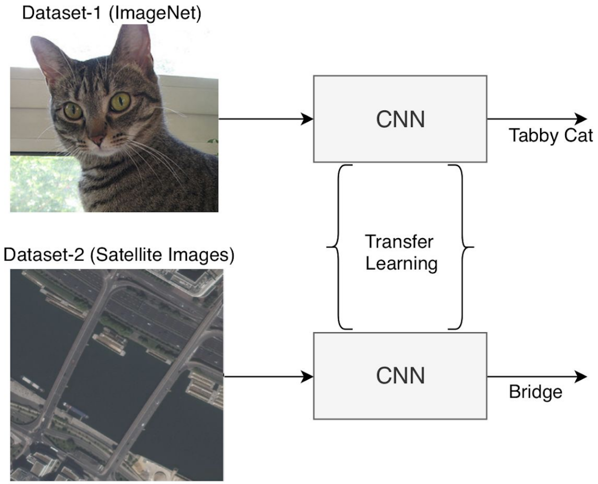
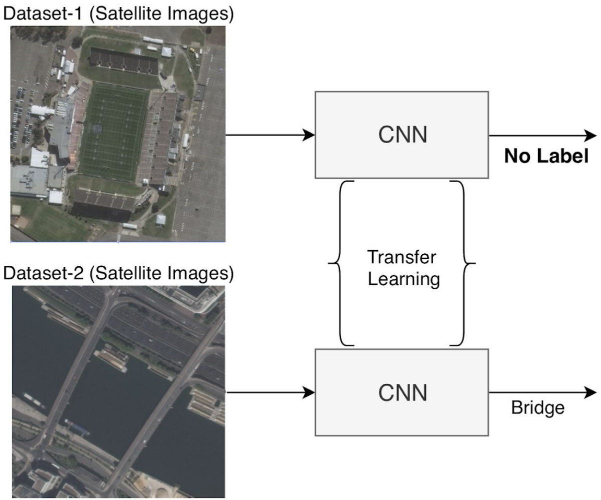
Exploring Large-Scale Pre-training for Satellite Images

Introduction

- Almost all of the state-of-the-art deep learning models rely on the following framework.
  - *Pre-train on ImageNet or another human labeled dataset.*
  - *Fine-tune on the target task.*





Learning from Instagram Images with Hashtags

- Mahajan et al. builds an image recognition dataset consisting of 3 billion images from Instagram.
- They label the images using the hashtags given by the users.
- Two sets of labels are used:
  - *ImageNet labels (1k)*
  - *WordNet synsets (17k)*
- Pre-training improves the recognition accuracy in the target task by *%5*.

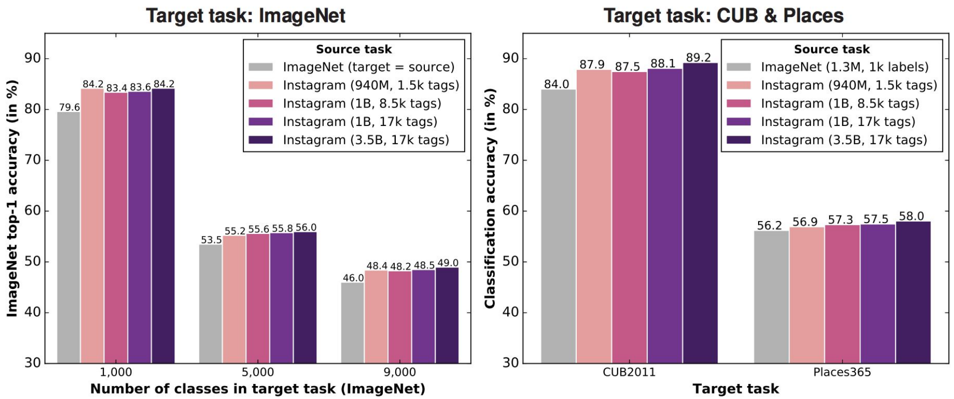
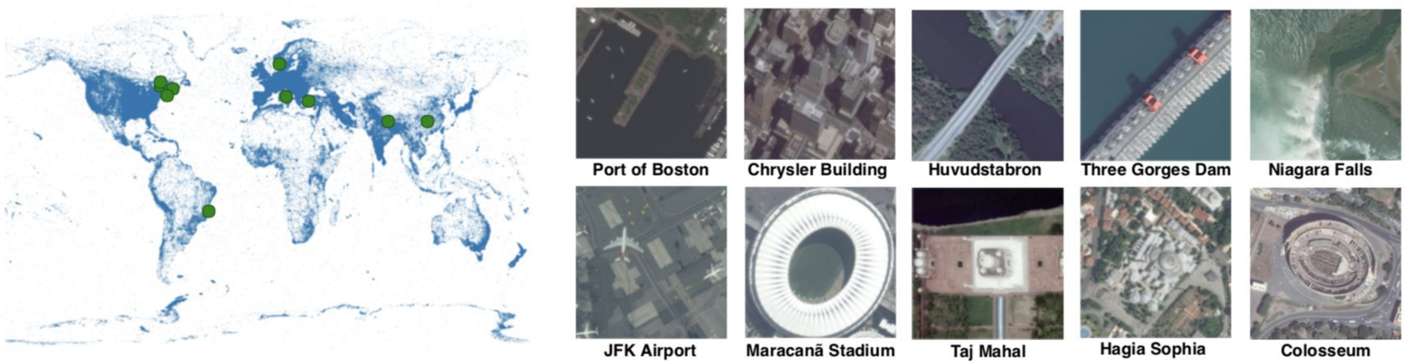


Fig. 1: Classification accuracy of ResNeXt-101 32×16d pretrained on IG-1B with different hashtag vocabularies (purple bars) on IN-{1k, 5k, 9k} (left) and CUB2011, Places365 (right). Baseline models (gray bars) are trained on IN-{1k, 5k, 9k} (left) and IN-1k (right), respectively. Full network finetuning is used. Higher is better.

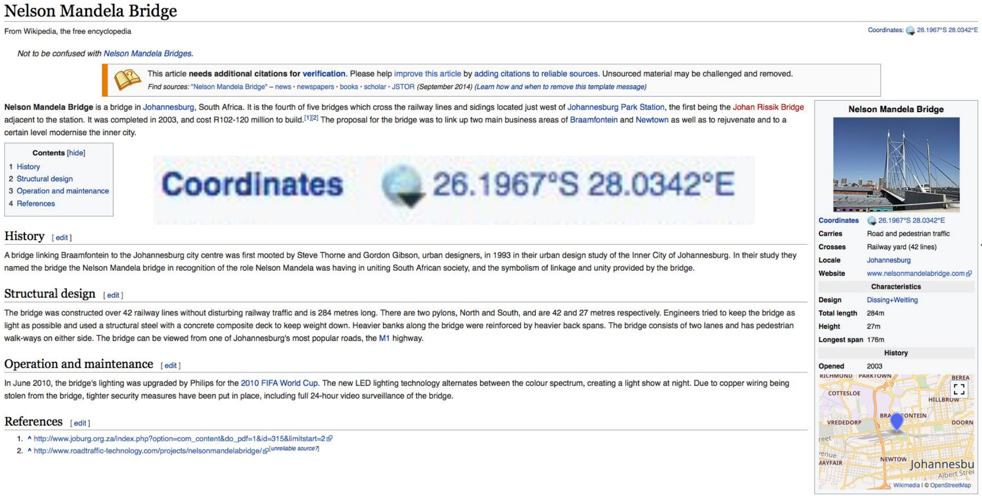
Learning from Satellite Images using Wikipedia Articles

- In its most recent 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 Images



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

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

RICHMOND PARKTOWN BEREA COTTESLOE HILLBROW VREDEDORP DOORN NEWTOW MAYFAIR Johannesbu Albert Stree
Wikimedia | © OpenStreetMap

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?]

