April 2026, ICLR
ML4RS
Rio, Brazil

Anthony Fuller, PhD student
Carleton Uni. / Vector Inst.
Ottawa, Canada

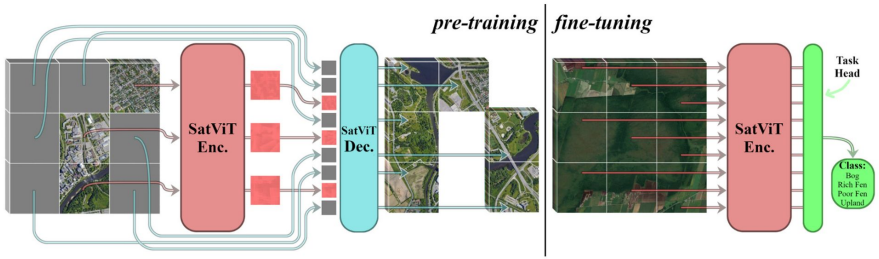
BAD TABLES

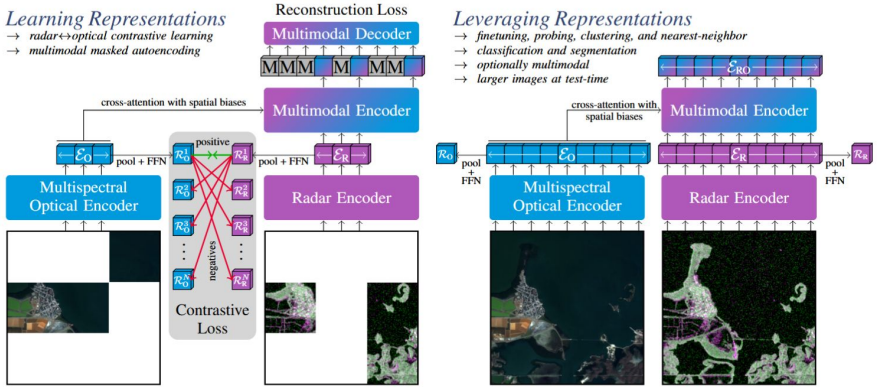
Why you shouldn't trust results tables in remote-sensing-foundation-model (RSFM) papers

Talk in a slide

- Despite the several dozen RSFMs—***all of which harness self-supervised learning*** **(SSL)**—we know very little about SSL for RS
- Since bigger models offer more accuracy, people control for it when assessing or making performance claims, e.g. ViT-B vs. ViT-B ✅, ViT-B vs. ViT-S ❌
  - This is what the computer vision community does, so surely we can...
- It mostly works in CV because "ViT-B" actually means: ViT-B, 16x16 patch size, 224x224 image size, 1 "channel group"
  - Typically, models in a table are trained on the same data and training length
- In RS "ViT-B" in a result table can vary: {1x1 → 16x16} patch size, {96x96 → 256x256} image size, {1 → 4} channel groups. We are secretly doing ViT-S vs. ViT-G ❌🙈❌
  - And there's more differences... training data, training length, etc.
  - These confounders invalidate direct comparisons between methods
- Advice: Standardize. We'll lose flexibility but gain clarity 🔎

My background in ML4RS (pt. 1 of 2)



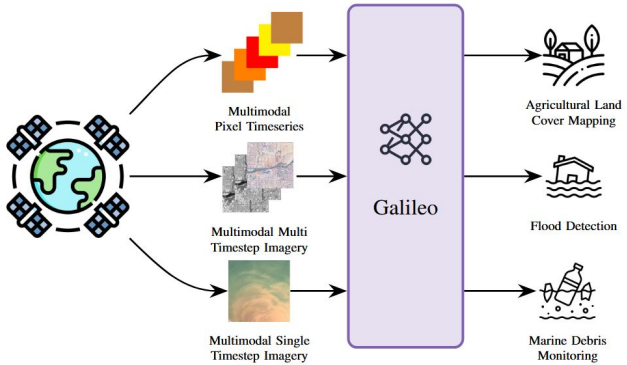


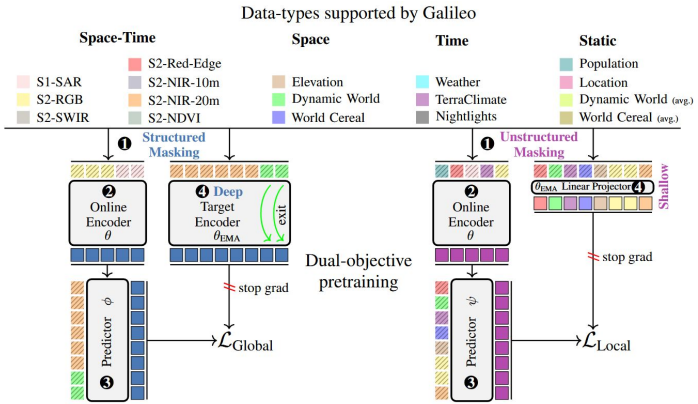
- SatViT: The First Self-Supervised Transformer for RS (concurrent with SatMAE)
  - Masked autoencoding on Sentinel-1 & 2
  - Continued in-domain pre-training (or was it "post-training"?)
- Authors: A. Fuller, K. Millard, and J.R. Green
- IEEE Geoscience and Remote Sensing Letters, 2022

- CROMA: Contrastive Radar-Optical Masked Autoencoders
  - 2D-ALiBi position encoding
  - Outperforms SatMAE on avg. +3.5% kNN
- Authors: A. Fuller, K. Millard, and J.R. Green
- NeurIPS 2023

