SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, Marc Rußwurm

Extracting **relevant** and **meaningful** characteristics of a location from satellite datasets is **challenging**.

Many location encoding techniques **underrepresent** regions and fail to generalize to unseen locations.

Contents

- Glossary
- Introduction
- CLIP
- SatCLIP
- Dataset
- Experiments
- Strengths and Weaknesses
- Key Takeaways

CLIP

CLIP **connects** text and images

Matches a batch of the same **image-caption** pairs.

A boy skateboarding

A dog skateboarding

A cat skateboarding

CLIP

CLIP **connects** text and images

Matches a batch of the same **image-caption** pairs.

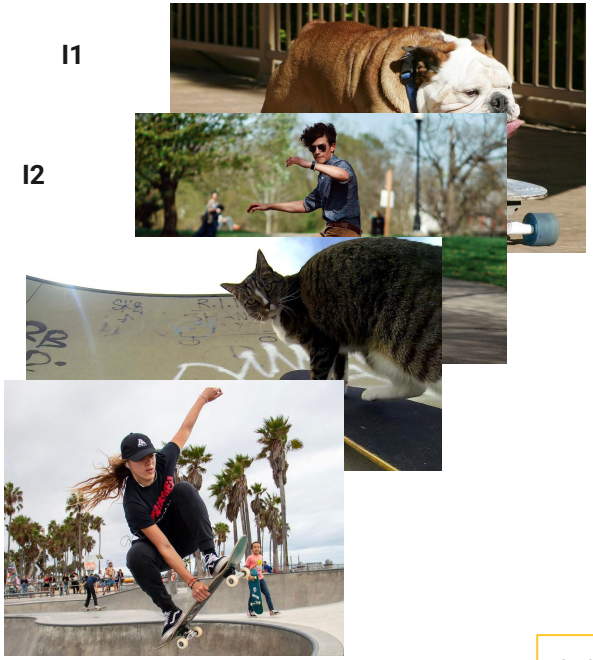
A boy skateboarding

A dog skateboarding

A cat skateboarding

CLIP

CLIP maximizes **cosine-similarity** for matching image-caption pairs and minimizes it for dissimilar pairs.



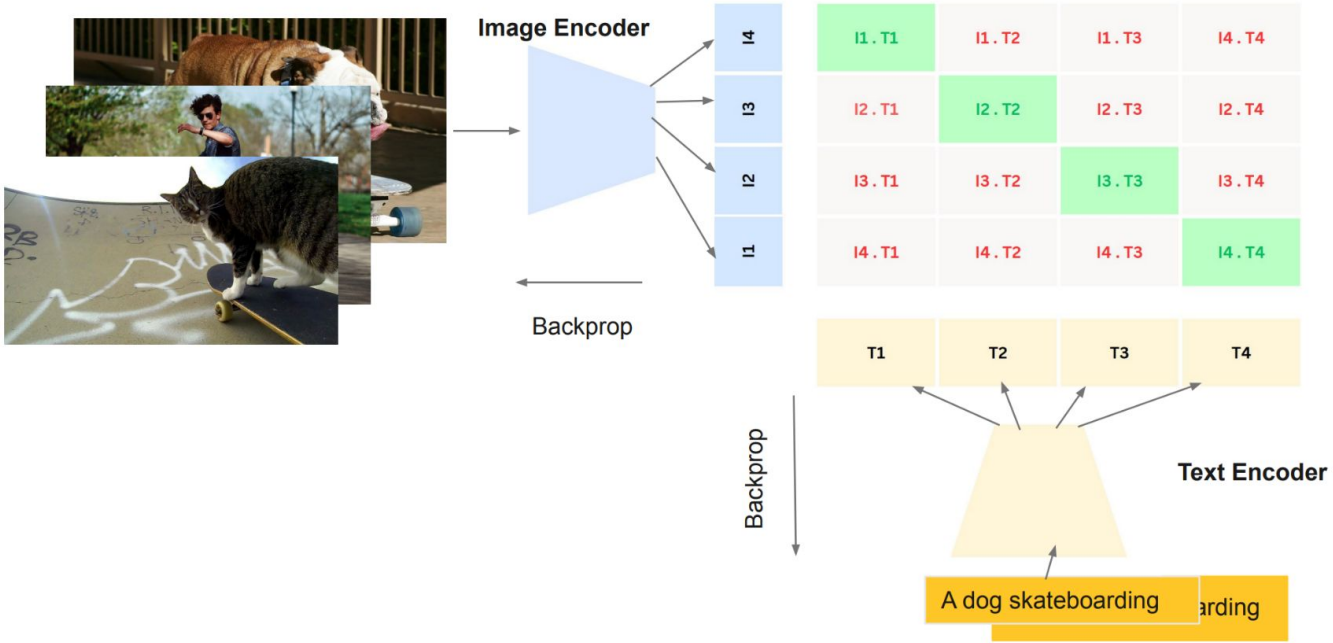
I1 . T1	I1 . T2	I1 . T3	I4 . T4
I2 . T1	I2 . T2	I2 . T3	I2 . T4
I3 . T1	I3 . T2	I3 . T3	I3 . T4
I4 . T1	I4 . T2	I4 . T3	I4 . T4

