

Learning to Map Anything, Anywhere, Anytime

Nathan Jacobs
Washington University in
St. Louis

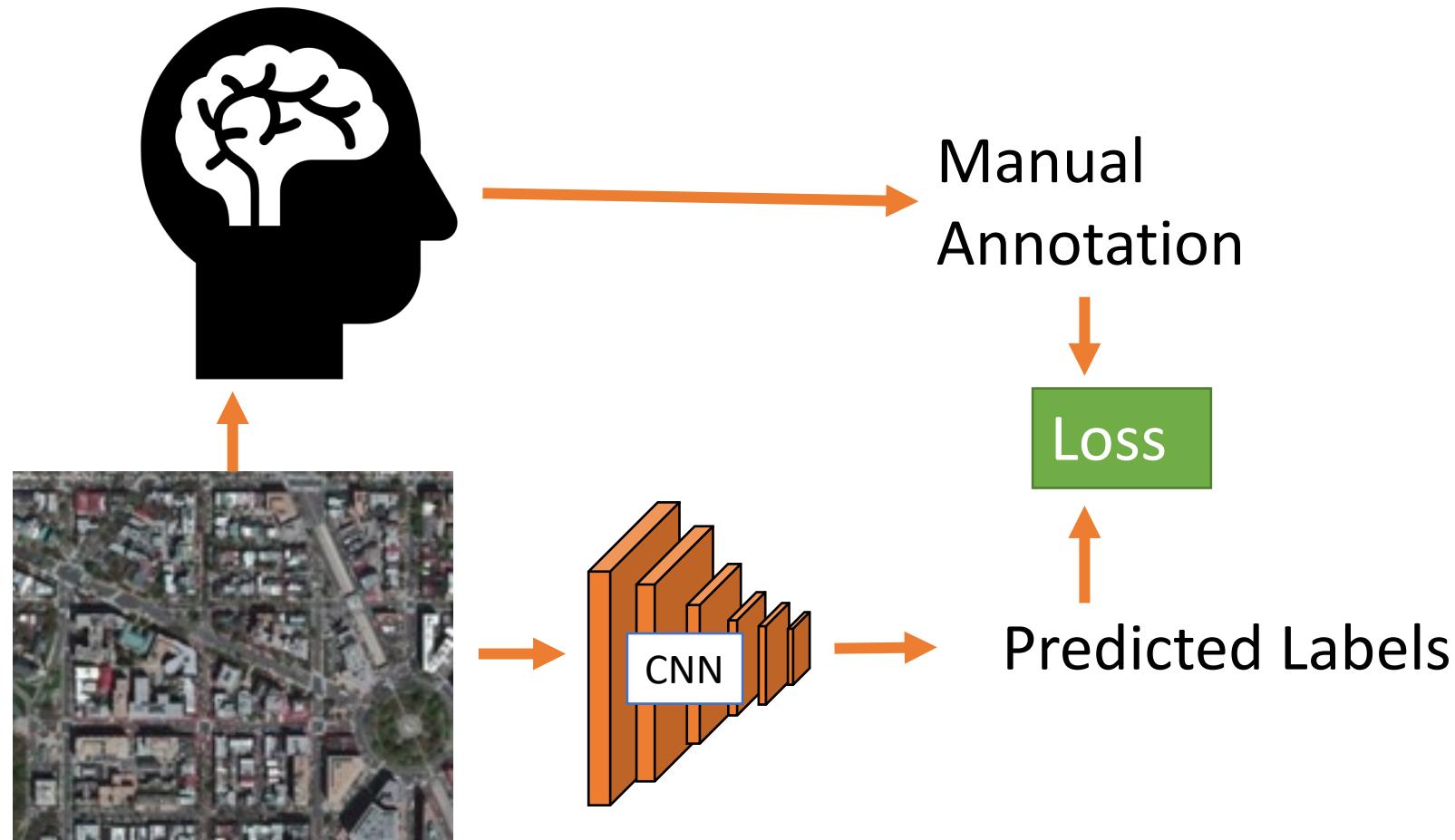
<https://mvrl.cse.wustl.edu/>



What is a map?

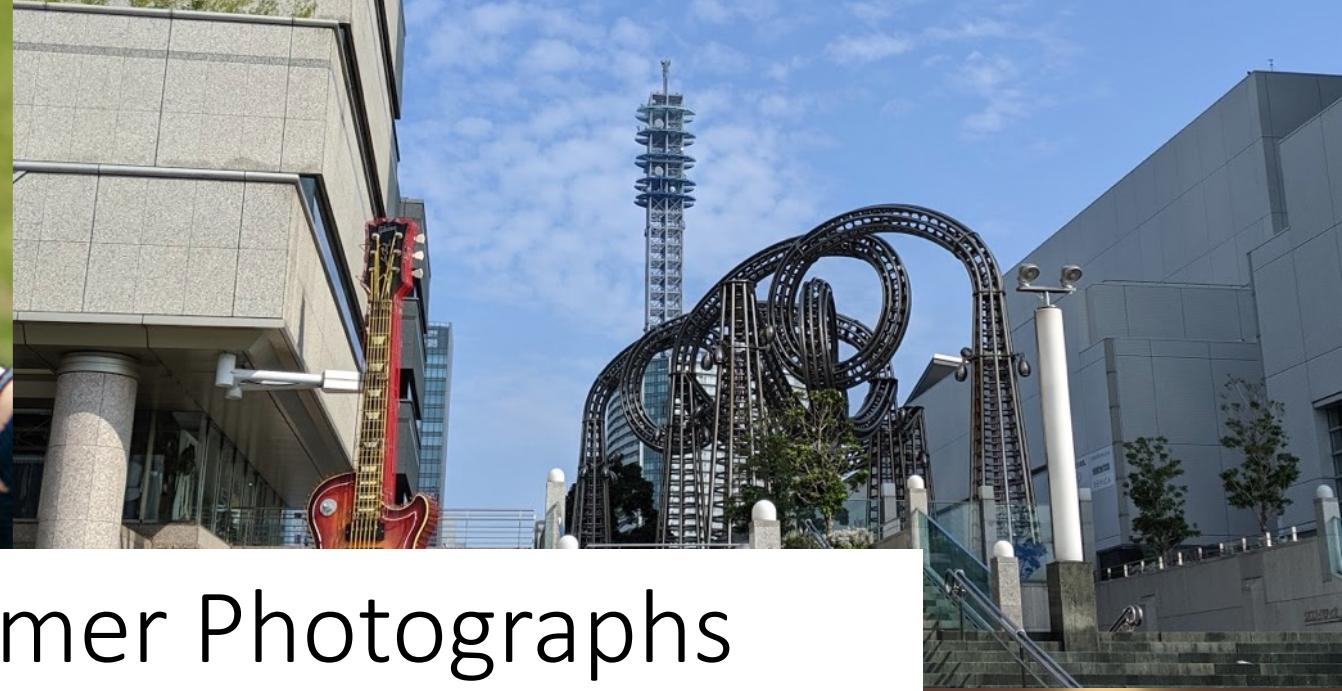


Standard Automated Approach: High-Quality Manual Annotations



What can you tell me about this building?





Billions of Consumer Photographs (Attribute Samples)



Image-Driven Mapping

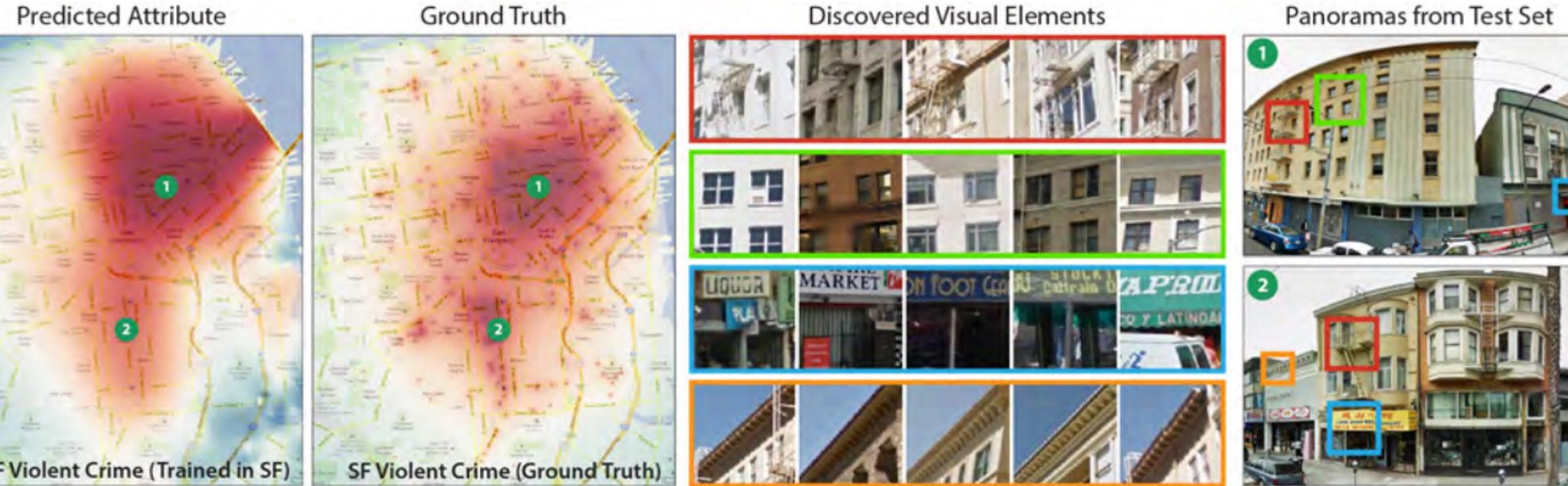
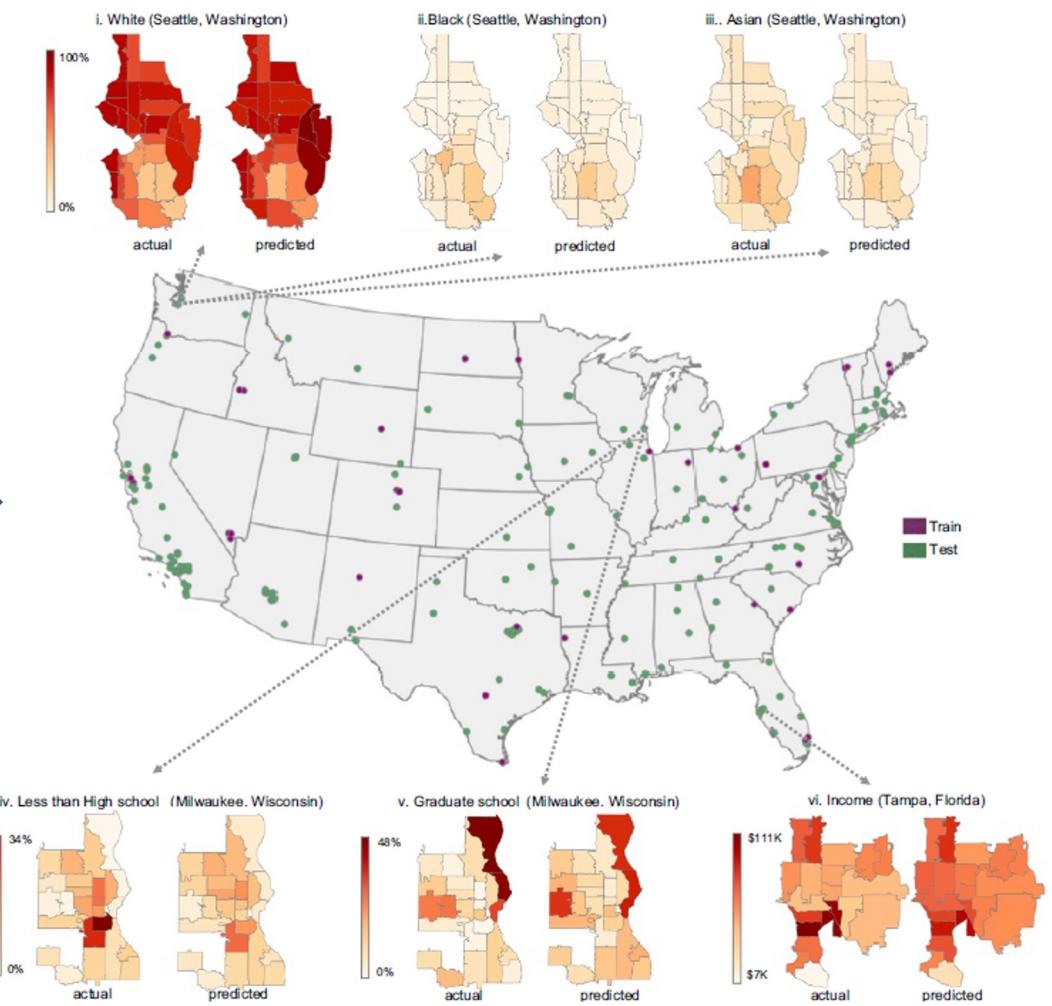
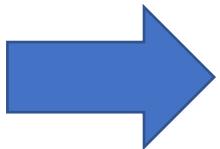
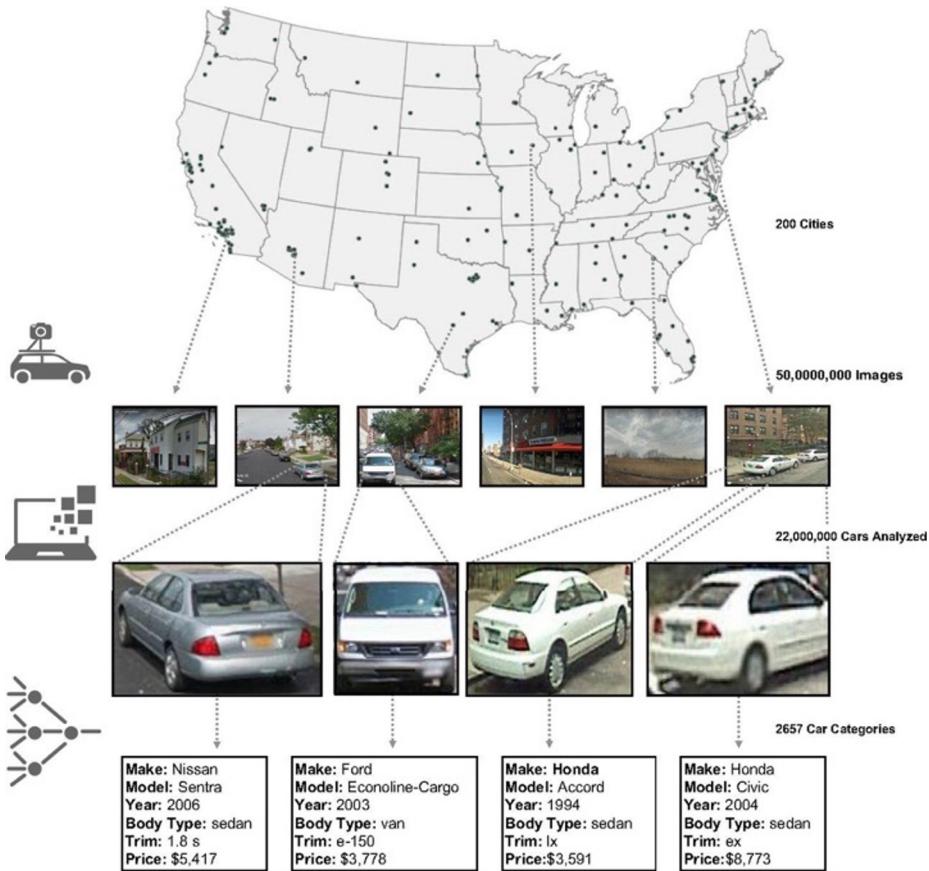
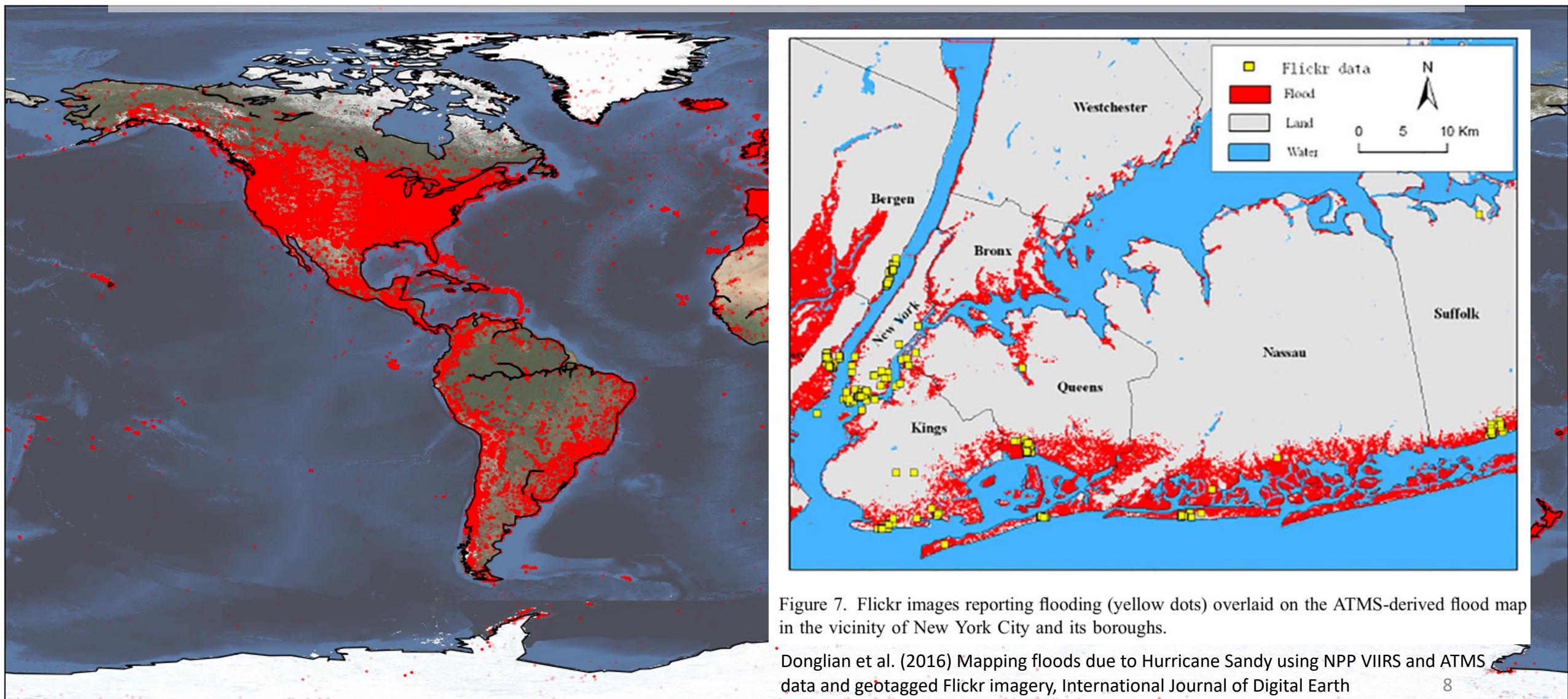


Image-Driven Mapping

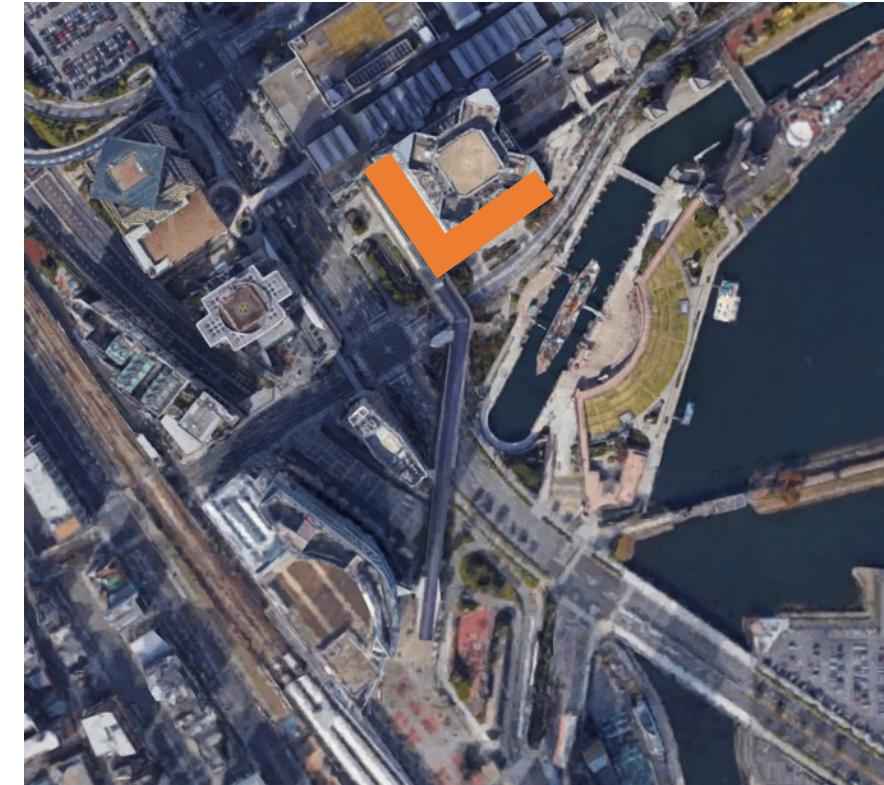


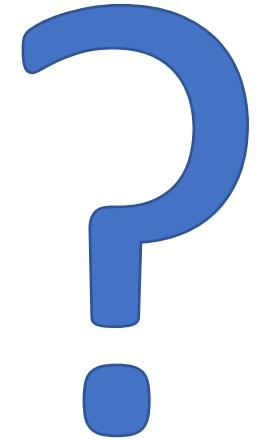
Gebru, Timnit, et al. "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States." PNAS (2017).