My background in ML4RS (pt. 2 of 2)

CROMA 🤝 Presto 🤝 Satlas



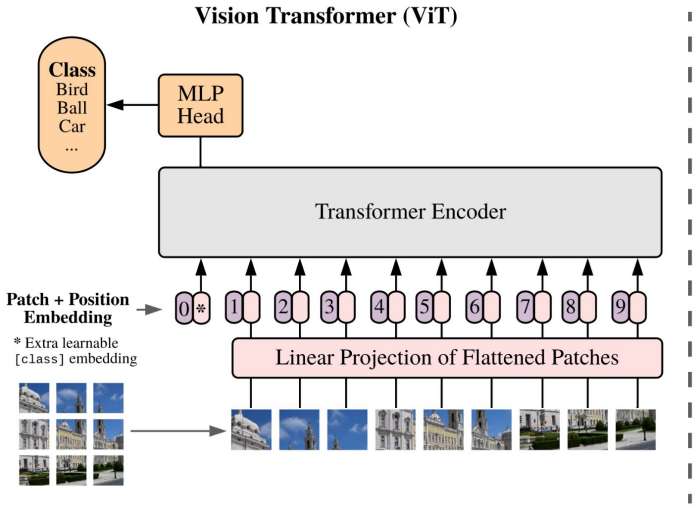


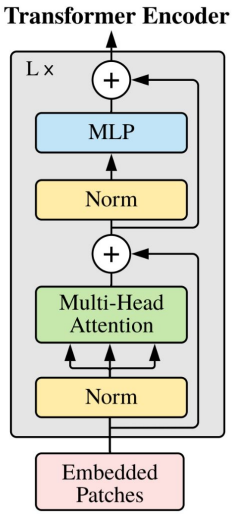
- Galileo: Highly Flexible and Multi-modal SSL for RS
- Authors: G. Tseng*, A. Fuller*, M. Reil, H. Herzog, P. Beukema, F. Bastani, J.R. Green, E. Shelhamer, H. Kerner†, and D. Rolnick†
- ICML 2025

- Outperforms CROMA on images and Presto on pixel-time series with a single model
- Different RS experts use different inputs, Galileo is the first model they can all share
- Galileo achieved its generality via novel global and local SSL tasks

Vision Transformers

Background on Vision Transformers (ViTs)



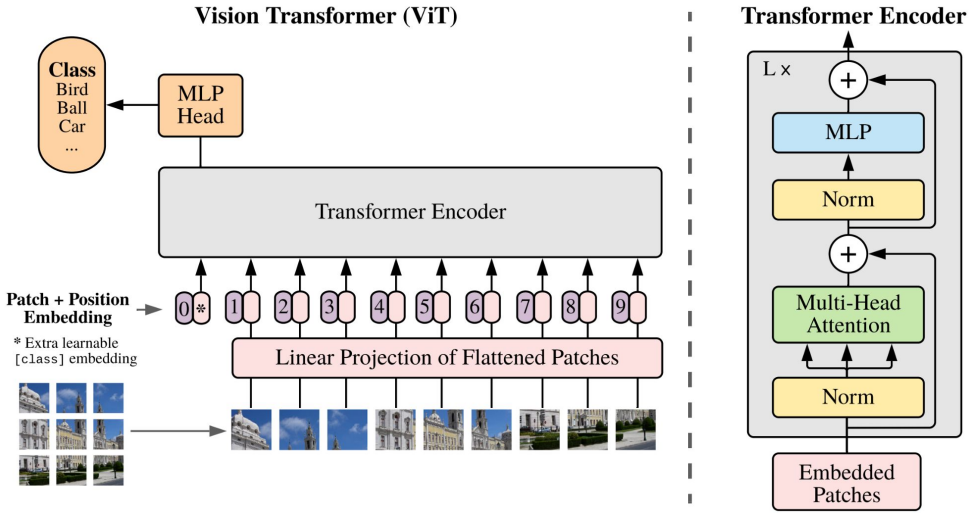


[figure credit: https://arxiv.org/abs/2010.11929]

- ViTs have N patches
  - In CV, patches are usually 16px per side
  - N scales inverse quadratically in patch size
- Patches are projected to D-dim vectors called tokens

- How do computational cost (e.g. FLOPs) and model parameters scale in D, L, N?
  - **Both contribute to model capacity, e.g.**
  - **DEIT3-B @384px > DEIT3-L @224px on IN-1K**

Background on Vision Transformers (ViTs)

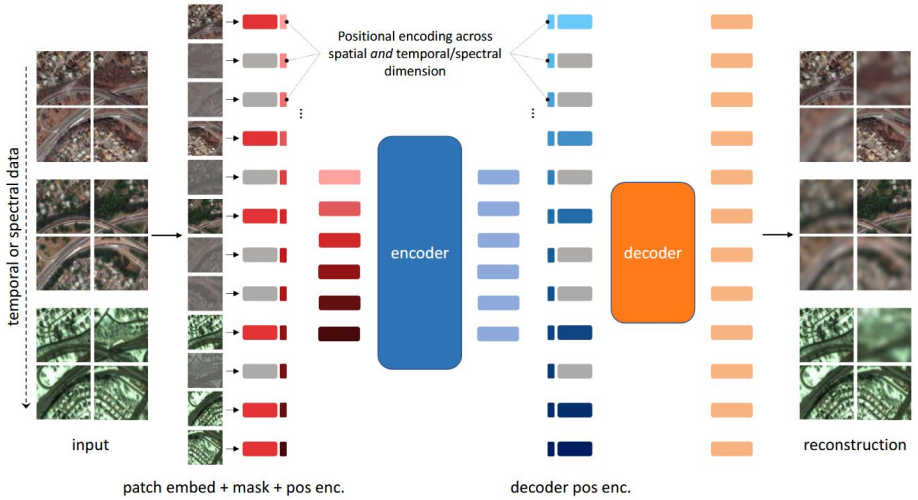


[figure credit: https://arxiv.org/abs/2010.11929]