Collect a high resolution image

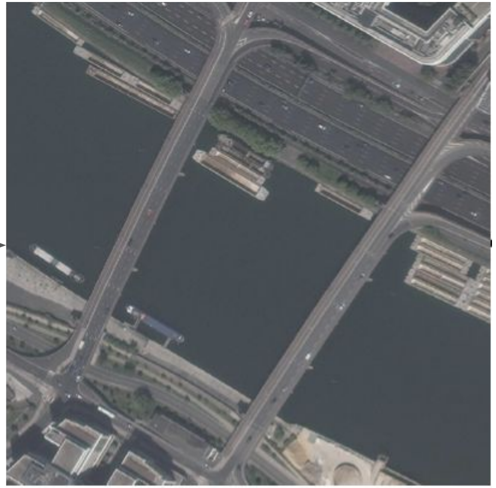
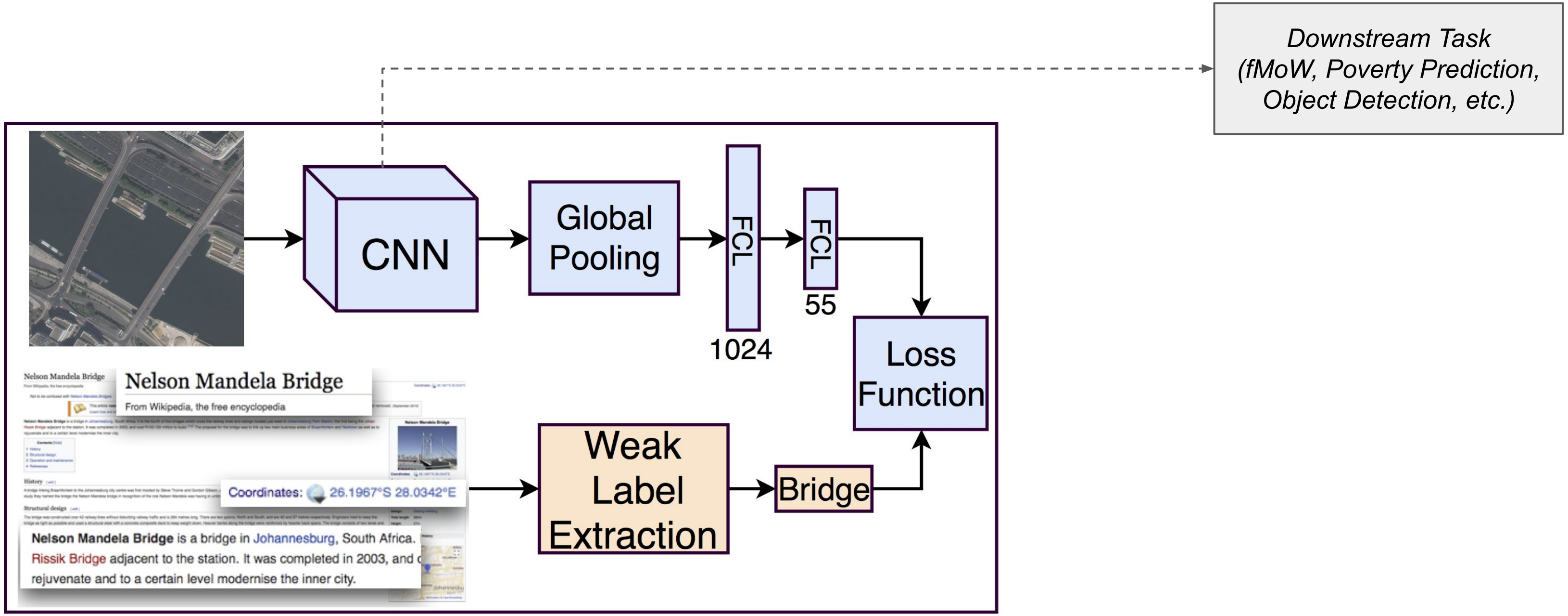


Image Collection

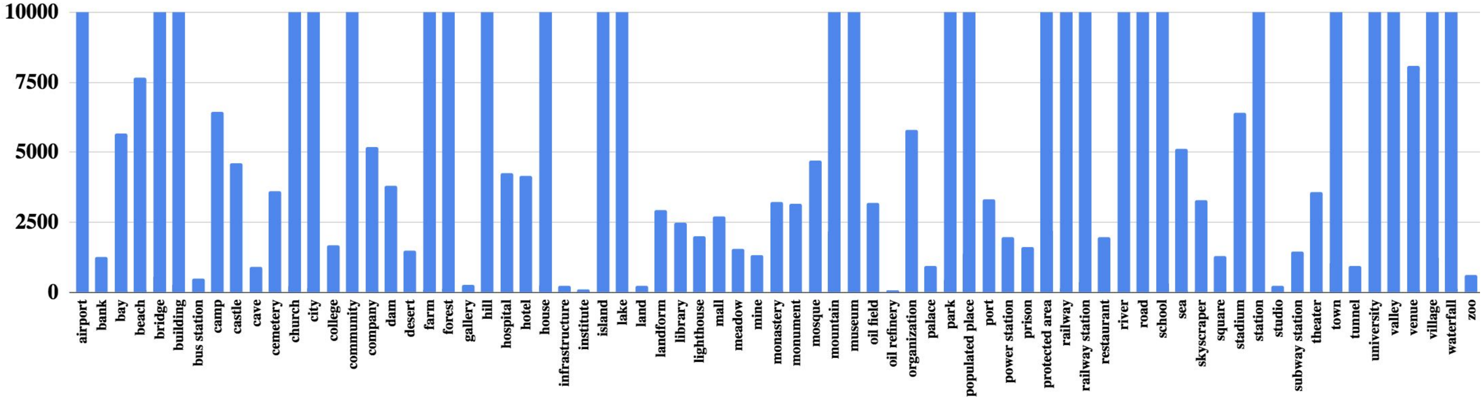
- We collect high resolution images from about *900k* coordinates worldwide.
- Images come from DigitalGlobe satellites and no filtering is applied to remove cloudy images.
- Grayscale images are kept and converted to RGB to add into our dataset.

Representation Learning using Weak Supervision



Post-processing the Weak Labels

- After the labeling step, we obtain labels from ***98 fine-level classes***.
- However, some labels such as ***culture, battle, event*** do not convey any visual information.
- Additionally, we remove labels that are represented by less than 100 samples, resulting in ***55 remaining labels***.



Flipped and Adversarial Label Noise

- Our crude method for labeling articles results in large amount of ***flipped*** and ***adversarial*** label noise.