A dog skateboarding	A boy skateboarding	A cat skateboarding	A girl skateboarding
T1	T2	T3	T4

More Specifically

CLIP maximizes cosine-similarity of the **representations/embeddings** of same image-caption pairs

Minimizes cosine-similarity of the **representations/embeddings** of different image-caption pairs

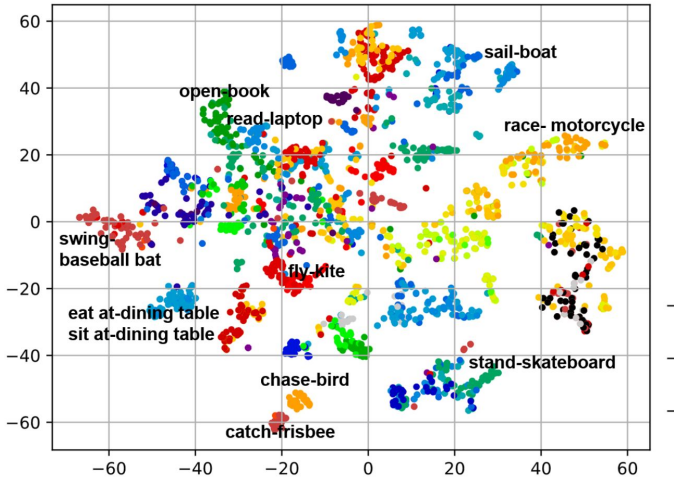


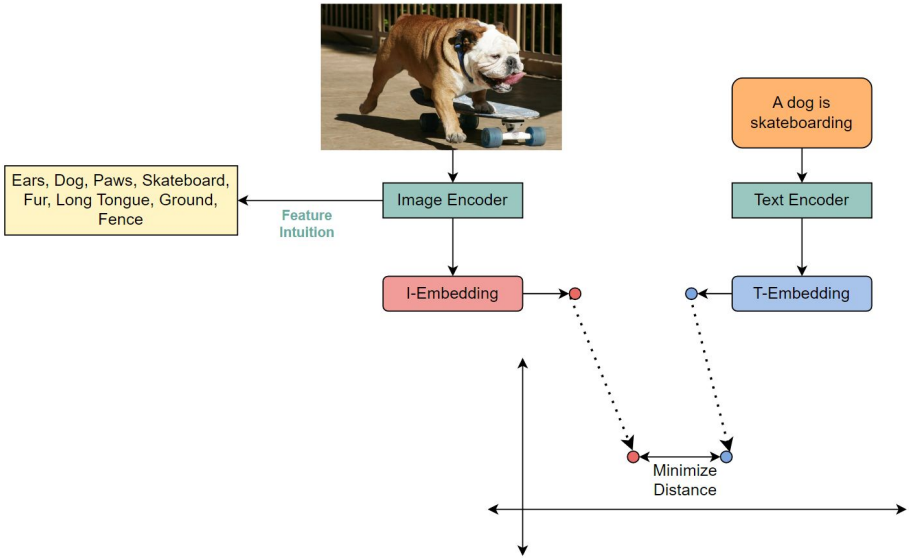
What does CLIP Learn?

