1.3M	100	-5.83	155.6	22.2

Sentinel-2

Pretraining Data Selection

- GeoPile
  - Variety of ground sample distances (GSDs)
  - Labeled and unlabeled datasets
  - Higher entropy and diverse

Method	# Images	Epochs	ARP ↑	Time ↓	CO_2 ↓
ImageNet-22k Sup.	14M	-	0.0	-	-
Sentinel-2 [30]	1.3M	100	-5.83	155.6	22.2
GeoPile	600k	200	0.92	133.3	19.0

Dataset	# Images	GSD	# Classes
NAIP [33]	300,000	1m	n/a
RSD46-WHU [29]	116,893	0.5m - 2m	46
MLRSNet [35]	109,161	0.1m - 10m	60
RESISC45 [9]	31,500	0.2m - 30m	45
PatternNet [48]	30,400	0.1m - 0.8m	38

Sentinel-2

GeoPile

Vanilla Continual Pretraining

- Validity Investigation
  - Initialize with ImageNet-22k weights
  - Pretraining with GeoPile (MIM)
  - Performance improvement

Method	# Images	Epochs	ARP ↑	Time ↓	CO_2 ↓
ImageNet-22k Sup.	14M	-	0.0	-	-
Sentinel-2 [30]	1.3M	100	-5.83	155.6	22.2
GeoPile	600k	200	0.92	133.3	19.0
GeoPile†	600k	200	1.24	133.3	19.0

	Data for pre-train	Weights initialization	Pre-train method
Sentinel-2	Sentinel-2 images	Random	MIM
GeoPile	GeoPile images	Random	MIM
GeoPile^{+}	GeoPile images	ImageNet-22k	MIM

Resulting model fine-tune on downstream datasets for performance evaluation (ARP)

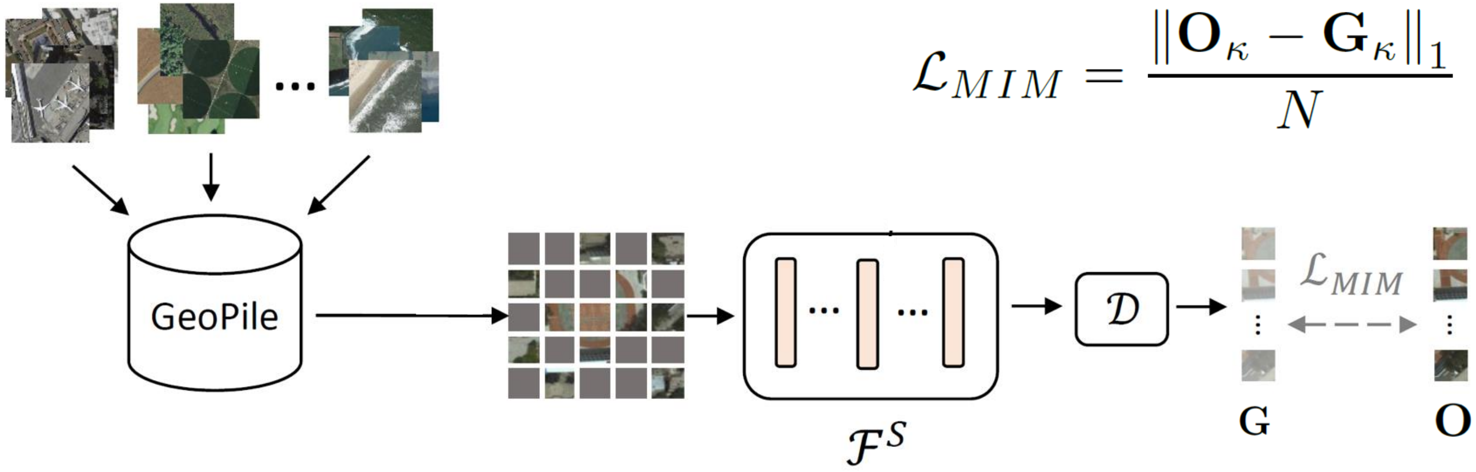
Vanilla Continual Pretraining

- Validity Investigation
  - Initialize with ImageNet-22k weights
  - Pretraining with GeoPile (MIM)
  - Performance improvement
- Longer Pretraining
  - Significantly more cost
  - Marginal gain

Method	# Images	Epochs	ARP ↑	Time ↓	CO_2 ↓
ImageNet-22k Sup.	14M	-	0.0	-	-
Sentinel-2 [30]	1.3M	100	-5.83	155.6	22.2
GeoPile	600k	200	0.92	133.3	19.0
GeoPile†	600k	200	1.24	133.3	19.0
GeoPile†	600k	800	1.45	533.2	76.0

How can we significantly improve performance while maintaining minimal compute and carbon footprint overhead?