Extracted Weak Label -> *‘School’*

Extracted Weak Label -> *‘Incident’*

Extracted Weak Label -> *County*

Extracted Weak Label -> *Town*

Representation Learning with Image to Text Matching

- Our crude method for labeling articles results in large amount of ***flipped*** and ***adversarial*** label noise.
- It is time-consuming and requires post-processing steps to reduce the label noise and handle ***class imbalance*** problem.
    - Merging labels results in class imbalance problem whereas not merging leads to large label noise.
- *Can we find a better way to learn representations using multi-modal data without even extracting the weak labels?*
    - ***Image to Text Matching***

Image to Text Matching (Wang et al. PAMI19)

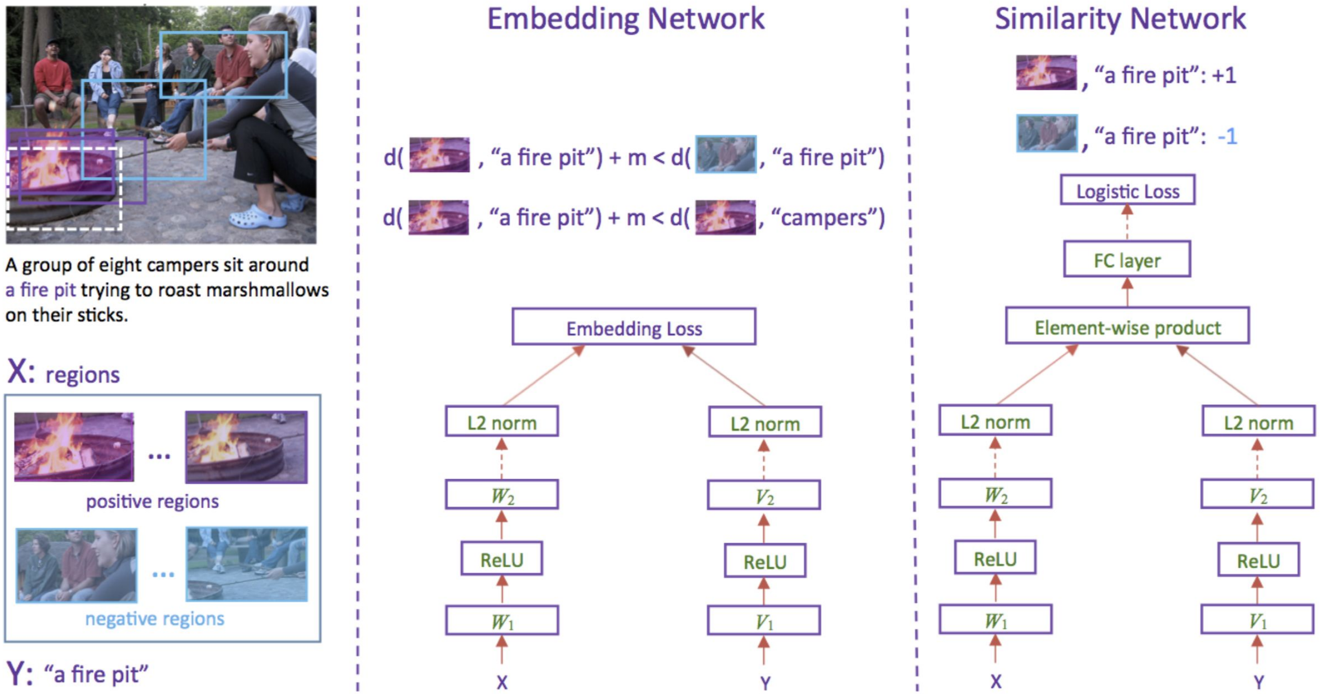
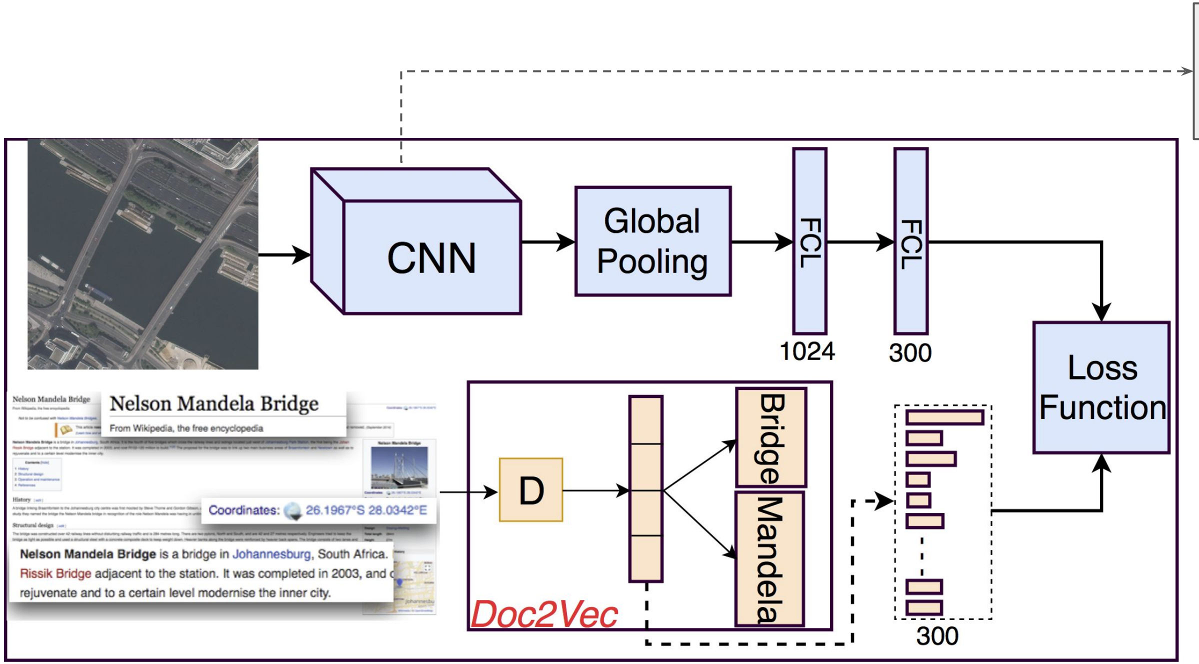
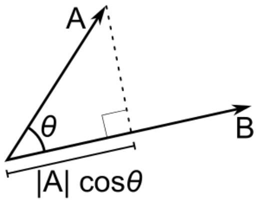


Image to Text Matching for Unsupervised Learning



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$



Flipped Label Noise

Extracted Weak Label -> *'INCIDENT'*

Iserbrook (ship)

Iserbrook was a general cargo and passenger brig built in 1853 at Hamburg (Germany) for *Joh. Ces. Godeffroy & Sohn*. It spent over twenty years as an immigrant and general cargo vessel, transporting passengers from Hamburg to South Africa, Australia and Chile, as well as servicing its owner's business in the Pacific. Later on, the vessel came into Australian possession and continued sailing for the Pacific trade. In 1878 it caught fire and was sunk the same year. At last, it was re-floated and used as a transport barge and hulk in Sydney until it sunk again and finally was blown up.