- FLOPs scale:
  - ⌒quadratically in D
  - ／linearly in L
  - ／linearly in N for linear layers (MLP/FFN and QKVO projections)
  - ⌒quadratically in N for attention

- Parameters scale:
  - ⌒quadratically in D
  - ／linearly in L
  - —**constant in N**

SatMAE introduced channel groups for multi-spectral data



[figure credit: https://arxiv.org/pdf/2207.08051]

- Channel groups are combinations of bands that are mapped to a single token
  - ViTs in CV group RGB together, so all have 1 channel group
- SatMAE uses 3 groups (10 channels) for multi-spectral Sentinel-2
  - This grows N by 3x
  - Often used for input flexibility (e.g. in Galileo), its performance effects are unknown

fMoW-Sentinel top-1 / top-5 acc.

SatMAE+Stack	ViT-Large	57.37/81.63
SatMAE+Group+IM	ViT-Large	59.30/82.81

[table 3 in SatMAE]

3.4x FLOPs

Problem

We know very little about SSL for RS, despite these (and more!)

Why don't we know much? Look at all the numbers 👀

Method	Arch.	m-EuroSat Top-1 Acc. Training % 100%	m-EuroSat Top-1 Acc. Training % 1%	m-BigEarthNet F1 Score Training % 100%	m-BigEarthNet F1 Score Training % 1%	m-So2Sat Top-1 Acc. Training % 100%	m-So2Sat Top-1 Acc. Training % 1%	m-Brick-Kiln Top-1 Acc. Training % 100%	m-Brick-Kiln Top-1 Acc. Training % 1%
SatMAE	ViT-Base	84.1	34.8	50.6	29.0	36.0	23.1	86.1	73.5
SatMAE++	ViT-Large	82.7	48.5	50.8	31.6	34.7	23.4	89.6	76.7
CROMA	ViT-Base	85.6	51.3	58.8	44.7	48.8	33.8	**92.6**	**85.1**
SoftCon	ViT-Small	89.8	27.2	**64.7**	43.3	51.1	31.4	89.2	77.8
DOFA-v1	ViT-Base	82.8	49.6	49.4	29.9	41.4	29.4	88.3	78.3
Satlas	Swin-Tiny	81.7	35.8	51.9	29.6	36.6	27.1	88.2	73.0
MMEarth	CNN-atto	81.7	30.0	58.3	39.6	39.8	25.1	89.4	79.7
DeCUR	ViT-Small	89.0	46.6	63.8	**49.6**	45.8	30.9	83.7	74.2
Prithvi 2.0	ViT-Large	80.2	48.0	49.4	28.8	29.5	26.1	87.9	80.6
AnySat	ViT-Base	82.2	47.1	54.9	33.7	39.8	29.0	85.3	72.0
Galileo	ViT-Nano	89.7	41.7	53.8	33.9	50.1	37.4	86.7	79.7
Galileo	ViT-Tiny	90.1	41.3	55.5	34.4	49.7	36.2	86.9	77.3
Galileo	ViT-Base	**93.0**	**56.6**	59.0	36.5	**54.8**	**43.2**	90.7	78.0

Method	Arch.	m-Cashew-Plant Training % 100%	m-Cashew-Plant Training % 1%	m-SA-Crop-Type Training % 100%	m-SA-Crop-Type Training % 1%	MADOS Training % 100%	MADOS Training % 1%	Sen1Floods11 Training % 100%	Sen1Floods11 Training % 1%	PASTIS Training % 100%	PASTIS Training % 1%
SatMAE	ViT-Large	30.8	22.7	24.8	16.9	55.6	13.2	N/A		29.6	11.5
SatMAE++	ViT-Large	29.6	23.3	25.7	16.8	49.9	12.7	N/A		30.5	12.0
CROMA	ViT-Base	31.8	26.8	**32.0**	18.3	64.2	**24.4**	78.9	77.6	44.4	18.5
SoftCon	ViT-Base	29.6	22.8	30.8	18.5	60.3	16.5	78.0	74.8	31.3	10.5
DOFA-v1	ViT-Large	27.7	23.3	25.4	16.8	51.6	19.1	78.1	77.4	29.8	13.4
Satlas	Swin-Tiny	25.1	18.6	23.4	16.2	45.9	12.4	N/A		28.0	10.9
MMEarth	CNN-atto	24.2	20.3	22.2	14.1	34.2	16.1	N/A		24.0	10.5
DeCUR	ViT-Small	26.2	22.8	21.5	15.3	54.8	16.6	74.5	72.2	22.4	11.0
Prithvi 2.0	ViT-Large	26.7	23.2	22.9	15.7	50.0	18.9	N/A		29.3	13.2
AnySat	ViT-Base	26.1	21.7	27.1	15.8	50.2	17.0	77.9	76.9	**46.2**	**23.5**
Galileo	ViT-Nano	24.4	24.5	19.7	14.5	54.8	13.9	78.6	77.1	17.5	13.1
Galileo	ViT-Tiny	27.4	27.9	22.5	17.1	60.8	17.5	78.0	77.9	28.1	16.9
Galileo	ViT-Base	33.0	**30.2**	30.1	**19.4**	**67.6**	14.7	**79.4**	**78.2**	39.2	18.7