Problem #1: Ground-Level Images are Unevenly Distributed

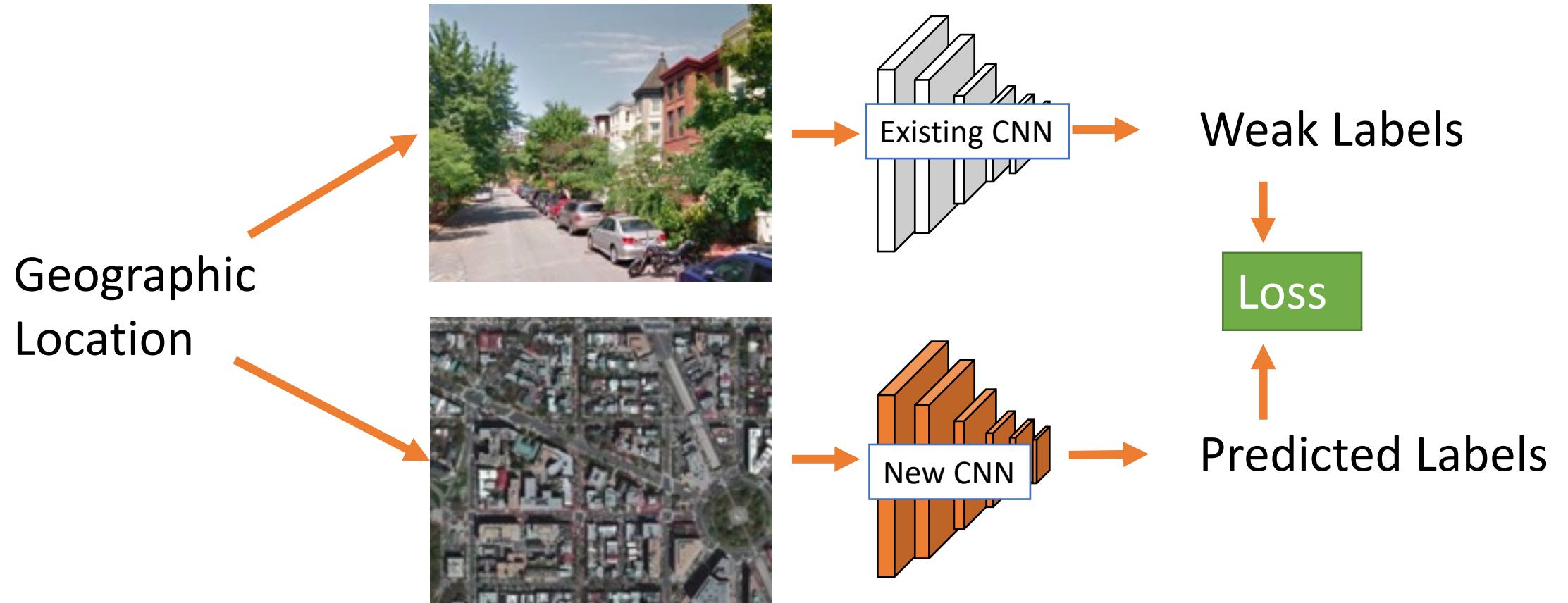


Problem #2: Boundary Delineation is Much Harder in Ground-Level Imagery





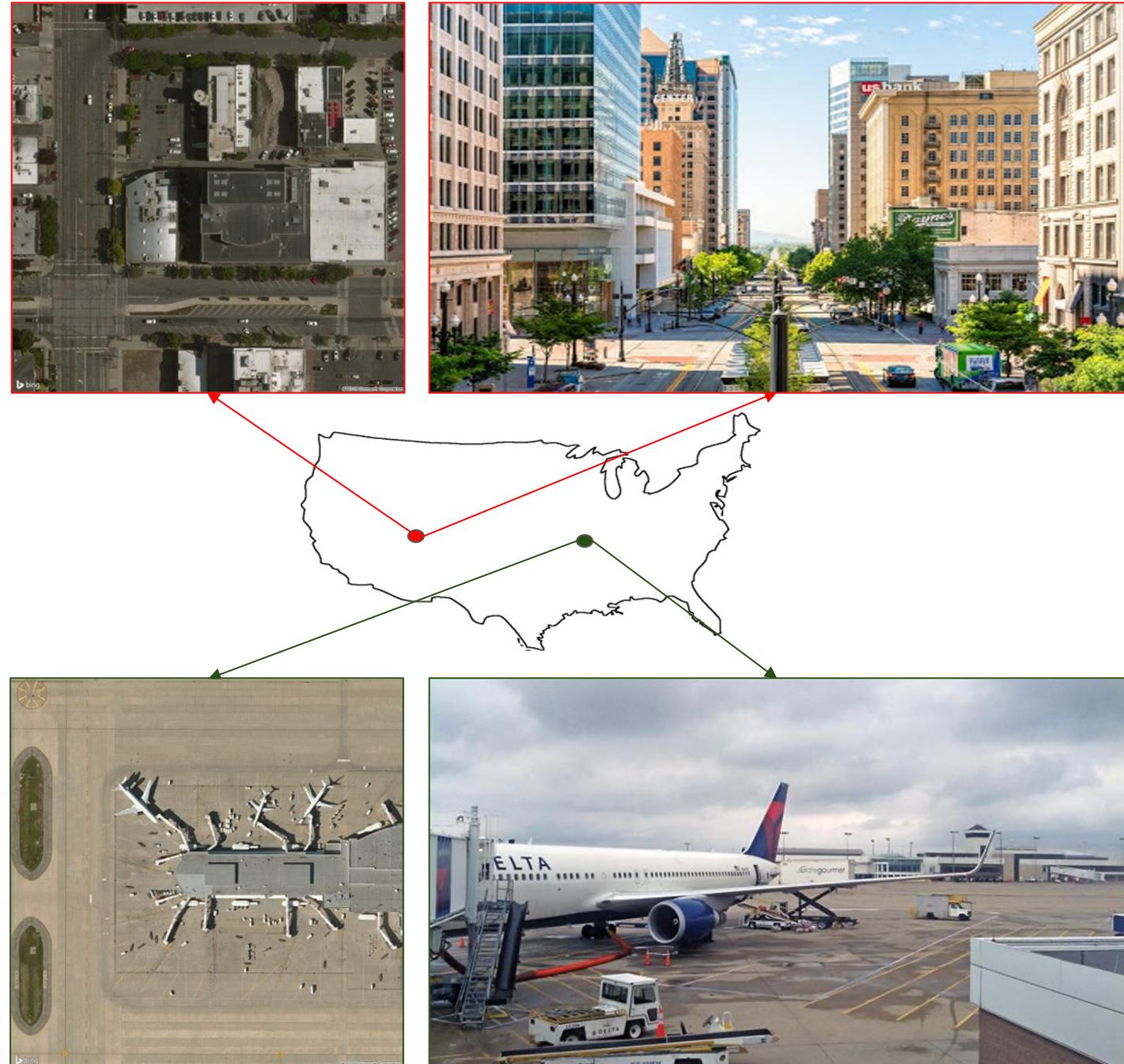
Idea: Use Consumer Photographs as a Weak Supervisory Signal (Cross-View Distillation)



$P(\text{attribute} \mid \text{OverheadImagery}(\text{location}))$

Motivation:

- 1) Strong models for extracting rich information from consumer photographs
- 2) Semantic overlap between ground-level and overhead images
- 3) Global coverage of overhead imagery



Example 1: Mapping Scene Categories

$P(\text{scene category} \mid \text{lat}, \text{lon})$



CVUSA: A Large Training Database of Ground-Level and Aerial Image Pairs

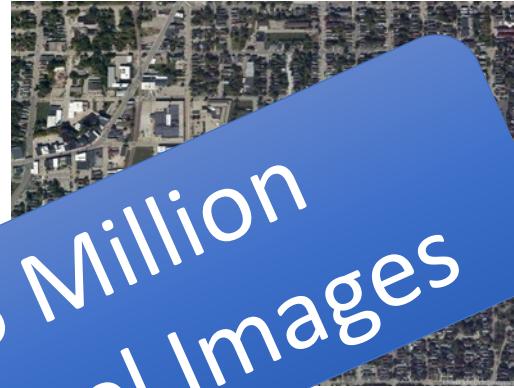
ground-level image



high-res overhead



med-res overhead



low-res overhead



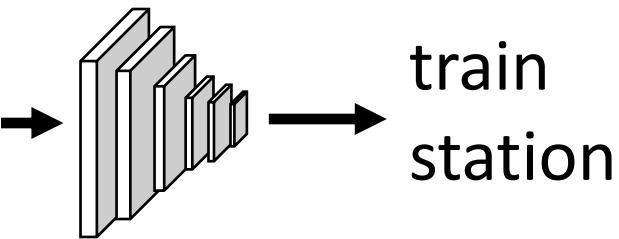
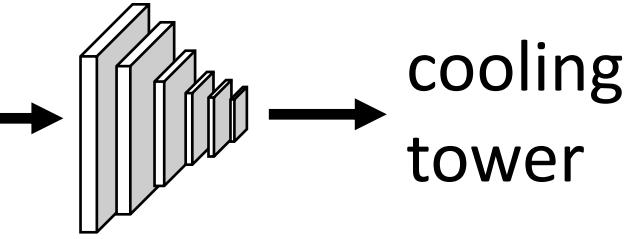
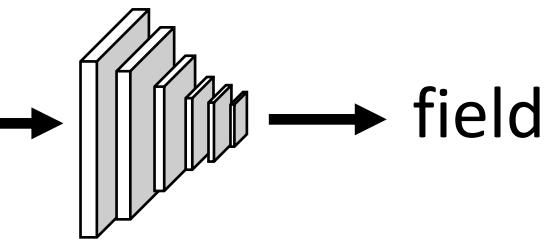
Over 1.5 Million
Ground-Level Images



:

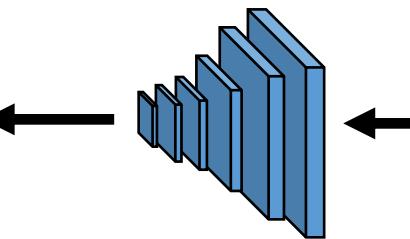
Learning to Predict Ground-Level Scene Categories from Overhead Imagery

Extract scene category

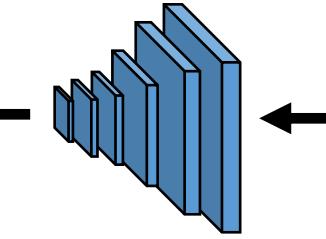


Optimize for maximum likelihood

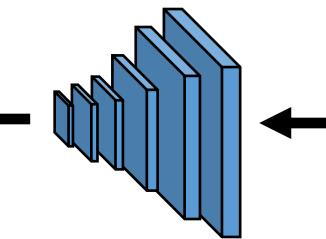
probably a
field



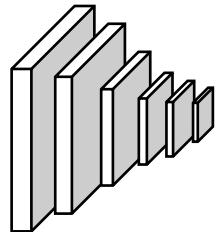
probably a
cooling tower



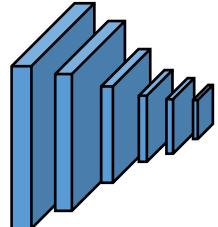
probably a
train station



Zero-Shot/Ad-Hoc Mapping

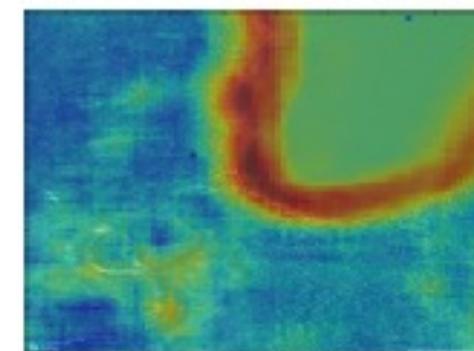
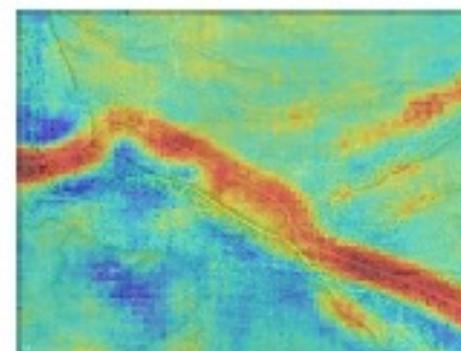
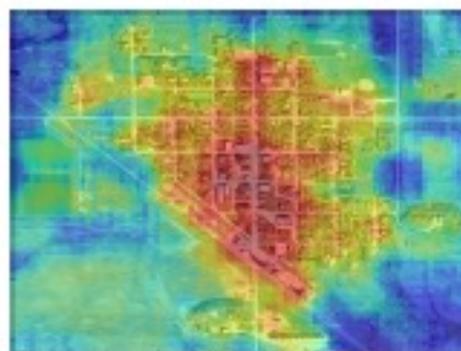
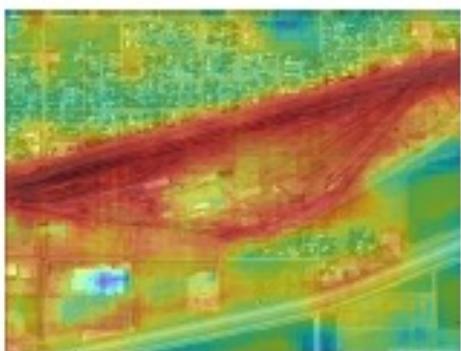
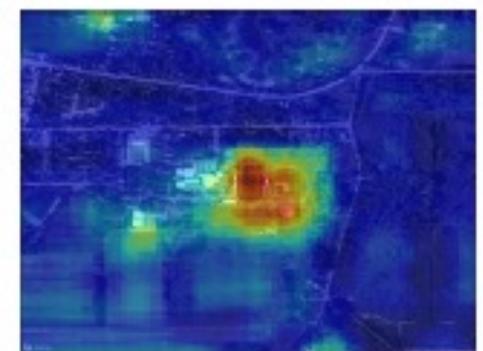
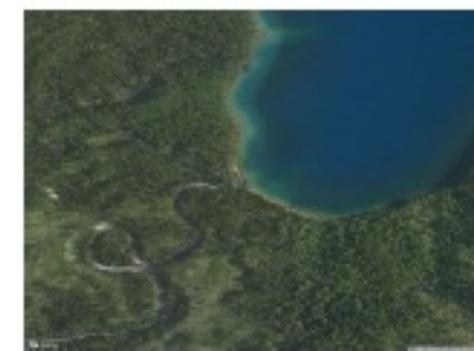


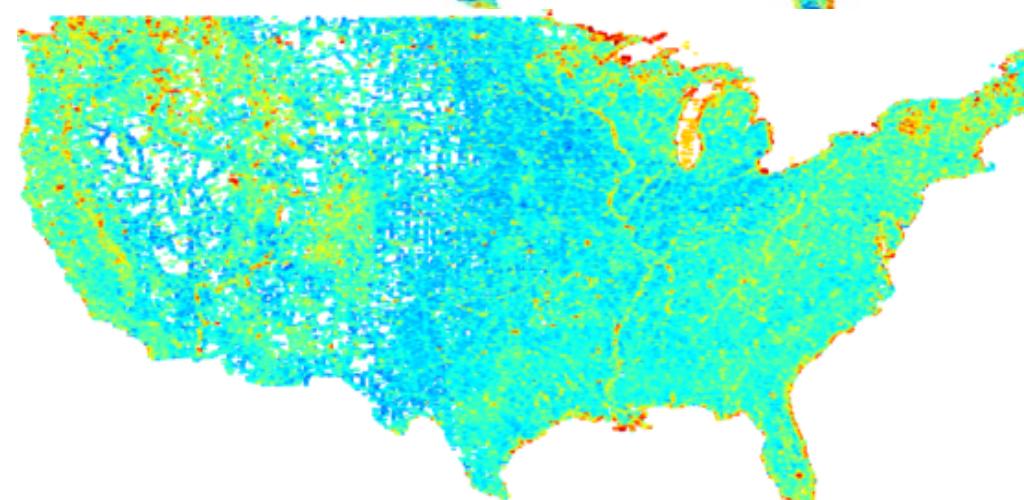
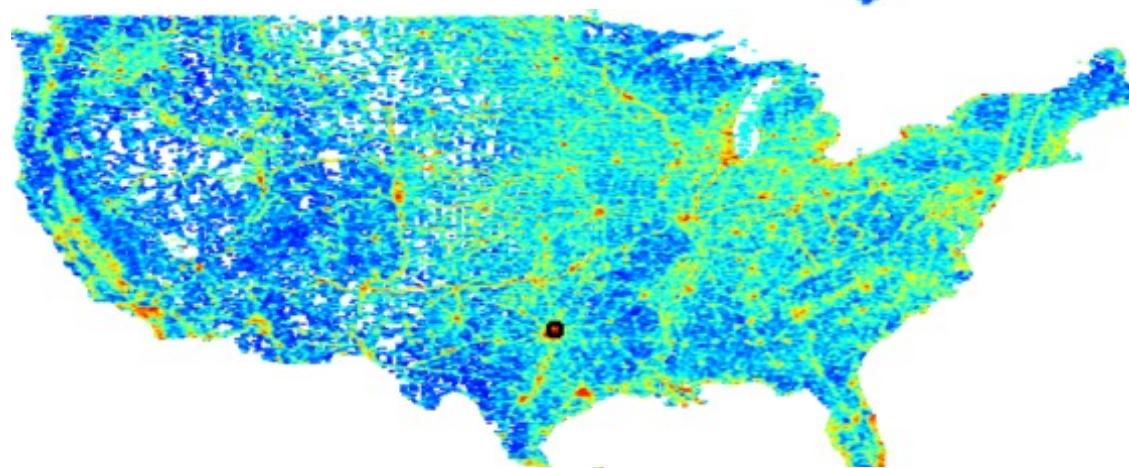
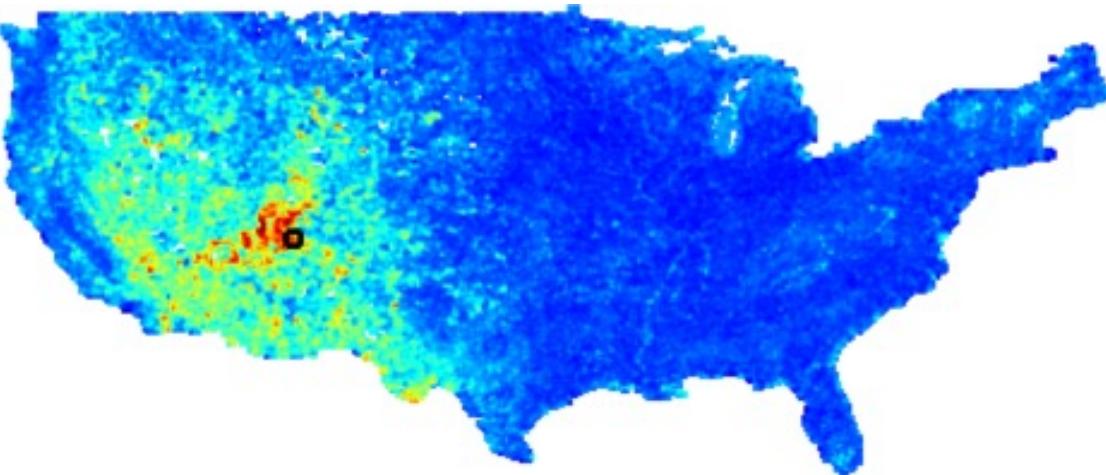
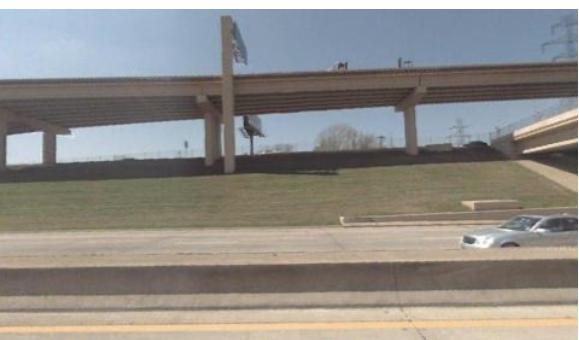
Description of
query image



Description of
location





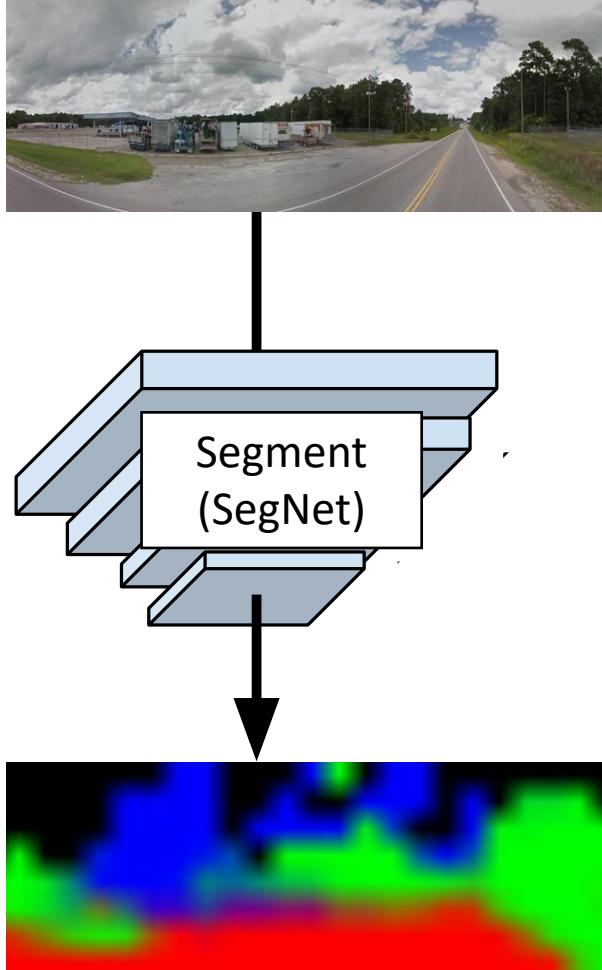


Example 2: “StreetView Anywhere”

