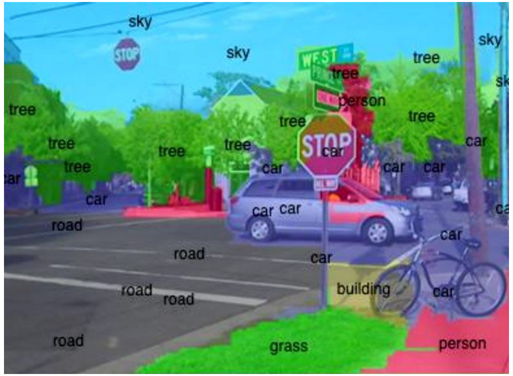
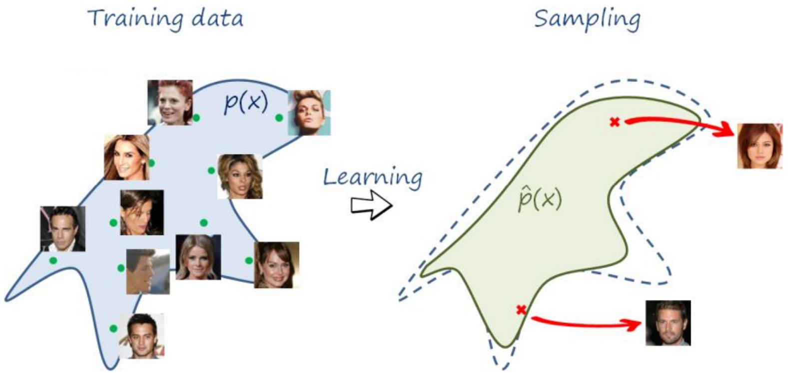
Stanford Artificial Intelligence Lab.



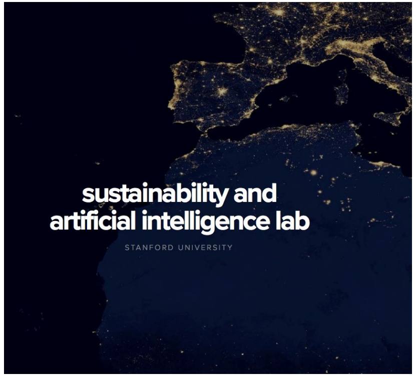
Efficient Computer Vision

Reinforcement Learning
Weakly Supervised Learning
Domain Adaptation
Unsupervised Learning
Adversarial Learning



Generative Models

Generative Adversarial
Networks



Machine Learning for Sustainability

Language Data
Multi-modal Data
Satellite Images.
Cellphone Data

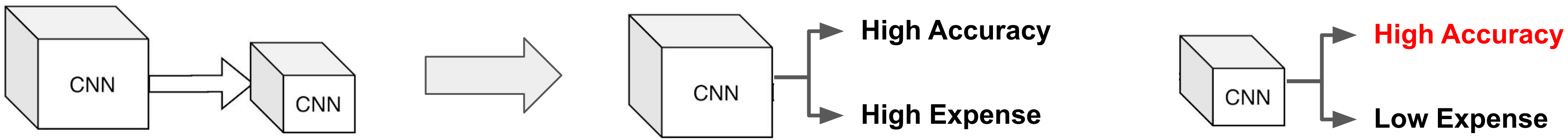
Learning Where and When to Zoom using Deep Reinforcement Learning

CVPR 2020

Burak Uzkent and Stefano Ermon

Department of Computer Science
Stanford University

Introduction - Runtime Efficiency



Model Compression

[Hinton et al. 2015, Han et al. 2015, Huang et al. 2018, Wu et al. 2018, Rastegari et al. 2016]

- Adaptive model compression ***maintains the accuracy*** of the complex models.

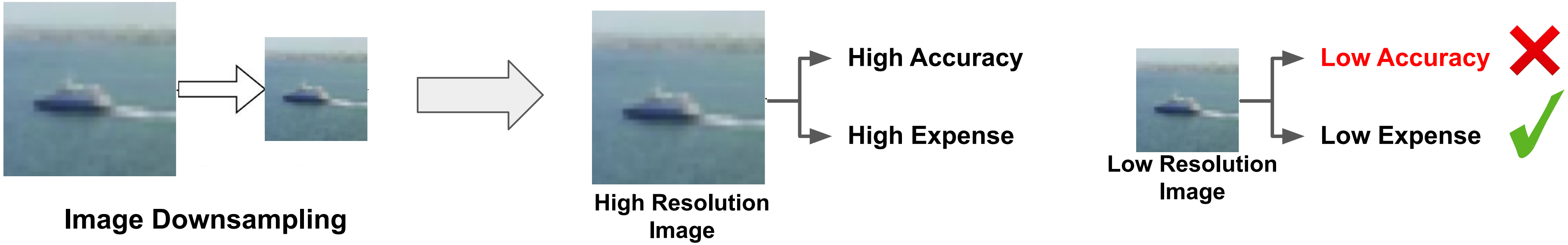


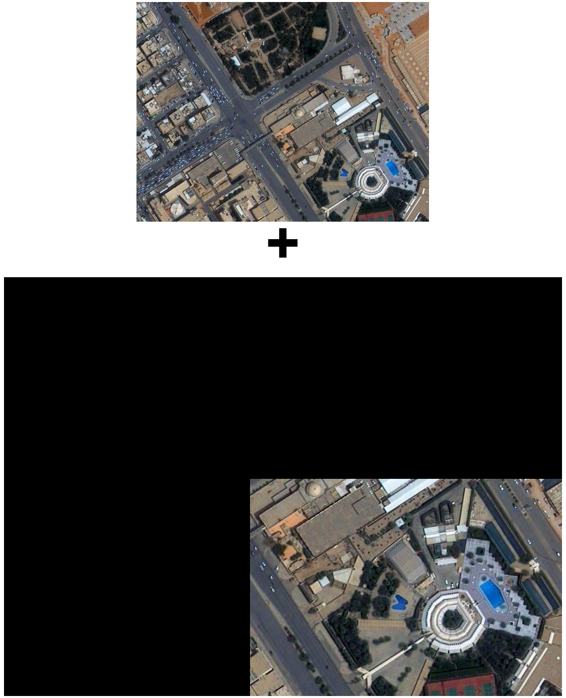
Image Downsampling

- There exists ***no adaptive compression*** technique on the image domain.