Construction and Description

The vessel was built for the Hamburg trading company *Joh. Ces. Godeffroy & Sohn*. At the time, the enterprise was operated by Johan César VI. Godeffroy who had large trading interests in the Pacific, focussing mainly on Copra, Coconut oil and luxuries like pearlshell. In the 1850s and 60s, the company was also strongly associated with emigration from Germany to Australia, especially to Adelaide and Brisbane.

The 240 ton Brig *Cesar & Helene* was built in 1855/56 in the Godeffroy shipyard at the Reiherstieg wharf. This vessel was just 30 tones larger and built one year after the *Iserbrook* for the same owners

In its original Hamburg registration (Bielbrief),

*The word "*Water*" is mentioned 10 times in the article.
*The word "*Sea*" is mentioned 11 times in the article
*The word "*Port*" is mentioned 11 times in the article

Flipped Label Noise

Extracted Weak Label -> *Event*

North Queensland Cowboys

The **North Queensland Cowboys** (Also known as the **North Queensland Toyota Cowboys** for sponsorship reasons) are an Australian professional rugby league football club based in Townsville, the largest city in North Queensland. They compete in Australia's premier rugby league competition, the National Rugby League (NRL) premiership. Since their foundation in 1995, the club has appeared in three grand finals (2005, 2015 and 2017) winning in 2015, and has reached the finals ten times. The team's management headquarters and home ground, the Willows Sports Complex, currently known as *1300SMILES Stadium* due to sponsorship rights, are located in the Townsville suburb of Kirwan.

The Cowboys were admitted to the premiership for the 1995 ARL season. They played in the breakaway Super League competition in 1997 before continuing to

North Queensland Cowboys

Club information	
Full name	North Queensland Cowboys Rugby League Football Club
Nickname(s)	Cowboys
Colours	*Primary:* Navy Grey *Secondary:* Yellow White
Founded	30 November 1992
Website	cowboys.com.au
Current details	
Ground(s)	Willows Sports Complex (*1300SMILES Stadium*) Townsville, Queensland (26,500)
CEO	Jeff Reibel (acting)
Coach	Paul Green

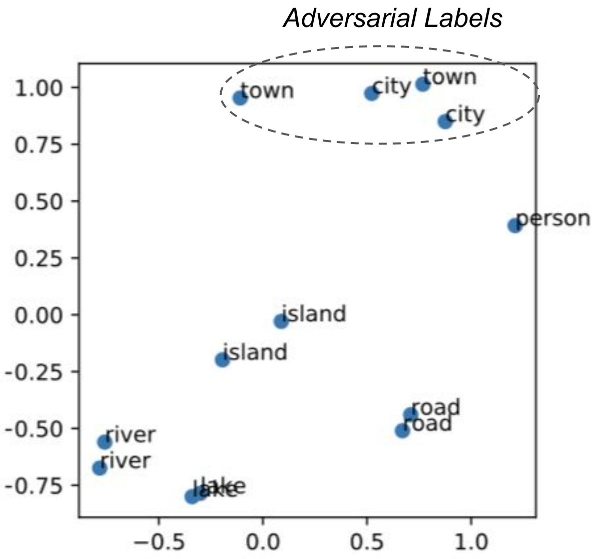
*The word "*Stadium*" is mentioned 19 times in the article.

Adversarial Label Noise



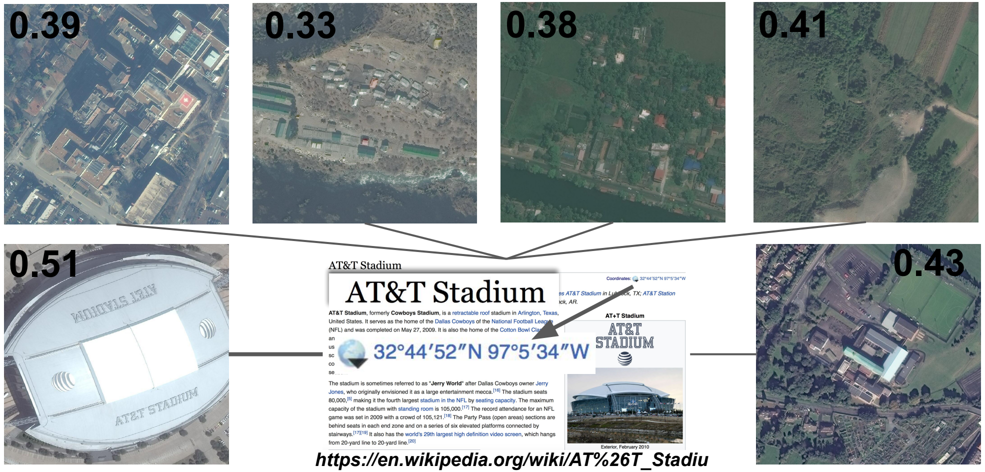

- A big part of the Wikipedia dataset consist of images that are not visually different but labeled into different categories such as *city, country, populated place*.
- Labeling satellite images are already difficult for humans. Doing crude labeling using the articles introduces large amount of *adversarial label noise*.
- *Image to text matching* method basically softens the loss function that penalizes the network.

Reducing Adversarial Label Noise using Image2Text Matching



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

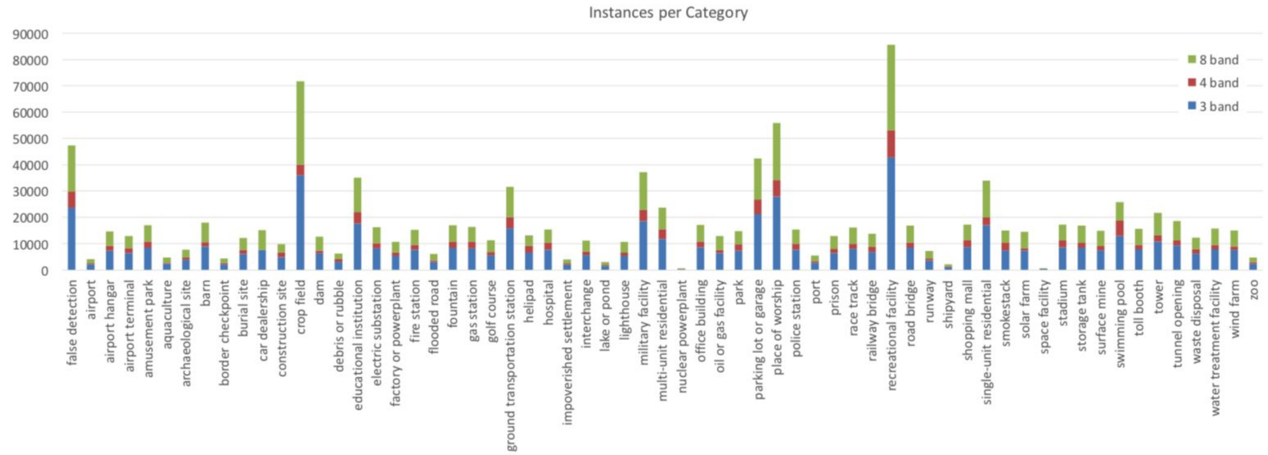
What is CNN Learning with Image2Text Matching?