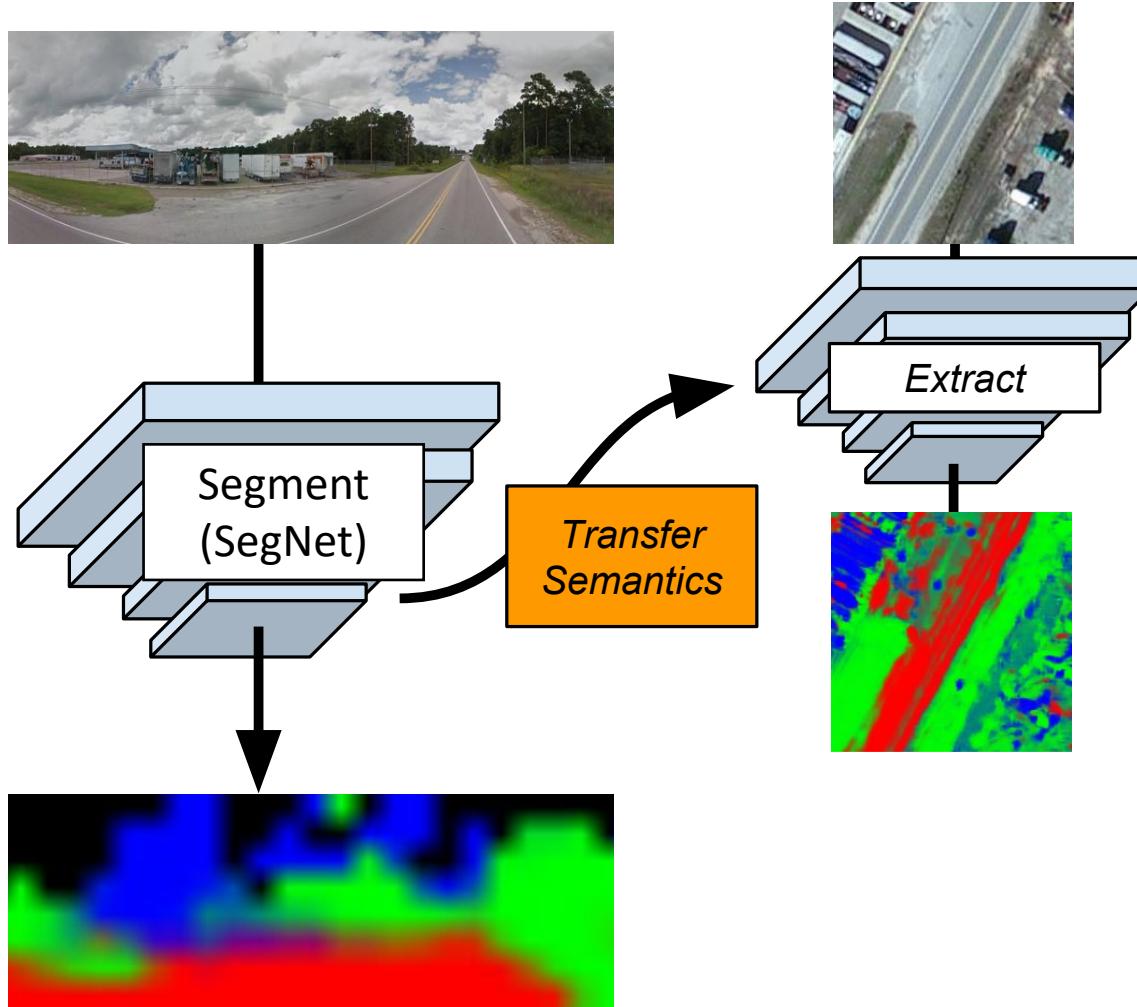
$$P(c, s | \theta, \phi, lat, lon)$$



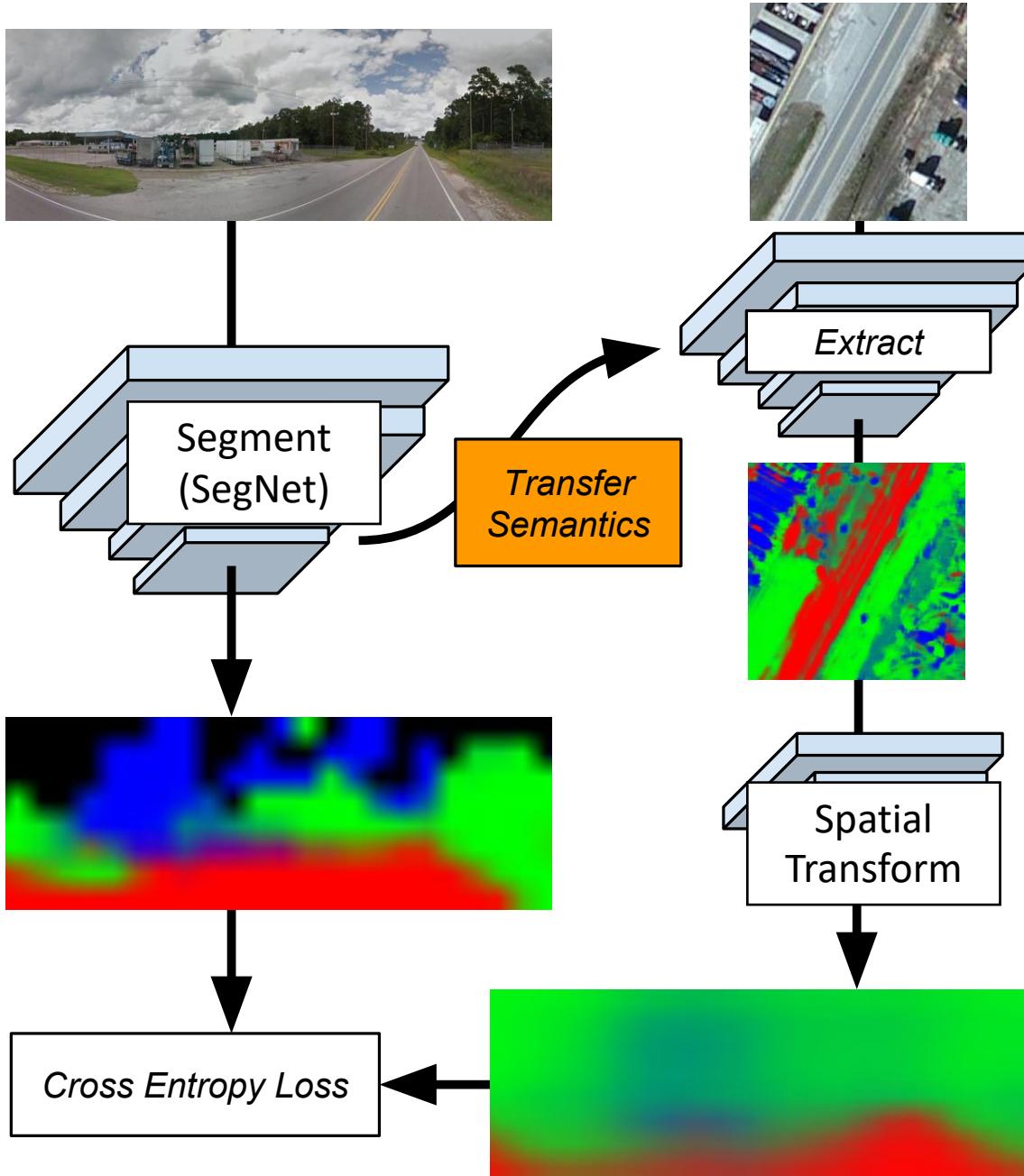
Similar Idea;
Richer Supervision

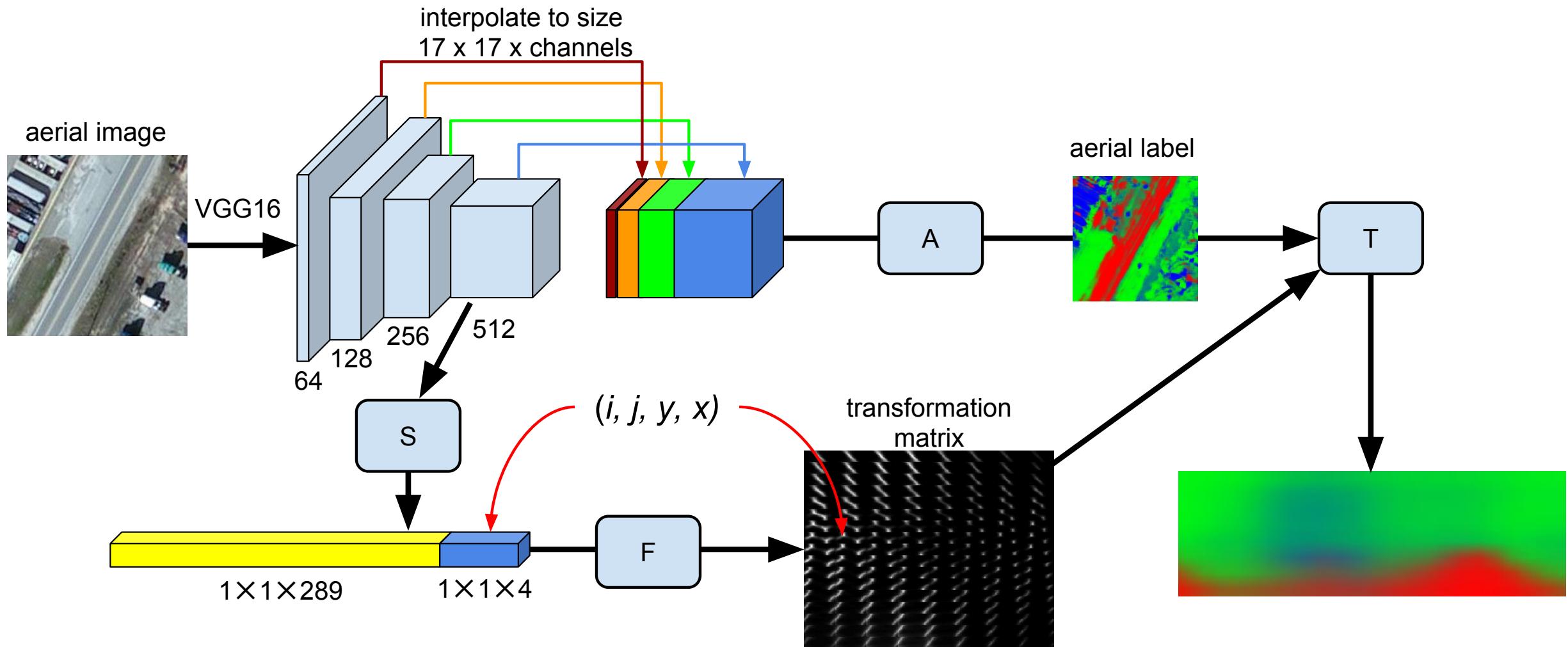


Similar Idea;
Richer Supervision

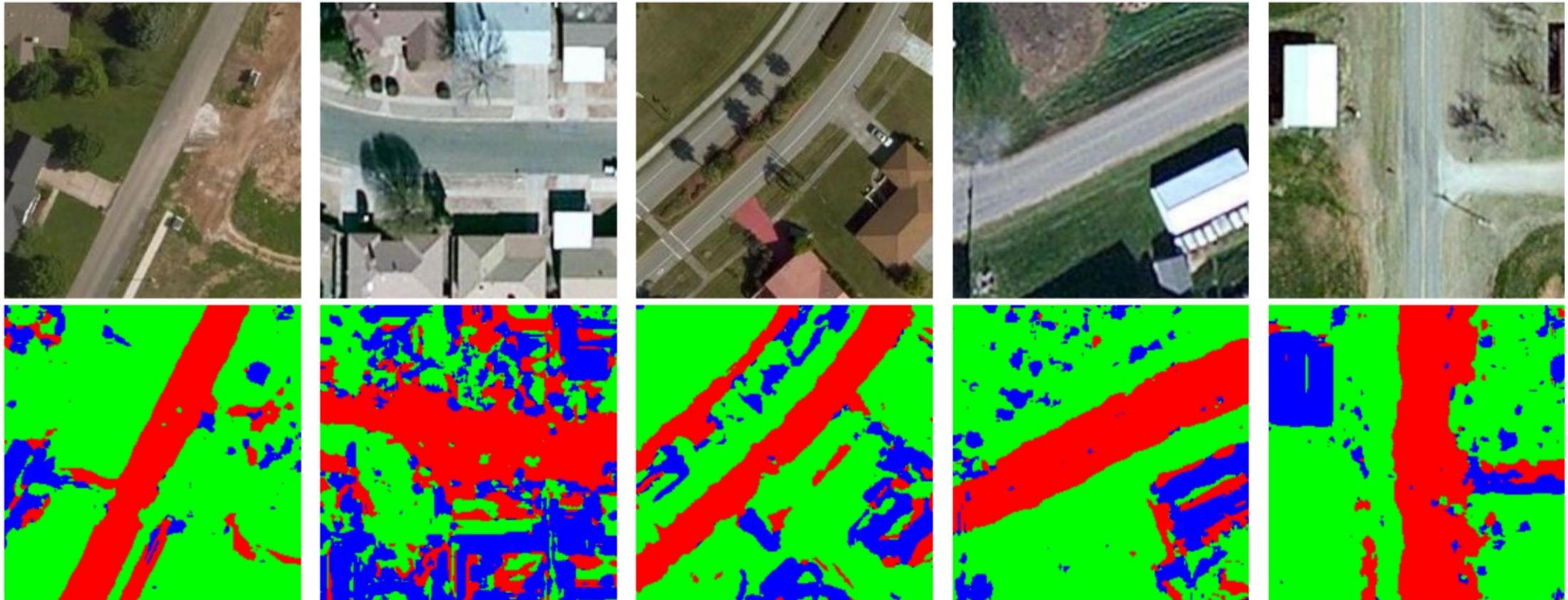


Similar Idea;
Richer Supervision

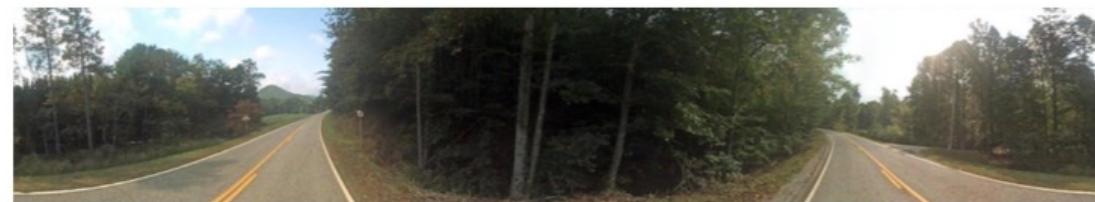




Segmentation without Labeled Satellite Imagery



Application: Synthesizing Ground-Level Images

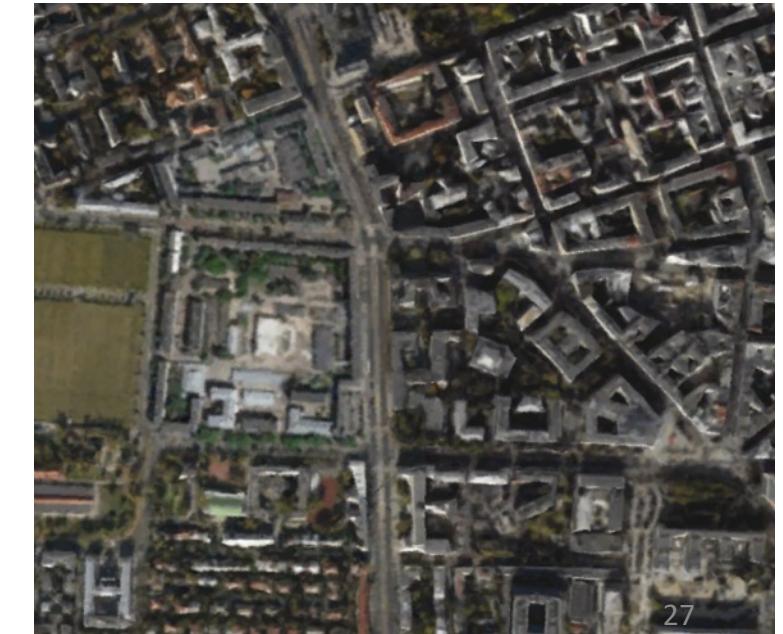
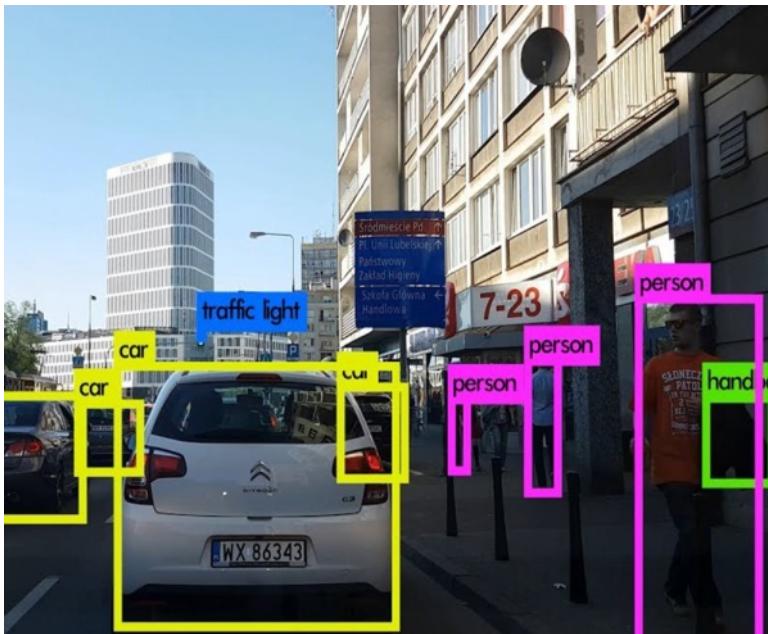
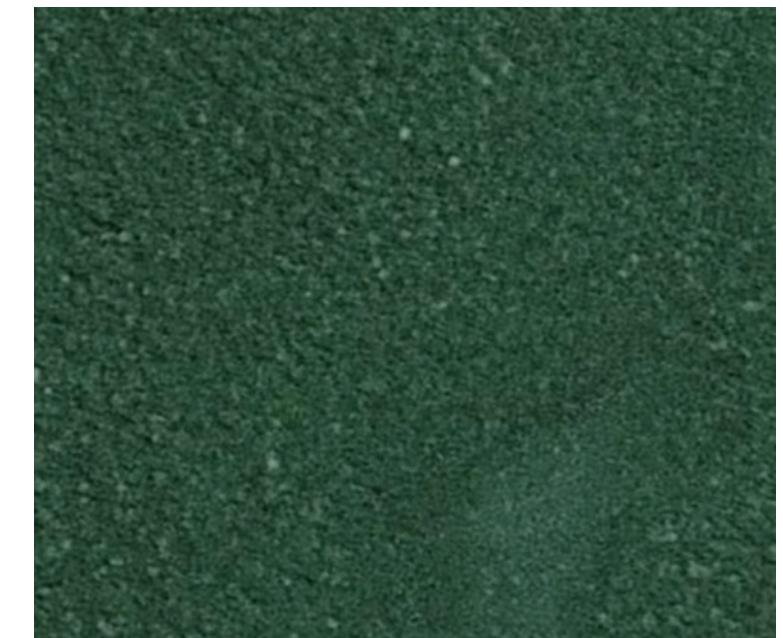
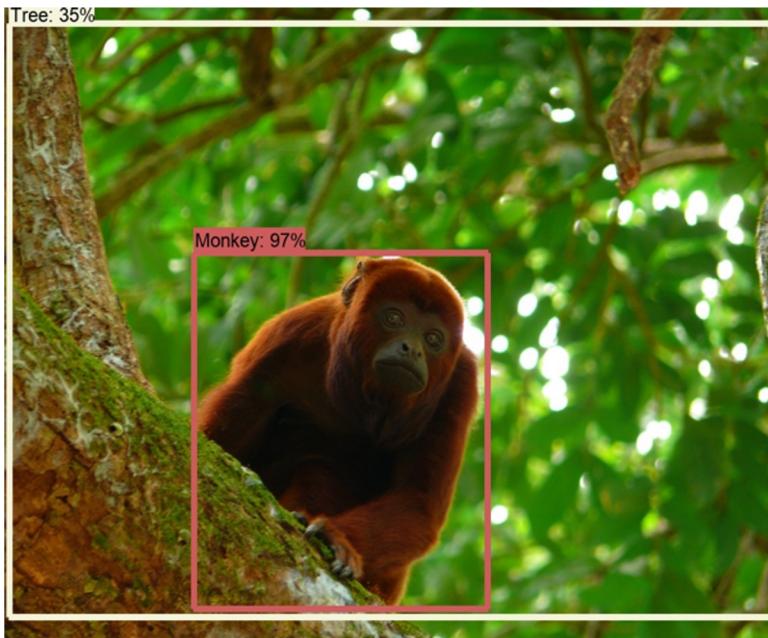


Example 3: Mapping Objects

$P(\text{object count} | \text{lat}, \text{lon})$



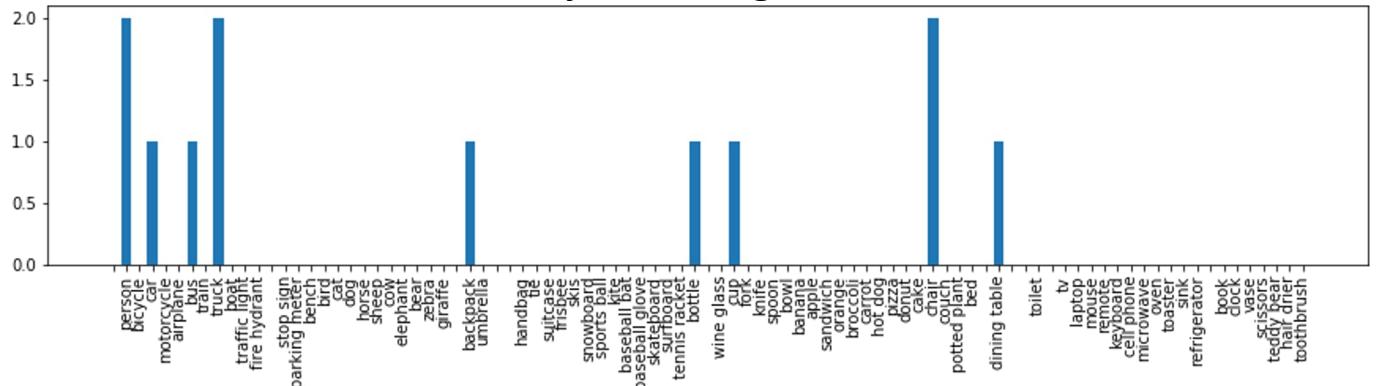
- 551,851 Geotagged Flickr Images (from CVUSA)
- Use Faster R-CNN to detect 91 Object Classes (from MS COCO)





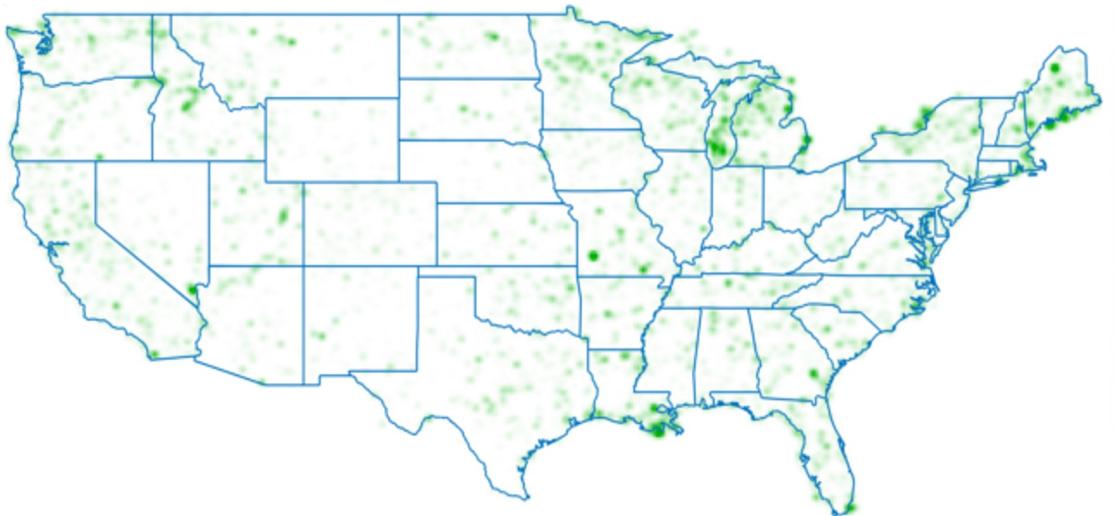
Detect
Objects

Object Histogram

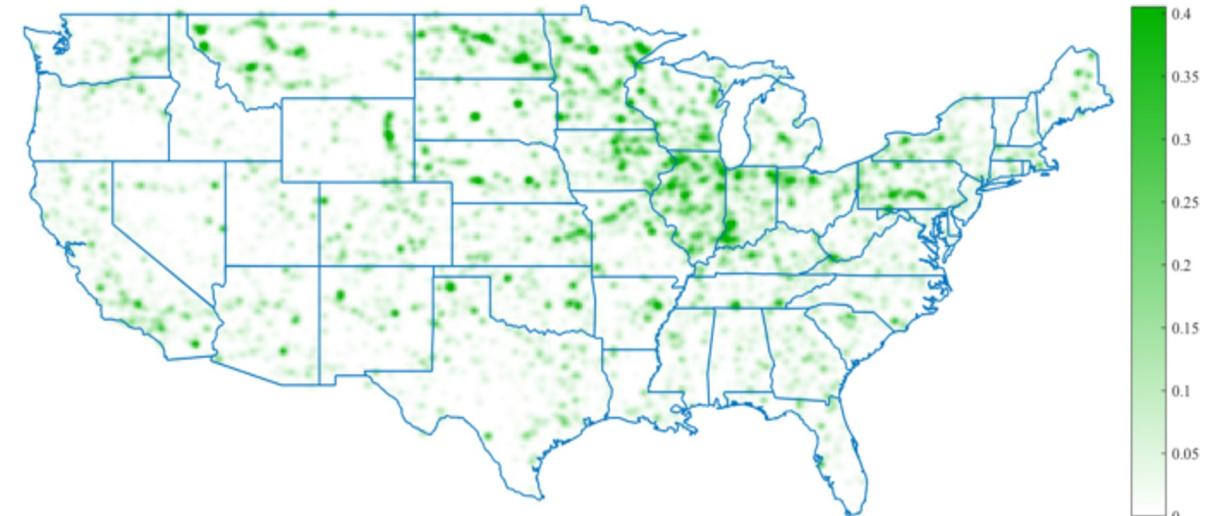


Class-Conditional Expectation of “Objects Per Image”

Boat



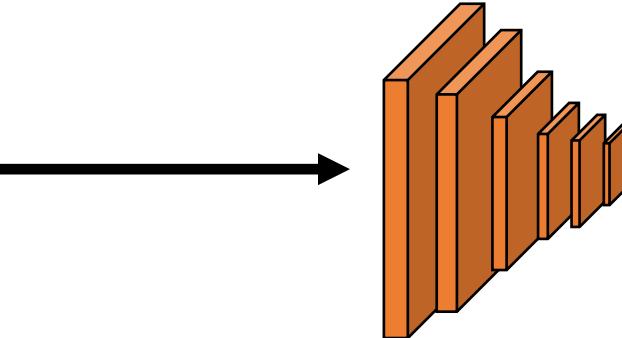
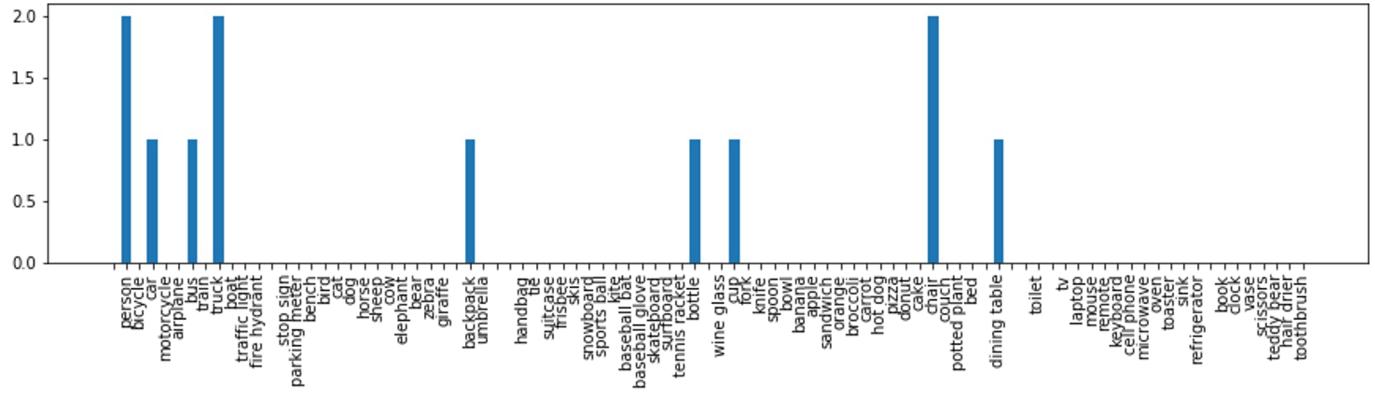
Train



Object Histogram

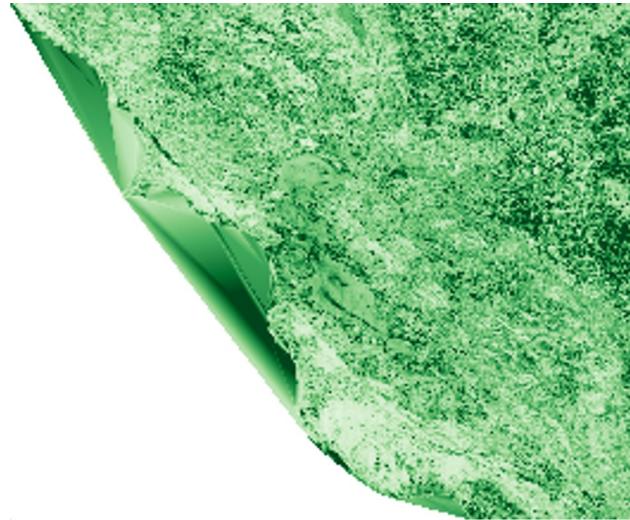


→
Detect
Objects

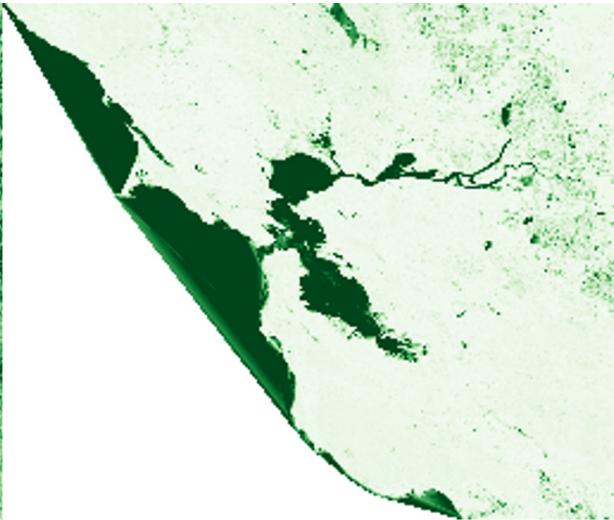


Maximize likelihood
(Independent
Poisson)

