Learning Representations of Satellite Images From Metadata Supervision

Jules Bourcier [1,2] Gohar Dashyan [2] Karteek Alahari [1] Jocelyn Chanussot [1]

[1] Inria [2] Preligens

EUROPEAN CONFERENCE ON COMPUTER VISION

Problem

Motivation

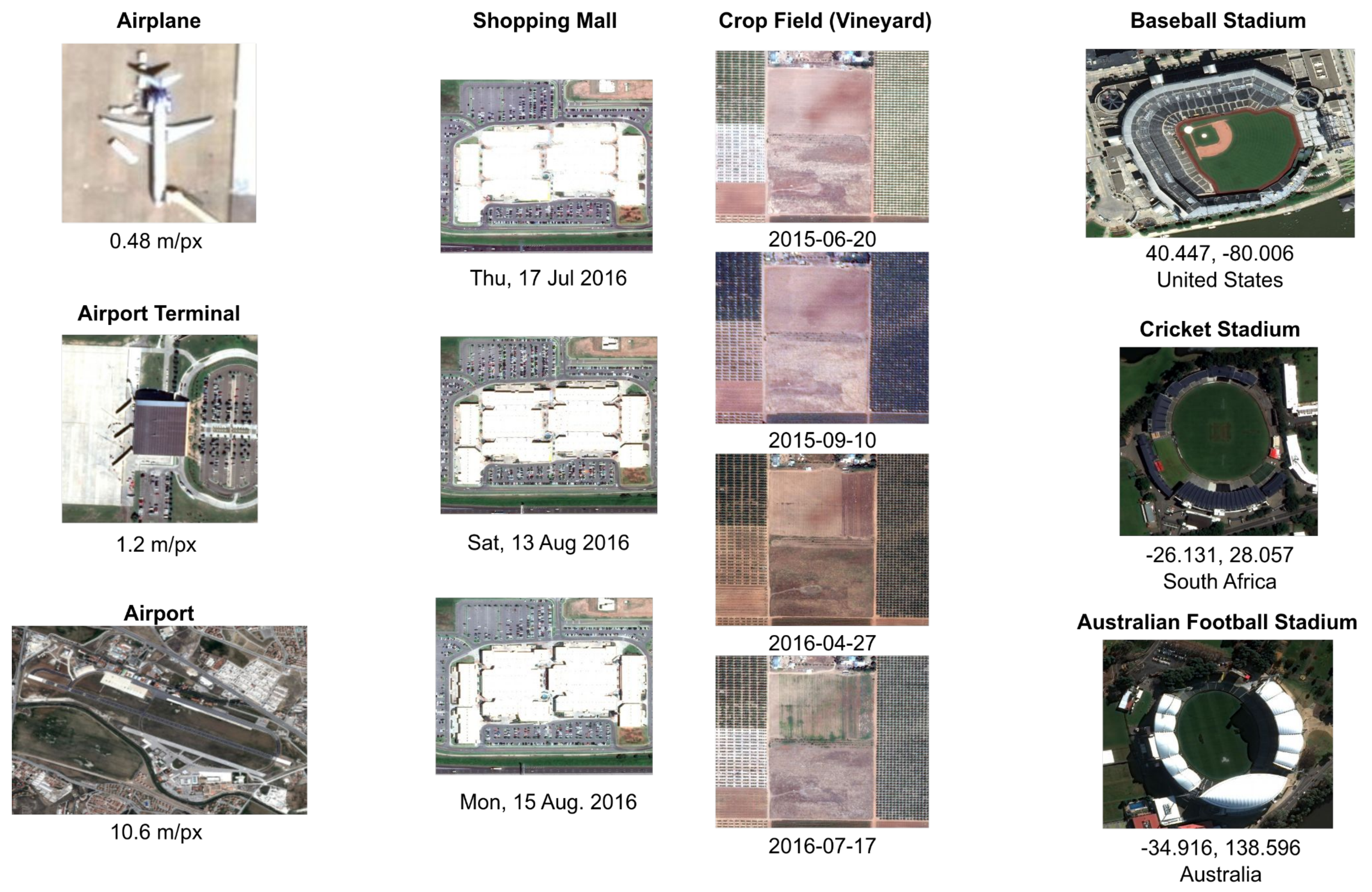
- Remote sensing is **data-rich** but **label-poor** ⇒ self-supervised learning is highly practical
- **Geospatial metadata** such as GSD, time and location give crucial info. about the **context** of an observation, and are **freely available**