Introduction - Remote Sensing

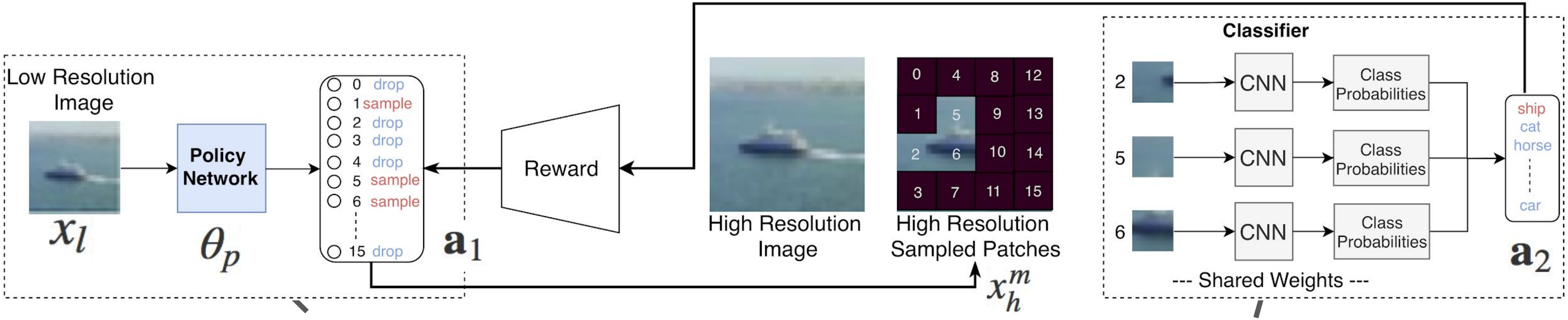
Cheap to Acquire
Low Accuracy

Expensive to Acquire
High Accuracy



Cheap to Acquire
High Accuracy

PatchDrop - Adaptive Solution



Policy Network

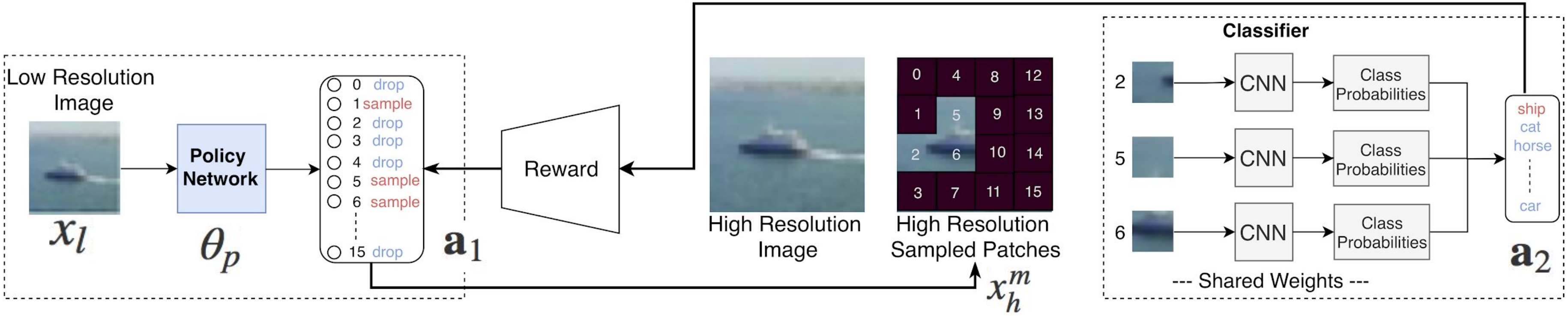
Classifier

Policies ⇨ $\pi_1(\mathbf{a_1}|x_l;\theta_p) = p(\mathbf{a_1}|x_l;\theta_p)$ $\quad$ $\pi_2(\mathbf{a_2}|x_h^m;\theta_{cl}) = p(\mathbf{a_2}|x_h^m;\theta_{cl})$

Actions ⇨ $\mathbf{a}_1 \in \{0,1\}^P$ $\quad$ $\mathbf{a}_2 \in \{0,1,\ldots N\}$

High Accuracy ✓ **Low Expense** ✓

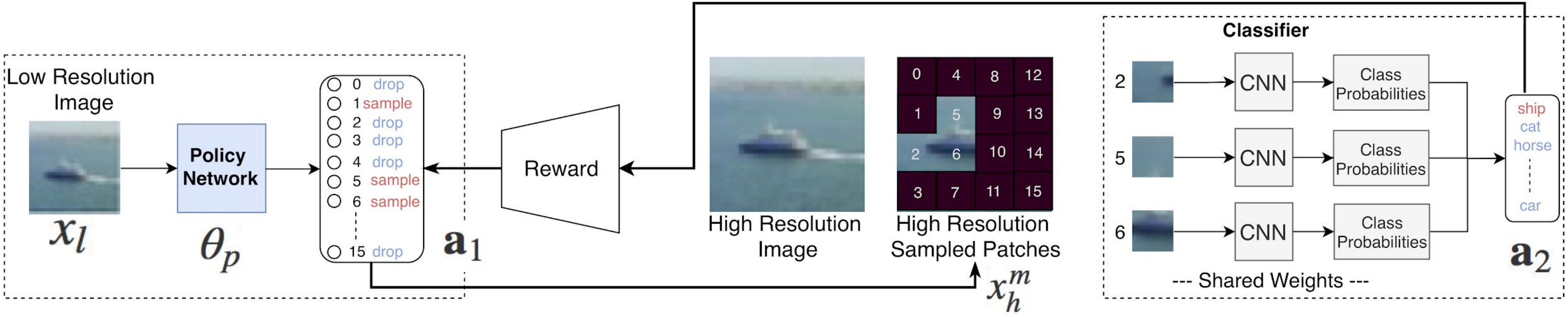
Modeling the Policy Network and Classifier



Patch Sampling Policy ⇨ $$\pi_1(\mathbf{a}_1|x_l,\theta_p) = \prod_{p=1}^{P} s_p^{\mathbf{a}_1^{\mathbf{p}}}(1-s_p)^{(1-\mathbf{a}_1^{\mathbf{p}})}$$

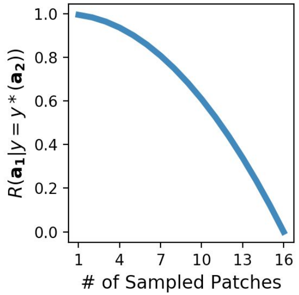
Classification Policy ⇨ $$\mathbf{a}_2 = softmax(f_{cl}(x_h^2;\theta_{cl}) + f_{cl}(x_h^5;\theta_{cl}) + f_{cl}(x_h^6;\theta_{cl}))$$

Modeling the Reward Function



Reward Function

$$R(\mathbf{a}_1, \mathbf{a}_2, y) = \begin{cases} 1 - \left(\frac{|\mathbf{a}_1|_1}{P}\right)^2 & \text{if } y = y^*(\mathbf{a}_2) \\ -\sigma & \text{Otherwise} \end{cases}$$

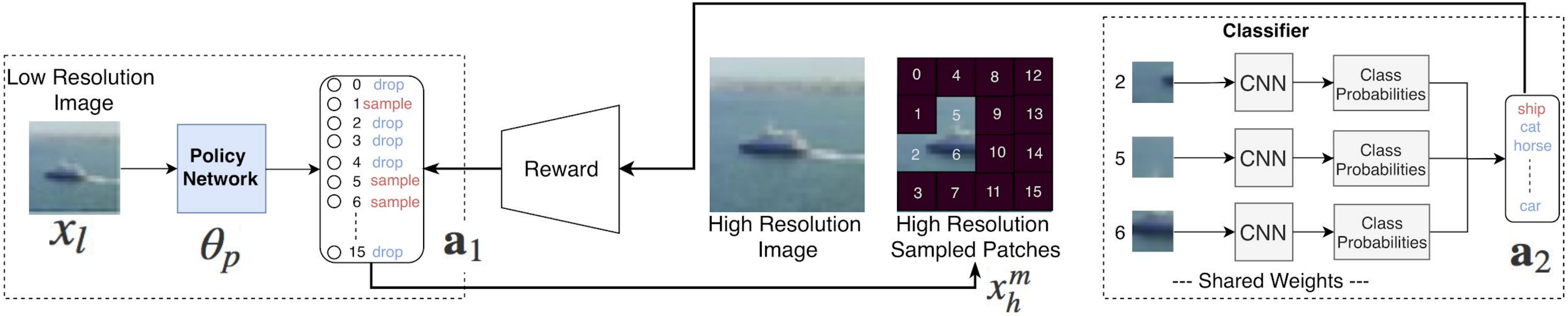


Cost Function

$$\nabla_{\theta_p} J = \mathbb{E}[R(\mathbf{a}_1, \mathbf{a}_2, y) \nabla_{\theta_p} \log \pi_{\theta_p}(\mathbf{a}_1 | x_l)]$$

Uses the Policy Gradient Algorithm

Modeling the Policy Network and Classifier



Cost Function

$$\nabla_{\theta_p} J = \mathbb{E}[R(\mathbf{a_1}, \mathbf{a_2}, y) \nabla_{\theta_p} \log \pi_{\theta_p}(\mathbf{a_1}|x_l)]$$

$$\nabla_{\theta_p} J = \mathbb{E}[A \sum_{p=1}^{P} \nabla_{\theta_p} \log(s_p \mathbf{a_1^p} + (1 - s_p)(1 - \mathbf{a_1^p}))]$$