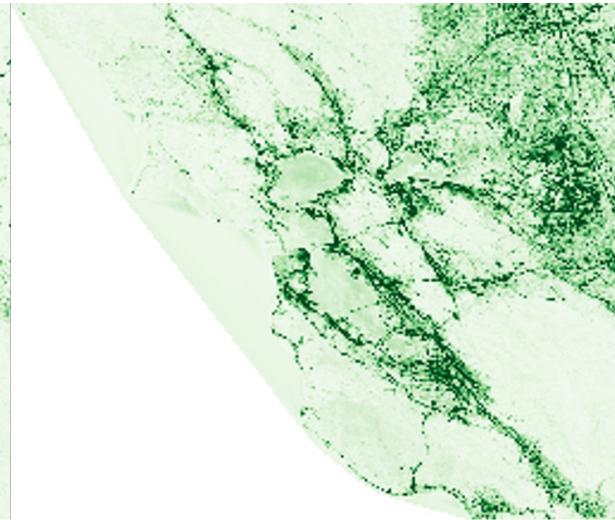
Satellite-Based Expectation of “Objects Per Image”



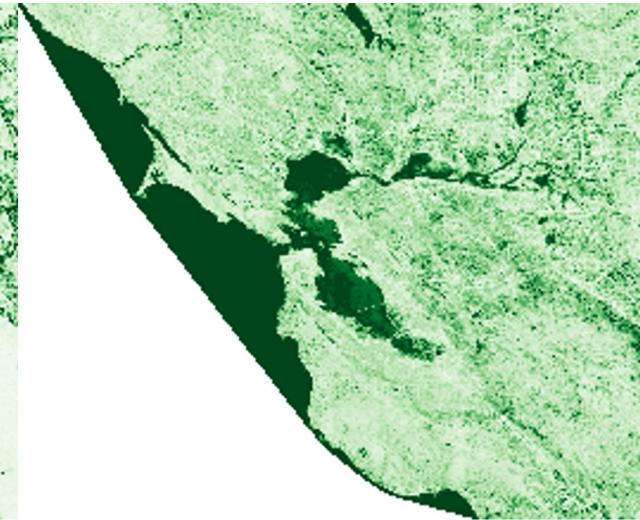
Person



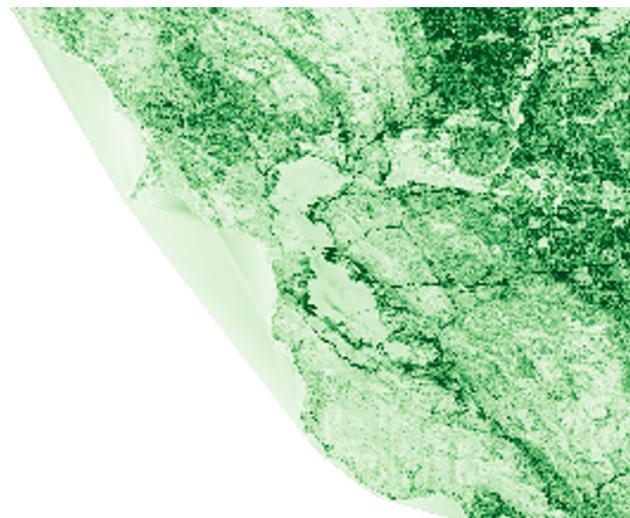
Boat



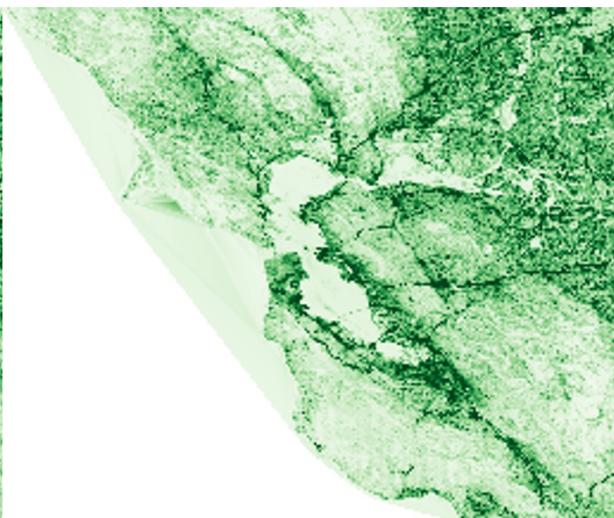
Train



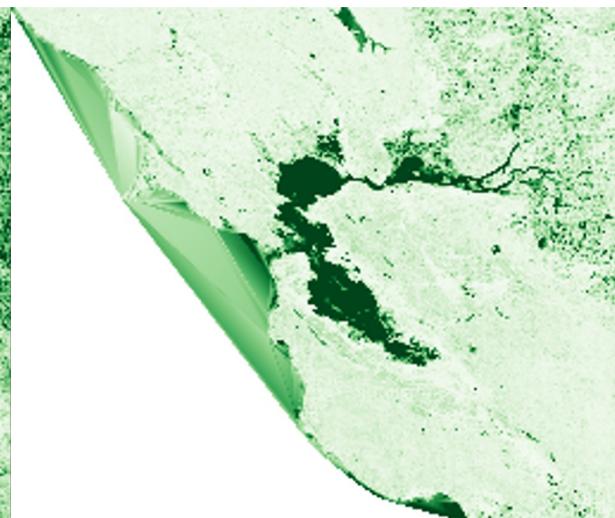
Surfboard



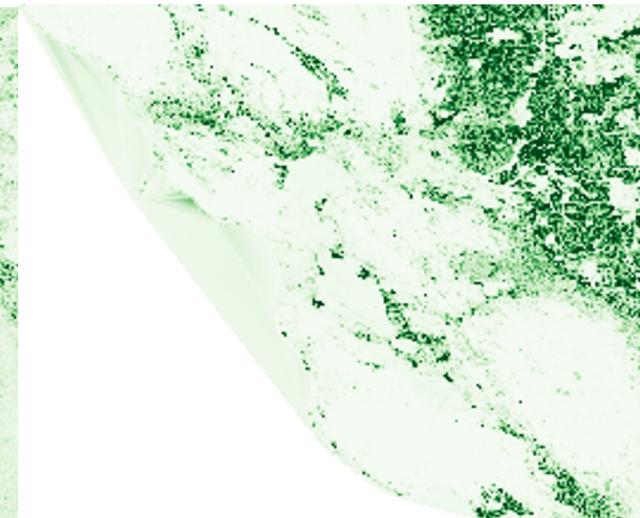
Truck



Car

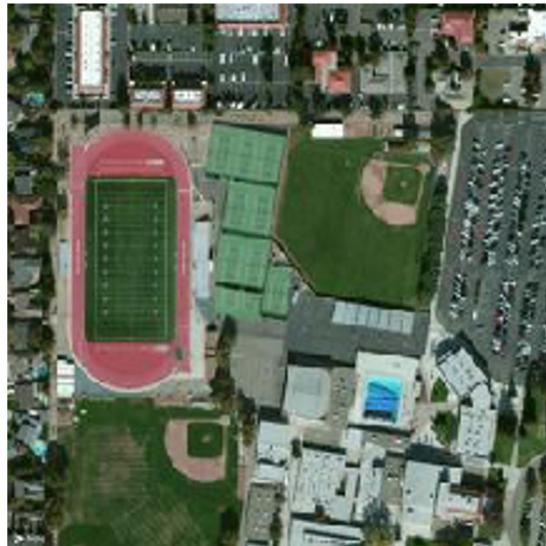


Bird



Airplane

Maximal Expectation Images



Person



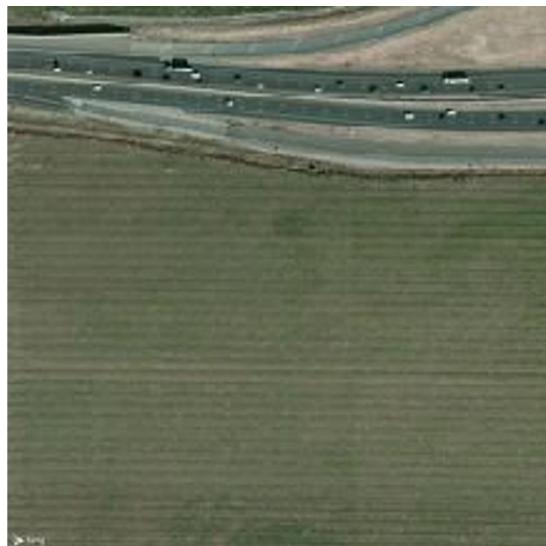
Boat



Train



Surfboard



Truck



Car



Bird



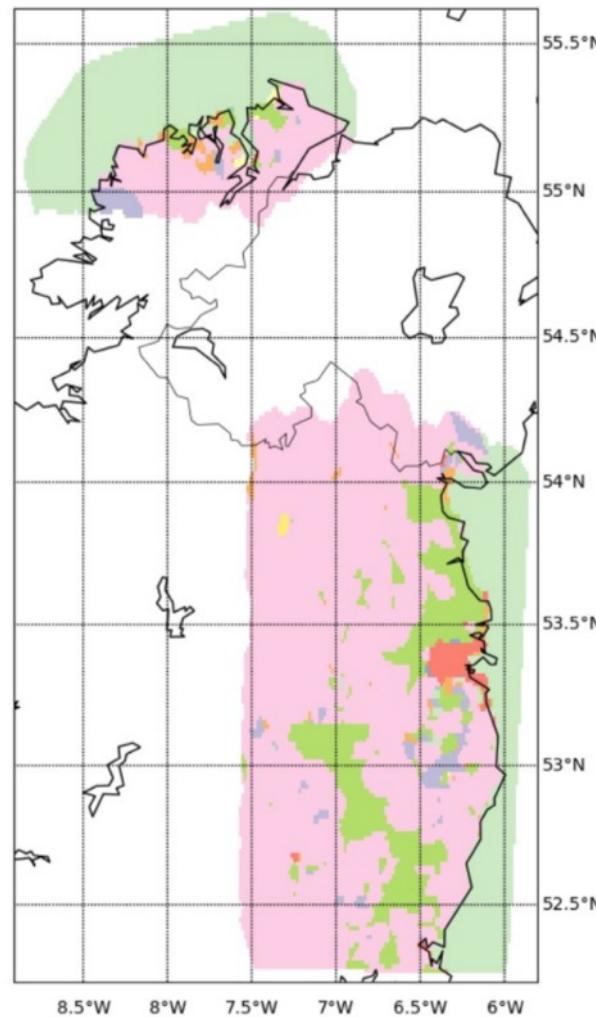
Airplane

Land use and object distributions are closely related

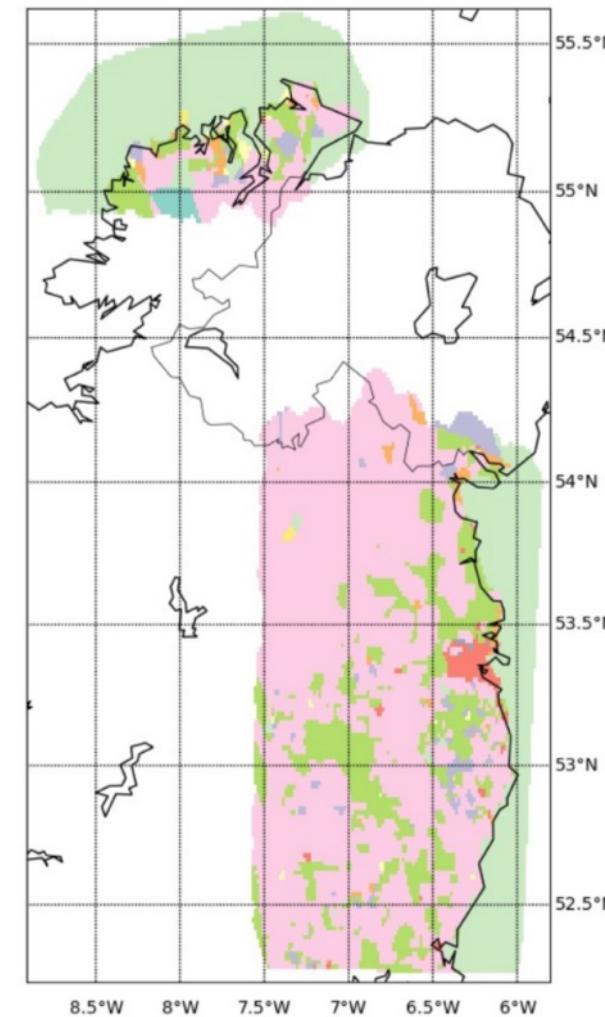
Satellite View



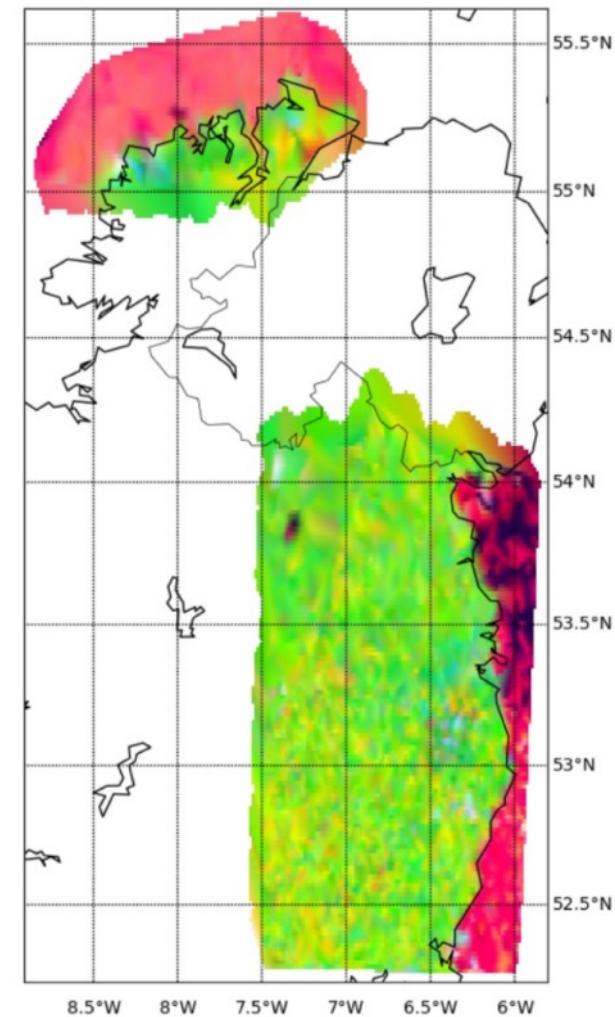
CORINE [1] Land Use



Predicted CORINE [1]



Learned Feature Map



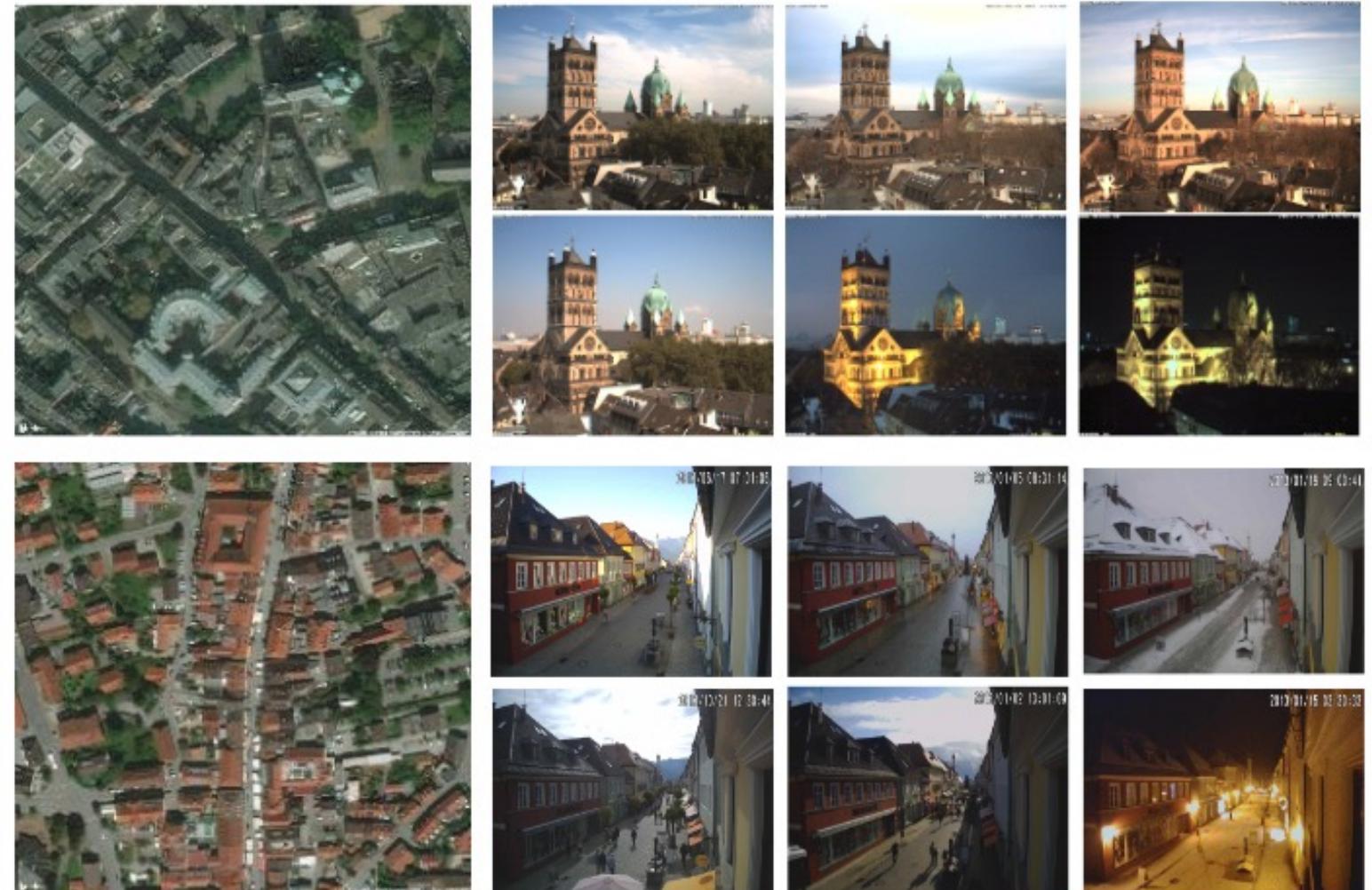
Example 4: Dynamic Geo-Temporal Scene Modeling

$$P(s_1, s_2, \dots | \text{time}, \text{date}, \text{lat}, \text{lon})$$



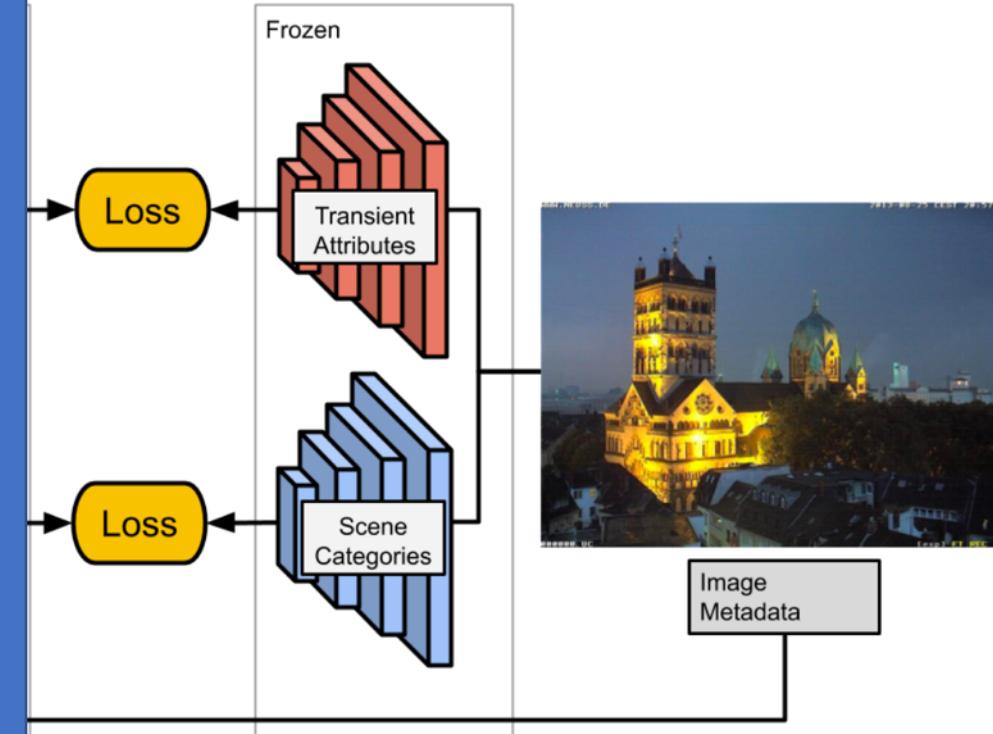
Cross-View Time (CVT) Dataset

- ~300k ground-level images
 - ~100k from 50 outdoor webcams
 - remainder captured by cell phone and shared via Flickr
- 25k reserved for evaluation
- .5 x .5 km overhead image for each (60 cm GSD)