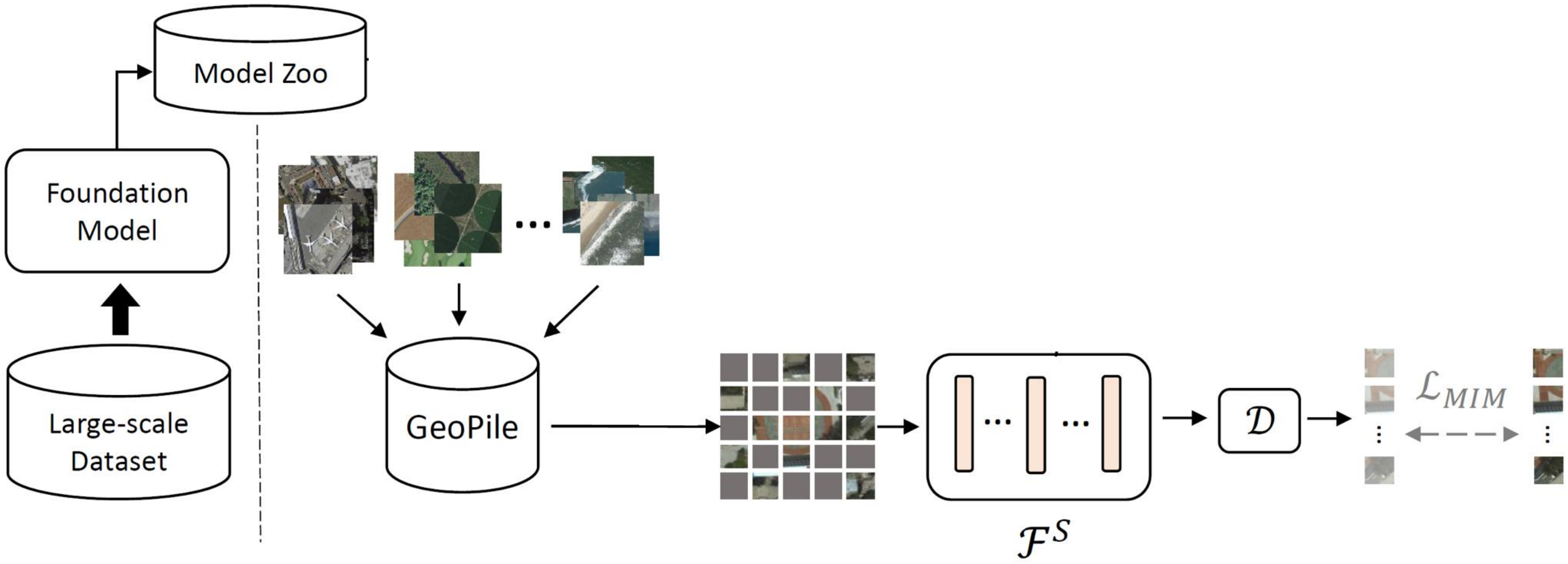
GFM Pretraining



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

- GeoPile dataset
  - Diverse pretraining data
- Student network
  - Randomly initialized
  - MIM objective

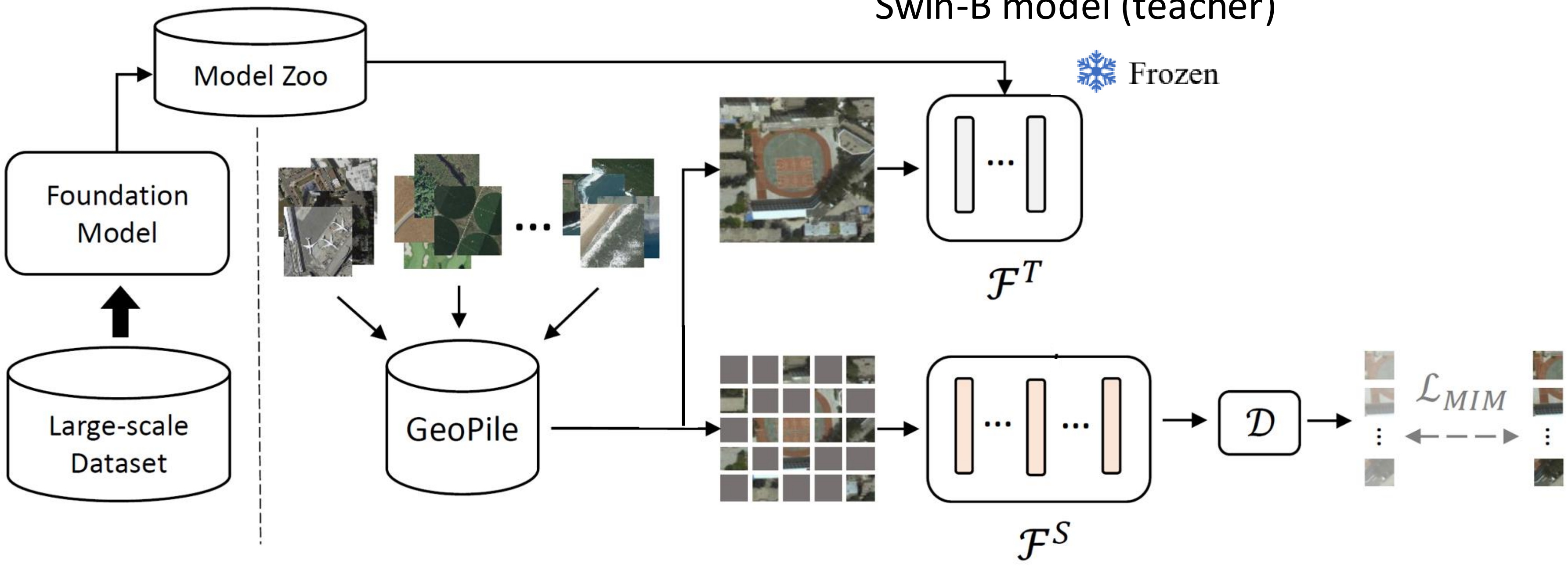
GFM Pretraining



- Leverage existing large-scale models

GFM Pretraining

ImageNet-22k pre-trained
Swin-B model (teacher)