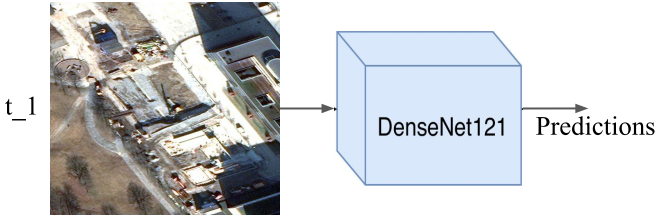
https://en.wikipedia.org/wiki/AT%26T_Stadium

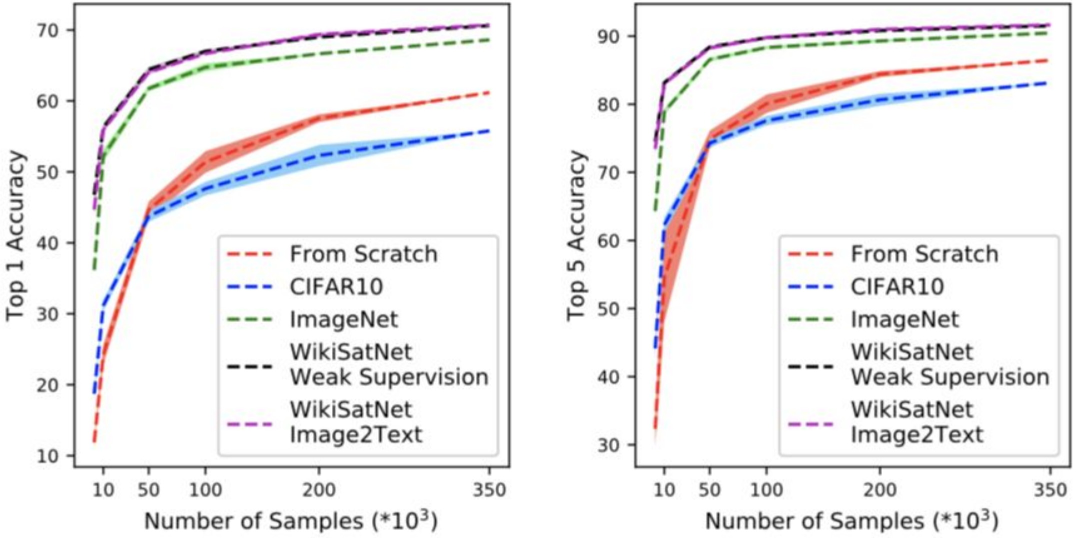
Target Task- fMoW

- We use the recently released functional map of the world (fMoW) dataset consisting high resolution DigitalGlobe images.
- It includes 83k, 15k, and 15k unique bounding boxes across 62 classes from the training, validation, and test sets.
- It also provides temporal views from each area.



Single View Reasoning on fMoW

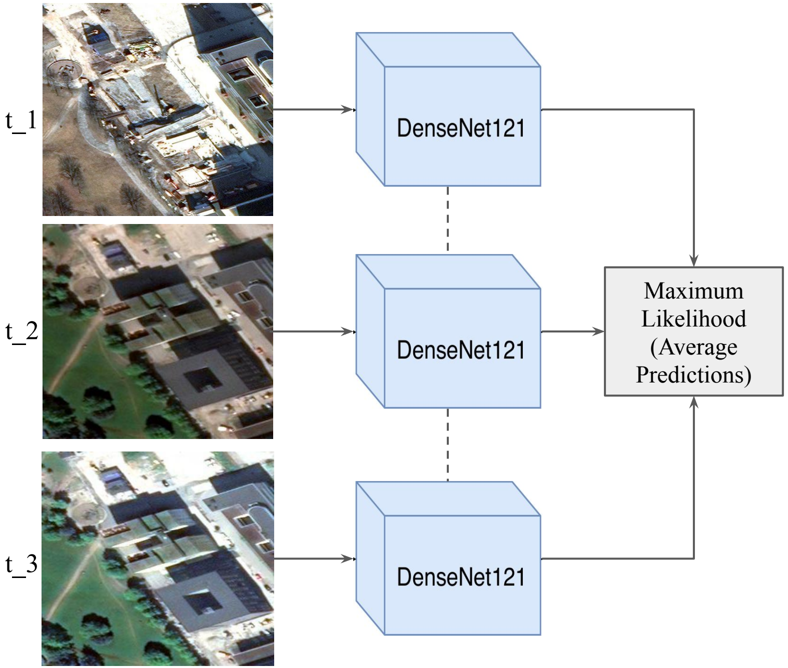




Gap decreases

Gap decreases

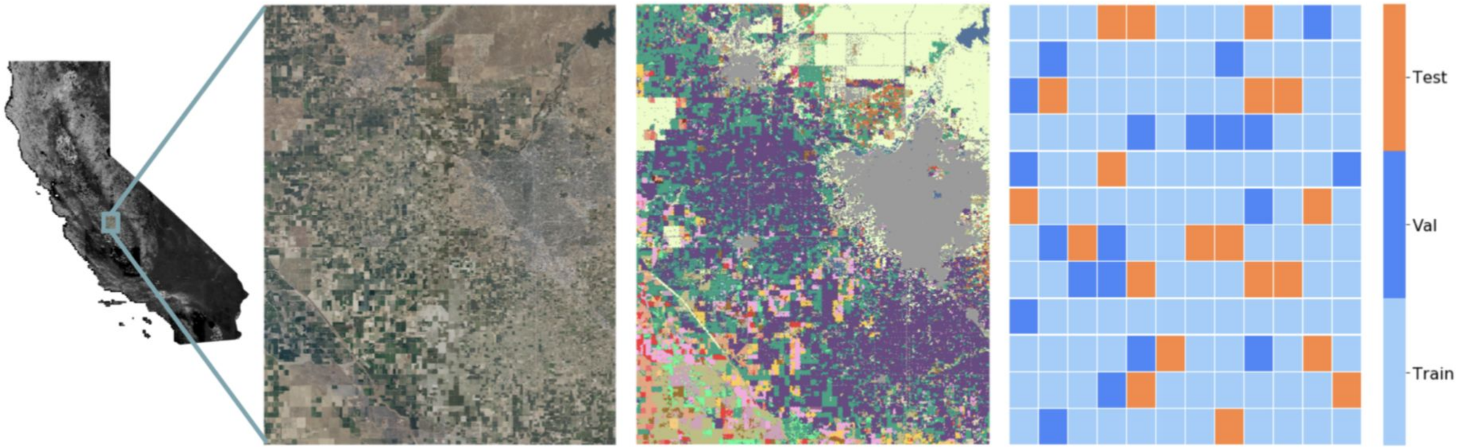
Temporal Reasoning on fMoW



<u>Full training set-350k samples</u>

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** (%)

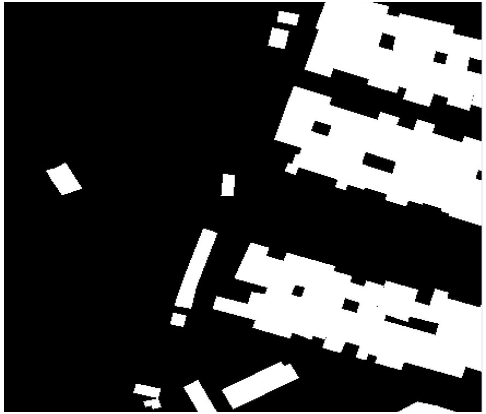
Target Task-Land Cover Classification



Model	CIFAR10	ImageNet	WikiSatNet *Weak Labels*	WikiSatNet *Image2Text*
Top 1 Acc.	42.01 (%)	40.11 (%)	46.16 (%)	**47.65** (%)
Top 5 Acc.	74.73 (%)	80.15 (%)	88.66 (%)	**88.77** (%)