Objective: Training CLIP integrates image characteristics into text embeddings, and vice versa.

Joint Embedding Space: Groups **semantically** similar images and text and pushes unrelated pairs within a shared space (D-dimension).

Outcome: Extensive connections are established between images and captions during Learning





SatCLIP

Satellite Images

?

Can we use CLIP here?

[78.95, -78.6]

[38.96, -42.8]

[66.89, 12.5]

Location Coordinates

Why SatCLIP?

Providing **Spatial information** improves model performance.

Spatial Signal

Lack of overlap between evaluation and training regions leads to poor generalization.

Location encoders inadequately represent and sample regions, hindering generalization.

Encoded locations don't have an implicit representation of satellite images.

Shortcomings

SATCLIP addresses these issues with semantically informed, globally representative embeddings obtained through contrastive pre-training.

Solution

SatCLIP

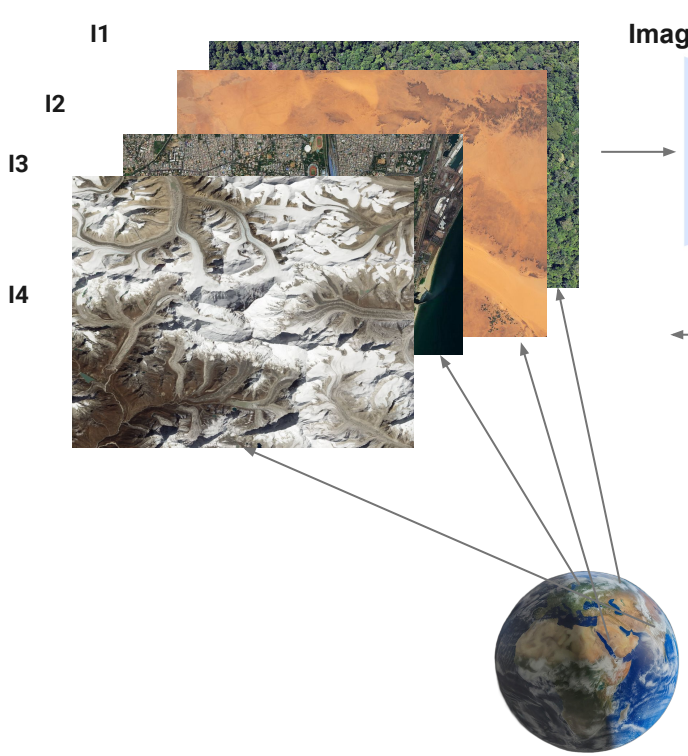


Image Encoder

I4
I3
I2
I1

Backprop

I1 . L1	I1 . L2	I1 . L3	I4 . L4
I2 . L1	I2 . L2	I2 . L3	I2 . L4
I3 . L1	I3 . L2	I3 . L3	I3 . L4
I4 . L1	I4 . L2	I4 . L3	I4 . L4

T → L

L1 L2 L3 L4

Backprop

Location Encoder

[78.95, -78.6] [38.96, -42.8] [66.89, 12.5] [-18.96, 23.9]