- Even *beautiful* tables like this include too many confounding variables to draw any conclusions about SSL algorithms in RS
- They only show Galileo models are (on average) more accurate than other RSFMs when all models are given equal downstream hyperparameter-tuning budgets and used at their default settings (i.e. patch sizes, image sizes, channel groups)

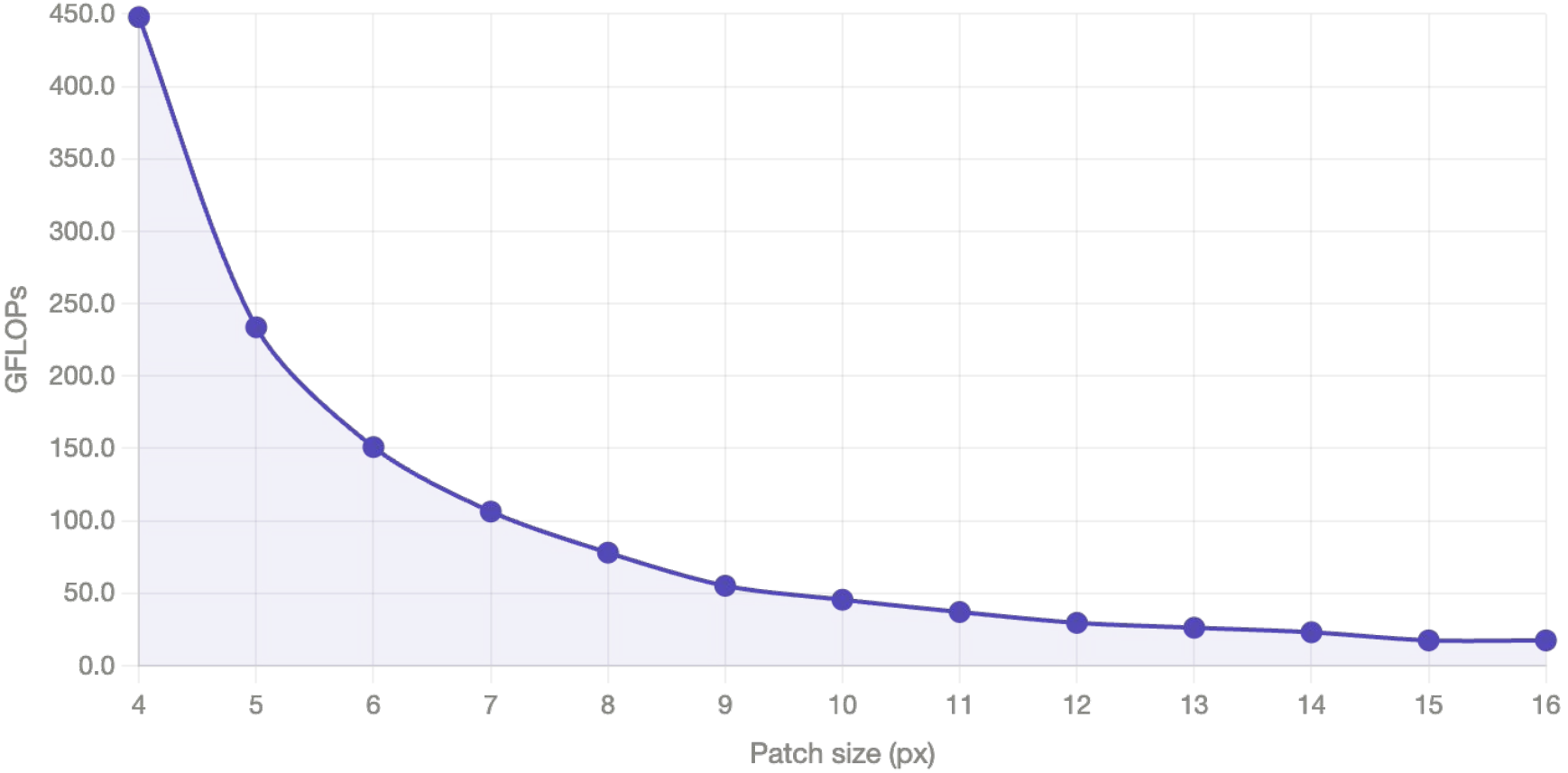
Confounding variables

- RSFMs can differ in:
  - Patch sizes: {1x1 px → 16x16 px}
  - Image sizes: {96x96 px → 256x256 px}
  - Channel groups: {1 → 4}