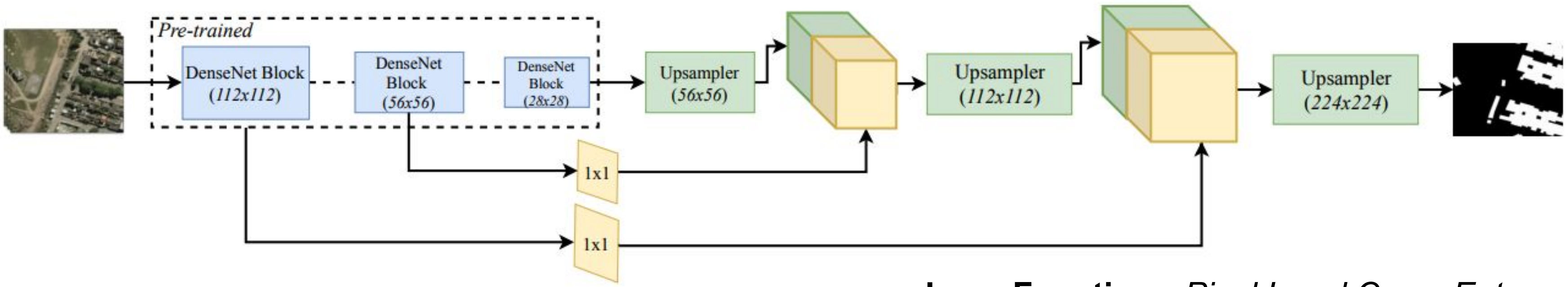
Target Task-Semantic Segmentation

- To quantify the learned representations on a different task, we use the SpaceNet Semantic Segmentation dataset.

- Overall, there are ***5000*** and ***2000*** training and test images from the RIO region for *building* class.

Architecture



Loss Function : *Pixel Level Cross Entropy*

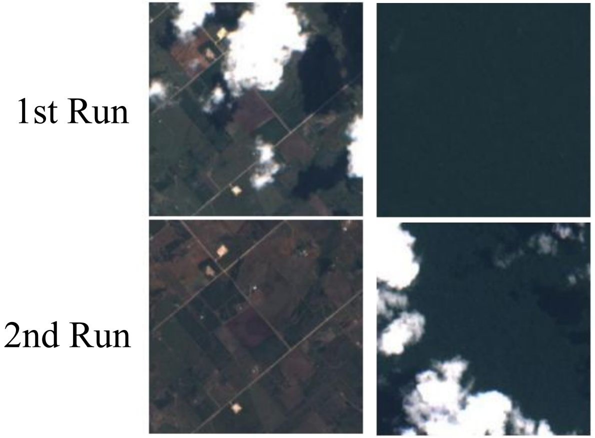
Model	From Scratch	ImageNet	WikiSatNet *Image2Text*
200 Samples	42.11 (%)	50.75 (%)	**51.70** (%)
500 Samples	48.98 (%)	54.63 (%)	**55.41** (%)
5000 Samples	57.21 (%)	59.63 (%)	**59.74** (%)

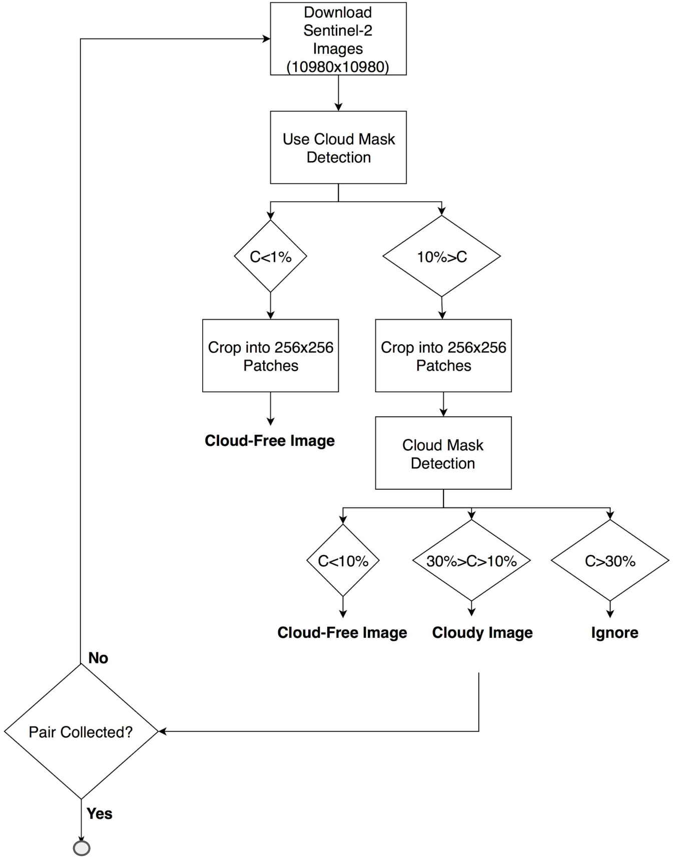
Cloud-Free Image Generation using Spatiotemporal Generative Networks

Introduction

- Clouds dominate satellite images as they can sometimes completely occlude the region.
- Mostly, when analyzing satellite images we simply generate cloud masks of the image, and discard the image.
- On the other hand, processing cloudy images with computer vision models can lead to wrong ground information collection.
- In this study, we propose a *Generative Adversarial Network* to generate cloud-free image conditioned on the cloudy images.