Questions

- Can we **use geospatial metadata as a source of supervision** for learning rich representations of satellite images ?
- How does metadata supervision compares to and **interplays with visual self-supervision** ?

Approach

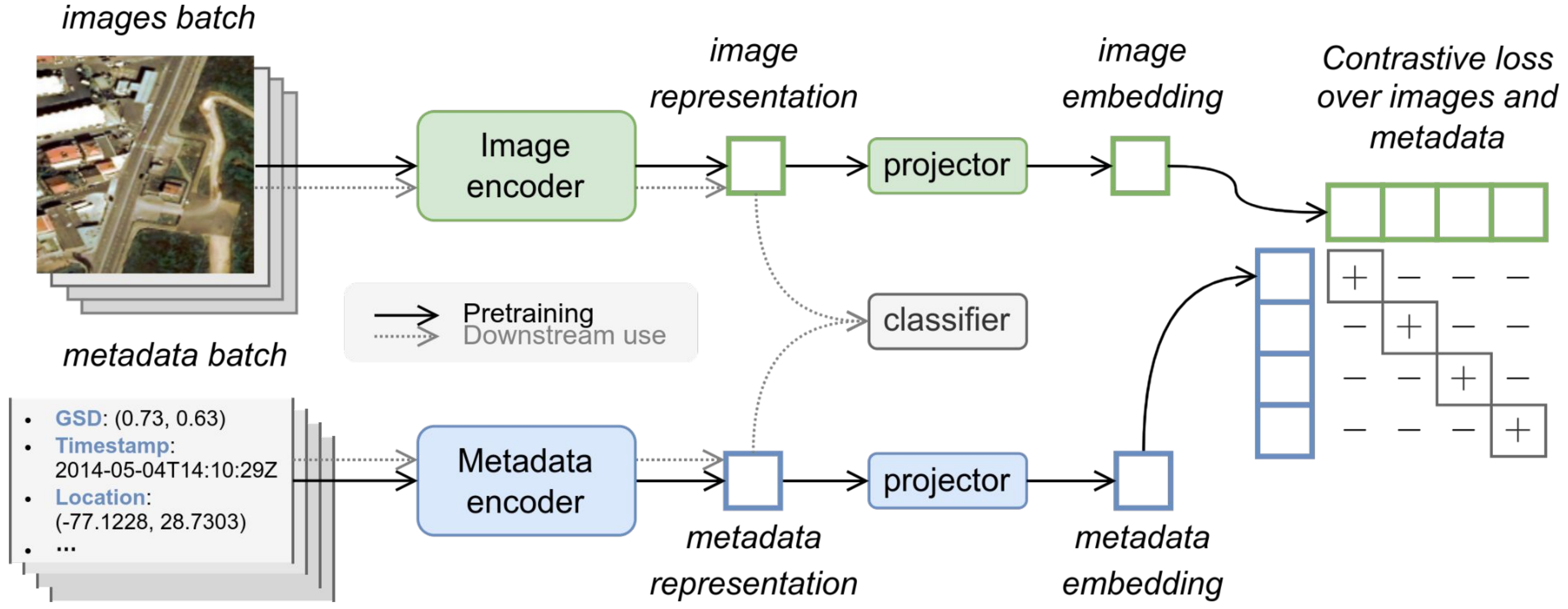
- Previous approaches predict metadata directly (Ayush et al. '21)
- Instead, see images and metadata as **two observation modalities**
- Model semantic interactions between images and metadata via **similarity of global features** of each modality