L1 L2 L3 L4

SatCLIP

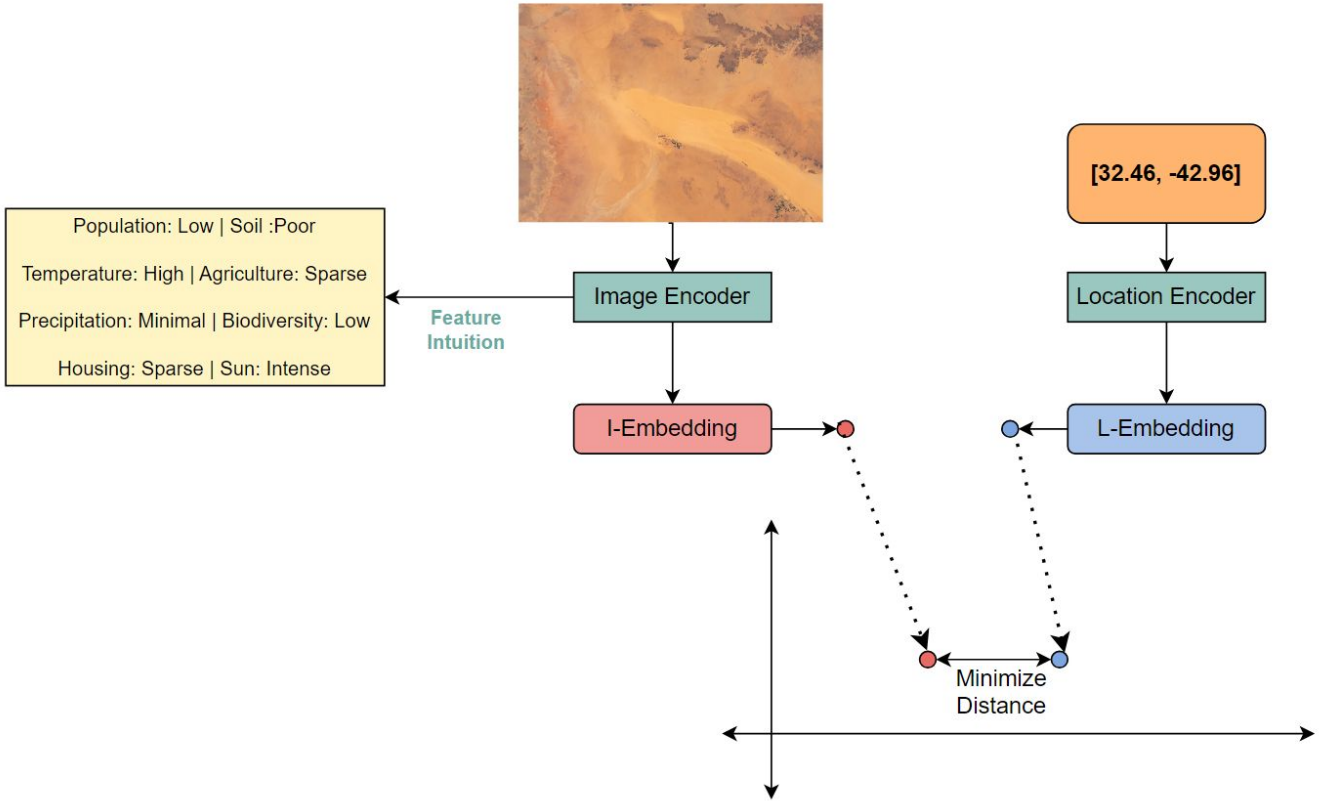
SatCLIP is a global, **generalized geo-location encoder** that learns an implicit representation of locations from multi-spectral satellite imagery.

It learns to **semantically group images** with the same locations together, i.e., images with similar geo-coordinates are close together.

Learns joint embedding of location and multi-spectral satellite imagery.