Advantage Function

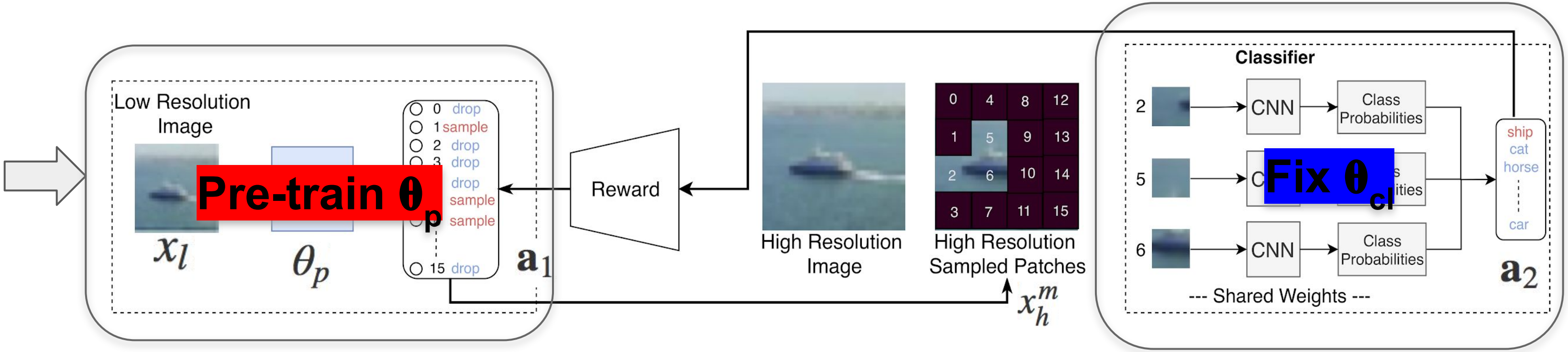
$$A(\mathbf{a_1}, \mathbf{a_2}, \hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2) = R(\mathbf{a_1}, \mathbf{a_2}, y) - R(\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, y)$$

Temperature Scaling

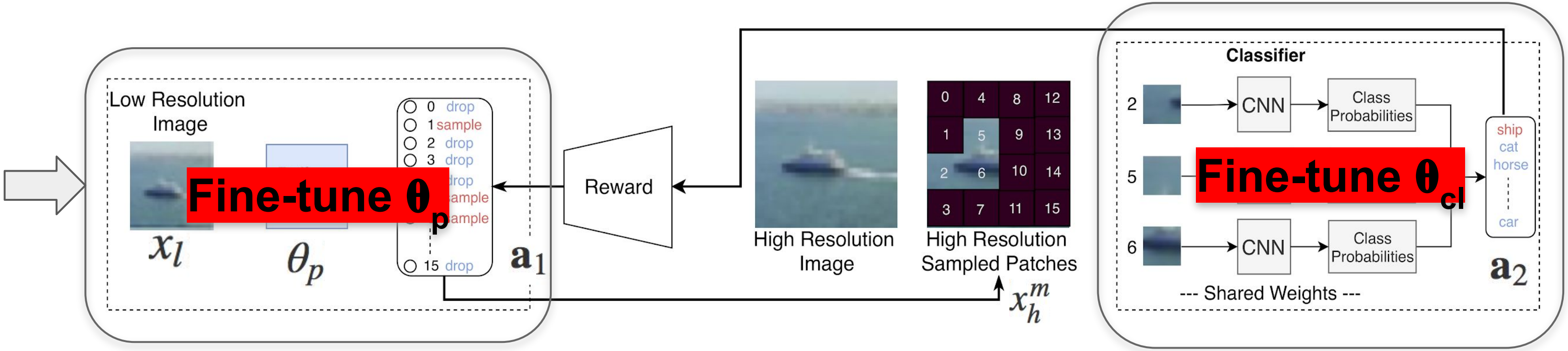
$$s_p = \alpha s_p + (1 - \alpha)(1 - s_p)$$

Training Protocol

Pre-training

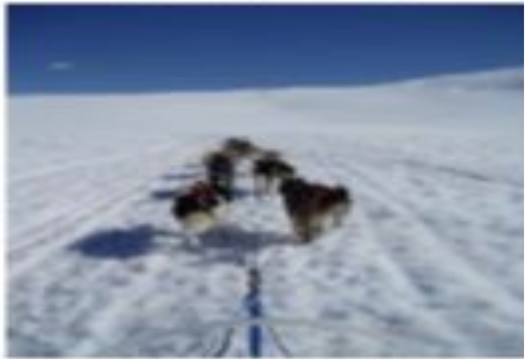


Fine-tuning



Experiments on ImageNet/CIFAR10

LR (56x56)



HR (224x224)

	CIFAR10			ImageNet		
	Acc. (%) (Pt)	Acc. (%) (Ft-1)	S	Acc. (%) (Pt)	Acc. (%) (Ft-1)	S
LR-CNN	75.8	75.8	0,0	58.1	58.1	0,0
SRGAN	78.8	78.8	0,0	63.1	63.1	0,0
KD	81.8	81.8	0,0	62.4	62.4	0,0
PCN	83.3	83.3	0,0	63.9	63.9	0,0
HR-CNN	92.3	92.3	16,16	76.5	76.5	16,16
Fixed-H	71.2	83.8	9,8	48.8	68.6	10,9
Fixed-V	64.7	83.4	9,8	48.4	68.4	10,9
Stochastic	40.6	82.1	9,8	38.6	66.2	10,9
STN	66.9	85.2	9,8	58.6	69.4	10,9
PatchDrop	**80.6**	**91.9**	**8.5,7.9**	**60.2**	**74.9**	**10.1,9.1**

Table 1: Experiments on ImageNet and CIFAR10

*We process ***about 45-50% fewer number of pixels*** than HR-CNN.

Experiments on functional map of the world (fMoW)

LR (56x56)

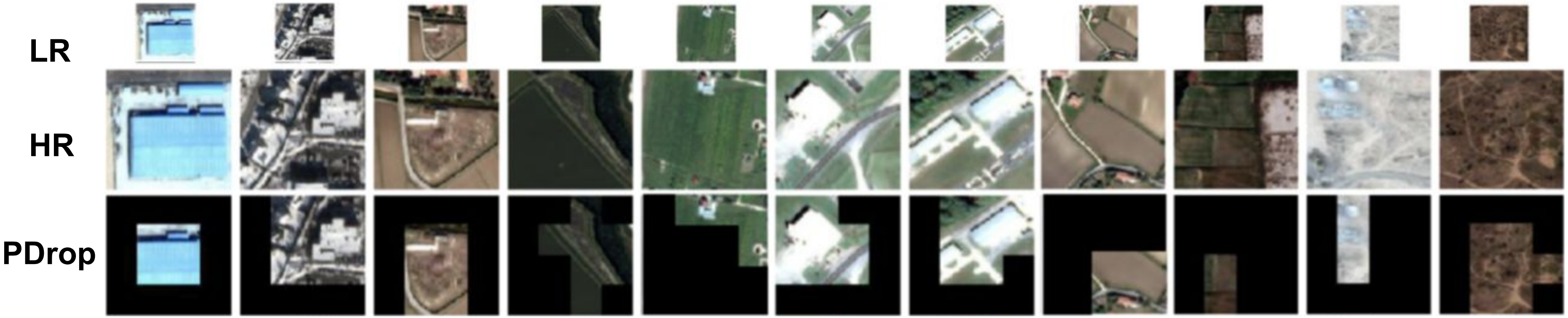
HR (224x224)

	Acc. (%) (Pt)	S	Acc. (%) (Ft-1)	S
LR-CNN	61.4	0	61.4	0
SRGAN	62.3	0	62.3	0
KD	63.1	0	63.1	0
PCN	63.5	0	63.5	0
HR-CNN	67.3	16	67.3	16
Fixed-H	47.7	7	63.3	6
Fixed-V	48.3	7	63.2	6
Stochastic	29.1	7	57.1	6
STN	46.5	7	61.8	6
PatchDrop	**53.4**	**7**	**67.1**	**5.9**

Table 2: Experiments on fMoW

*We use ***about 60% less # of pixels*** than HR-CNN
*We can save about 100,000 dollars when performing a vision task using HR satellite images at global scale.