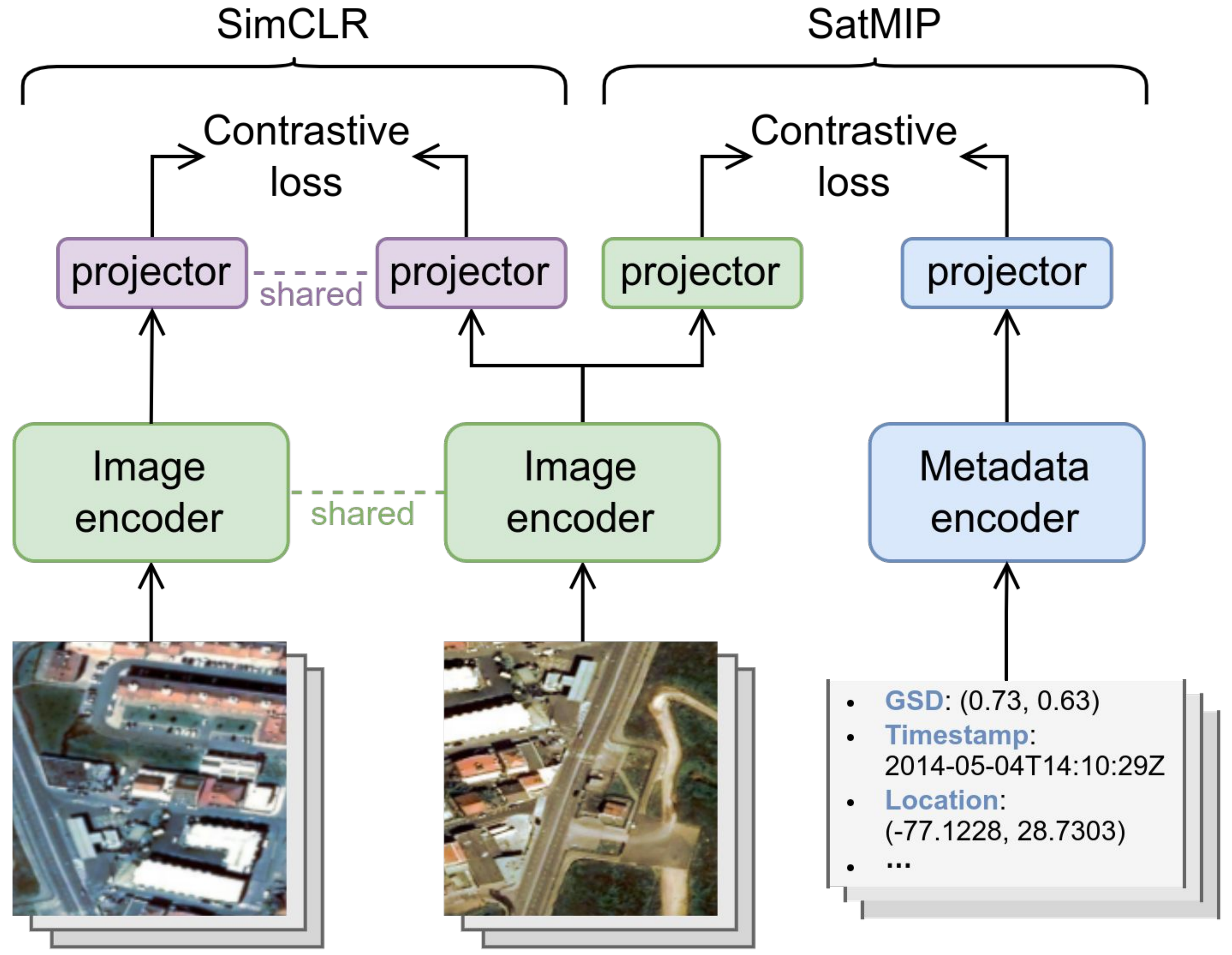
(a) GSD (b) Time (c) Location

Satellite Metadata-Image Pretraining

- **SatMIP** learns a **multimodal joint embedding** between images and metadata with a **contrastive** loss
- Encoding metadata with a **textual or tabular** Transformer (Gorishniy et al. '21)
- After pretraining: enables downstream **visual** classification or **bimodal classification** on image + metadata features