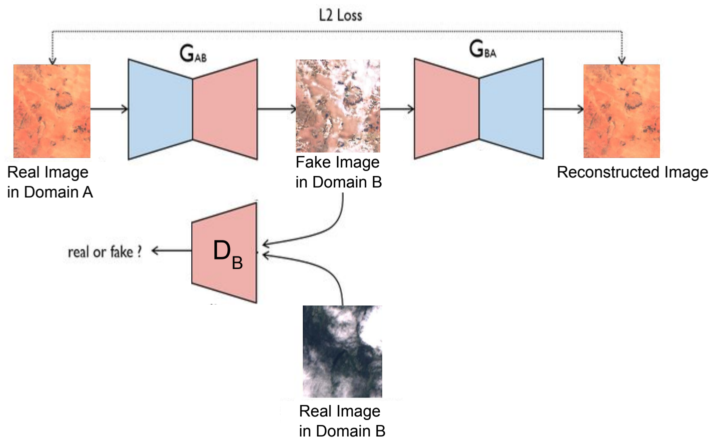
Framework to Build Paired Dataset

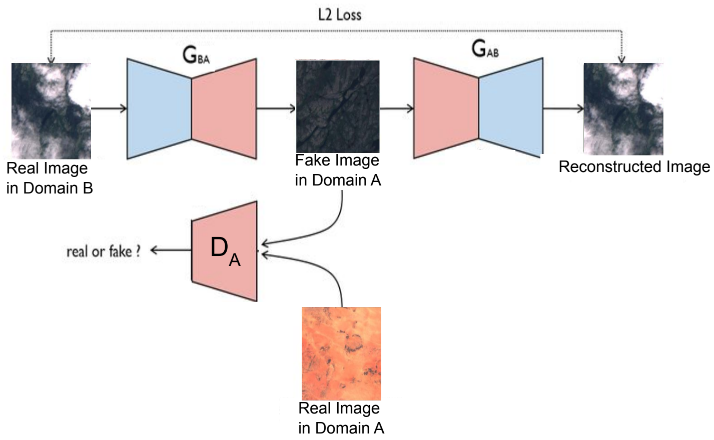




Building Paired Dataset using CycleGan

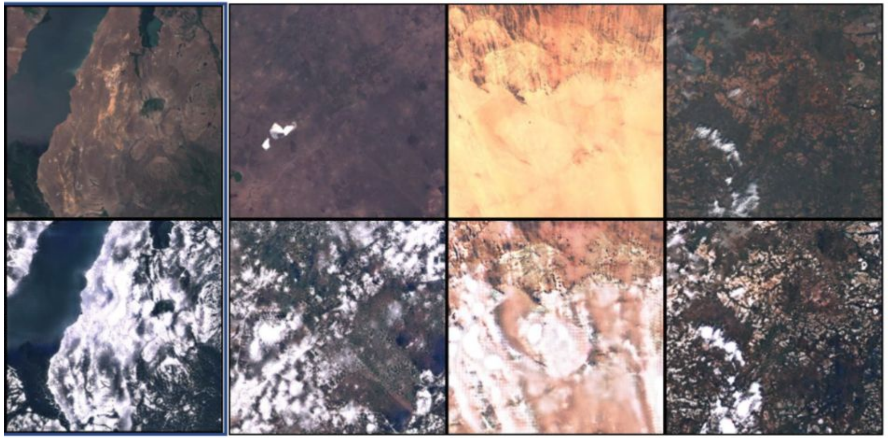
- At the end of first iteration, we collect *97640* cloudy or cloud-free image from a point at time t.
- We can use *CycleGan* to generate cloudy image given cloud-free image, and vice versa.