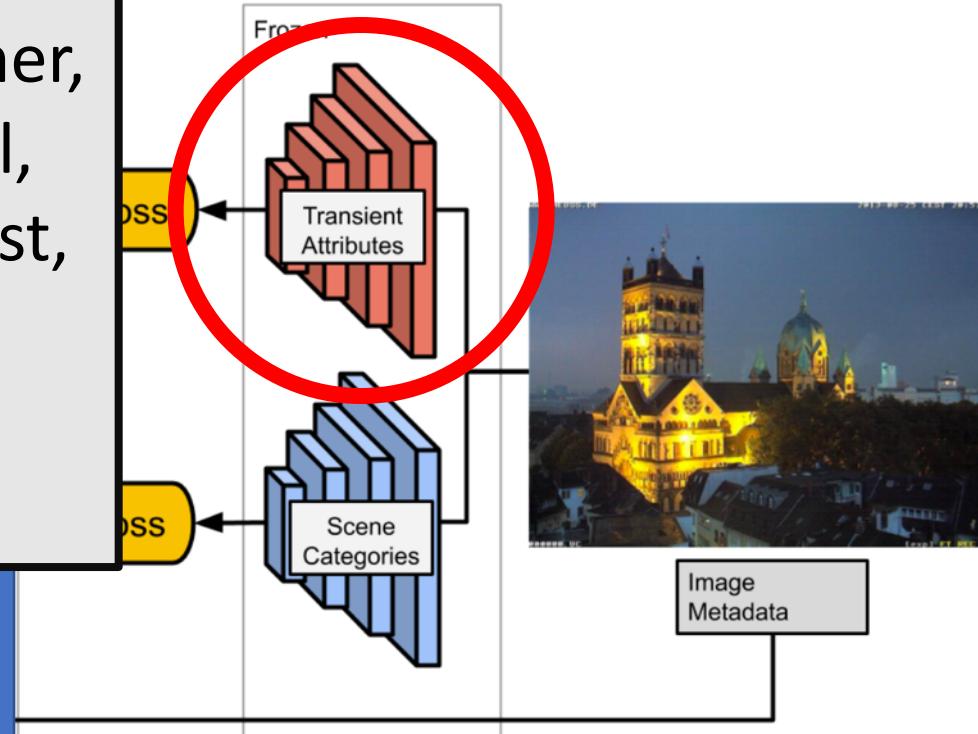




Trainable Neural Network

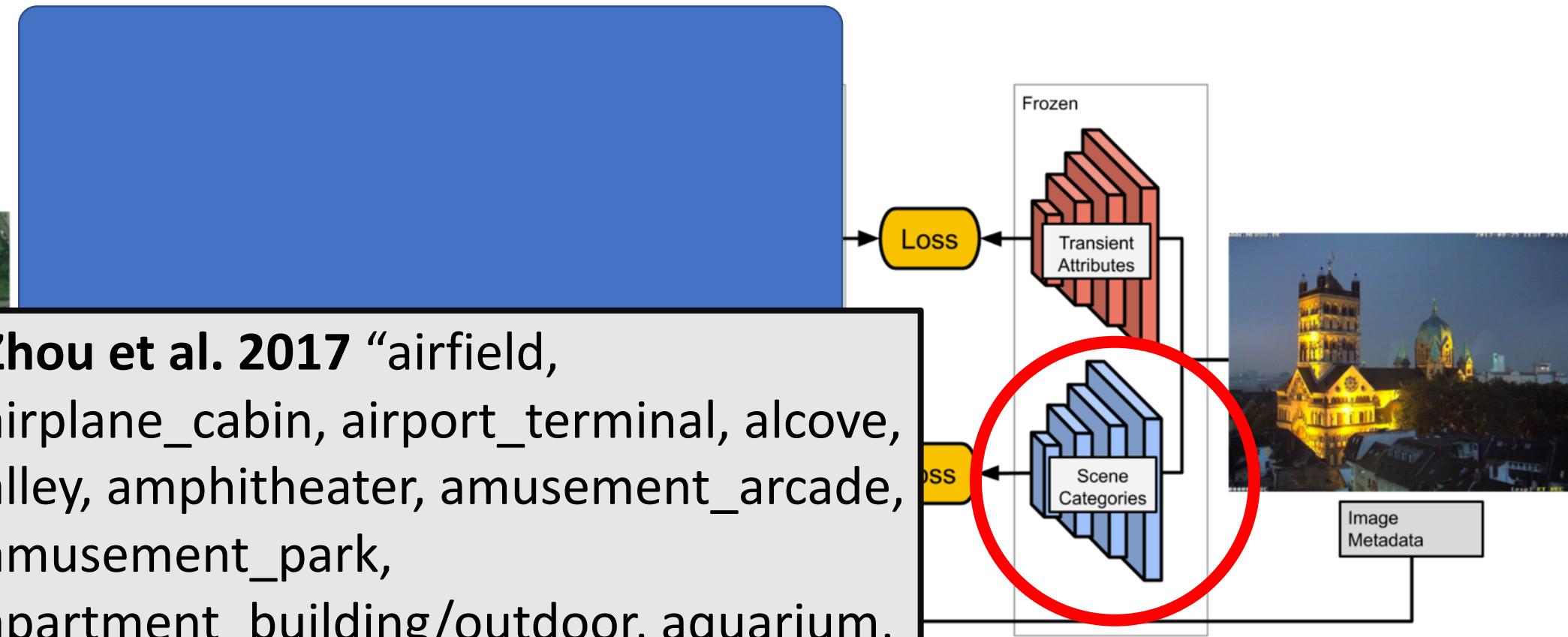


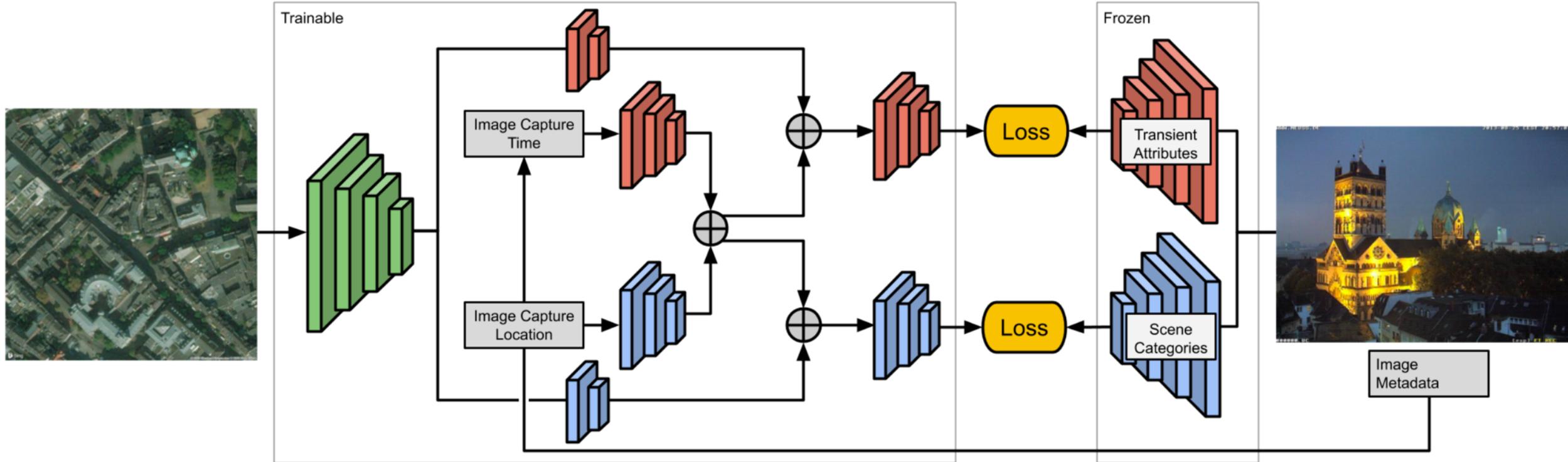
Laffont et al. 2014 “dirty, daylight, night, sunrise/sunset, dawn/dusk, sunny, clouds, fog, storm, snow, warm, cold, busy, beautiful, flowers, spring, summer, autumn, winter, glowing, colorful, dull, rugged, midday, dark, bright, dry, moist, windy, rain, ice, cluttered, soothing, stressful, exciting, sentimental, mysterious, boring, gloomy, lush”



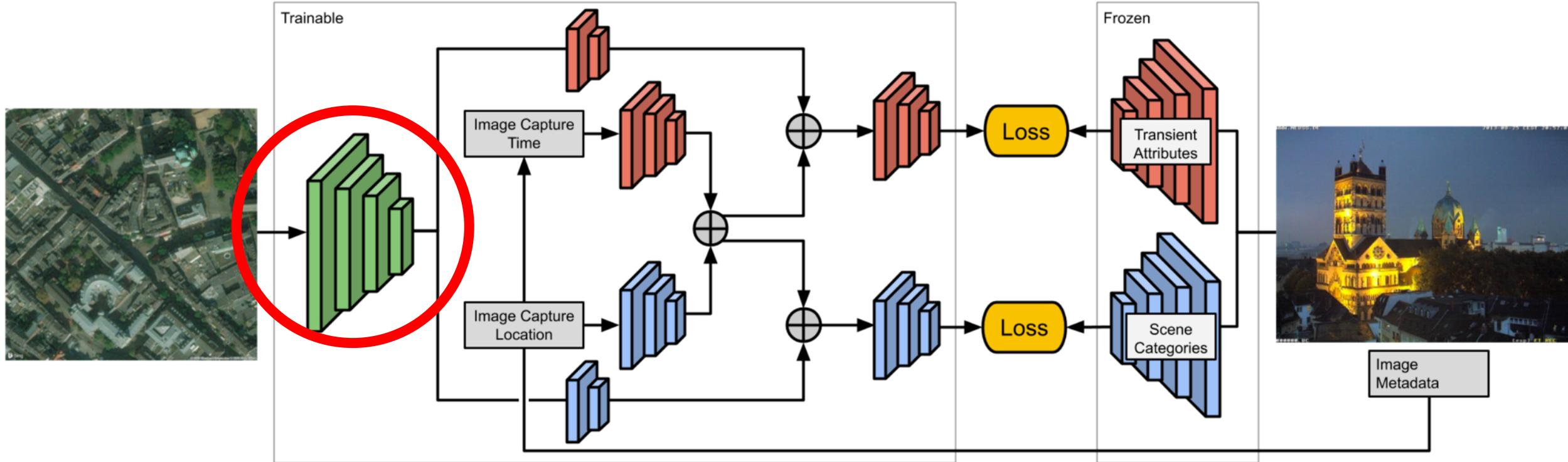


Zhou et al. 2017 “airfield, airplane_cabin, airport_terminal, alcove, alley, amphitheater, amusement_arcade, amusement_park, apartment_building/outdoor, aquarium, aqueduct, arcade, arch, archaeological_excavation, archive, arena/hockey, ...”

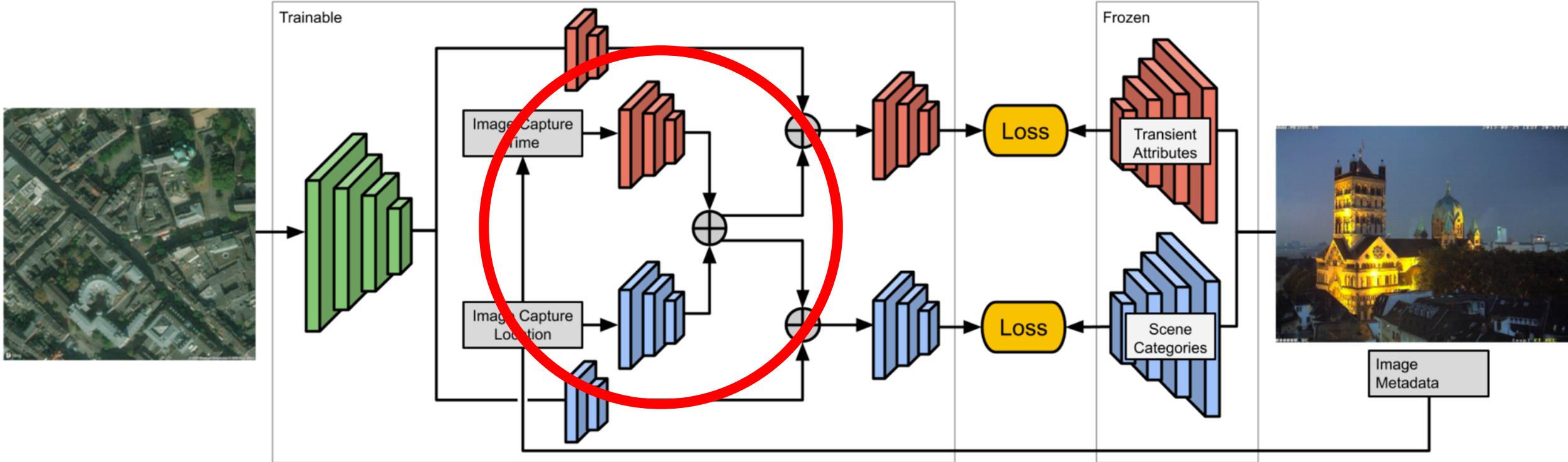




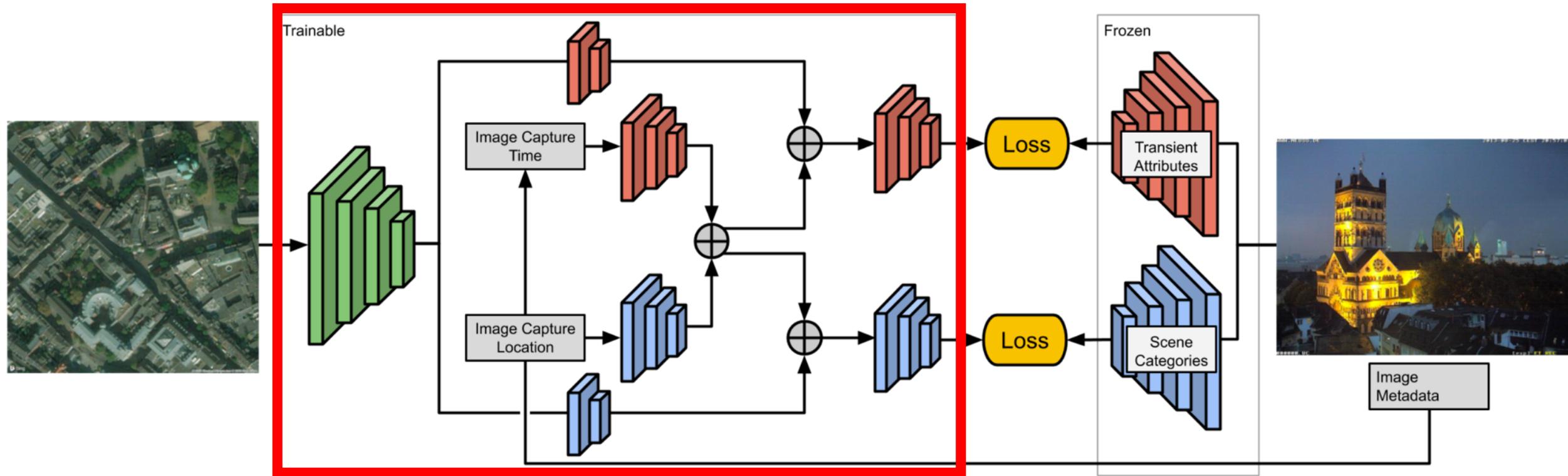
Extracts features, related to visual appearance, at high spatial resolution.



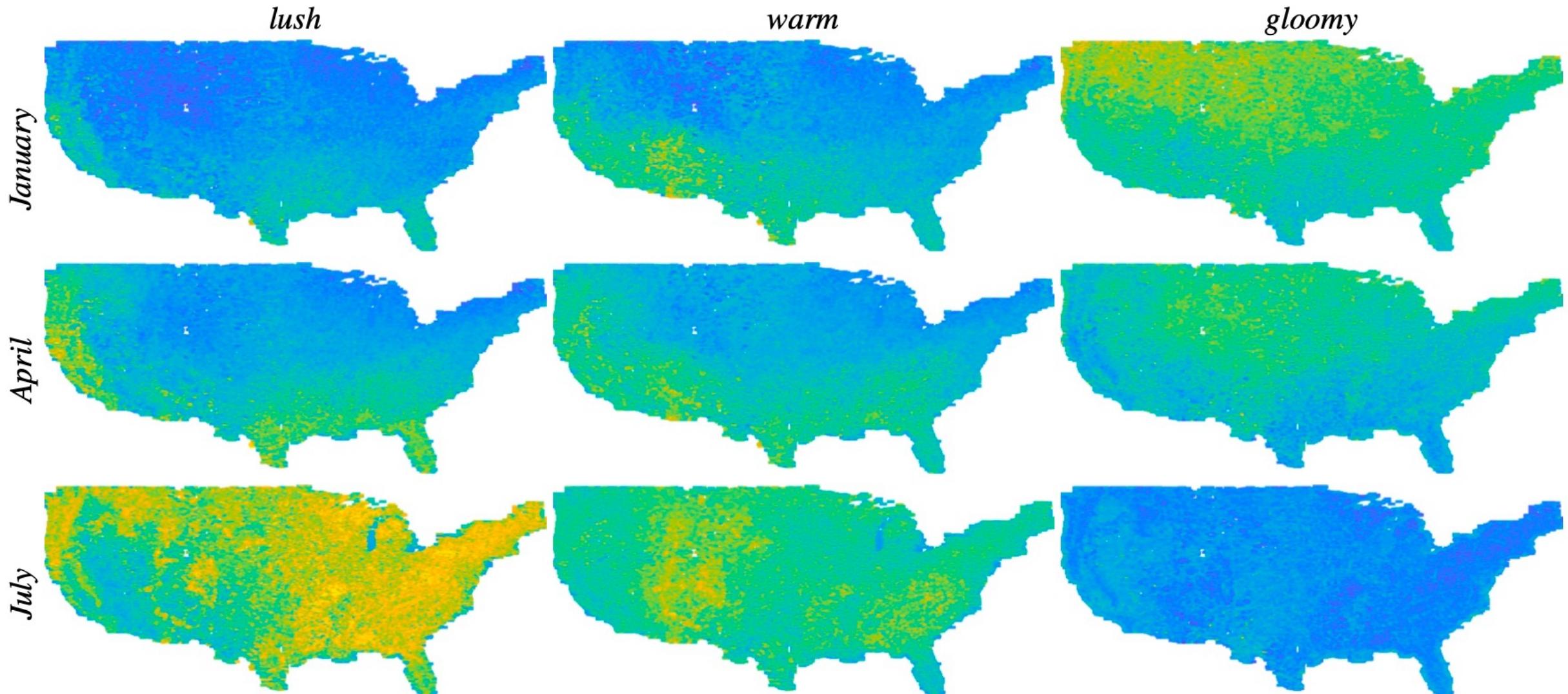
Captures **low resolution, global trends** in visual appearance.



This is a dynamic geo-temporal scene model!

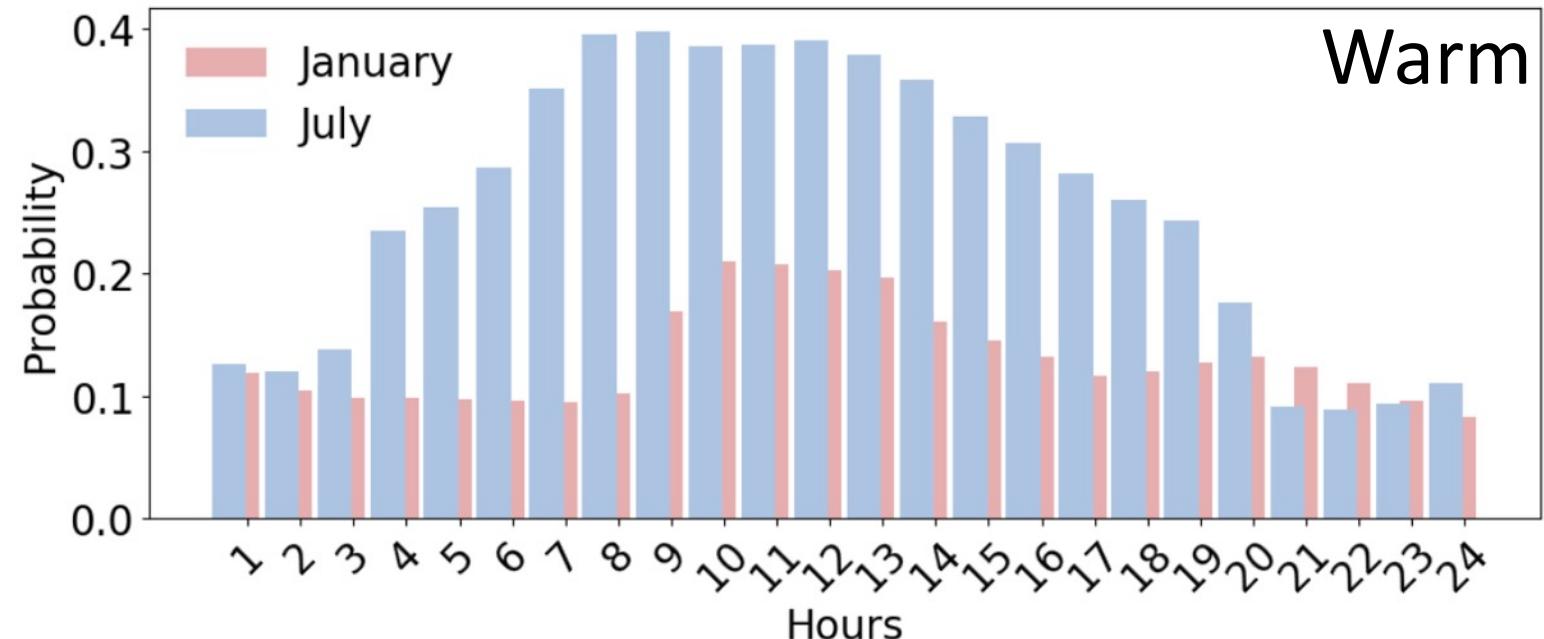
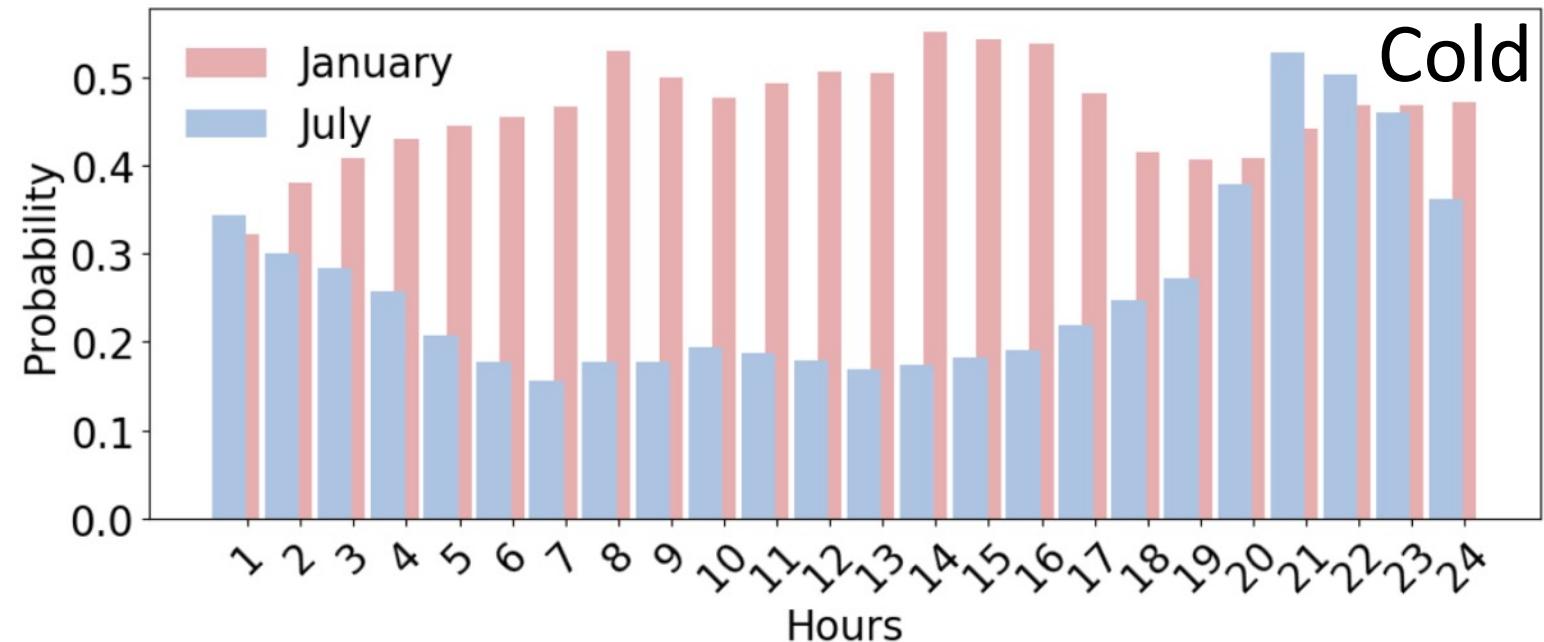