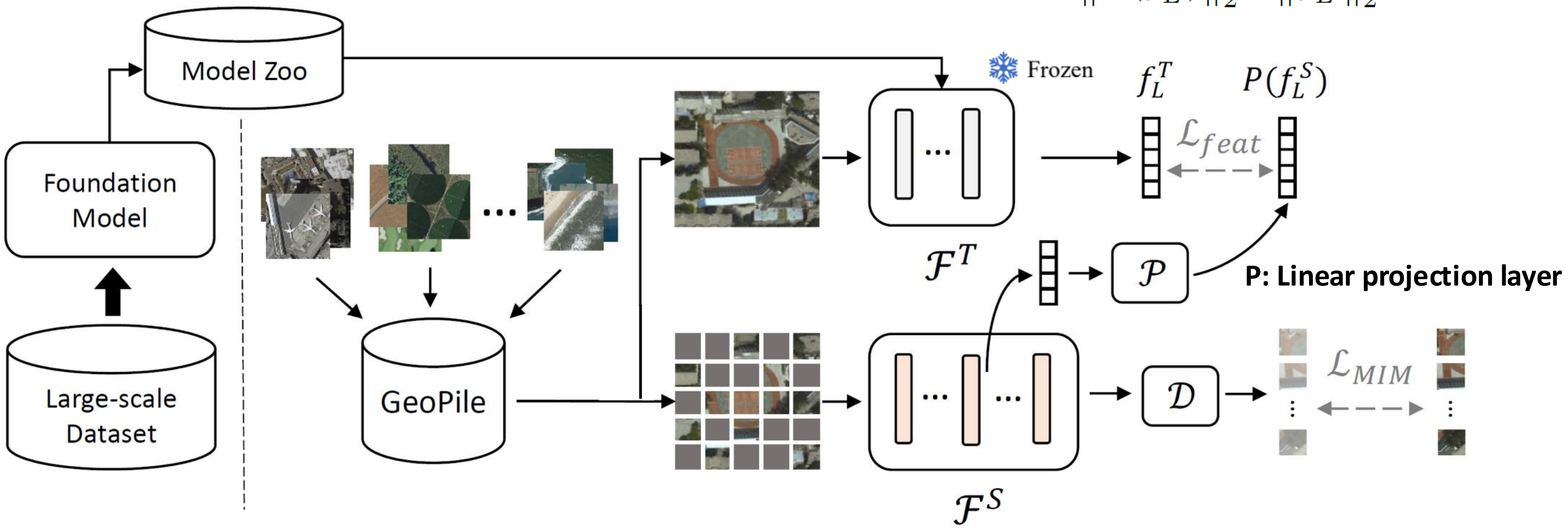


- Leverage existing large-scale models

- Teacher network
  - Readily available ImageNet-22k
  - Frozen

GFM Pretraining

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



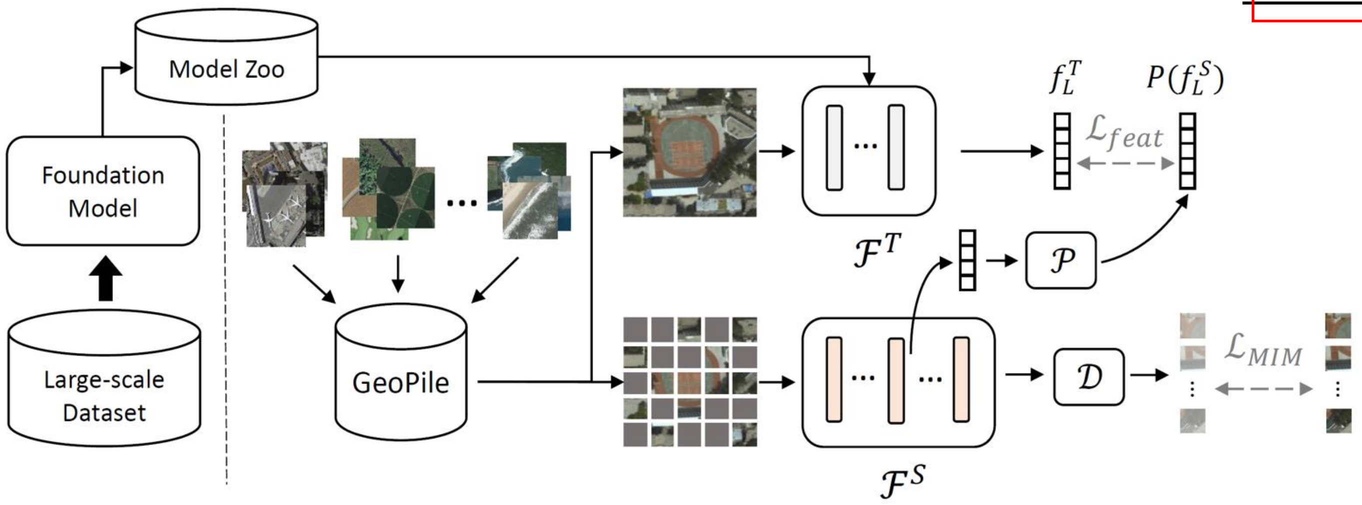
- Leverage existing large-scale models

- Masked feature guidance
  - Intermediate features (L^{th} layer feature)
  - Cosine similarity

GFM Pretraining

- Multi-objective continual training paradigm
  - Guide and accelerate learning ($\mathcal{L}_{feat}$)
  - Freedom to acquire valuable in-domain features ($\mathcal{L}_{MIM}$)