- Combining metadata and image self- supervision gives **SatMIPS**
- **Multitask** learning: share image encoder and jointly optimize SatMIP and SimCLR losses (Chen et al. '20)
- Using image **view coupling** for efficiency



- *Analogous LIP methods:*
  - SatMIP ~ CLIP (Radford et al. '21)
  - SatMIPS ~ SLIP (Mu et al. '22)

Geospatial Metadata

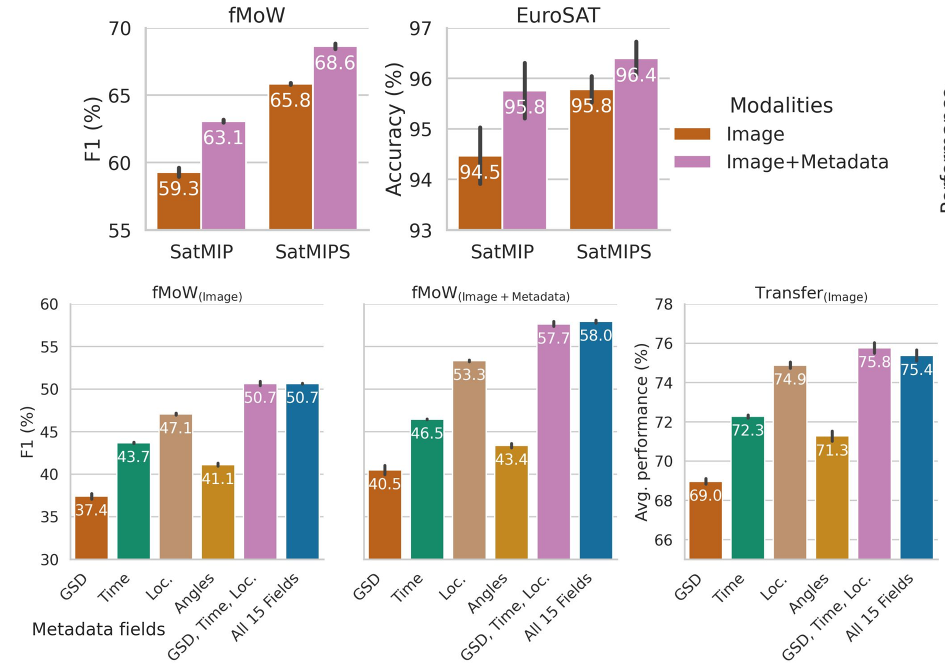
- Up to **15 metadata fields** from the environment and the sensor
- **Heterogeneous** fields: numerical (e.g. GSD), or categorical (e.g. sensor name)