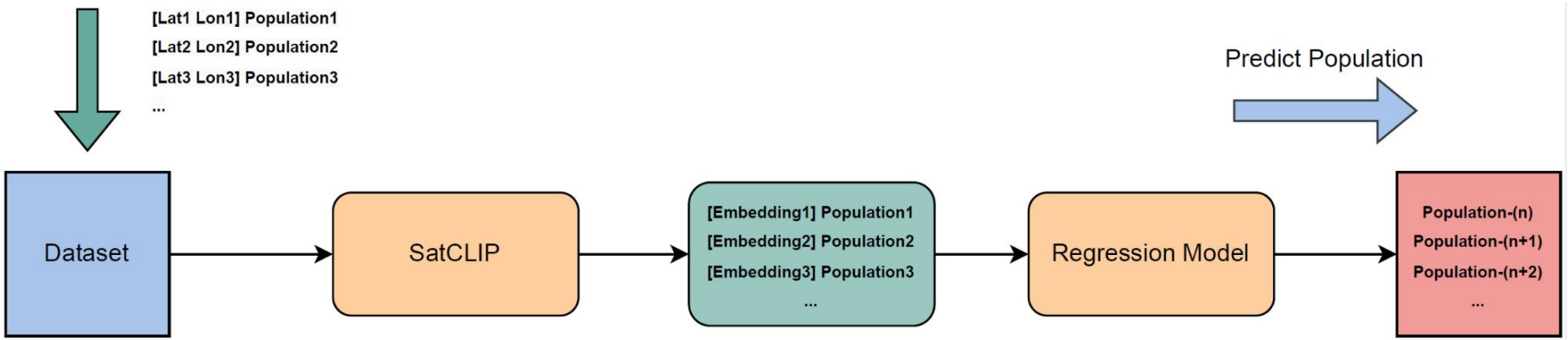
Image-location pairs are uniformly sampled across the **global landmass**.

What does SatCLIP learn?



Downstream Task

Population Density Prediction



Dataset

The S2-100K dataset comprises 100,000 multi-spectral (12-channel) satellite images sourced from Sentinel-2 through the Microsoft Planetary Computer.

Data from Sentinel-2 spans from **Jan 1, 2021, to May 17, 2023**.

The dataset is uniformly sampled across **landmasses** and exclusively consists of cloud-free images.

Experiments

Experiments

Downstream Adaptation: Do SatCLIP embeddings generalize across a diverse set of geospatial modeling tasks?

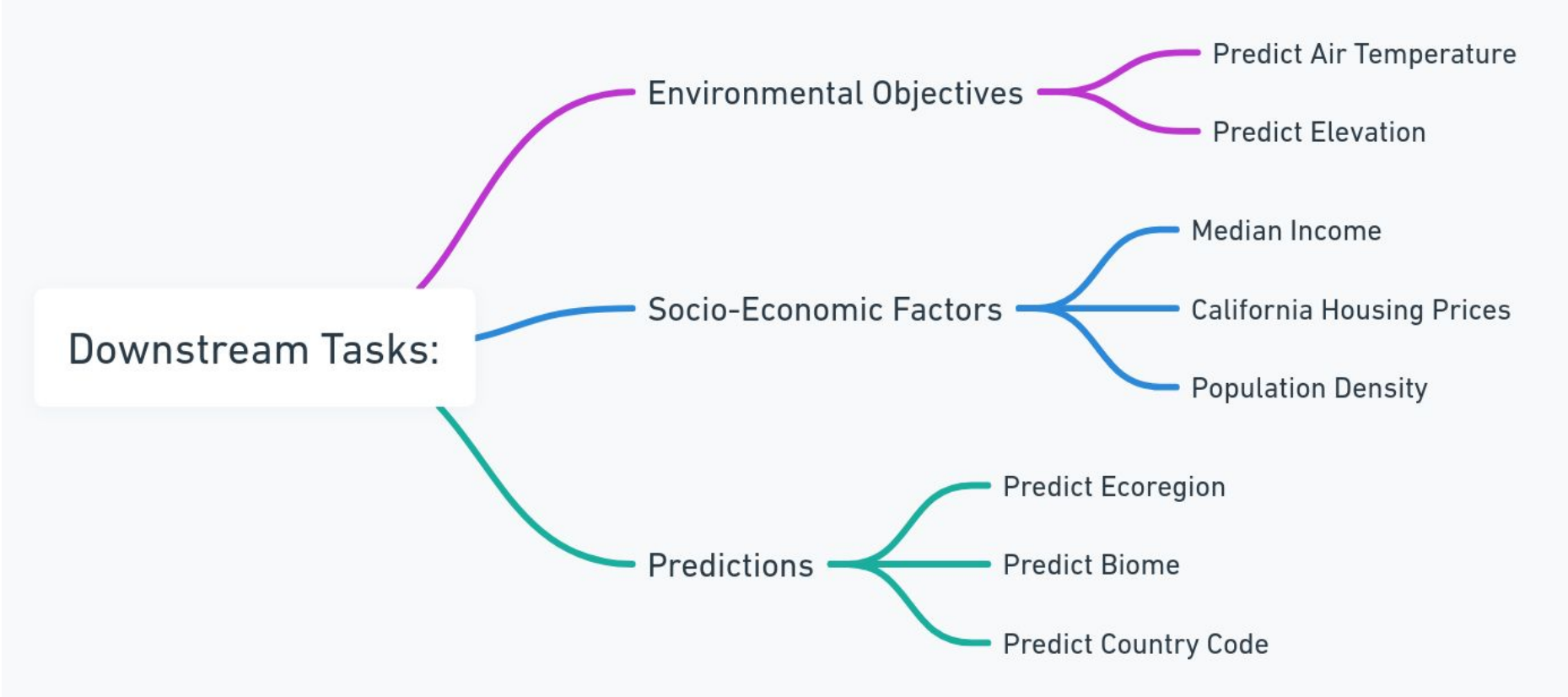
Geographic Adaptation: Do they adapt to Geographic domain shift?