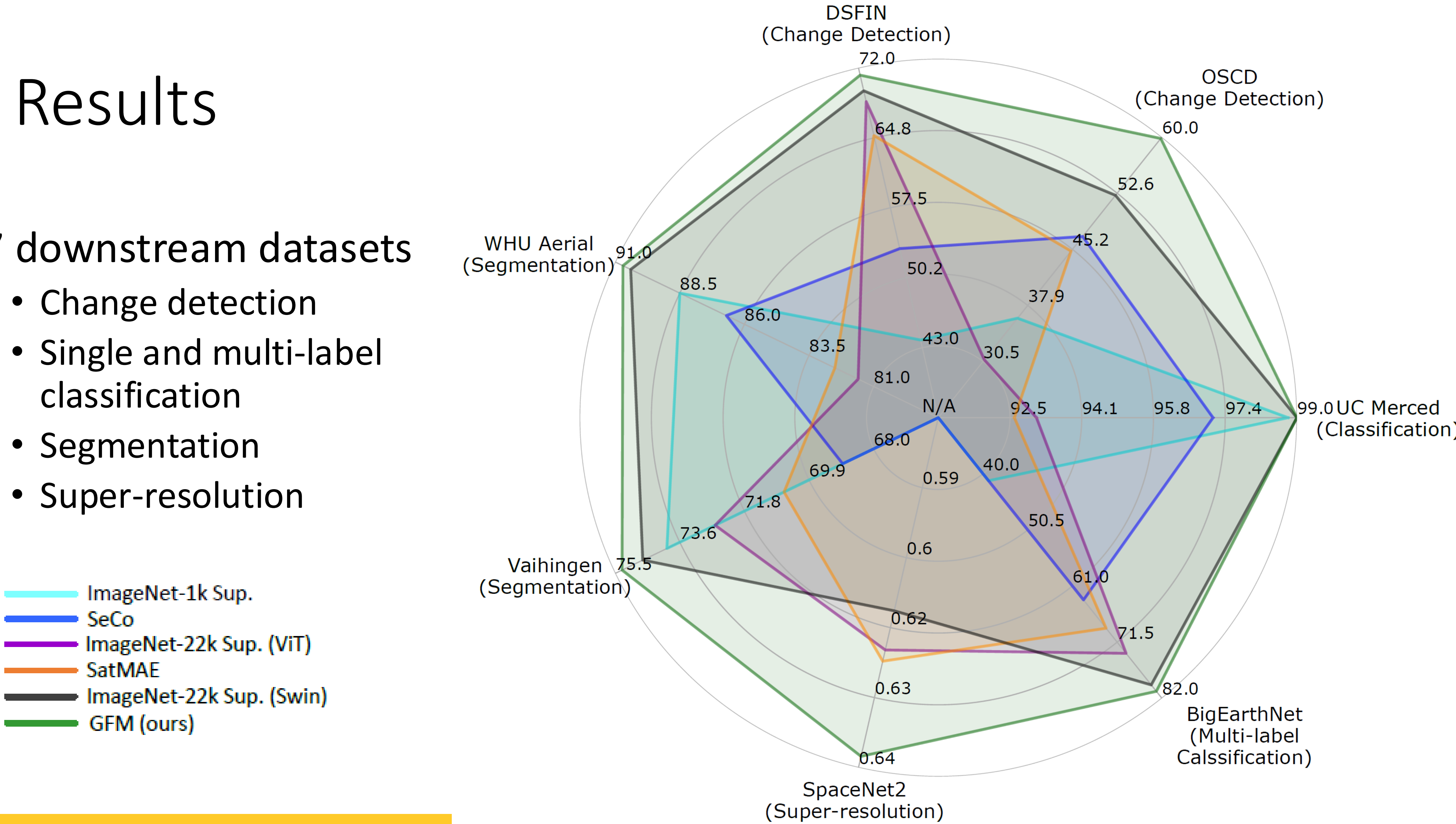
Method	# Images	Epochs	ARP ↑	Time ↓	CO_2 ↓
ImageNet-22k Sup.	14M	-	0.0	-	-
Sentinel-2 [30]	1.3M	100	-5.83	155.6	22.2
GeoPile	600k	200	0.92	133.3	19.0
GeoPile†	600k	200	1.24	133.3	19.0
GeoPile†	600k	800	1.45	533.2	76.0
GFM	600k	100	3.31	93.3	13.3



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

Results

- 7 downstream datasets
  - Change detection
  - Single and multi-label classification
  - Segmentation
  - Super-resolution



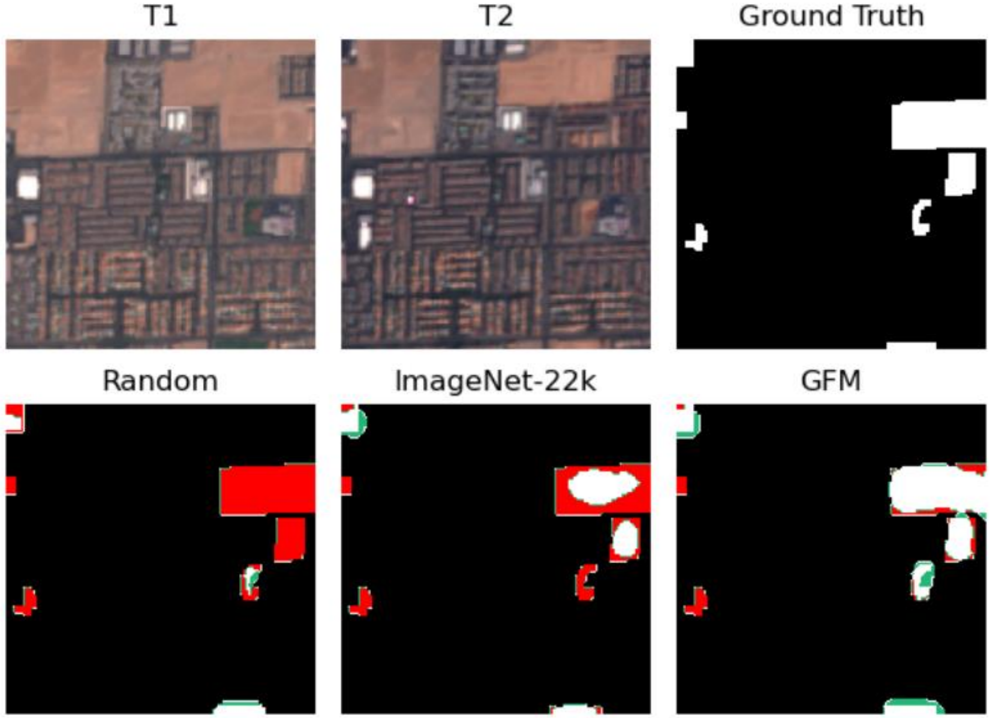
Change Detection

- Onera Satellite Change Detection
  - Sentinel-2 imagery
    - 10m – 60m GSD
  - Pixel-level change
    - Urban developments