Qualitative Results - fMoW

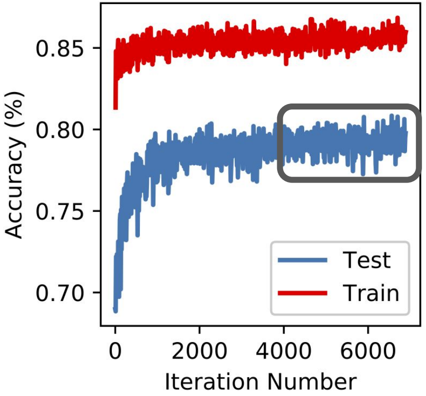


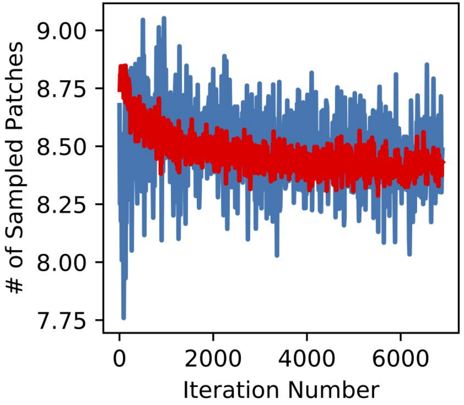
LR -> 56x56 pixels

HR -> 224x224 pixels

Impact of Joint Fine-tuning on CIFAR10

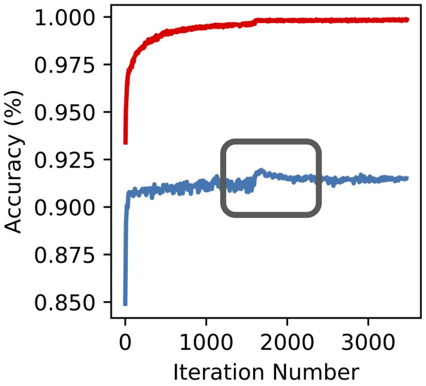
Pretraining

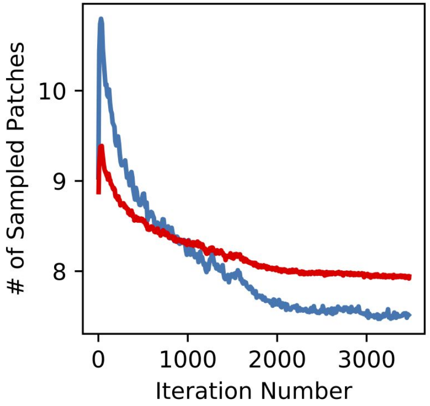




Acc(%): 80.6
S : 8.5

Joint Fine-tuning

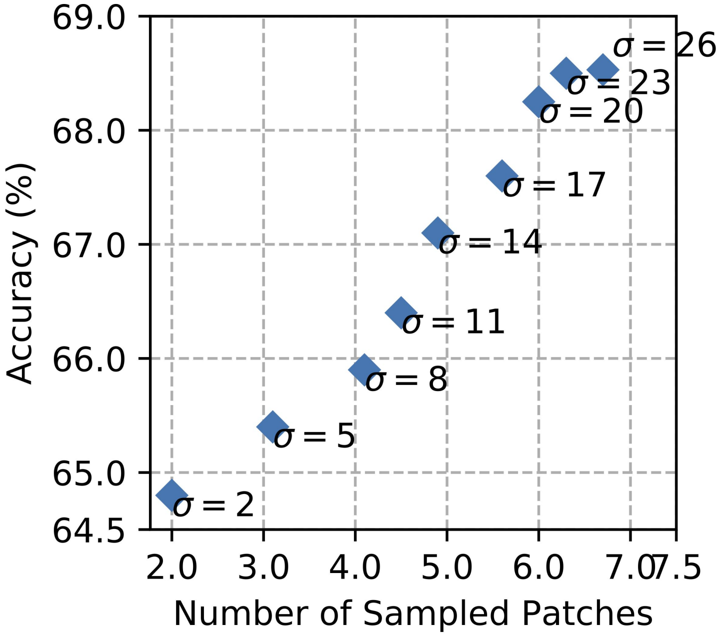




Acc(%): 91.9 ↑ ***11.3***
S : 7.9 ↓ ***0.6***

Reward Function (CIFAR100)

$$R(\mathbf{a}_1, \mathbf{a}_2, y) = \begin{cases} 1 - (\frac{|\mathbf{a_1}|_1}{P})^2 & \text{if } y = y^*(\mathbf{a}_2) \\ -\sigma & \text{Otherwise} \end{cases}$$



Run-time Efficiency

	CIFAR10	CIFAR100	fMoW	ImageNet
LR-CNN	4.4 M	4.4 M	240 M	240 M
HR-CNN	69.1 M	69.1 M	3.8 B	3.8 B
Fixed-H	39 M	43 M	1.7 B	2 B
Fixed-V	39 M	43 M	1.7 B	2 B
Stochastic	39 M	43 M	1.7 B	2 B
STN	41.2 M	46.7 M	2 B	2.3 B
PatchDrop	**40.1** M	**45.4** M	**1.9** B	**2.2** B

Table 3: Run-time efficiency (FLOPS) on four different benchmarks.

*Patchwise inference ***reduce computational complexity by 40-50%*** without changing the underlying CNN structure.

Brendel, Wieland, and Matthias Bethge. "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet." *arXiv preprint arXiv:1904.00760* (2019).

Conclusions

- We proposed an ***adaptive, conditional*** method to process adaptive number of pixels with convolutional neural networks.
- With the proposed method, on average we use up to ***50% less*** *number of pixels* and this leads to:
  - **40-50% less** run-time FLOPs.
  - **less dependency** on high resolution images (can be cost-saving in some application domains.)
- We extended the problem to object detection in large images and show that we can reduce the dependency on using HR images for object detection.

Reducing Dependency on Labels on Remote Sensing Images

Learning to Interpret Satellite Images using Wikipedia Articles

IJCAI 2019

*Burak Uzkent, *Evan Sheehan, *Chenlin Meng, **David Lobell, **Marshall Burke, *Stefano Ermon

*Department of Computer Science, Stanford University

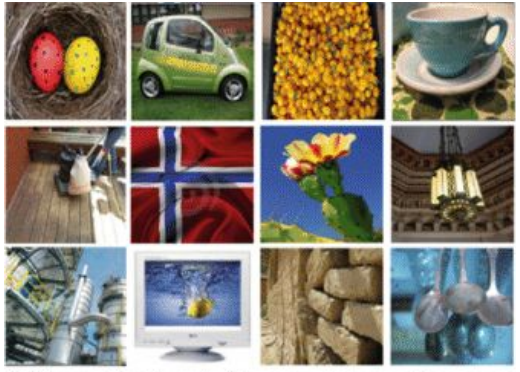
**Department of Earth Science, Stanford University

Introduction

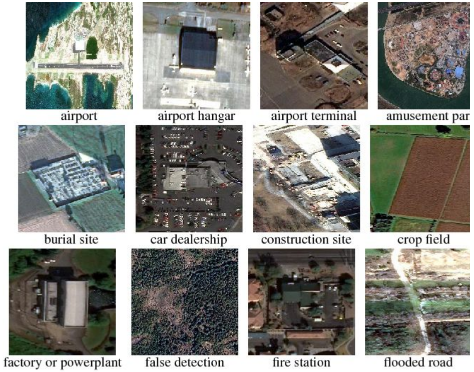
- Training Convolutional Neural Networks are usually done by:
  - *First pre-training on **ImageNet Dataset**.*
  - *And then fine-tuning on the **Target Dataset**.*
- This procedure can be very useful for:
  - **Faster convergence** in target dataset training
  - **Improved downstream accuracy** for small-size target datasets.
- However, pre-training on ImageNet can be less helpful when the shift between ImageNet and target dataset distribution is ***large***.

Motivation

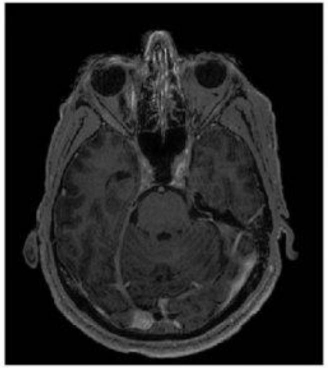
- In some applications, i.e. *remote sensing and medical images*, data distribution is very different from ImageNet's one.