Mapping Transient Visual Attributes

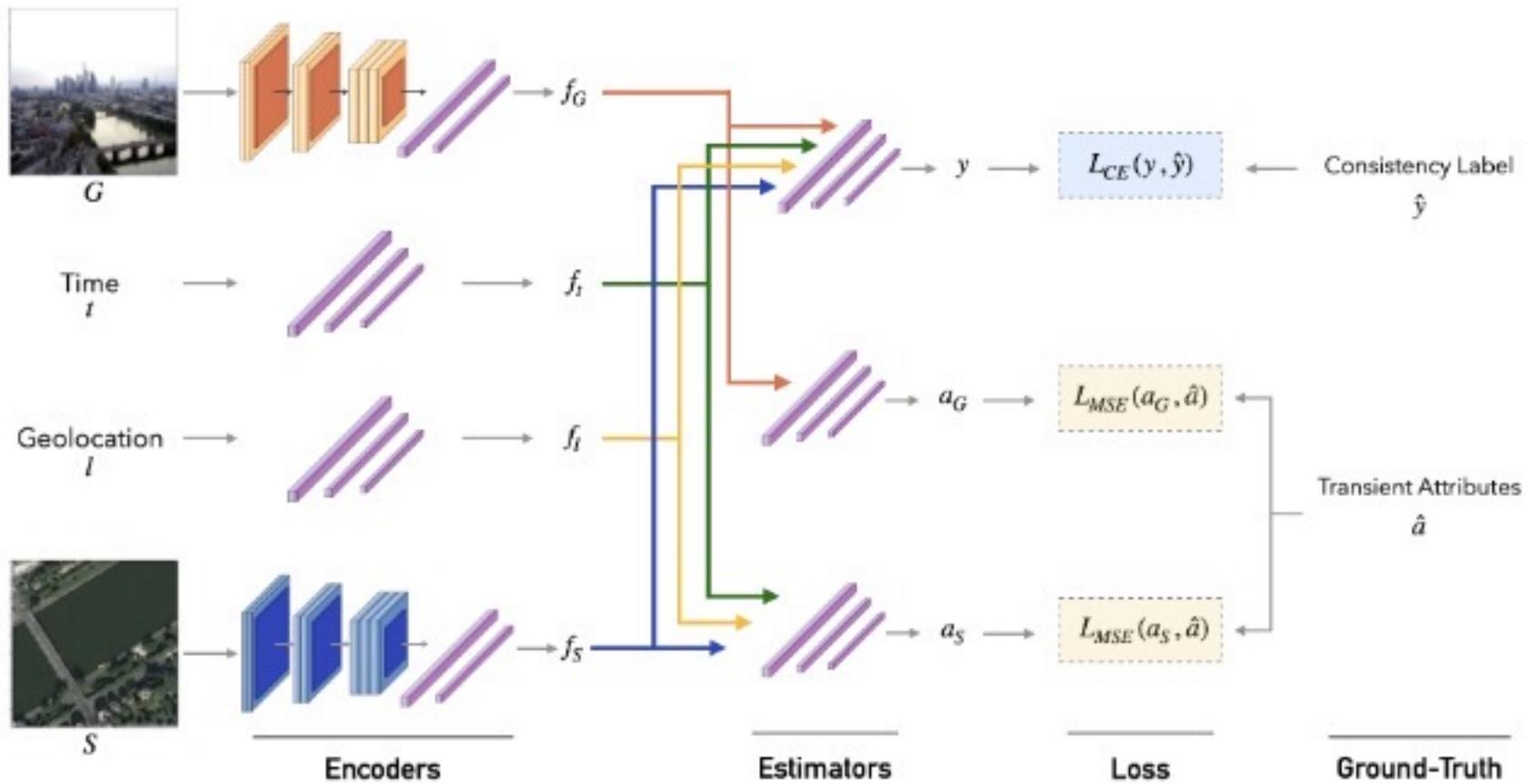


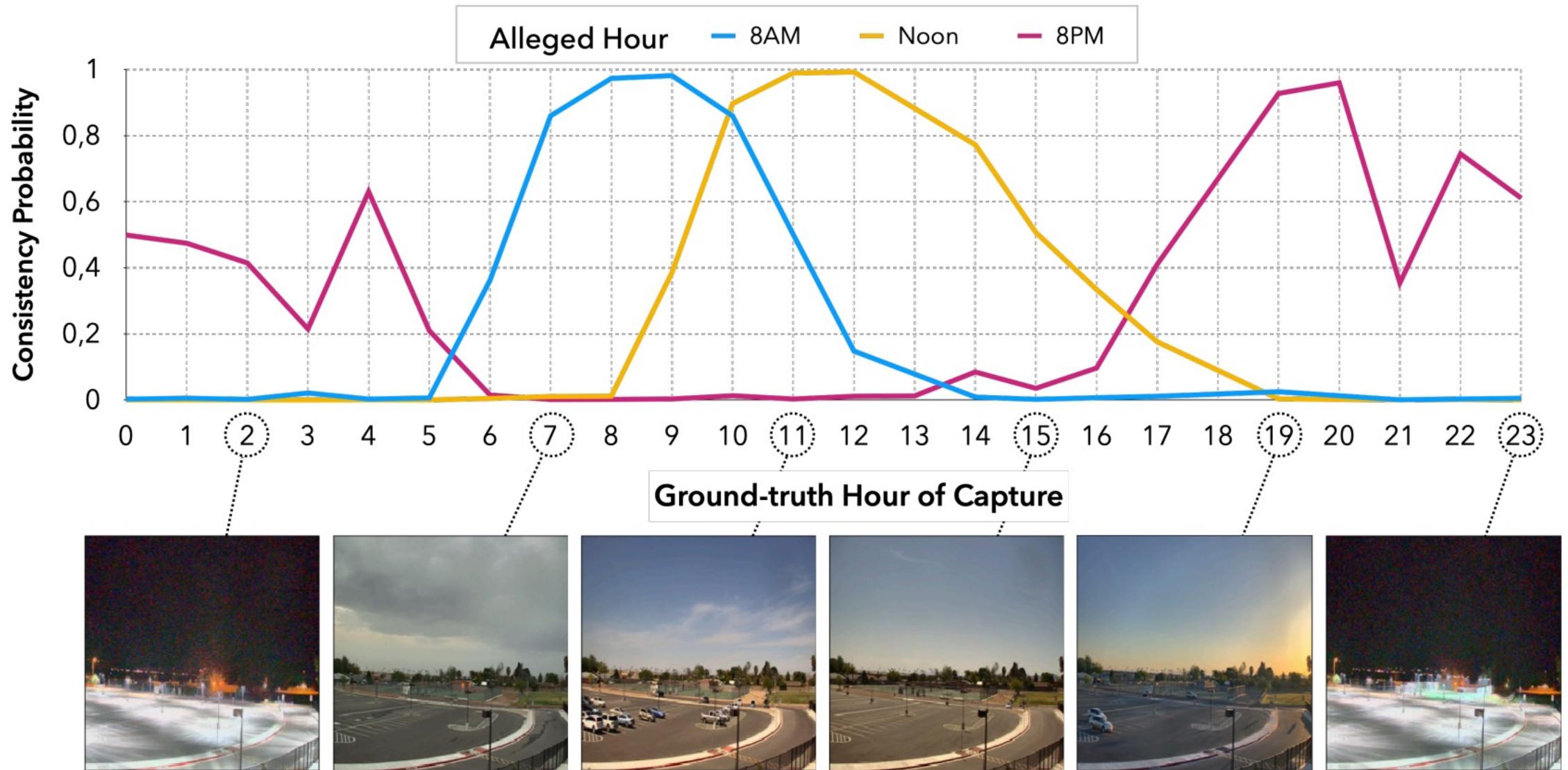


Latitude: 47.367
Longitude: 8.55



Discriminative Training for Media Forensics



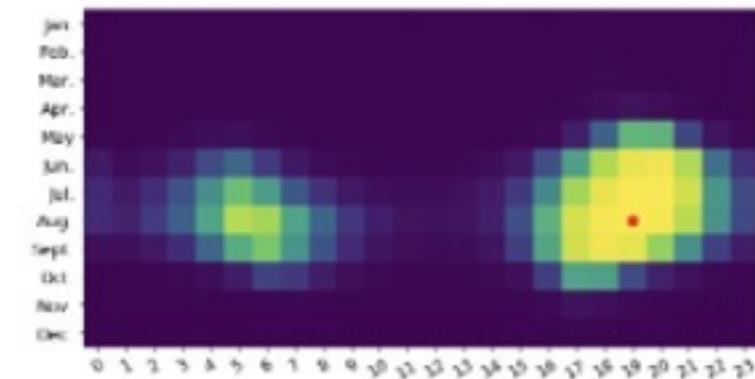
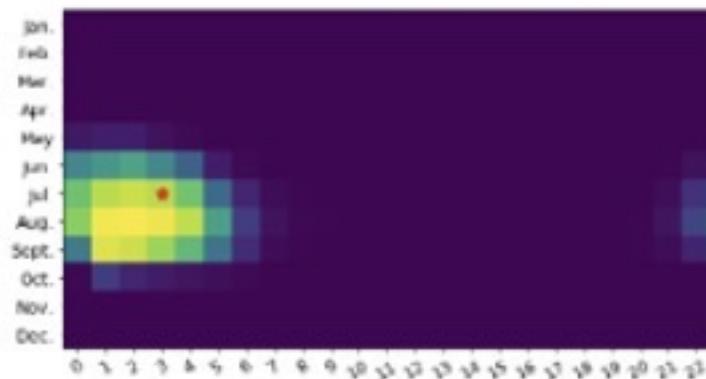
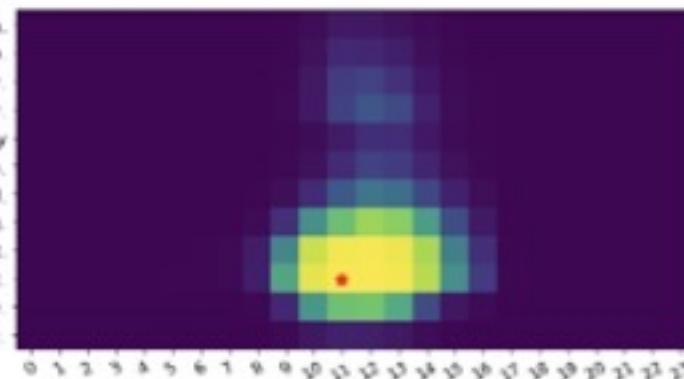


(a) Same location recorded in April under different hours of the day.

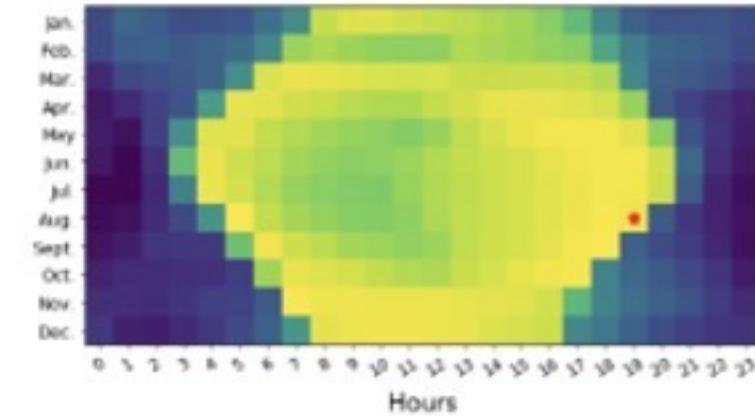
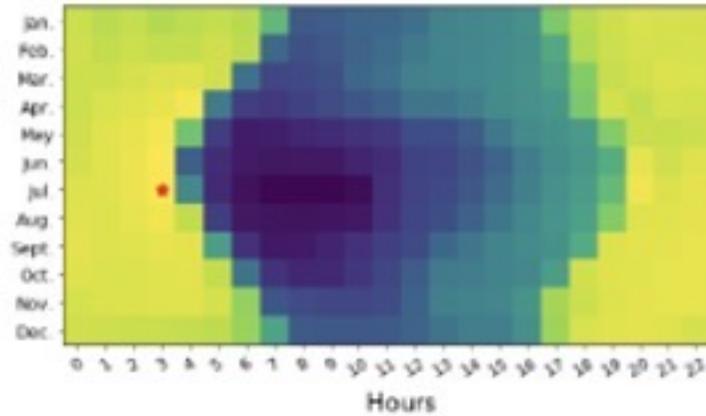
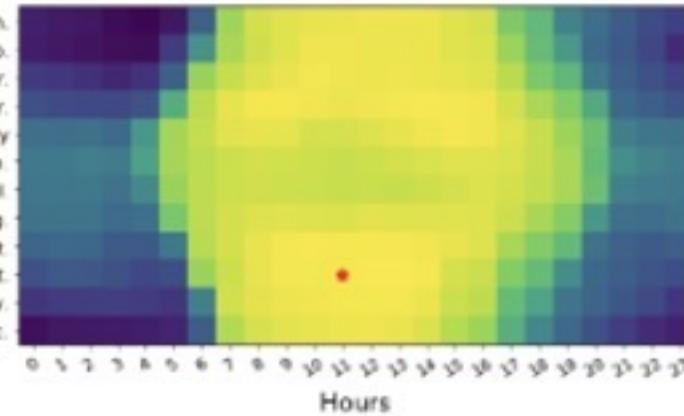
Enables Weak Timestamp Estimation and Verification



Proposed Method



Salem et al. [24]



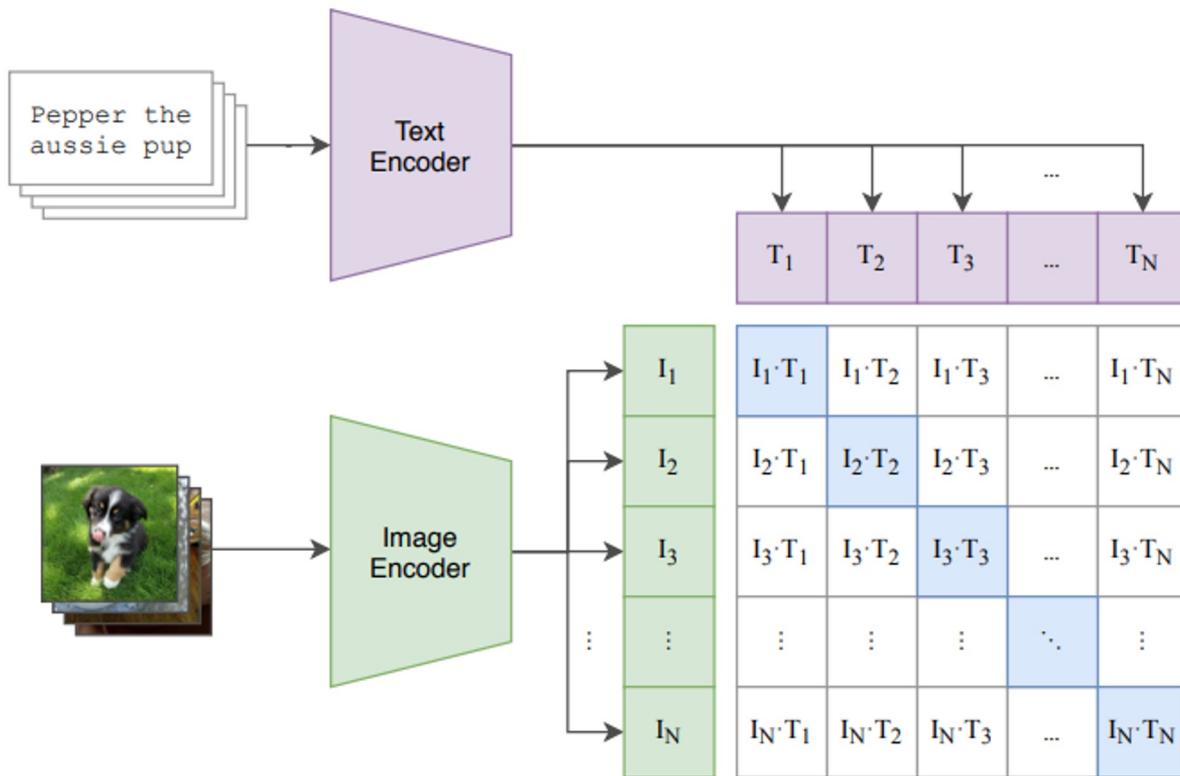
Example 6: Text-Driven Mapping

$P(\text{text} \mid \text{lat}, \text{lon}, \text{time})$

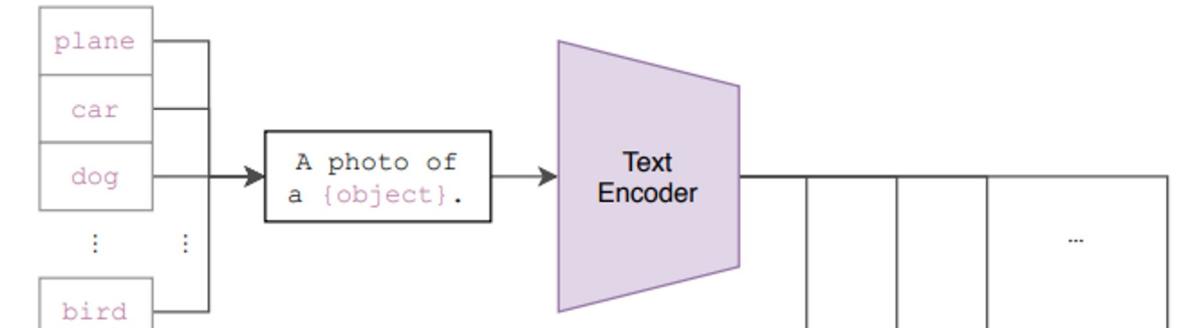


Contrastive Language Image Pretraining (CLIP)

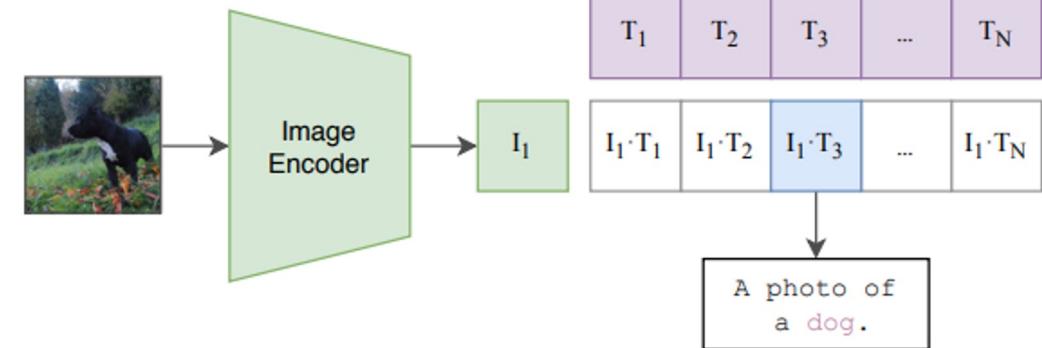
(1) Contrastive pre-training

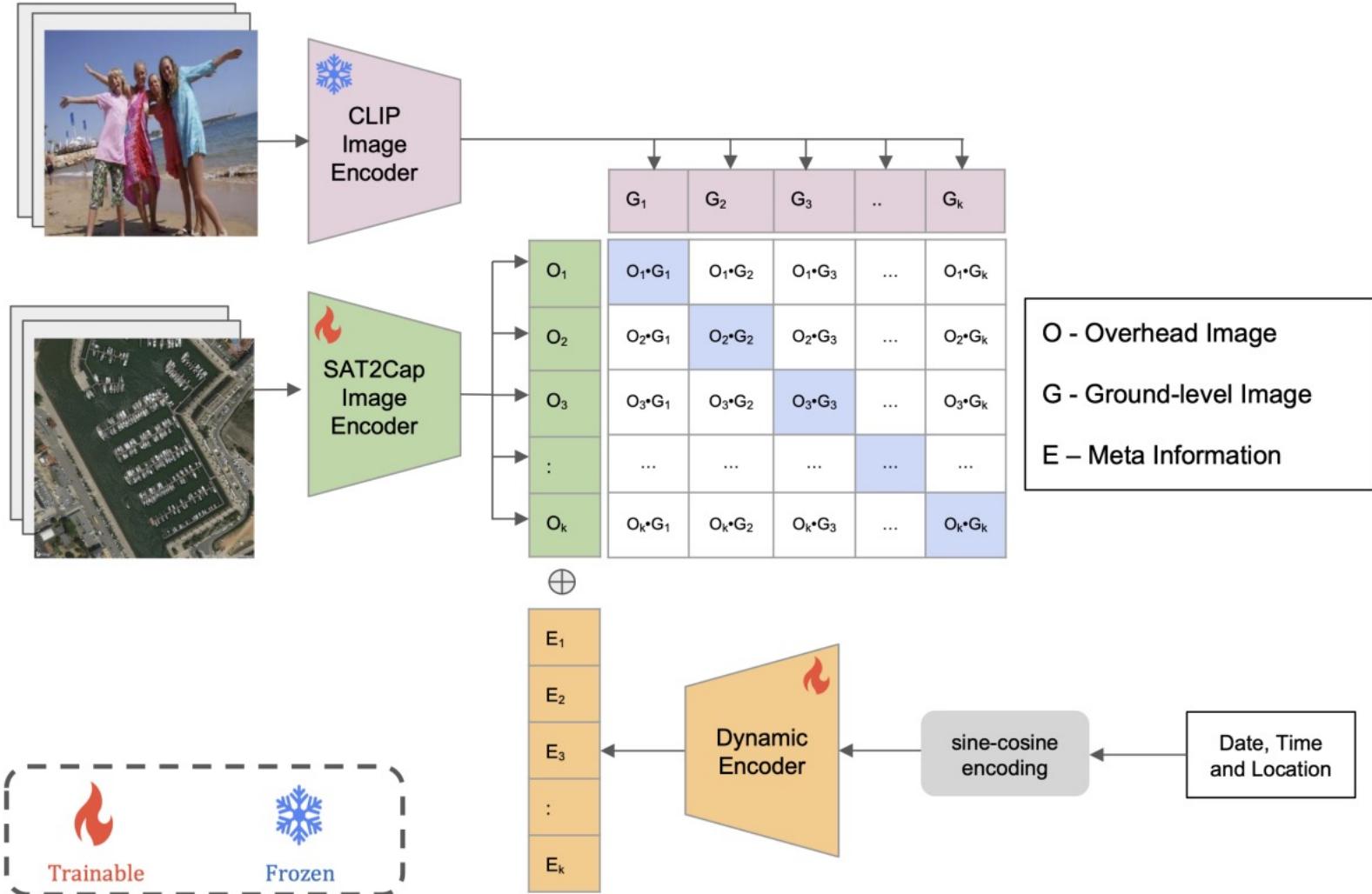


(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

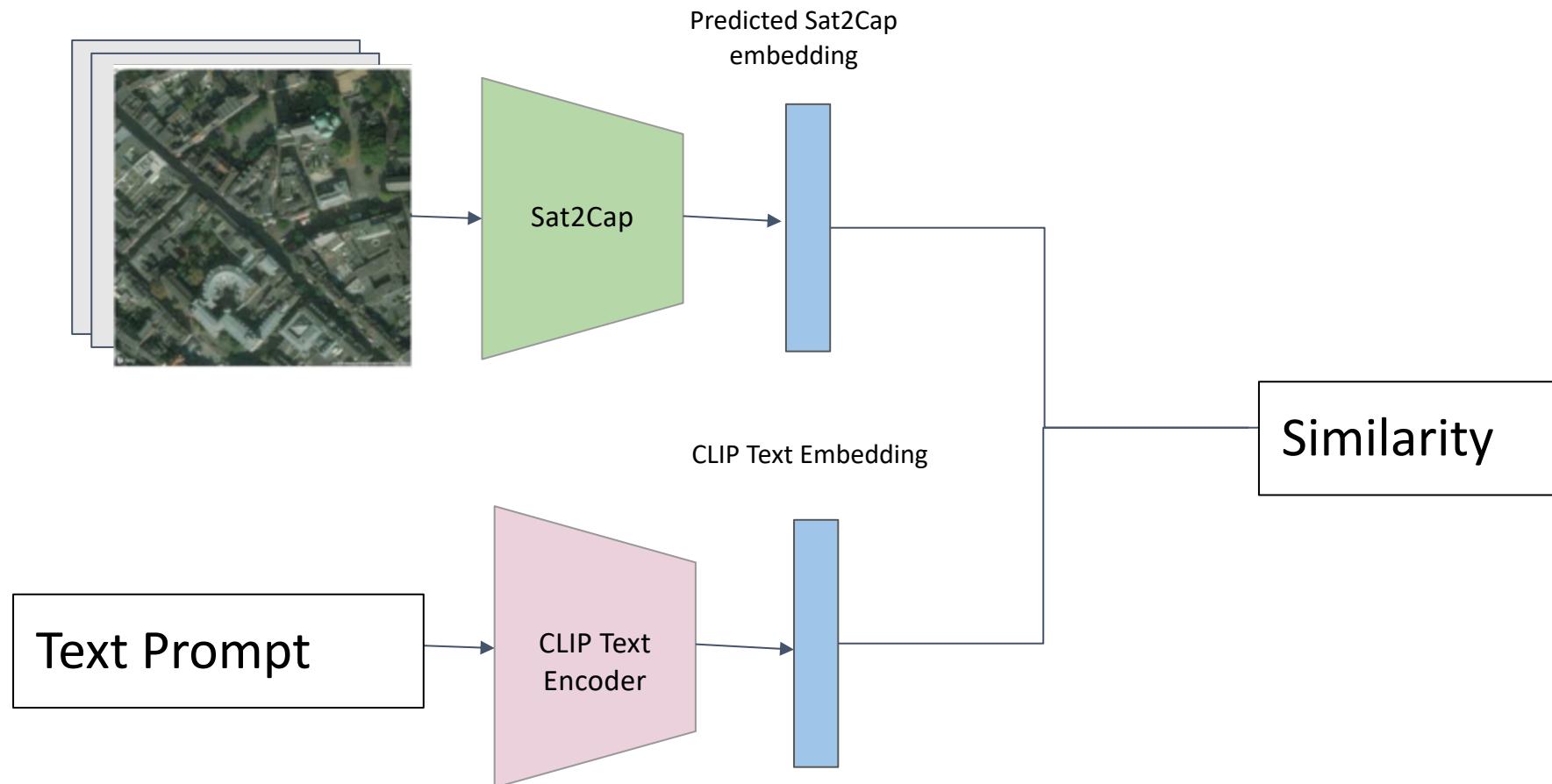




Cross-Modality Retrieval Performance

Method				Overhead2Ground (10K)			Ground2Overhead (10K)		
Model	Dynamic Encoder	Dropout	Meta Information	R@5↑	R@10↑	Median-R↓	R@5↑	R@10↑	Median-R↓
CLIP	-	-	-	0.007	0.013	1700	0.108	0.019	2857
ours	✗	✗	✗	0.398	0.493	15	0.356	0.450	11
	✓	✗	✗	0.322	0.413	34	0.254	0.343	20
	✓	✗	✓	0.368	0.467	23	0.298	0.398	13
	✓	✓	✗	0.467	0.564	13.5	0.366	0.462	7
	✓	✓	✓	0.493	0.591	12	0.390	0.482	6

Zero-Shot Textual Mapping



Zero-Shot Textual Mapping

(a) "People playing sports"



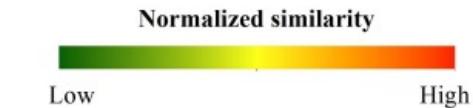
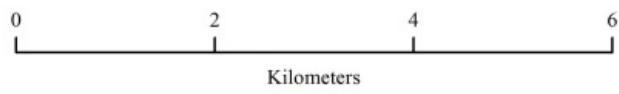
(b) "People with animals"