Change Detection

- Onera Satellite Change Detection
  - Sentinel-2 imagery
    - 10m – 60m GSD
  - Pixel-level change
    - Urban developments

Method	Precision ↑	Recall ↑	F1 ↑
ResNet50 (ImageNet-1k) [20]	**70.42**	25.12	36.20
SeCo [30]	65.47	38.06	46.94
MATTER [1]	61.80	57.13	59.37
ViT (ImageNet-22k) [15]	48.34	22.52	30.73
SatMAE [10]	48.19	42.24	45.02
Swin (random)[27]	51.80	47.69	49.66
Swin (ImageNet-22k)[27]	46.88	59.28	52.35
GFM	58.07	**61.67**	**59.82**



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

Change Detection

- Onera Satellite Change Detection
  - Sentinel-2 imagery
    - 10m GSD
- DSFIN
  - WorldView-3 and GeoEys-1
    - 1m GSD
- Results
  - GFM improves in both datasets
  - SatMAE does not

Onera Satellite Change Detection

Method	Precision ↑	Recall ↑	F1 ↑
ResNet50 (ImageNet-1k) [20]	**70.42**	25.12	36.20
SeCo [30]	65.47	38.06	46.94
MATTER [1]	61.80	57.13	59.37
ViT (ImageNet-22k) [15]	48.34	22.52	30.73
SatMAE [10]	48.19	42.24	45.02
Swin (random)[27]	51.80	47.69	49.66
Swin (ImageNet-22k)[27]	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 [30]	39.68	81.02	53.27
ViT (ImageNet-22k) [15]	70.77	66.34	68.49
SatMAE [10]	70.45	60.29	64.98
Swin (random)[27]	57.97	62.06	59.94
Swin (ImageNet-22k)[27]	67.11	72.33	69.62
GFM	**74.83**	67.98	**71.24**

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 [30]	97.1	82.6	63.6
ViT (ImageNet-22k)[15]	93.1	84.7	73.6
SatMAE [10]	92.6	81.8	68.9
Swin (random)[27]	66.9	80.6	65.7
Swin (ImageNet-22k) [27]	**99.0**	85.7	79.5
GFM	**99.0**	**86.3**	**80.7**

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

Segmentation

- 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 [30]	86.7	68.9
ViT (ImageNet-22k) [15]	81.6	72.6
SatMAE [10]	82.5	70.6
Swin (random) [27]	88.2	67.0
Swin (ImageNet-22k) [27]	90.4	74.7
GFM	**90.7**	**75.3**

Super-resolution

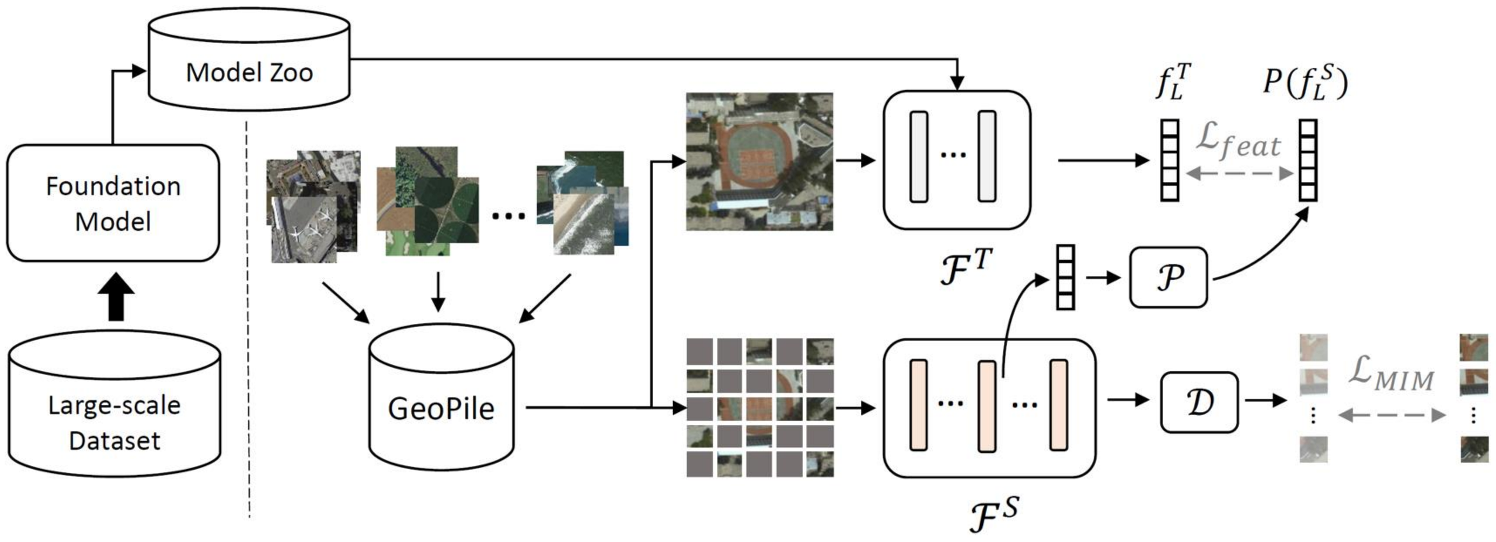
- SpaceNet2
  - 1.24m 8-band input
  - Generate 0.3m pan-sharpened equivalent

Method	PSNR ↑	SSIM ↑
ViT (ImageNet-22k)[15]	**23.279**	0.619
SatMAE [10]	22.742	0.621
Swin (random) [27]	21.825	0.594
Swin (ImageNet-22k) [27]	21.655	0.612
GFM	22.599	**0.638**

- Baseline comparisons
  - SatMAE lags behind
  - GFM continues to improve

Conclusion

- Consistent improvement across downstream tasks.
- More than 8× reduction in total training time and carbon impact in comparison to SOTA
- Sustainable and effective geospatial pretraining