ImageNet



Satellite

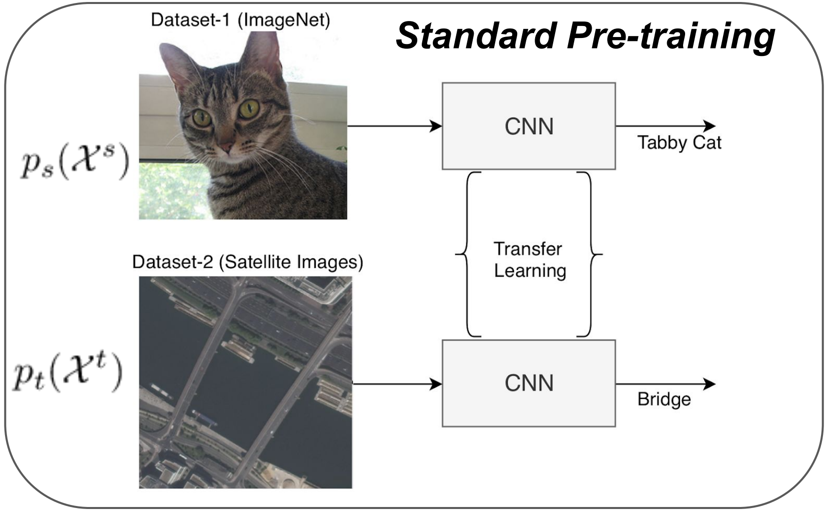


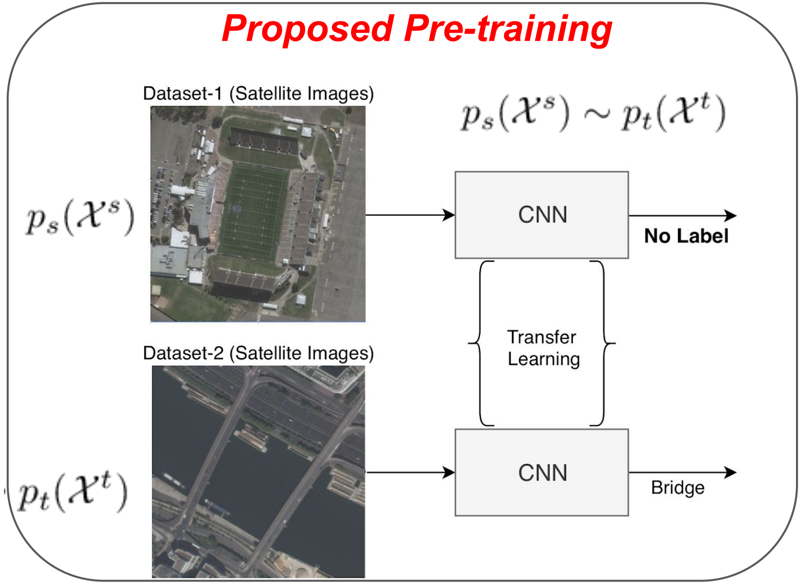
MRI

- In these cases, it is beneficial to do pre-training on a similar distribution dataset. [Zhang et al. Arxiv20]

Proposed Method

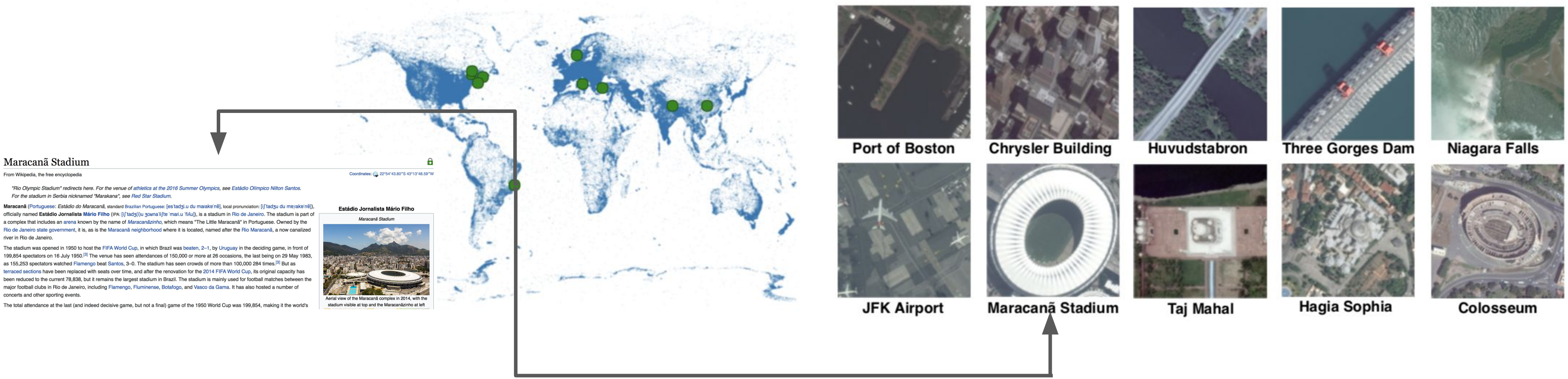
- In this study, we propose a method to efficiently pre-train a CNN on dataset with satellite images.





Learning from Satellite Images using Wikipedia Articles

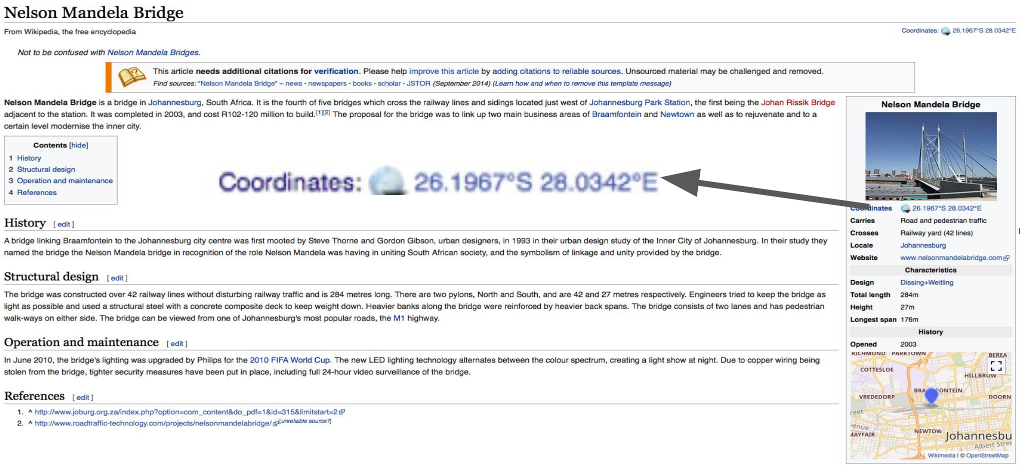
- In its latest dump, Wikipedia contains **~5 million articles (English)** and **~1 million articles** are geo-referenced.



Scatter plot of the distribution of geo-tagged Wikipedia articles together with corresponding high resolution images.

Pairing Articles to Satellite Images - WikiSatNet

$$\mathcal{D} = \{(c_1, x_1, y_1), (c_2, x_2, y_2), \cdots, (c_N, x_N, y_N)\}$$


Nelson Mandela Bridge

From Wikipedia, the free encyclopedia

Coordinates: 26.1967°S 28.0342°E

Not to be confused with Nelson Mandela Bridges.

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "Nelson Mandela Bridge" – news · newspapers · books · scholar · JSTOR (September 2014) (Learn how and when to remove this template message)

Nelson Mandela Bridge is a bridge in Johannesburg, South Africa. It is the fourth of five bridges which cross the railway lines and sidings located just west of Johannesburg Park Station, the first being the Johan Rissik Bridge adjacent to the station. It was completed in 2003, and cost R102-120 million to build.[1][2] The proposal for the bridge was to link up two main business areas of Braamfontein and Newtown as well as to rejuvenate and to a certain level modernise the inner city.

Contents [hide]

Coordinates: 26.1967°S 28.0342°E

History [edit]

A bridge linking Braamfontein to the Johannesburg city centre was first mooted by Steve Thorne and Gordon Gibson, urban designers, in 1993 in their urban design study of the Inner City of Johannesburg. In their study they named the bridge the Nelson Mandela bridge in recognition of the role Nelson Mandela was having in uniting South African society, and the symbolism of linkage and unity provided by the bridge.