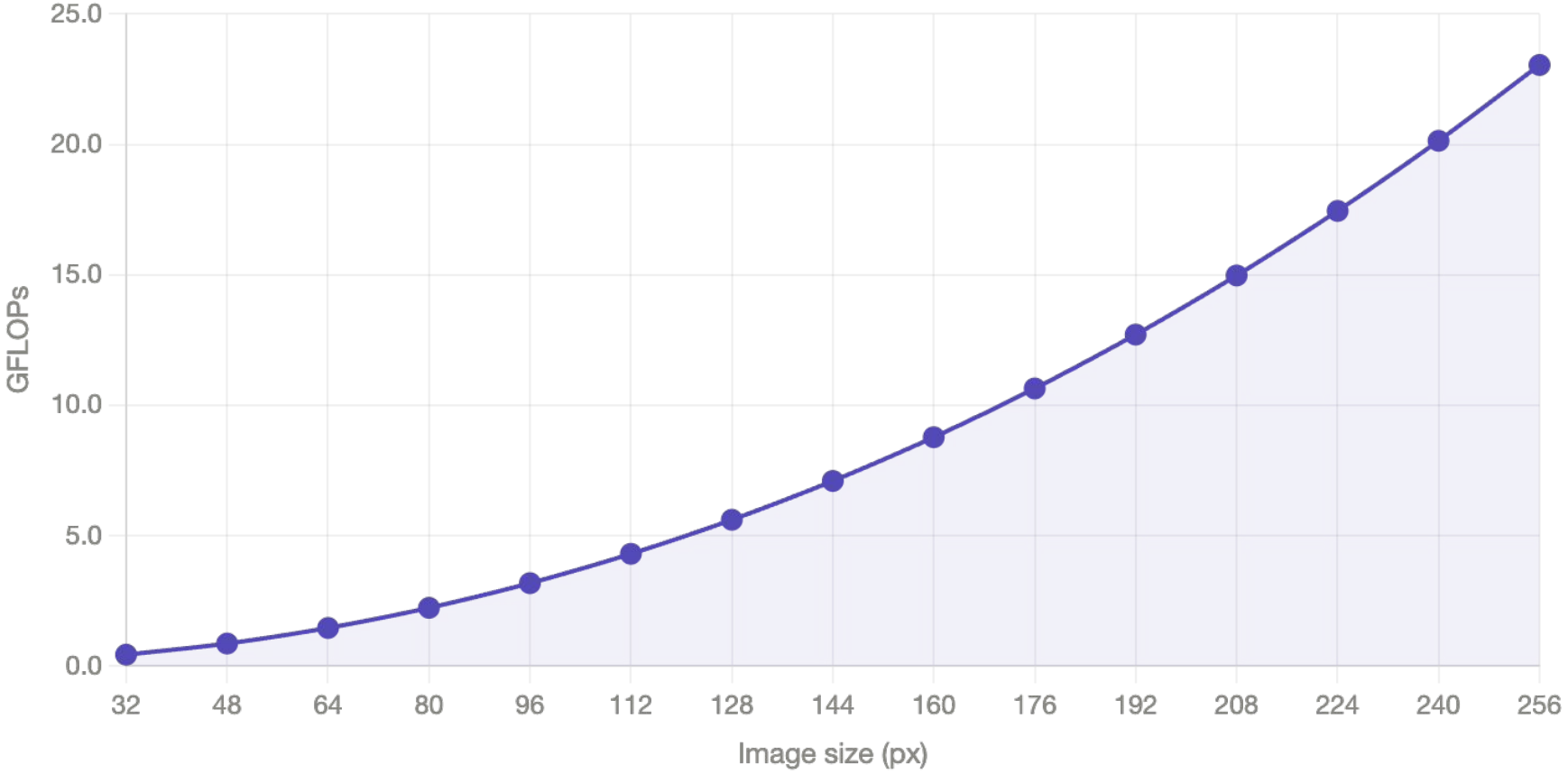
- Since models with more capacity offer more accuracy, then we must control for capacity to learn about accuracy
- If you think controlling for model size is important
  - E.g. ViT-B vs. ViT-B ✅, ViT-B vs. ViT-S ❌
- Then in ML4RS, ViT-B vs. ViT-B ❌ (because they probably have different settings)

- Lets see the FLOPs

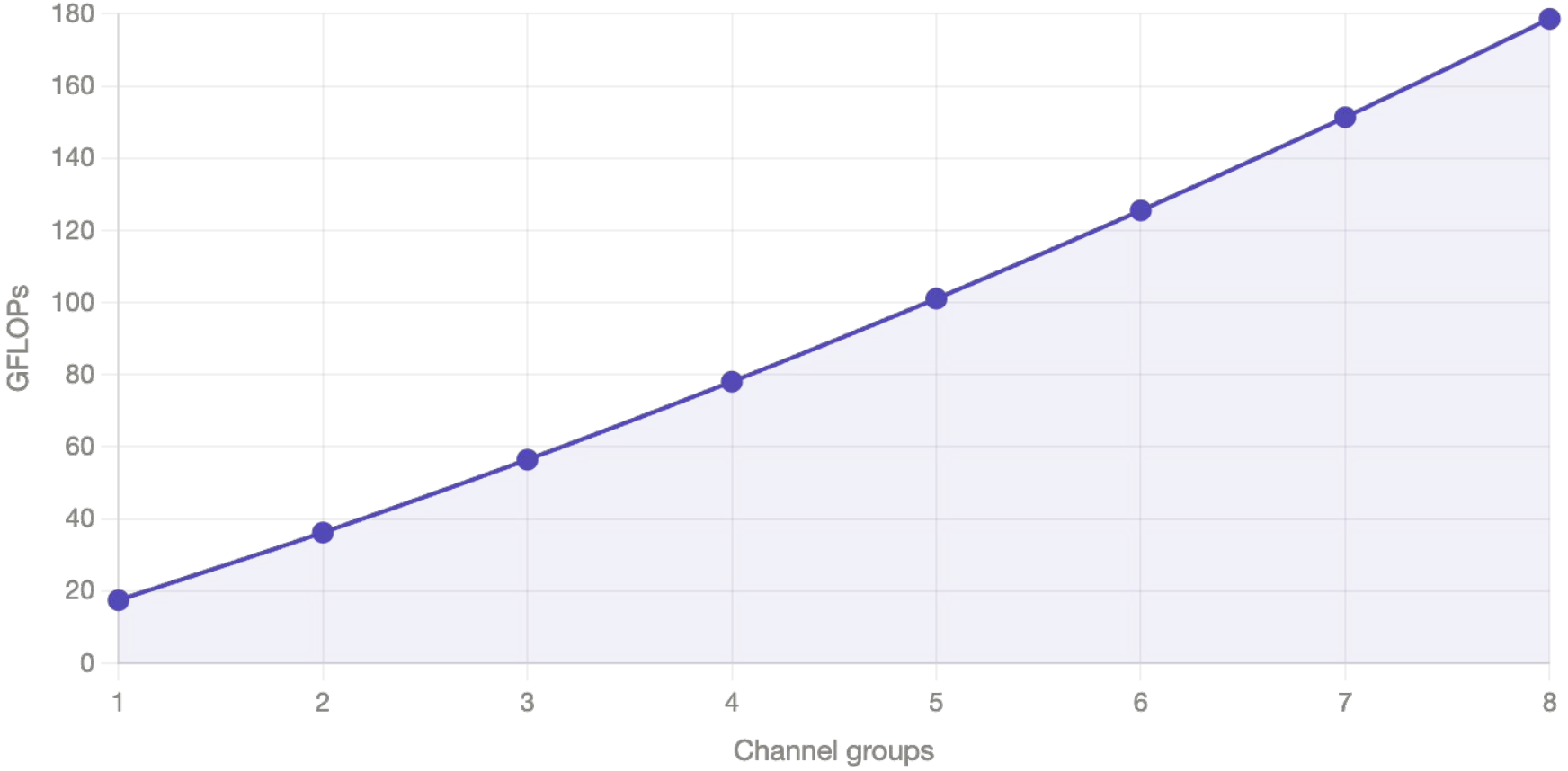
How FLOPs scale with patch size (everything else constant)