Code and models are available at https://github.com/mmendiet/GFM

BiomedGPT

BiomedGPT: A generalist vision–language foundation model for diverse biomedical tasks

Kai Zhang[1], Rong Zhou[1], Eashan Adhikarla[1], Zhiling Yan[1], Yixin Liu[1], Jun Yu[1], Zhengliang Liu[2], Xun Chen[3], Brian D. Davison[1], Hui Ren[4], Jing Huang[5,6], Chen Chen[7], Yuyin Zhou[8], Sunyang Fu[9], Wei Liu[10], Tianming Liu[2], Xiang Li[4*], Yong Chen[5,11,12,13], Lifang He[1*], James Zou[14,15], Quanzheng Li[4], Hongfang Liu[9], and Lichao Sun[1*]

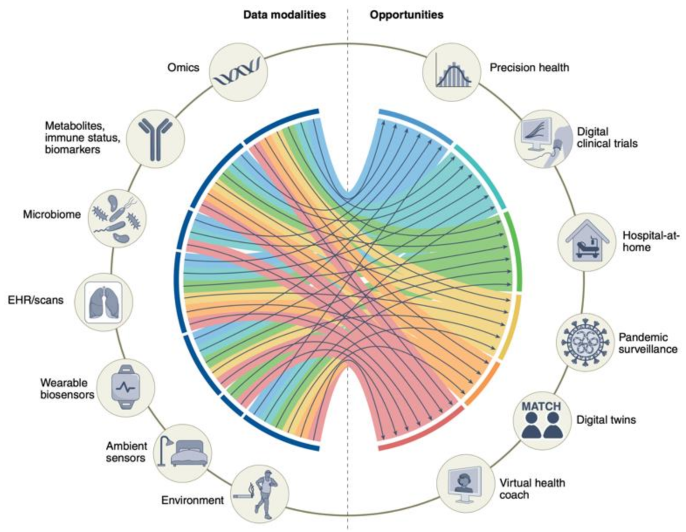
[1]*Department of Computer Science and Engineering, Lehigh University, PA, United States*
[2] *School of Computing, University of Georgia, GA, United States*
[3]*Samsung Research America, CA, United States*
[4]*Department of Radiology, Massachusetts General Hospital and Harvard Medical School, MA, United States*
[5]*Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, PA, United States*
[6]*PolicyLab, Children's Hospital of Philadelphia, PA, United States*
[7]*Center for Research in Computer Vision, University of Central Florida, FL, United States*
[8]*Department of Computer Science and Engineering, University of California, Santa Cruz, CA, United States*
[9]*McWilliams School of Biomedical Informatics, UTHealth Houston, TX, United States*
[10]*Department of Radiation Oncology, Mayo Clinic, AZ, United States*
[11]*The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, PA, United States*
[12]*Penn Institute for Biomedical Informatics (IBI), PA, United States*
[13]*Leonard Davis Institute of Health Economics, PA, United States*
[14]*Department of Biomedical Data Science, Stanford University School of Medicine, CA, United States*
[15]*Department of Computer Science, Stanford University, CA, United States*

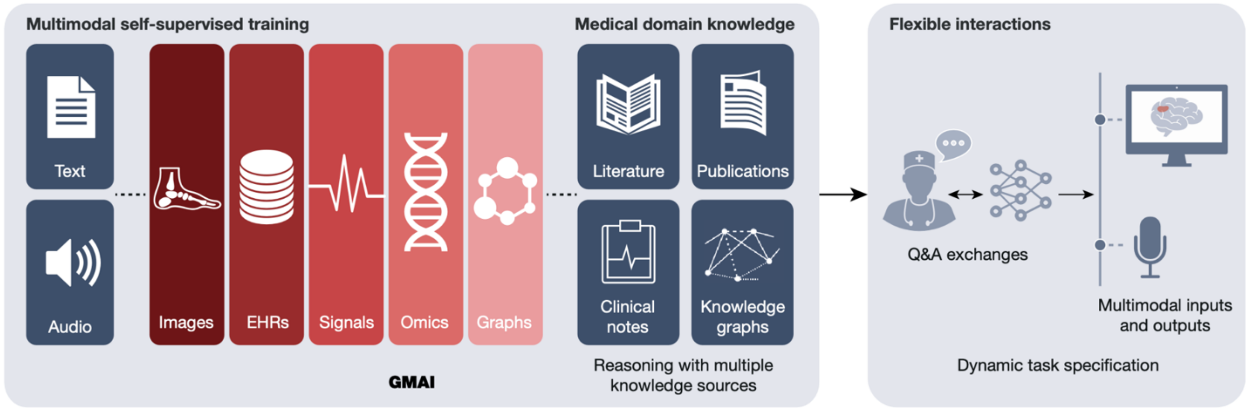
Zhang, Kai, et al. "A generalist vision–language foundation model for diverse biomedical tasks." *Nature Medicine* (2024): 1-13.

Most medical AI models are the specialist

- The limited amount of accessible high-quality annotated biomedical data.