Structural design [edit]

The bridge was constructed over 42 railway lines without disturbing railway traffic and is 284 metres long. There are two pylons, North and South, and are 42 and 27 metres respectively. Engineers tried to keep the bridge as light as possible and used a structural steel with a concrete composite deck to keep weight down. Heavier banks along the bridge were reinforced by heavier back spans. The bridge consists of two lanes and has pedestrian walk-ways on either side. The bridge can be viewed from one of Johannesburg's most popular roads, the M1 highway.

Operation and maintenance [edit]

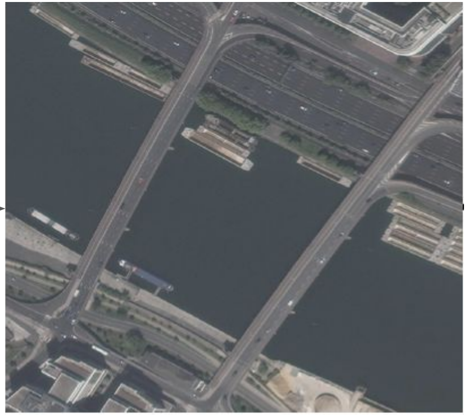
In June 2010, the bridge's lighting was upgraded by Philips for the 2010 FIFA World Cup. The new LED lighting technology alternates between the colour spectrum, creating a light show at night. Due to copper wiring being stolen from the bridge, tighter security measures have been put in place, including full 24-hour video surveillance of the bridge.

References [edit]

1. ^ http://www.joburg.org.za/index.php?option=com_content&do_pdf=1&id=315&limitstart=2
2. ^ http://www.roadtraffic-technology.com/projects/nelsonmandelabridge/ [unreliable source?]

Nelson Mandela Bridge	
Coordinates	26.1967°S 28.0342°E
Carries	Road and pedestrian traffic
Crosses	Railway yard (42 lines)
Locale	Johannesburg
Website	www.nelsonmandelabridge.com
Characteristics	
Design	Dissing+Weitling
Total length	284m
Height	27m
Longest span	176m
History	
Opened	2003

Pair to an overhead image

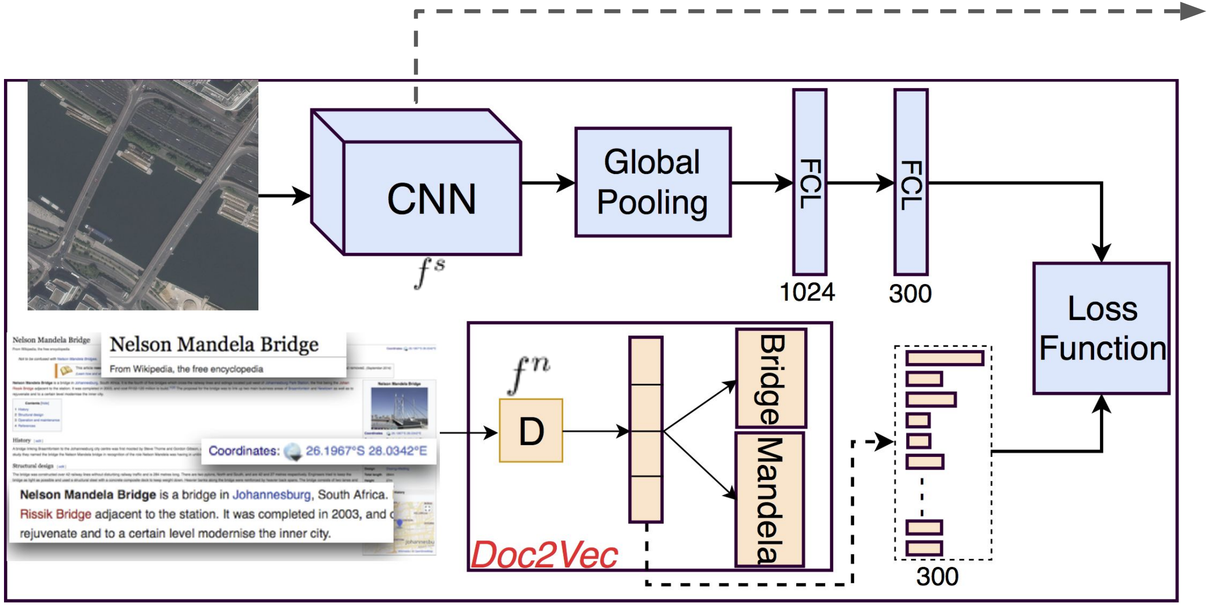


Coordinates: 26.1967°S 28.0342°E

Gomez, L., Patel, Y., Rusiñol, M., Karatzas, D. and Jawahar, C.V., 2017. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4230-4239).

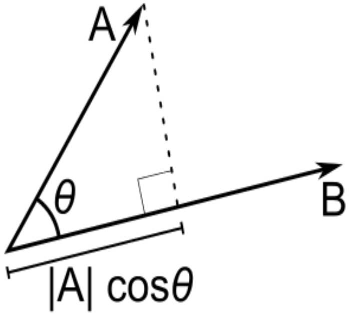
Representation Learning with Image2Text Matching

Downstream Task
(Image Recognition
Object Detection
Semantic Segmentation)



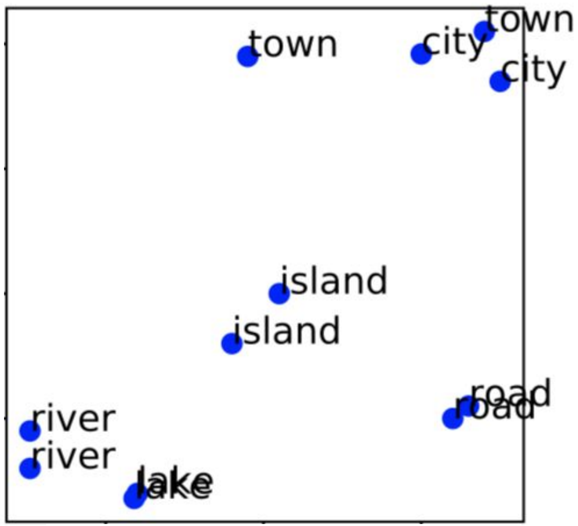
Loss Function

$$cos(\theta) = \frac{f^s(x_i^s;\theta^s)^T f_n(y_i;\theta^n)}{\|f^s(x_i^s;\theta^s)\|_2 \, \|f^n(y_i;\theta^n)\|_2}$$



***An automatic approach.**

Analyzing Doc2Vec Model



City - Middletown, Connecticut
City - Milton, Georgia
Lake - Timothy Lake
Lake - Tinquilco Lake
Town - Mingona Township, Kansas
Town - Moon Township, Pennsylvania
Road - Morehampton Road, Dublin
Road - Motorway M10 Pakistan
River - Motru River
River - Mousam River
Island - Aupaluktok Island
Island - Avatanak Island

***Articles with similar content are projected to the similar latent space.**

Pre-training Experiments (Image2Text)

- We use DenseNet with 121 layers to parameterize the CNN.