Adaptability

Are they fine-grained enough to capture spatial variations?

Spatial Variation

Downstream Tasks

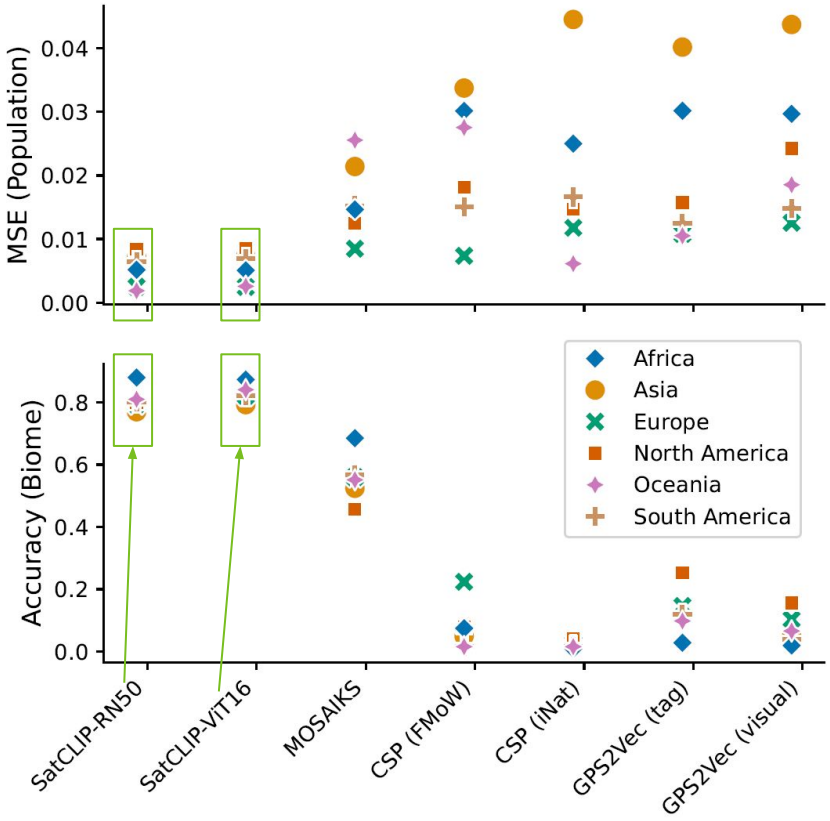


Downstream Performance

Table 2. **Downstream task performance using SatCLIP ($L = 40$) vs. baseline location embeddings.** We report average test set MSE and accuracy ± 1 standard deviation across 10 independently initialized MLP training runs.