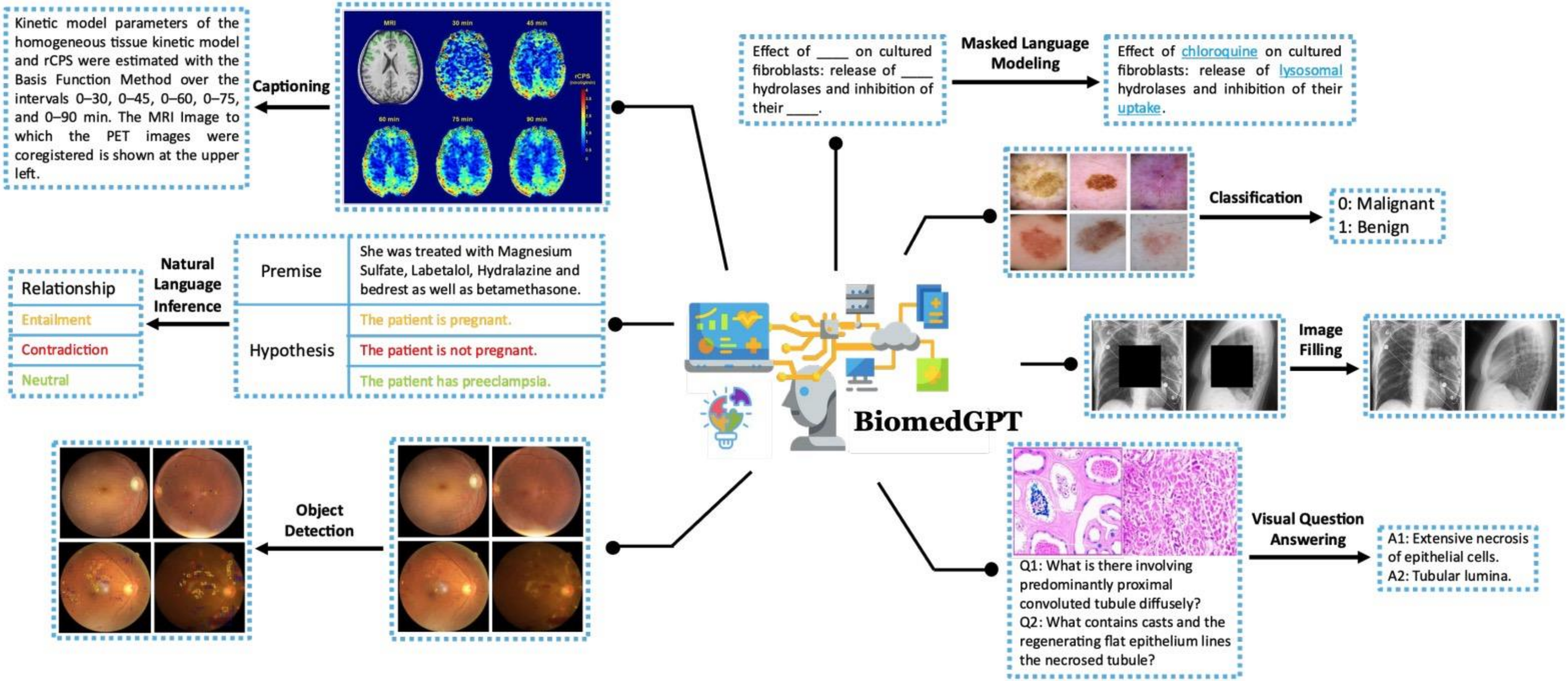
However, the increasing availability of biomedical data could set the early stage for the development of **Generalist AI** solutions that capture the complexity within biomedicine.



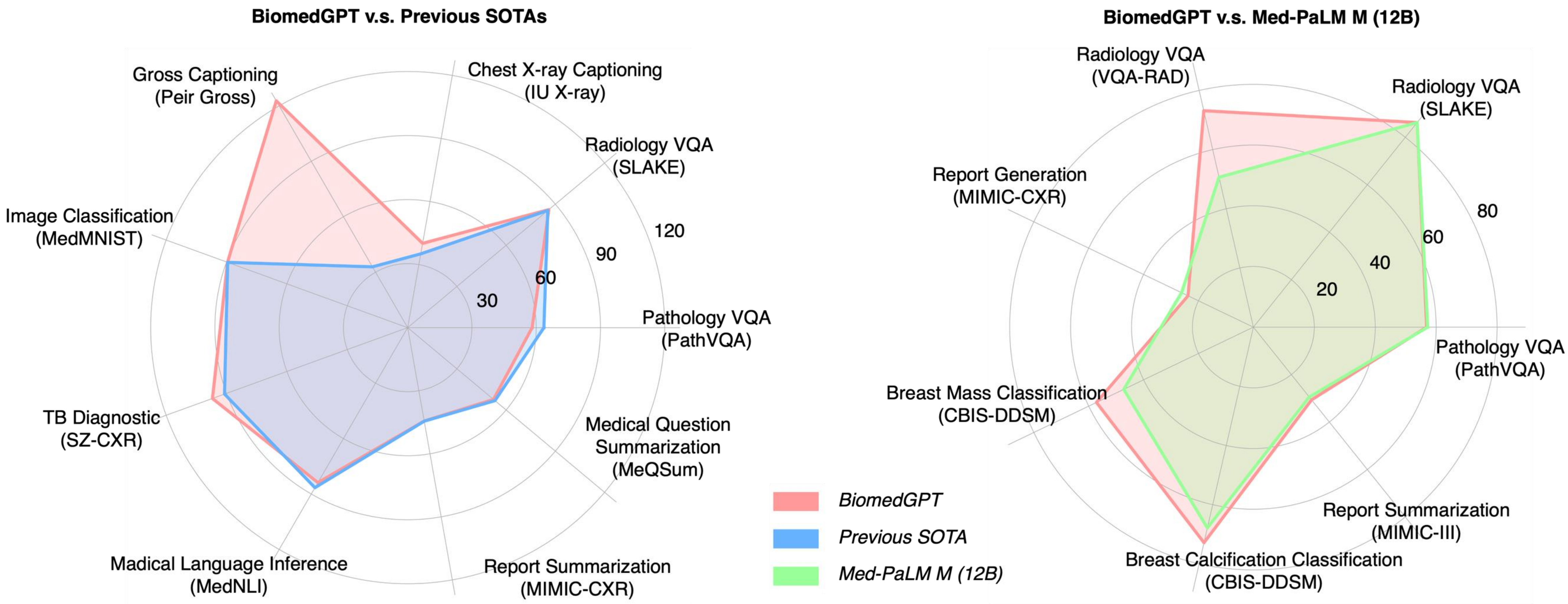


Acosta, Julián N., et al. "Multimodal biomedical AI." Nature Medicine 28.9 (2022): 1773-1784.
Moor, Michael, et al. "Foundation models for generalist medical artificial intelligence." Nature 616.7956 (2023): 259-265.