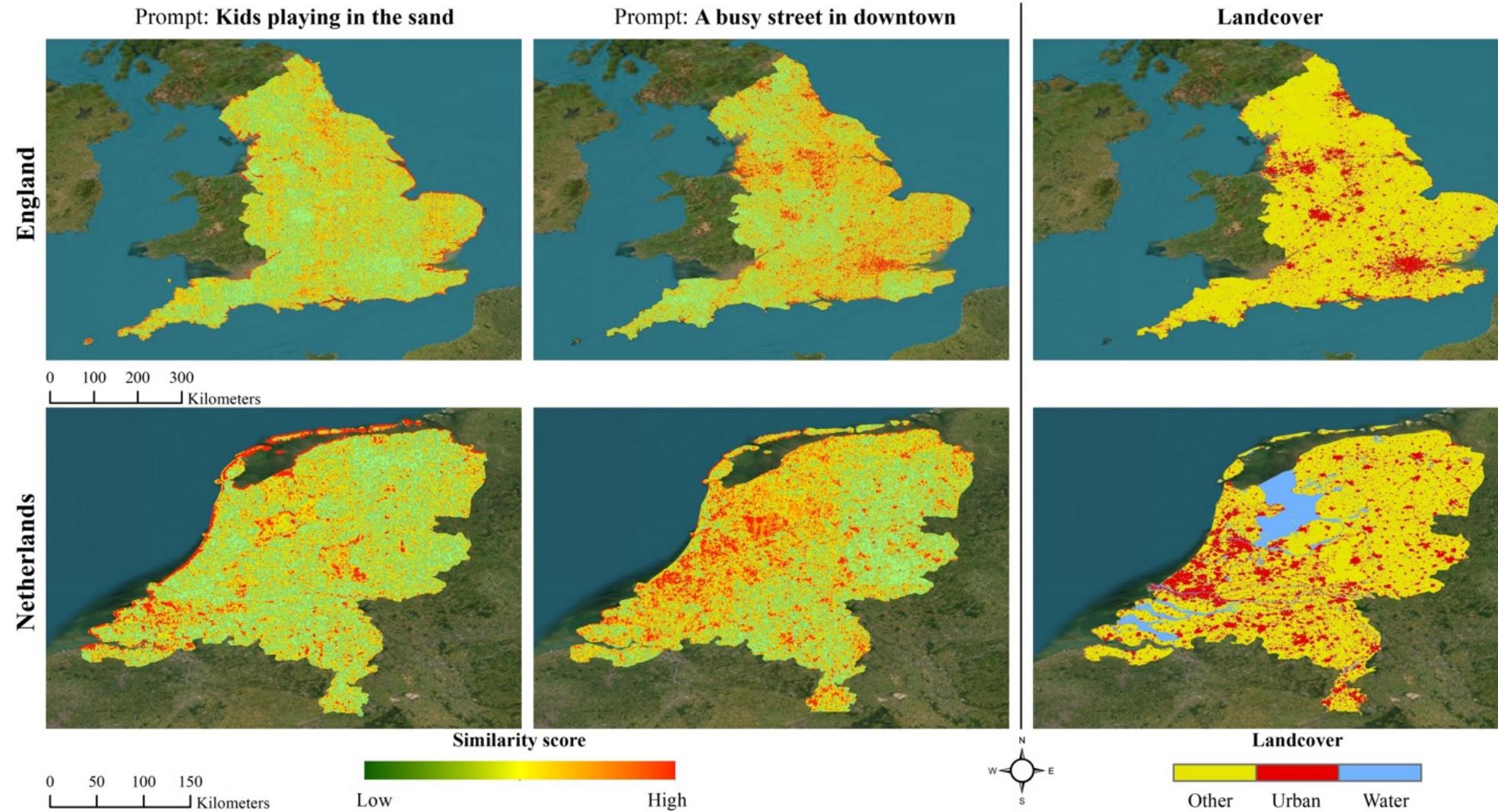
(c) "An amusement park"



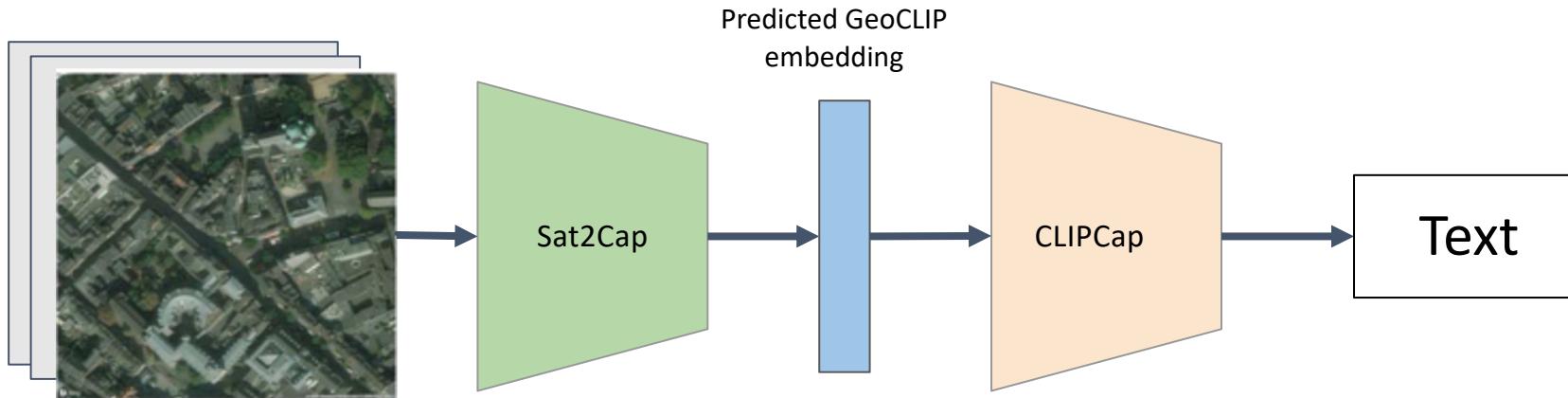
(d) "A historical museum"

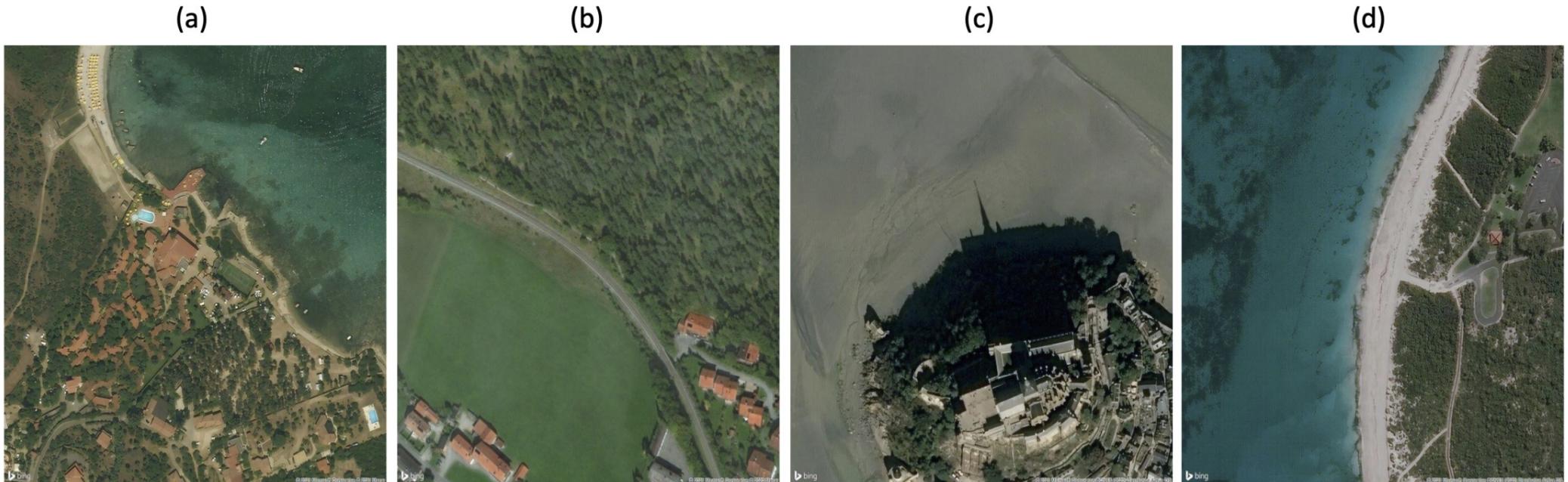


Zero-Shot Textual Mapping

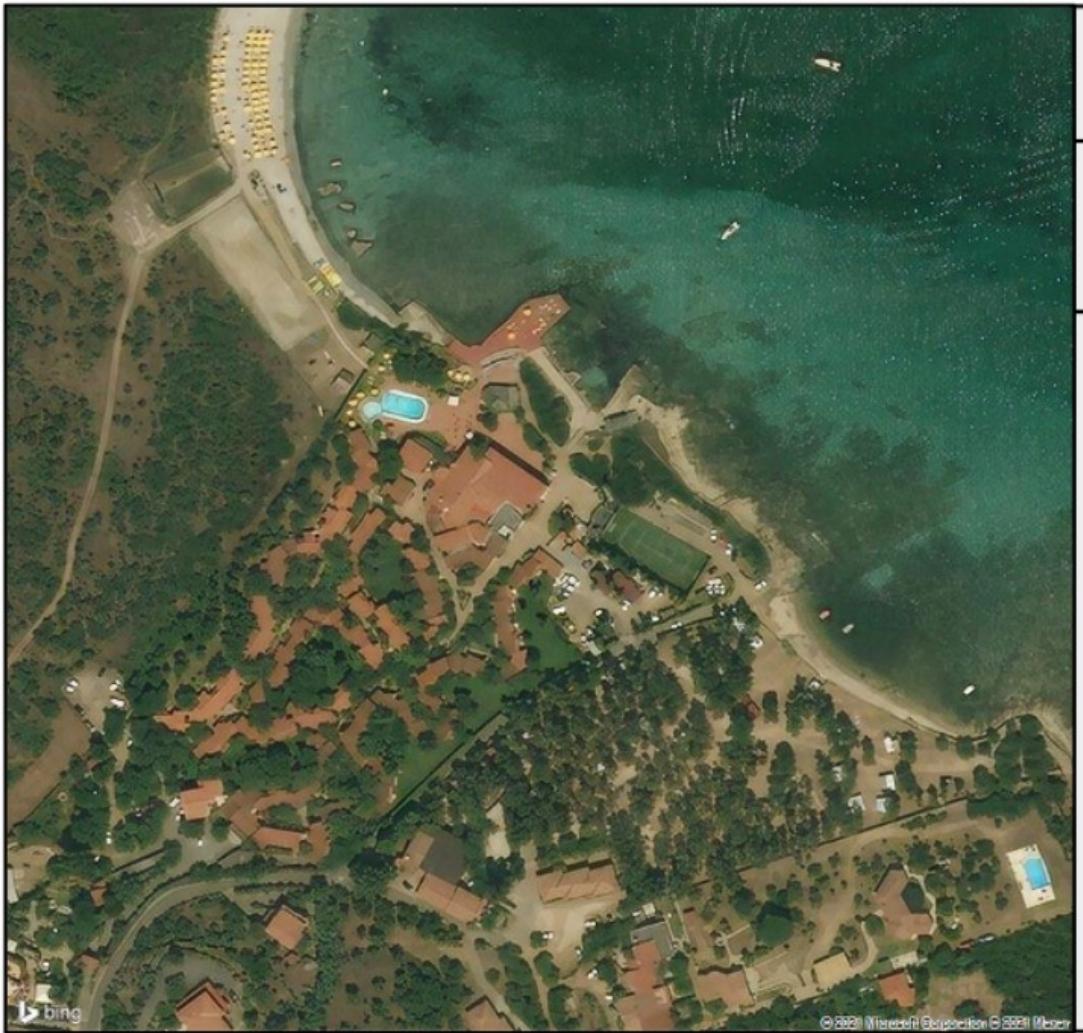


Satellite to Text Prediction: Using Sat2Cap and CLIPCap





	CLIP	(a) "aerial view of a beach"	(b) "house m from the center with internet, air conditioning, parking."	(c) "aerial view of an island"	(d) "aerial view of the property"
May		"sea facing apartment with swimming pool, terrace in a quiet residential area"	"beautiful mountain landscape with a green meadow and old wooden fence."	"Medieval Castle on the coast"	"kite on the beach at sunset"
Ours		"sailboat on the sea in winter"	"Frosty Winter Morning in the mountains"	"Medieval Castle on a winters day"	"jetski on the beach at sunset"
Jan					



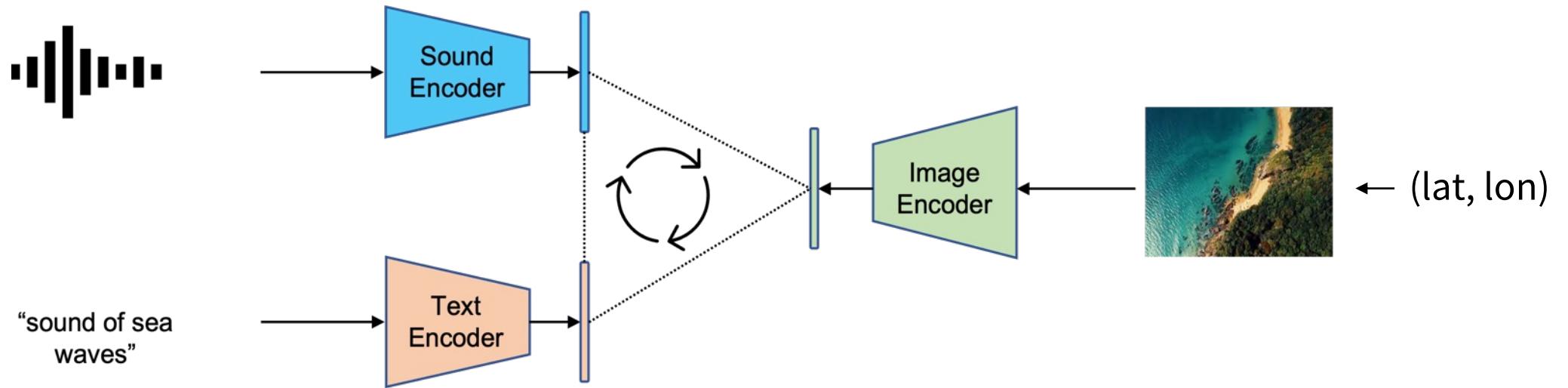
	Model	Date/Time	Description
	CLIP	-	<i>"aerial view of a beach"</i>
Ours		May 20 08:00 am	<i>"property image sea facing apartment with swimming pool, terrace in a quiet residential area."</i>
		Dec 20 10:00 am	<i>"Sailboat on the sea in winter"</i>
		Dec 20 05:00 pm	<i>"Person on the beach at night"</i>
		Dec 20 11:00 pm	<i>"Nighttime on the beach"</i>

Example 7: Mapping Sounds

$P(\text{sound} \mid \text{lat}, \text{lon})$



GeoCLAP Approach Overview

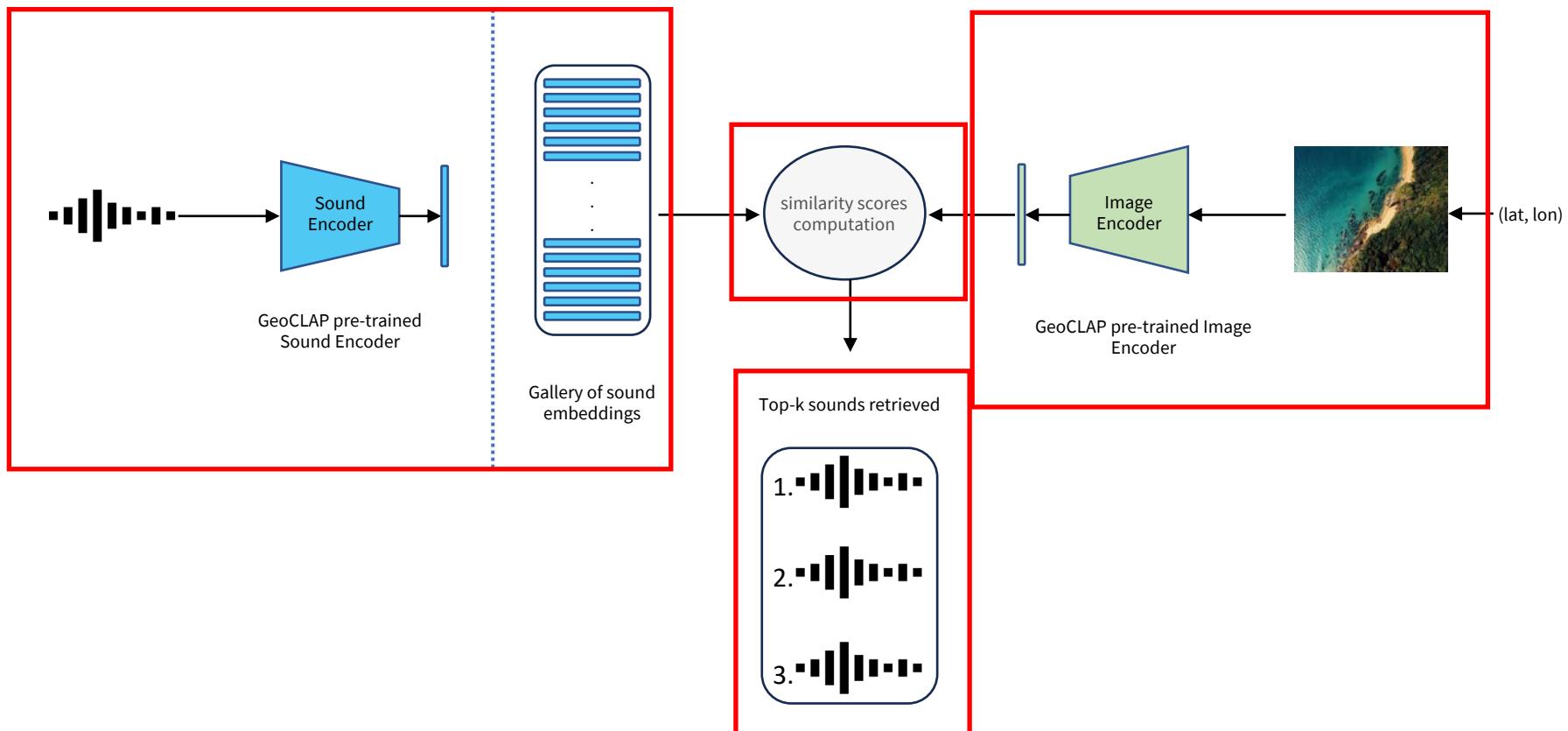


GeoCLAP is trained using the **SoundingEarth dataset** using contrastive loss between three pairs of modalities.

$$l = \frac{1}{2N} \sum_{k=1}^N \left(\log \frac{\exp((E_k^m \cdot E_k^n)/\tau_{mn})}{\sum_{j=1}^N \exp((E_k^m \cdot E_j^n)/\tau_{mn})} + \log \frac{\exp((E_k^n \cdot E_k^m)/\tau_{mn})}{\sum_{j=1}^N \exp((E_k^n \cdot E_j^m)/\tau_{mn})} \right)$$

$$\text{loss} = l(E^{\text{audio}}, E^{\text{text}}) + l(E^{\text{audio}}, E^{\text{image}}) + l(E^{\text{image}}, E^{\text{text}})$$

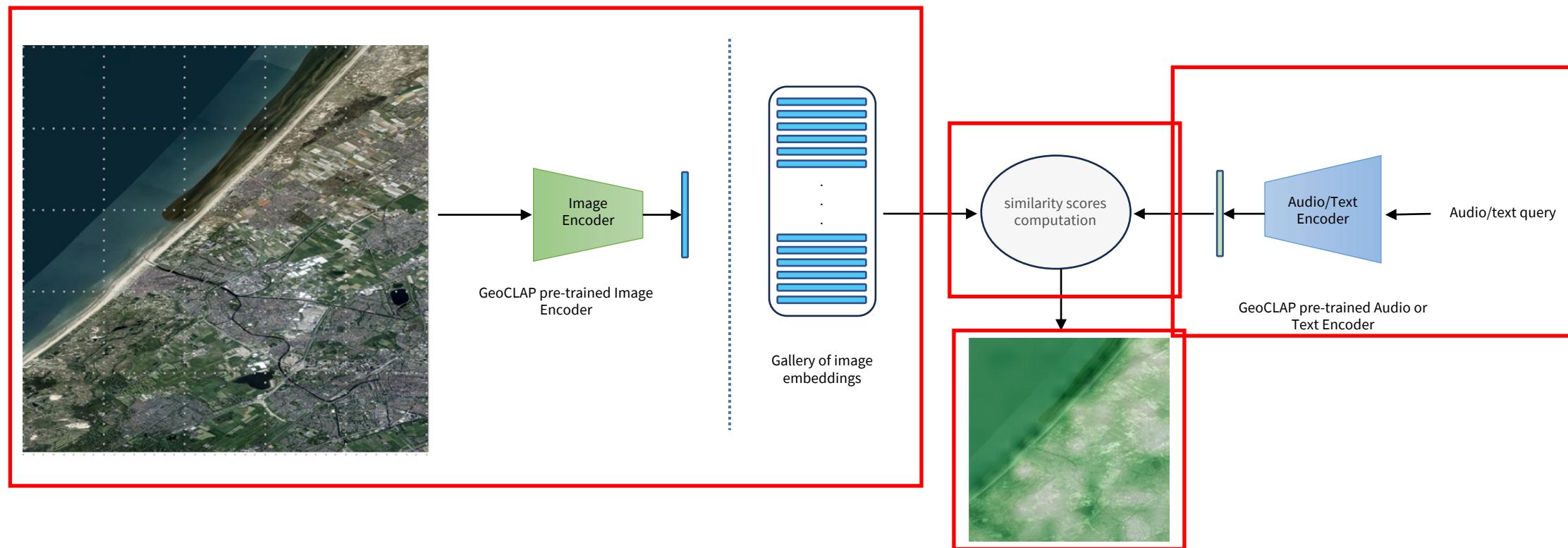
Satellite Image to Sound Retrieval



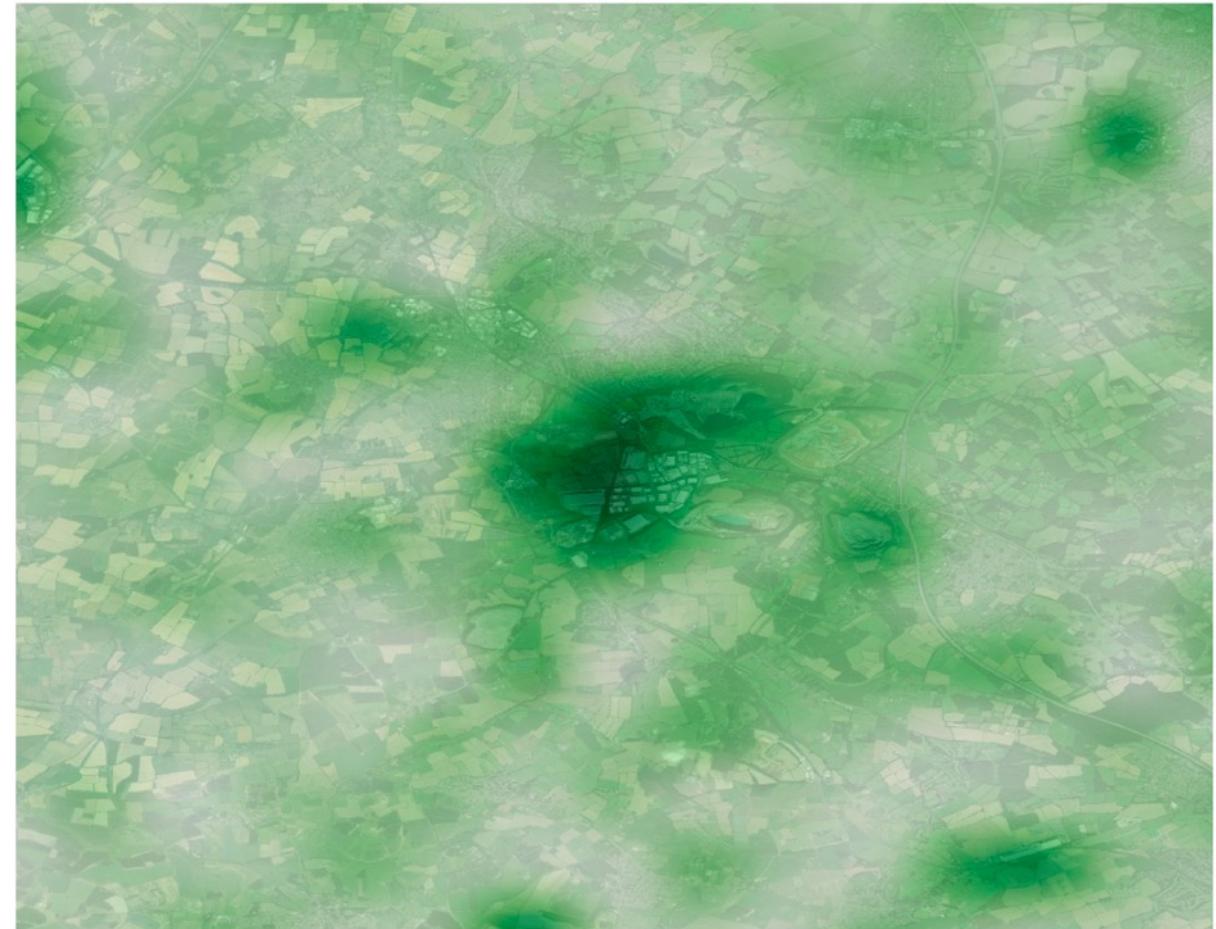
Cross-Modal Retrieval Performance

Experiment	Image Encoder	Text-Audio Encoder	Method		Image2Sound		Sound2Image	
			Text	Address	R@100	Median-R	R@100	Median-R
Baseline [1]	ResNet18	ResNet18	✗	✗	0.256	814	0.250	816
ours	SATMAE	L-CLAP-frozen	✗	✗	0.352	360	0.348	369
ours	SATMAE	L-CLAP-frozen	✓	✗	0.328	428	0.325	428
ours	SATMAE	L-CLAP-frozen	✗	✓	0.298	546	0.295	544
ours	SATMAE	L-CLAP-frozen	✓	✓	0.317	439	0.311	443
ours	SATMAE	L-CLAP	✗	✗	0.384	230	0.385	237
ours	SATMAE	L-CLAP	✓	✗	0.423	172	0.419	175
ours	SATMAE	L-CLAP	✗	✓	0.432	166	0.431	167
ours	SATMAE	L-CLAP	✓	✓	0.434	159	0.434	167