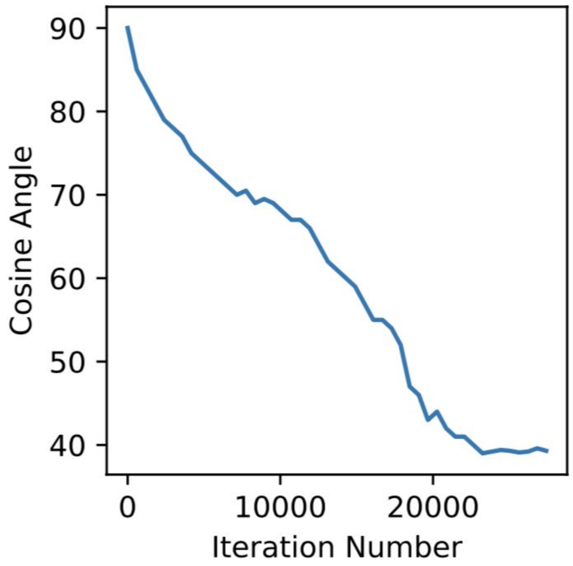


Figure 1: Training Loss

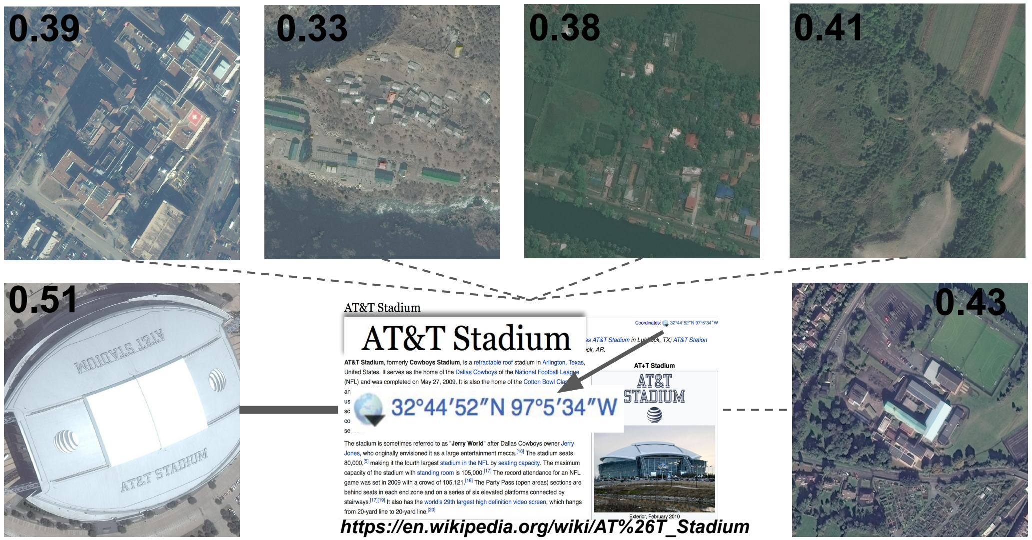
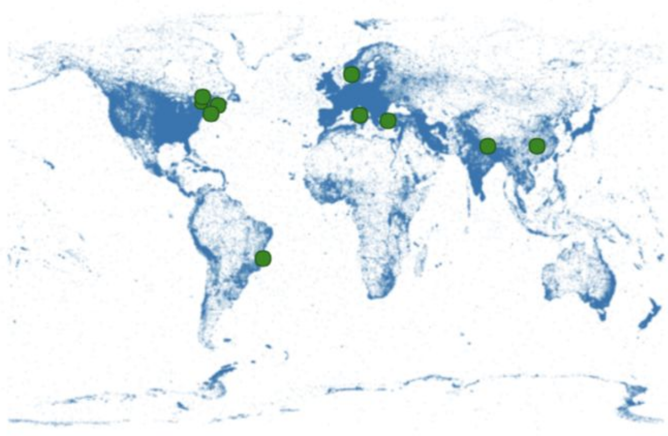


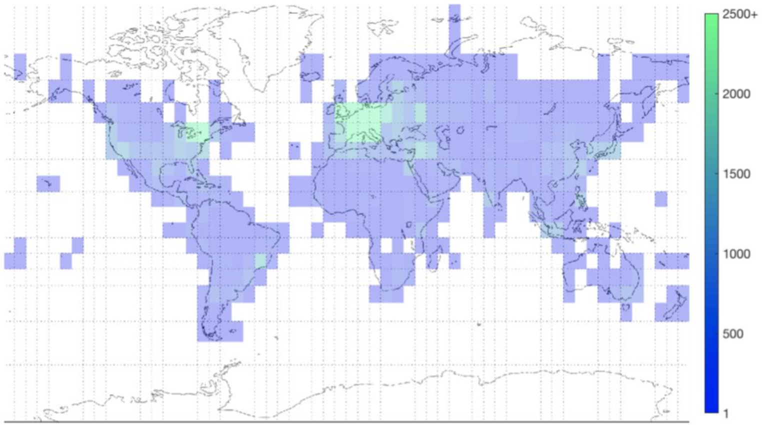
Figure 2: Cosine distance between positive and negative pairs

Target Task - functional Map of the World (fMoW)

- It includes 350k, 50k, 50k samples across 62 classes from the training, validation, and test sets.



Pre-training Dataset (WikiSatNet)

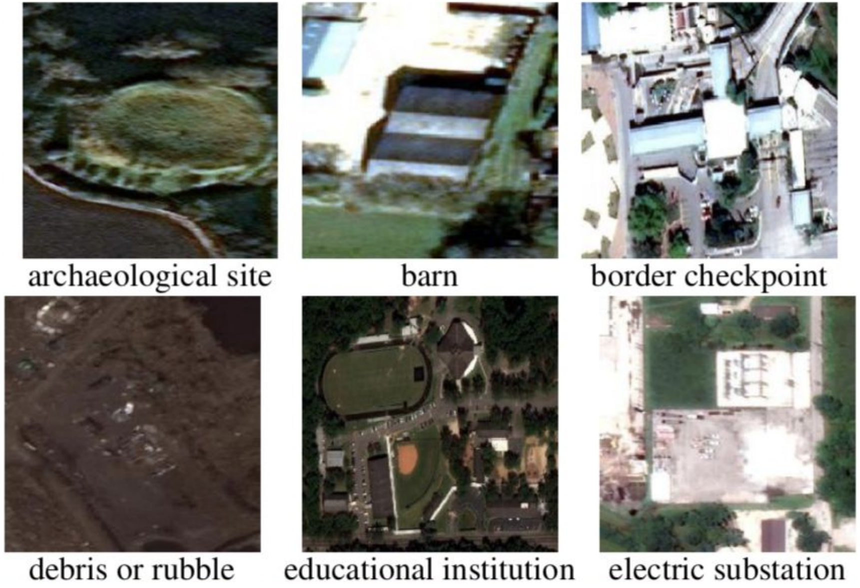


Target Dataset (fMoW)

Christie, Gordon, Neil Fendley, James Wilson, and Ryan Mukherjee. "Functional map of the world." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172-6180. 2018.

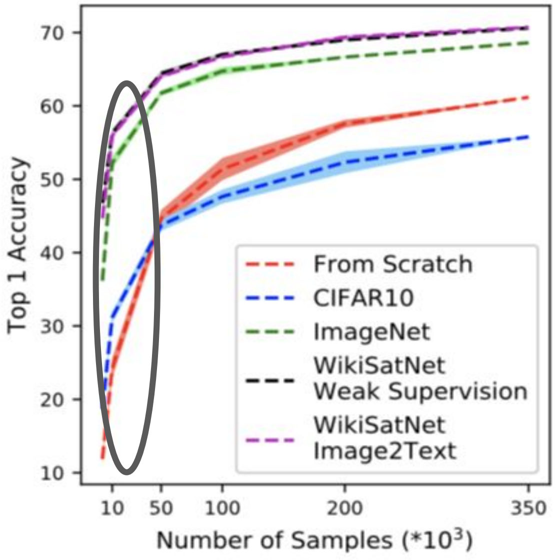
Examples from Target Dataset

Pre-training Dataset (WikiSatNet)



Target Dataset (fMoW)

Image Classification on fMoW



*Gap decreases w.r.t sample complexity

Model	CIFAR10	ImageNet	WikiSatNet *Weak Labels*	WikiSatNet *Image2Text*
F1 Score (*Single View*)	55.34 (%)	64.71 (%)	66.17 (%)	**67.12** (%)
F1 Score (*Temporal Views*)	60.45 (%)	68.73 (%)	71.31 (%)	**73.02** (%)

Table 1: F1 scores of pre-training methods on fMoW's test set.

*We achieve similar accuracy with the *trained from scratch model* when using 10 times less amount of labeled samples.