A snapshot: how powerful and generalist BiomedGPT is



A snapshot: how powerful and generalist BiomedGPT is

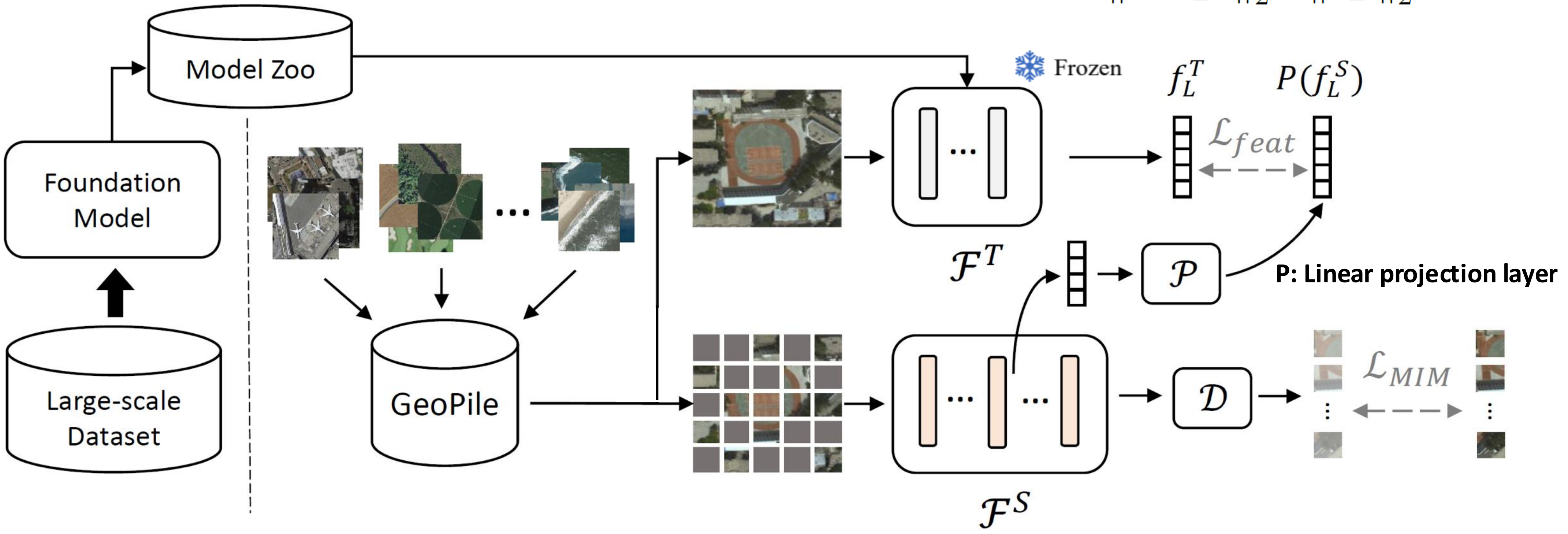


A generalist model – the same model with the same set of weights, without finetuning, can excel at a wide variety of tasks.

How the GFM training paradigm can be applied here?

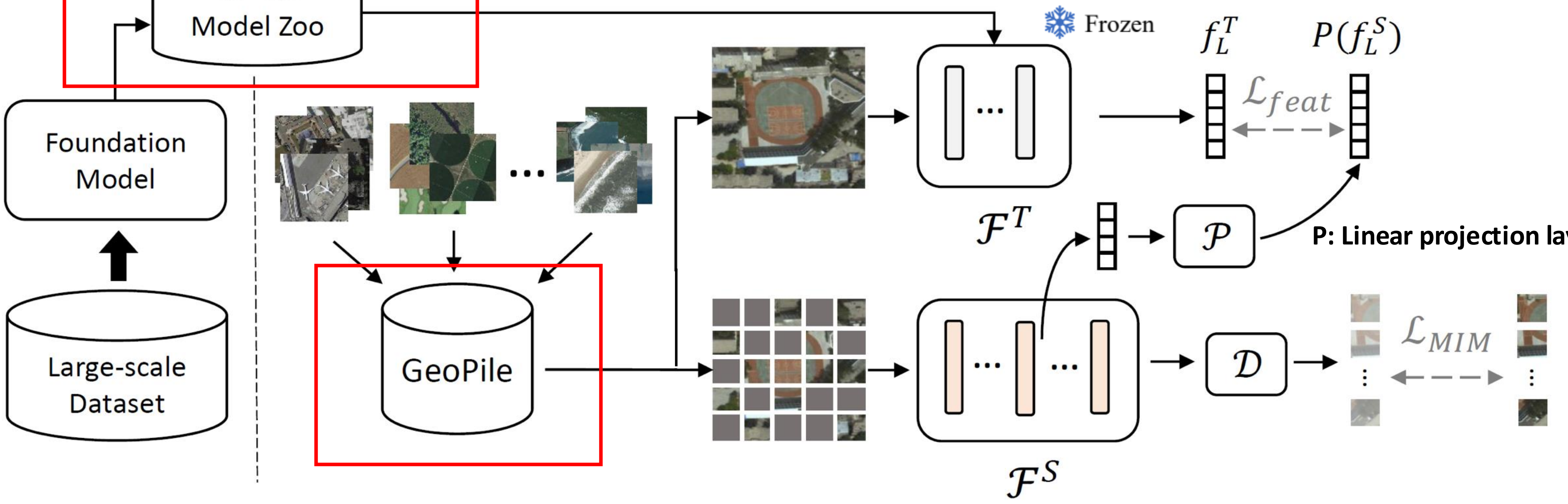
GFM Pretraining

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



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

cosine similarity



P: Linear projection layer

Announcement

- No class next week (2/25 & 2/27)

Thank you!