Task ↓ Data →	SatCLIP-RN50 (S2-100K)	SatCLIP-ViT16 (S2-100K)	CSP (FMoW)	CSP (iNat)	GPS2Vec (tag)	GPS2Vec (visual)	MOSAIKS (Planet)
Regression	MSE ↓						
Air temperature	$\mathbf{0.27 \pm 0.03}$	$\mathbf{0.25 \pm 0.02}$	2.81 ± 1.11	4.71 ± 1.78	2.37 ± 0.00	2.92 ± 0.01	4.61 ± 6.05
Median income	$\mathbf{0.71 \pm 0.16}$	$\mathbf{0.67 \pm 0.01}$	1.39 ± 0.07	1.35 ± 0.03	1.06 ± 0.00	1.31 ± 0.00	1.31 ± 0.07
Cali. housing	2.42 ± 0.12	2.62 ± 0.28	5.67 ± 0.00	5.68 ± 0.01	$\mathbf{1.64 \pm 0.15}$	2.20 ± 0.14	4.30 ± 0.11
Elevation	$\mathbf{0.15 \pm 0.00}$	$\mathbf{0.15 \pm 0.01}$	0.80 ± 0.05	1.11 ± 0.06	1.11 ± 0.01	1.17 ± 0.00	0.98 ± 0.01
Population	$\mathbf{0.48 \pm 0.01}$	$\mathbf{0.50 \pm 0.02}$	1.69 ± 0.16	1.72 ± 0.28	1.99 ± 0.00	2.28 ± 0.00	1.45 ± 0.05
Classification	% Accuracy ↑						
Countries	$\mathbf{96.00 \pm 0.14}$	95.77 ± 0.14	77.78 ± 1.66	82.11 ± 1.72	70.35 ± 0.06	67.80 ± 0.03	76.16 ± 0.50
iNaturalist	$\mathbf{66.03 \pm 0.54}$	$\mathbf{65.98 \pm 0.61}$	56.73 ± 0.83	60.47 ± 0.56	58.78 ± 0.48	53.27 ± 0.78	56.73 ± 0.80
Biome	$\mathbf{94.41 \pm 0.14}$	$\mathbf{94.27 \pm 0.15}$	75.81 ± 1.53	73.18 ± 5.58	69.69 ± 0.06	68.29 ± 0.11	79.61 ± 0.42
Ecoregions	$\mathbf{91.67 \pm 0.15}$	$\mathbf{91.61 \pm 0.22}$	76.87 ± 1.27	78.43 ± 1.71	68.46 ± 0.06	67.26 ± 0.02	70.48 ± 0.21

Downstream Performance



SatCLIP outperforms in 8/9 downstream tasks and separately across all continents

Geographic Adaptation

Table 3. **Geographic adaptation capabilities of SatCLIP ($L = 40$) vs. baseline location embeddings to new geographic areas with no ($*$) or very few ($\dagger$) samples from the held-out test continent.** We report average test set MSE and accuracy in % ± 1 standard deviation across 10 independently initialized MLP fine-tuning runs.