Geography-Aware Self-Supervised Learning

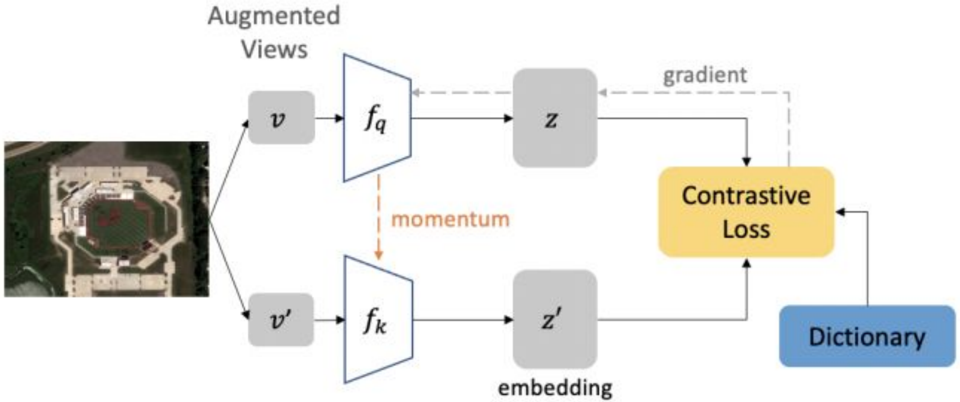
*Kumar Ayush, *Burak Uzkent, *Chenlin Meng, **David Lobell, **Marshall Burke, *Stefano Ermon

*Department of Computer Science, Stanford University

**Department of Earth Science, Stanford University

Unsupervised Learning with Contrastive Loss

- The task is to learn representations without any supervision.
- Unsupervised learning has seen tremendous growth with the contrastive learning.



Contrastive Loss Function

$$L_z = -\log \frac{\exp(z \cdot \hat{z}/\lambda)}{\exp(z \cdot \hat{z}/\lambda) + \sum_{j=1}^{N} \exp(z \cdot k_j/\lambda)}$$

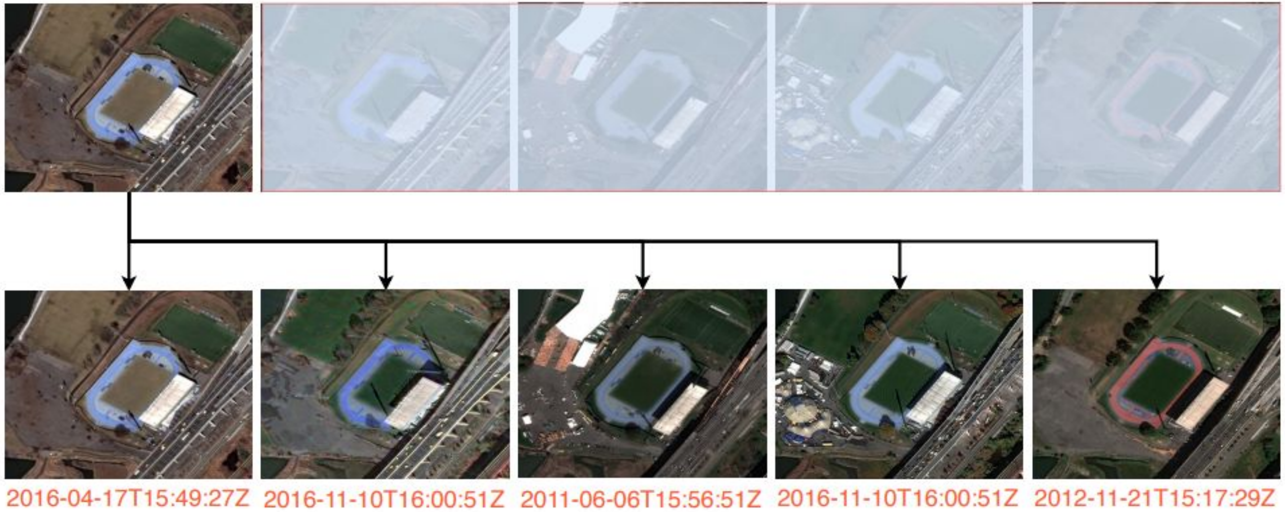
Remote Sensing Images with Metadata

- Remote sensing images come with metadata information which can be used to improve unsupervised learning.

"gsd":	2.10264849663	2.06074237823	1.9968634	2.2158575	1.24525177479	1.4581833	1.2518295
"img_width":	2421	2410	2498	2253	4016	3400	4003
"img_height":	2165	2156	2235	2015	3592	3041	3581
"country_code":	IND	IND	IND	IND	IND	IND	IND
"cloud_cover":	6	0	1	0	0	2	0
“timestamp":	2015-11-02 T05:44:14Z	2016-03-09 T05:25:30Z	2017-02-02 T05:47:02Z	2017-02-27 T05:24:30Z	2015-04-09 T05:36:04Z	2016-12-28 T05:57:06Z	2017-04-12 T05:51:49Z

***Such meta-data for remote sensing images is free and comes with every image.**

Contrastive Learning with Temporal Positives



Contrastive Loss Function

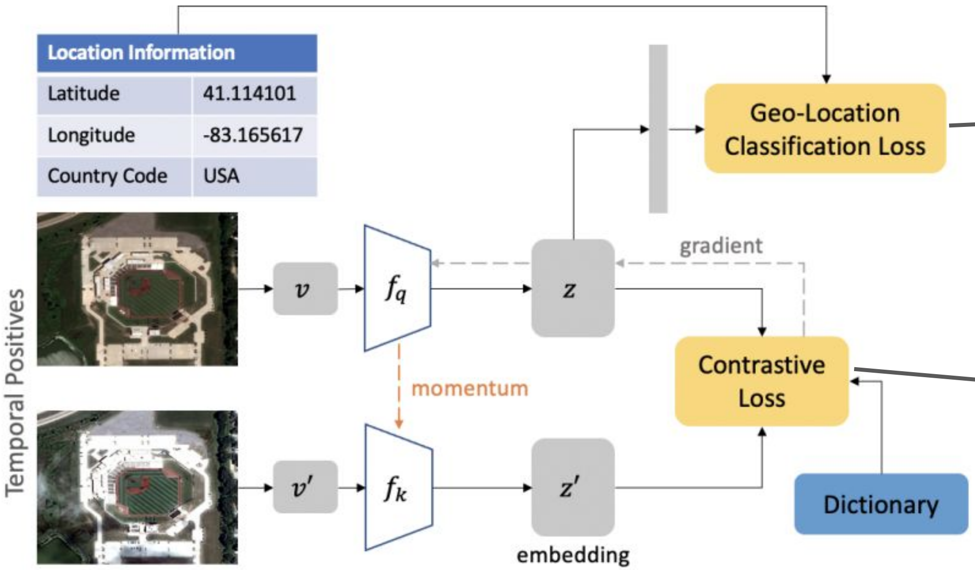
$$L_z = -\log \frac{\exp(z \cdot \hat{z}/\lambda)}{\exp(z \cdot \hat{z}/\lambda) + \sum_{j=1}^{N} \exp(z \cdot k_j/\lambda)}$$

→

Contrastive Loss Function with Temporal Positives

$$L_{z_i^{t_1}} = -\log \frac{\exp(z_i^{t_1} \cdot z_i^{t_2}/\lambda)}{\exp(z_i^{t_1} \cdot z_i^{t_2}/\lambda) + \sum_{j=1}^{N} \exp(z_i^{t_1} \cdot k_j/\lambda)}$$

Incorporating Geo-location Classification



Geo-location Classification Loss

$$L_g = -\sum_{i=1}^{K} p(c_i = k) \log(\hat{p}(c_i = k | f_c(z_i^t))$$

Contrastive Loss Function with Temporal Positives

$$L_{z_i^{t_1}} = -\log \frac{\exp(z_i^{t_1} \cdot z_i^{t_2}/\lambda)}{\exp(z_i^{t_1} \cdot z_i^{t_2}/\lambda) + \sum_{j=1}^{N} \exp(z_i^{t_1} \cdot k_j/\lambda)}$$

Final Loss Function

$$\arg\min_{\theta_q, \theta_k, \theta_c} L_f = \alpha L_{z^{t_1}} + \beta L_g$$

Experiments on fMoW

- The fMoW dataset consists of 350k training and 53k validation images.
- We perform linear probing on the same dataset to evaluate the representations.

	Backbone	Accuracy ↑ (100 Epochs)	Accuracy ↑ (200 Epochs)
Sup. Learning*	ResNet50	69.05	69.05
Geoloc. Learning*	ResNet50	52.40	52.40
MoCo-V2	ResNet50	58.32	60.69
MoCo-V2+Geo	ResNet50	63.65	64.07
MoCo-V2+TP	ResNet50	**67.15**	**68.32**
MoCo-V2+Geo+TP	ResNet50	65.77	66.33

THANKS!
ANY QUESTIONS?