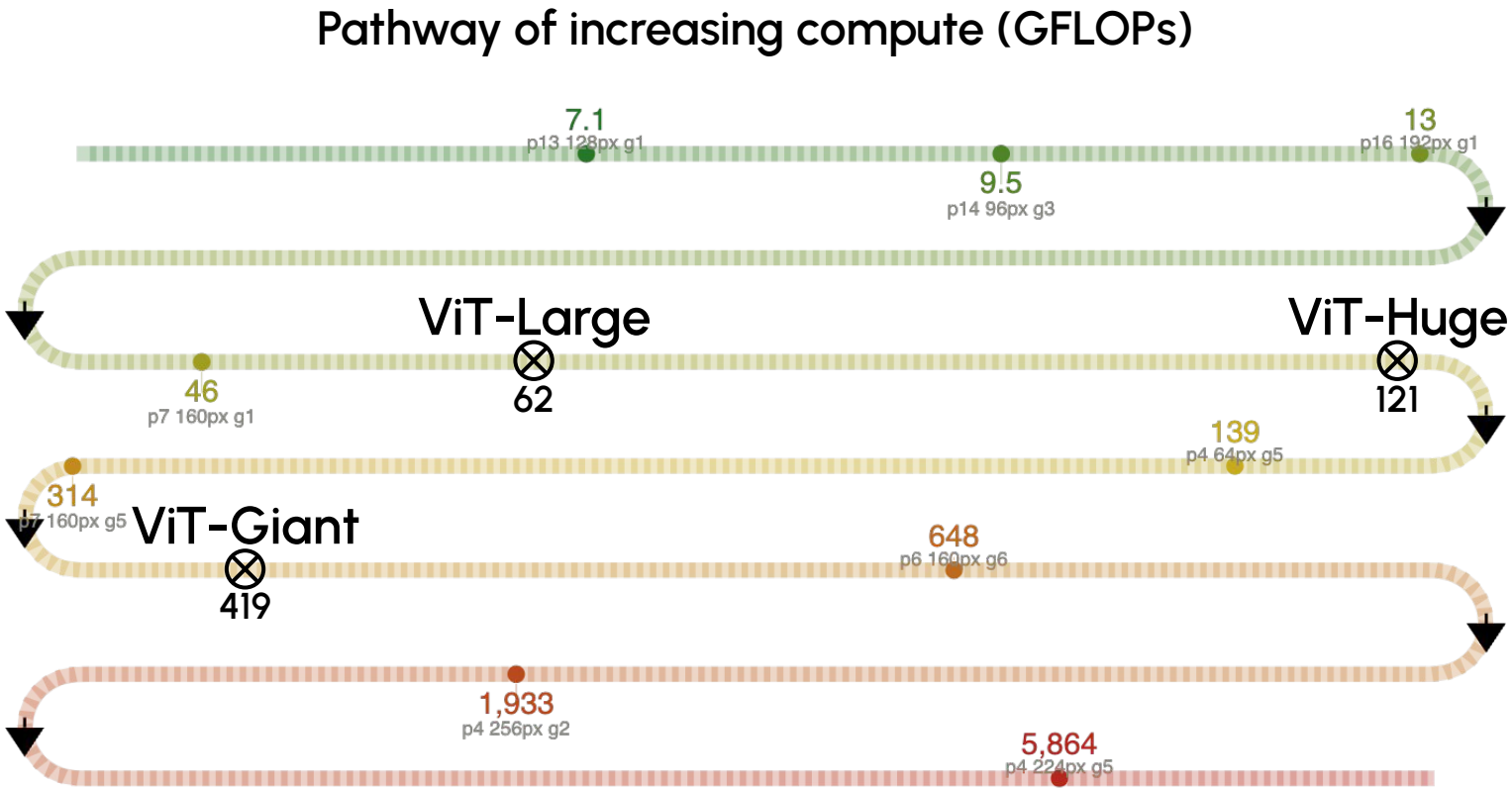
How FLOPs scale with image size (everything else constant)