Zero-Shot Soundscape Mapping

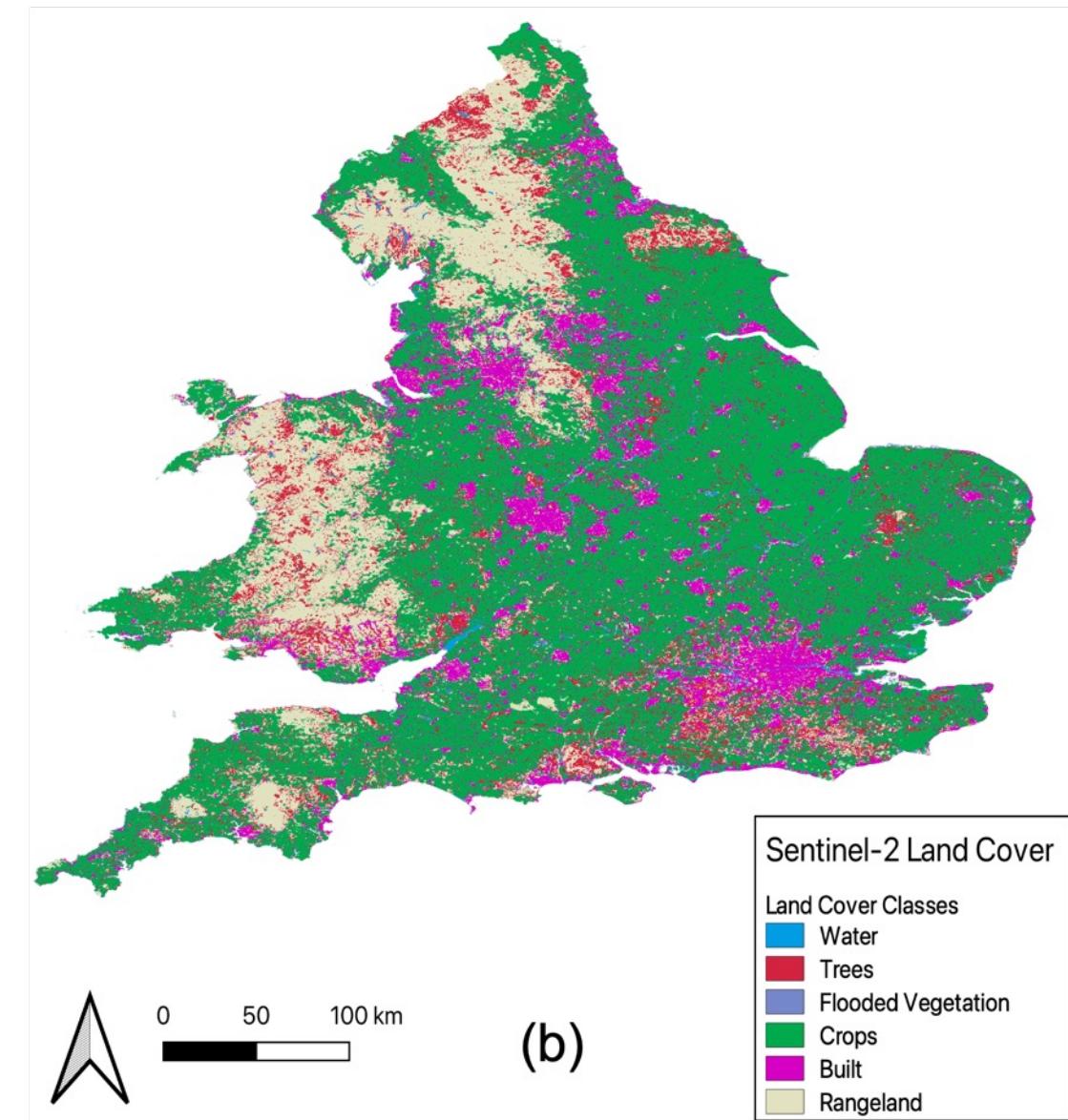
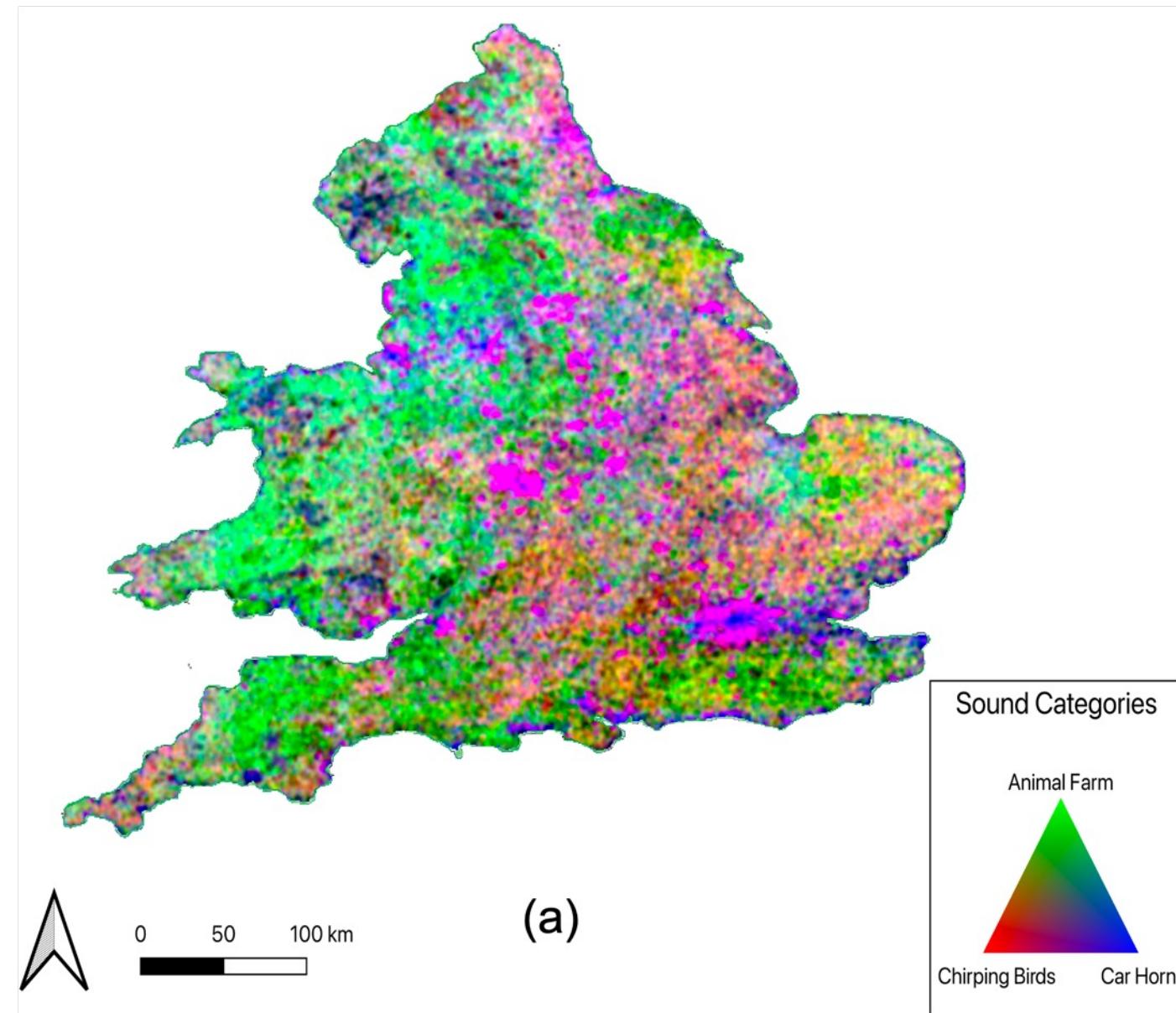


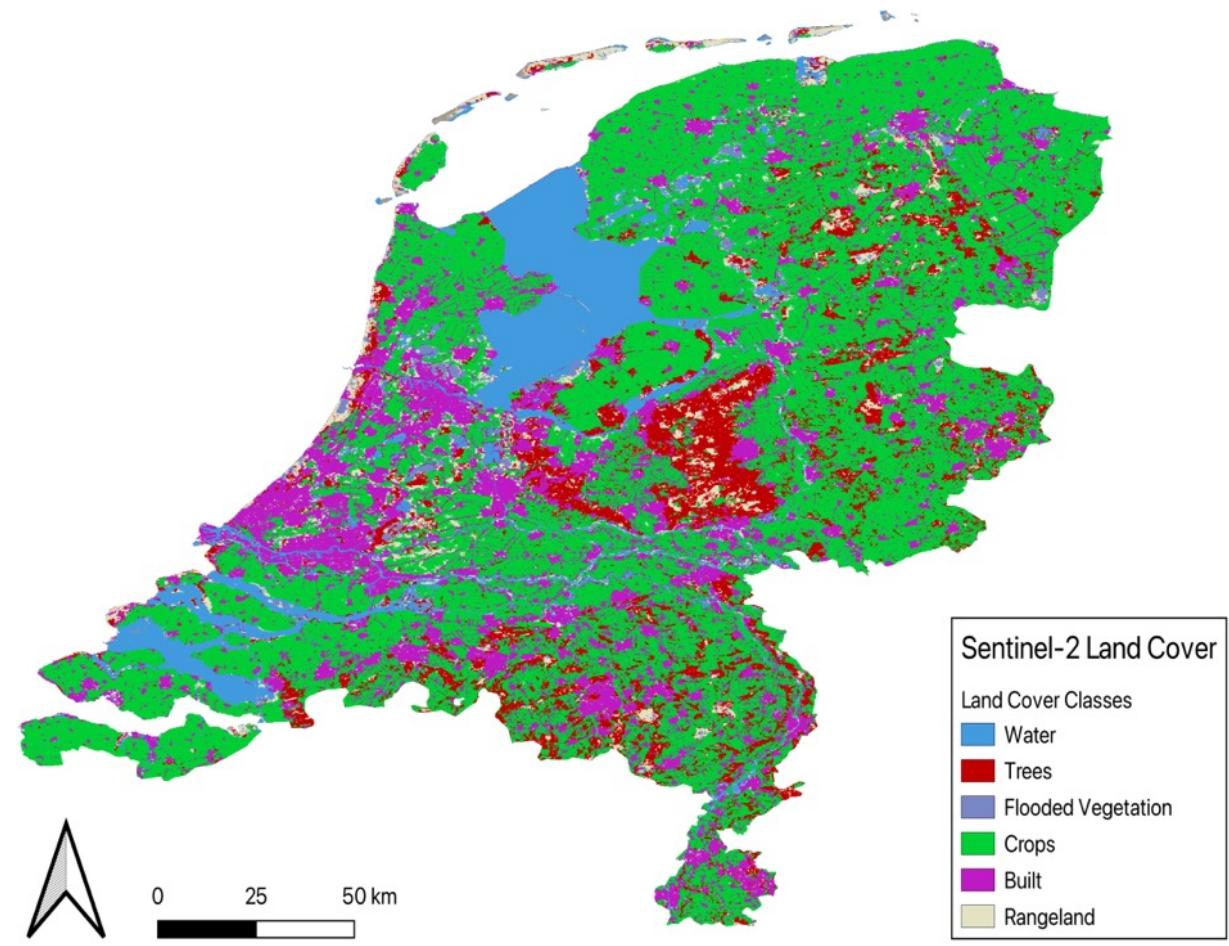
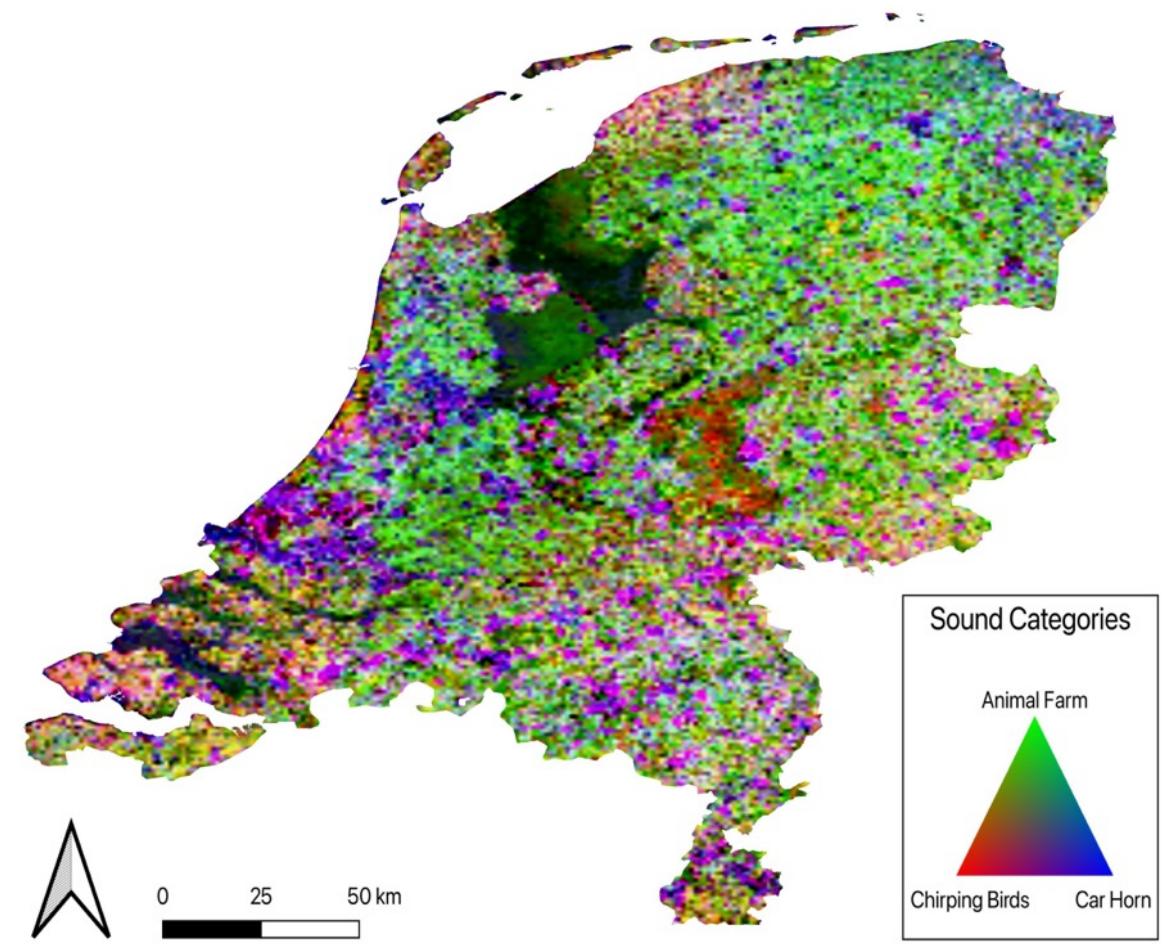
“the sound of a factory”



green: more probable, white: less probable







Summary: Cross-View Distillation

- Capable of mapping a wide variety of attributes
- Little need for manually annotated datasets
- Many opportunities for digging deeper:
 - Combining into a unified framework
 - Noise and uncertainty modeling
 - Integrating with applications
 - Overcoming inherent biases in the datasets



Near/Remote Sensing Models

- Distillation models are limited by the information extractable from the overhead satellite imagery
- Idea: combine overhead and ground-level imagery

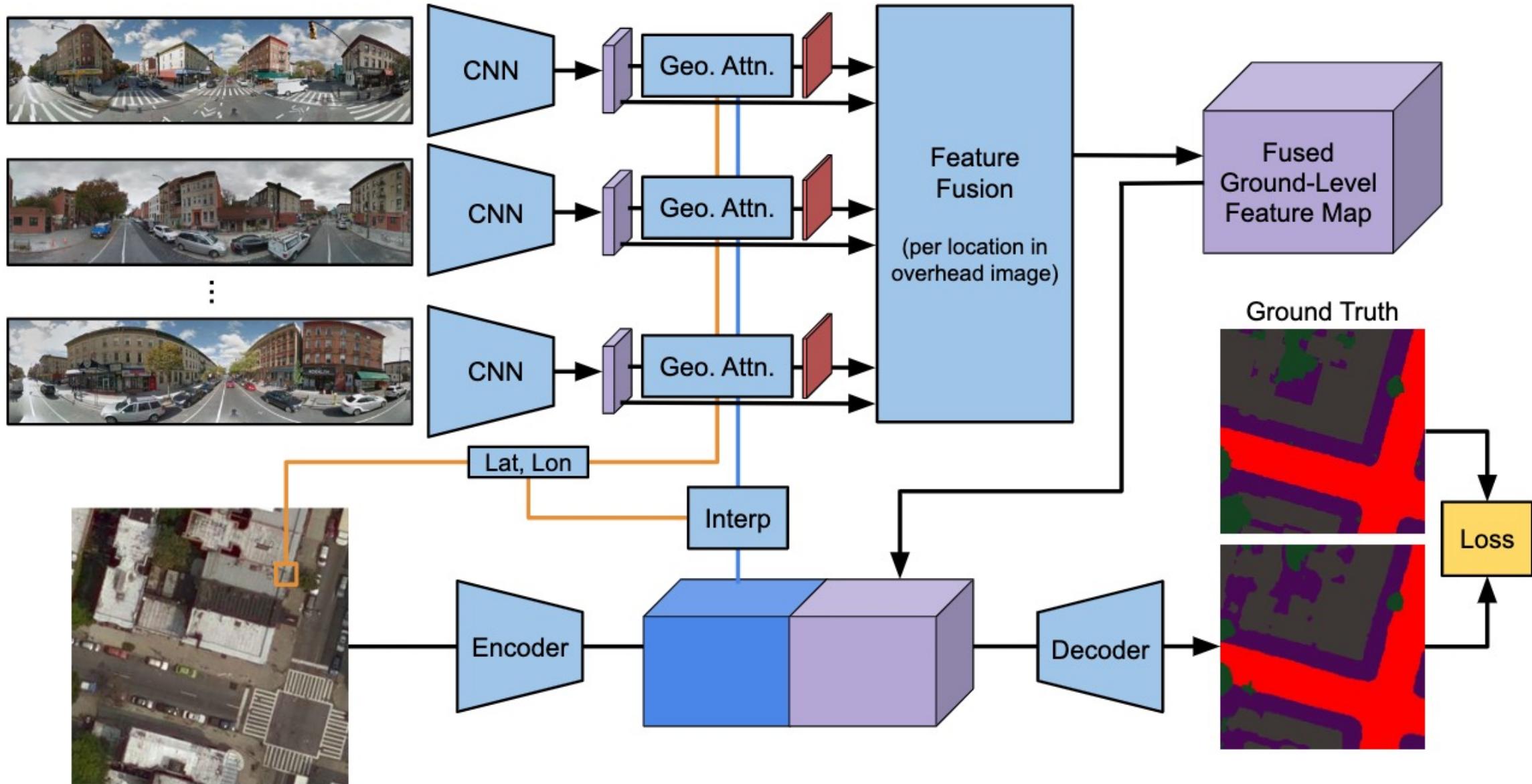


What can you tell me about this building?





Near/Remote Sensing using an Attention Architecture



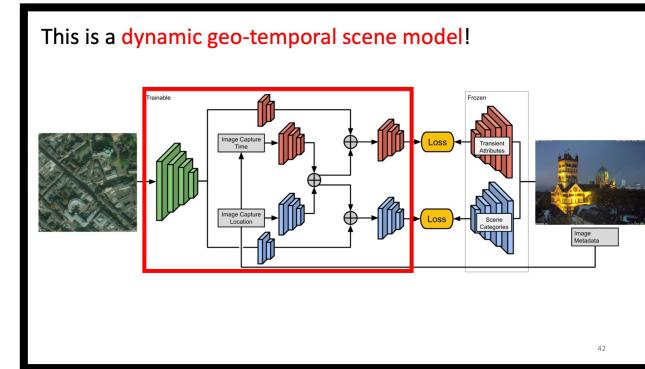
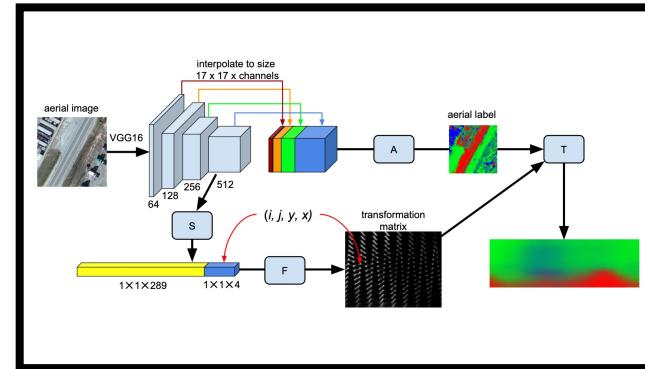
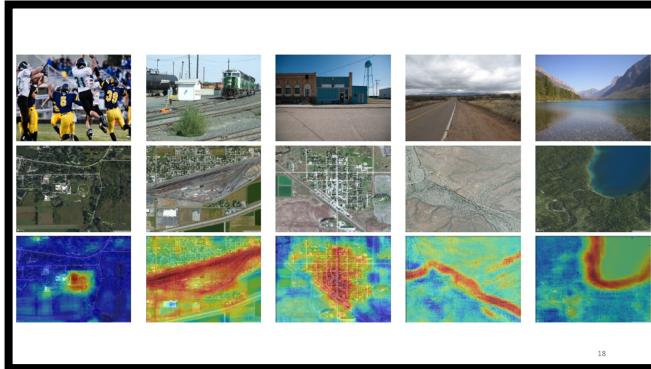
Results

Table 1. Brooklyn evaluation results.

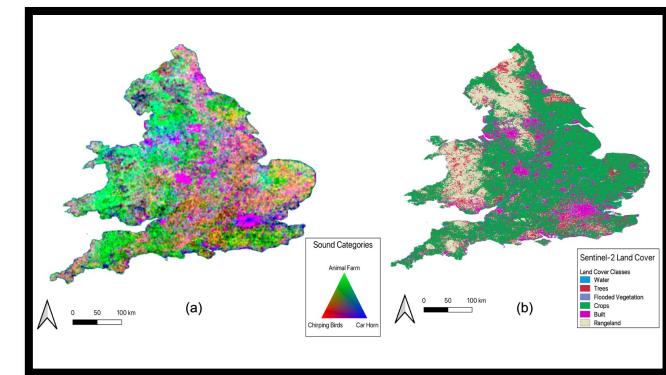
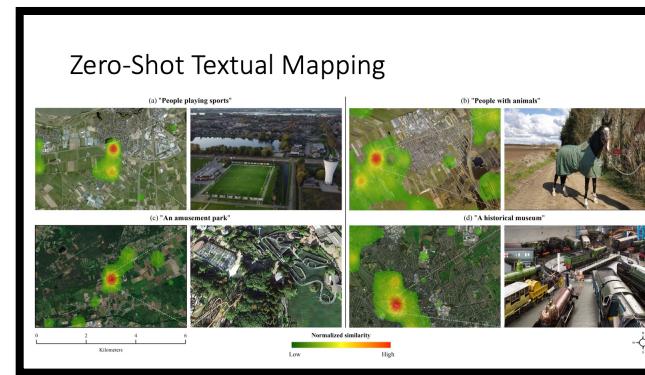
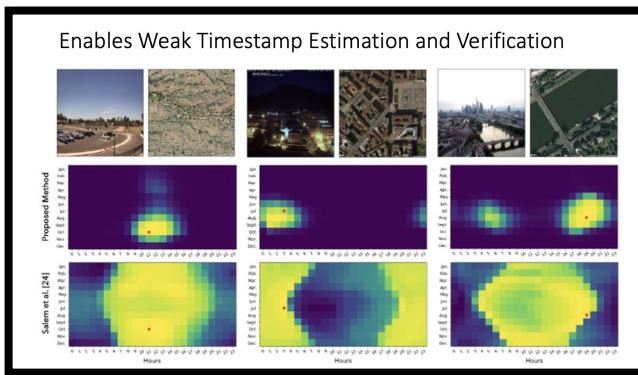
	Land Use		Age		Function		Land Cover		Height	
	mIOU	Acc	mIOU	Acc	mIOU	Acc	mIOU	Acc	RMSE	RMSE log
Workman <i>et al.</i> [55]	45.54%	77.40%	23.13%	43.85%	14.59%	44.88%				
Cao <i>et al.</i> [5]	48.15%	78.10%								
proximate	49.82%	75.30%	36.68%	56.48%	12.13%	43.81%	38.27%	67.63%	4.440	1.031
remote	40.30%	72.98%	16.40%	34.43%	4.50%	34.53%	69.48%	86.71%	3.260	0.785
ours	69.24%	86.82%	51.70%	70.34%	27.40%	60.31%	74.59%	88.10%	2.845	0.747

Contributions

- Consumer photographs are a strong source of supervision for remote sensing
- It's possible to map attributes for which it's hard to obtain ground truth
- Overhead and ground-level images can be merged to make better fine-grained predictions



Thanks! Questions?



More Info: <https://mvrl.cse.wustl.edu/>

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1553116. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Additional Funding Acknowledgements: Google Faculty Research Award, IARPA (Finder), AWS Research Education Grant, NVIDIA Hardware Donation.