Test Continent	SatCLIP-RN50 (S2-100K)	SatCLIP-ViT16 (S2-100K)	CSP (FMoW)	CSP (iNat)	GPS2Vec (tag)	GPS2Vec (visual)	MOSAIKS (Planet)
Asia							
Air Temp.* MSE ↓	1.50 ± 0.10	$\mathbf{1.26 \pm 0.15}$	3.06 ± 1.24	5.07 ± 4.45	16.70 ± 16.50	4.15 ± 0.58	10.56 ± 11.82
Elevation*	3.28 ± 0.09	$\mathbf{2.06 \pm 0.28}$	4.73 ± 0.29	4.98 ± 0.13	5.09 ± 0.06	4.92 ± 0.02	4.22 ± 0.23
Pop. Density*	2.82 ± 0.22	$\mathbf{1.94 \pm 0.15}$	4.53 ± 0.38	7.10 ± 1.14	4.84 ± 0.13	5.54 ± 0.03	3.35 ± 0.43
Countries† % Acc. ↑	14.29 ± 1.62	$\mathbf{19.17 \pm 2.82}$	1.22 ± 0.05	1.28 ± 0.01	1.12 ± 0.00	0.92 ± 0.02	1.56 ± 0.47
iNaturalist*	17.67 ± 0.32	20.91 ± 0.77	19.85 ± 0.55	$\mathbf{21.49 \pm 0.85}$	17.52 ± 0.38	18.11 ± 0.34	16.14 ± 0.42
Biome*	30.26 ± 3.00	16.44 ± 1.21	1.98 ± 0.62	3.00 ± 2.60	1.76 ± 0.04	2.79 ± 0.19	$\mathbf{37.81 \pm 4.47}$
Ecoregions†	8.46 ± 0.79	$\mathbf{10.86 \pm 1.19}$	1.55 ± 0.17	1.41 ± 0.14	1.49 ± 0.03	1.48 ± 0.00	1.36 ± 0.10
Africa							
Air Temp.* MSE ↓	$\mathbf{2.17 \pm 0.33}$	$\mathbf{1.79 \pm 0.50}$	3.35 ± 2.30	2.65 ± 4.19	6.44 ± 0.03	5.99 ± 0.19	13.32 ± 13.27
Elevation*	0.81 ± 0.06	$\mathbf{0.57 \pm 0.04}$	0.66 ± 0.04	1.05 ± 0.26	$\mathbf{0.54 \pm 0.03}$	$\mathbf{0.55 \pm 0.01}$	0.85 ± 0.20
Pop. Density*	2.99 ± 0.23	$\mathbf{1.96 \pm 0.22}$	3.64 ± 0.49	3.10 ± 0.39	3.19 ± 0.06	3.25 ± 0.01	$\mathbf{2.03 \pm 0.12}$
Countries† % Acc. ↑	$\mathbf{8.95 \pm 1.04}$	$\mathbf{10.22 \pm 1.62}$	0.47 ± 0.01	0.45 ± 0.04	0.47 ± 0.01	0.45 ± 0.00	0.48 ± 0.00
iNaturalist*	5.22 ± 0.26	6.23 ± 0.47	6.63 ± 0.57	$\mathbf{8.65 \pm 0.52}$	7.47 ± 0.53	6.85 ± 0.39	5.18 ± 0.38
Biome*	33.77 ± 2.69	12.34 ± 1.75	0.94 ± 0.00	1.09 ± 0.48	1.29 ± 0.04	1.17 ± 0.21	$\mathbf{49.86 \pm 1.57}$
Ecoregions†	$\mathbf{13.54 \pm 2.06}$	$\mathbf{12.91 \pm 1.63}$	0.90 ± 0.00	0.94 ± 0.04	0.88 ± 0.01	0.90 ± 0.00	0.92 ± 0.12
# wins of 14 total tasks	3	**10**	0	2	1	1	3

Strengths

First paper that creates a SOTA general purpose pretrained geo-location encoder with global coverage.

First paper that **investigates** location encoders for out-of-domain tasks and global generalization.

The model and code are open-source and available for use in any downstream task.

Weaknesses

The foundational model doesn't reach its full potential due to training on a small dataset (CLIP - 400 million data points).

Consequently, the representations created lack fine granularity at high resolution.

The authors did not incorporate other modalities such as text in **geotagged posts** and **Points of Interest Data**.

The authors did not include **spatio-temporal** data.

Key Takeaways

SOTA generalized location encoders can be created via **uniform sampling** of geo-data points and **Contrastive Learning**.

The SatCLIP Objective integrates features from the image modality into the location modality.

SatCLIP generalizes well to **geo-distribution shifts** and offers zero-shot/one-shot capabilities for downstream tasks.

SatCLIP location embeddings enable the accomplishment of any out-of-domain location-based task.

Questions?