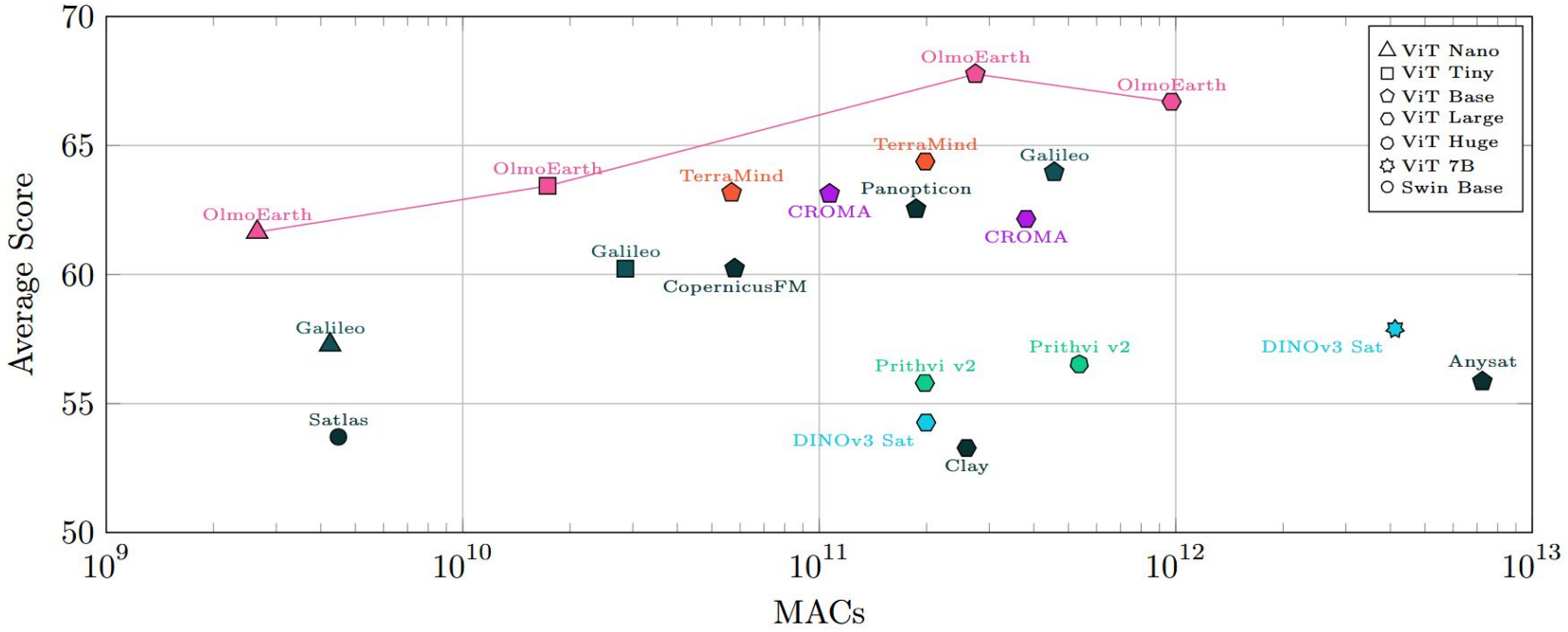
How FLOPs scale with channel groups (everything else constant)



ViT-Base models with varying patch size (p), image size (px), and channel groups (g)



Then lets control for FLOPs...



[figure credit: https://arxiv.org/abs/2511.13655]

- Most papers do *not* have these compute vs. accuracy plots—good on OlmoEarth team!
- Better? A little...

More Problems

More confounders

- These models differ in:
  - Pre-training data
    - Space and time distributions
    - Number of samples
    - Use of pseudo-labels (e.g. Dynamic World)
    - Modalities
  - Pre-training compute
    - E.g. 10 epochs vs. 100 epochs vs 1000 epochs
    - And this matters!

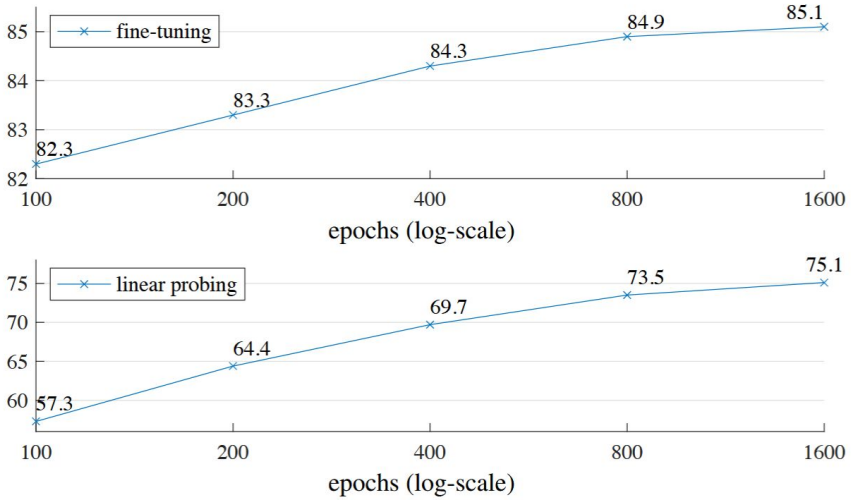


Figure 7. **Training schedules**. A longer training schedule gives a noticeable improvement. Here each point is a full training schedule. The model is ViT-L with the default setting in Table 1.