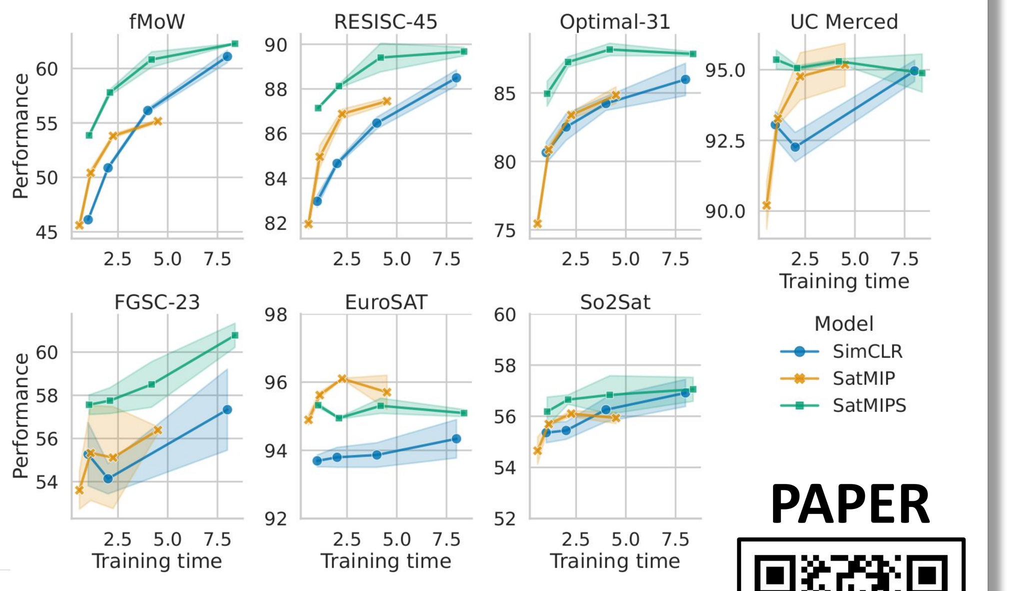
```
                  gsd: [11.4546, 9.1344]
   multispectral_gsd: 41.1896
          pixel_size: [1.03e-04, 8.23e-05]
           timestamp: 2015-09-21T15:30:08Z
            location: [-73.310776, -3.785814]
            utm_zone: 18M
        country_code: PER
         cloud_cover: 14
      scan_direction: Reverse
         wavelengths: [661, 545, 477]
target_azimuth_angle: 39.12
   sun_azimuth_angle: 77.28
 sun_elevation_angle: 70.44
     off_nadir_angle: 27.70
     sensor_platform: GEOEYE01
```

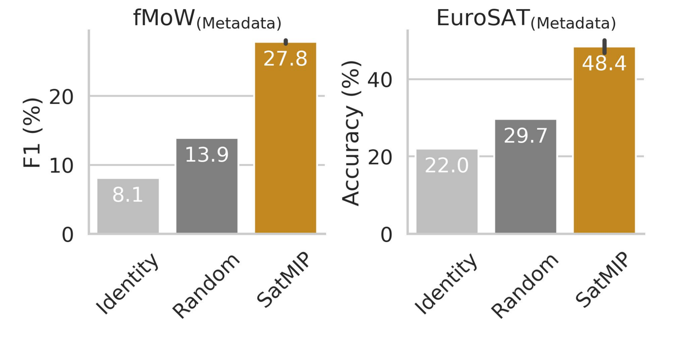
Results

- Pretraining on fMoW-RGB (Christie et al. '18), evaluation on 7 remote sensing image classification datasets, with kNN and linear probing.

Method	fMoW F1	Avg. transfer performance	Training time	Memory /GPU
SimCLR	$61.1_{\pm 0.6}$	$79.7_{\pm 0.7}$	1.0	1.0
SatMIP	$55.2_{\pm 0.2}$	$79.3_{\pm 0.4}$	**0.56**	**0.62**
SatMIPS	$\mathbf{62.3}_{\pm 0.0}$	$\mathbf{80.9}_{\pm 0.4}$	1.05	1.11







PAPER

CODE

- SatMIP yields a **meaningful pretext task**: competitive with SimCLR
- **Computationally efficient**: SatMIP trains 44% faster than SimCLR
- **Synergistic supervision sources**: SatMIPS outperforms SimCLR and converges faster (similar accuracy with about 2× less pretraining epochs)
- **Complementary features:** bimodal classification outruns visual-only classification
- **Location** is the most useful field, **combining multiple fields** constructively improves accuracy