[figure credit: https://arxiv.org/pdf/2111.06377]

Fine, let's do ablations and controlled experiments, which controls for model compute, training data, and training length

Ablations from Galileo:

Table 7. Deep targets combined with structured space-time masking excels in **global** feature extraction. Segmentation tasks are gray-ed to focus on classification with our global task. We measure % top-1 accuracy via kNN.

masking strategy	target enc. computation	loss function	MADOS	Floods	CropH.	EuroSat
space+time	varied	PatchDisc$_B$	58.91	76.92	88.72	89.50
random	varied	PatchDisc$_B$	11.71	69.62	82.12	17.40
random+space+time	varied	PatchDisc$_B$	22.87	71.62	76.53	66.30
space+time	0	PatchDisc$_B$	61.73	76.66	85.79	86.90
space+time	6	PatchDisc$_B$	63.83	76.93	88.17	89.20
space+time	12	PatchDisc$_B$	60.35	77.19	87.30	87.90
space+time	varied	MSE	62.35	76.78	86.02	87.20
space+time	varied	PatchDisc	25.74	71.68	75.30	62.50

Table 8. Deep-shallow contrastive learning combined with unstructured random masking excels in **local** feature extraction. Classification tasks are gray-ed to focus on segmentation with our local task. We measure % mIoU (↑) of linear prediction on frozen features.

masking strategy	target enc. computation	loss function	MADOS	Floods	CropH.	EuroSat
random	0	PatchDisc	71.48	77.39	86.77	86.90
random+space+time	0	PatchDisc	68.63	77.82	85.31	88.80
space+time	0	PatchDisc	62.25	77.22	86.82	87.00
random	6	PatchDisc	58.53	75.66	76.58	65.40
random	12	PatchDisc	11.65	72.60	71.92	27.50
random	varied	PatchDisc	8.25	68.89	77.83	18.40
random	0	MSE	65.34	77.09	86.71	87.40
random	0	PatchDisc$_B$	70.12	77.26	85.27	88.20

Table 9. Our dual-objective algorithm excels on both classification and segmentation, and is more consistent than our single-objective algorithms. MADOS and Sen1Floods11 (% mIoU) via linear probing. CropHarvest and EuroSat (% top-1 acc.) via kNN.

global loss	local loss	share predictors	target context	MADOS	Floods	CropH.	EuroSat
PatchDisc$_B$	PatchDisc	no	all	64.37	77.33	87.72	89.70
PatchDisc	PatchDisc	no	all	67.79	77.66	87.87	91.00
PatchDisc$_B$	PatchDisc	no	dec.	63.54	76.95	86.98	89.30
PatchDisc	PatchDisc	no	dec.	36.98	74.21	85.49	83.30
PatchDisc	PatchDisc	no	dec.+enc.	63.41	77.36	85.87	89.30
PatchDisc	PatchDisc	yes	all	67.04	78.23	85.23	88.50
PatchDisc$_B$	PatchDisc$_B$	no	all	67.88	77.08	86.61	89.50
MSE	MSE	no	all	62.36	77.17	86.28	88.70

- Better? A little...

More Problems

Ablations rarely tune and controlled experiments are hard

	Case (default)	Ablation	batch size	cost	Classification (mAP)			Segmentation (mIoU)			
					R	O	RO	R	O	RO	Avg
	all default		7.2k	1.0×	78.2	84.5	84.8	40.8	56.0	56.5	66.8
❶	objectives (both)	MAE-only	7.2k	1.0×	−10.4	−6.0	−5.6	−9.2	−8.4	−5.1	−7.5
		contrast-only	14k	0.6×	−3.2	−3.0	—	−2.9	−3.7	—	—

- I used the same learning rate for CROMA vs. MAE-only vs. contrast-only
  - These are very different algorithms that have different optimal learning rates
- I knew this at the time but couldn't afford more runs!
- (Small changes are probably fine without tuning)

- And controlled experiments require:
  - Lots of compute to run and fairly tune prior methods on your data
  - Correct implementations of prior methods

Solution

Standardize and simplify (pt. 1 of 2)

- All models in a table should use:
  - Pre-training data: SSL4EO
  - Image size: 120x120 px
  - Time steps: 4 (amount available in SSL4EO)
  - Channel groups: 3
  - Patch size: 8x8 px
  - Model sizes: ViT-T, ViT-S, ViT-B
  - Epochs: 100, 300, optional 600

- More rules:
  - Tune your pre-training learning rates as much as you want
  - Use standard benchmarks and recipes (e.g. resize all inputs to 120x120 px)
  - Only worry about ablating the core components of ***your*** method

- If industry wants to use different data, more timesteps, more modalities, train for longer, or chant incantations while training on their 256xH100s...

LET THEM

Standardize and simplify (pt. 2 of 2)

- If your model doesn't beat OlmoEarth++, that's fine!
- Your model must only beat others that play by these rules

- Yes, it'll be less multi-modal and have less timesteps than you may want
  - This means no weather variables, lidar inputs, or geographic coordinates in *academic* SSL models

But the experimental science of SSL for RS will *begin* and we'll have...

GOOD TABLES