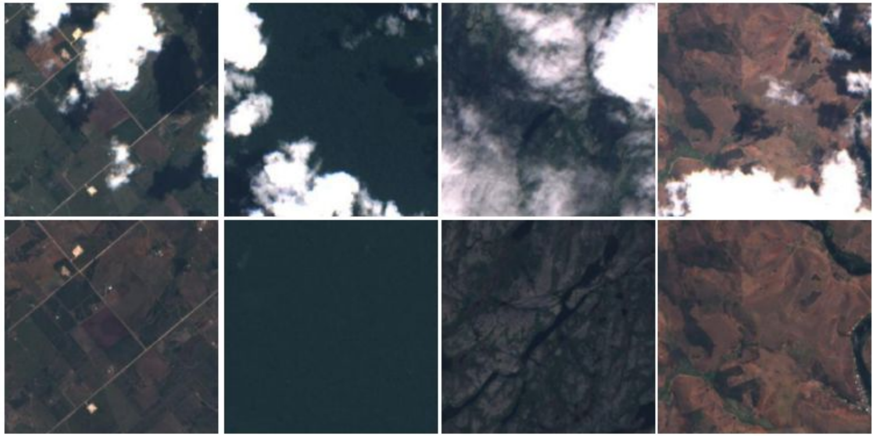


Visual Examples

CycleGan generated pairs

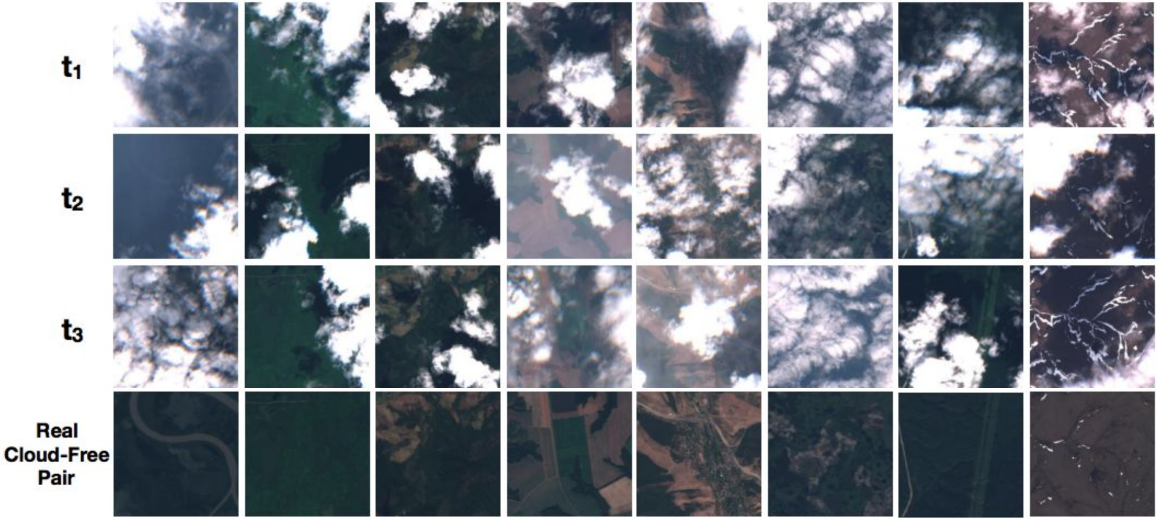


Pairs from real dataset

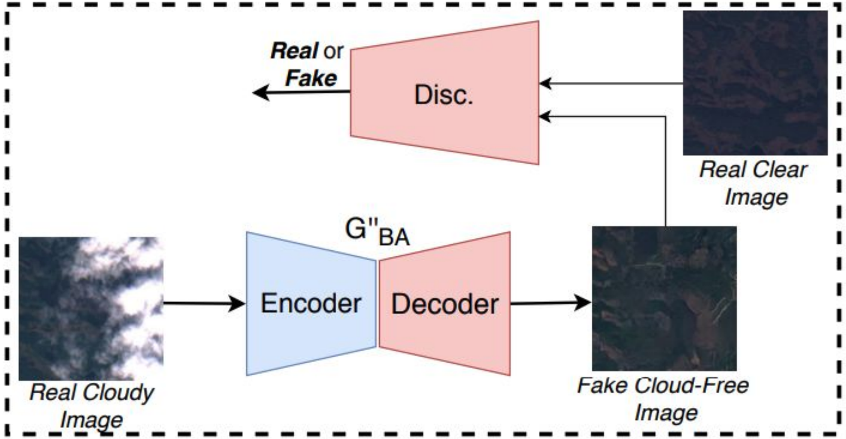


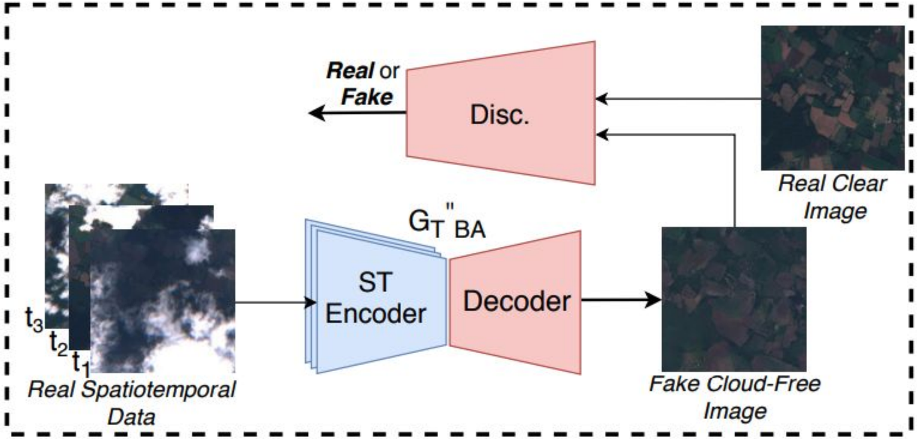
Collecting Spatiotemporal Dataset

- To build a spatiotemporal dataset, we simply collect images from the same points at the previous time periods until we find *three cloudy* and *one cloud-free image* from the same area.

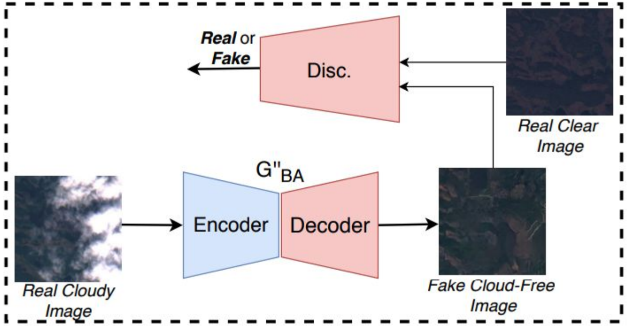


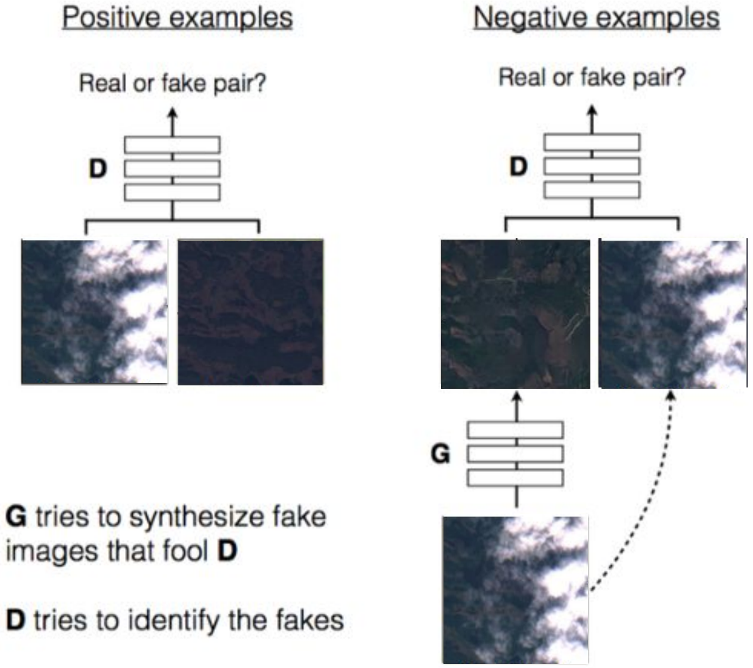
Spatial-only and Spatiotemporal Methods



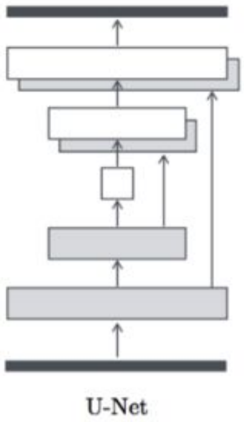


Pix2Pix for Paired Spatial-only Dataset



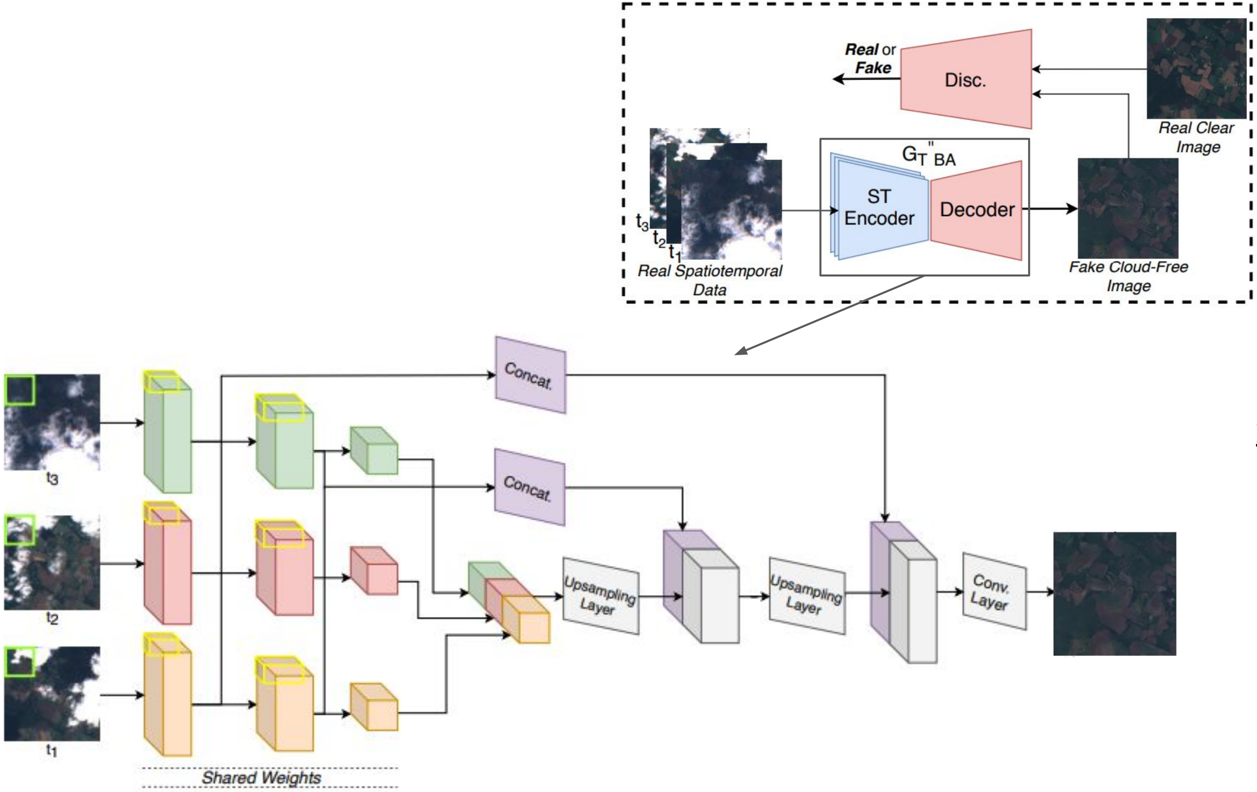


Discriminator



Generator

Spatiotemporal GANs - (STGAN-Branched U-Net)



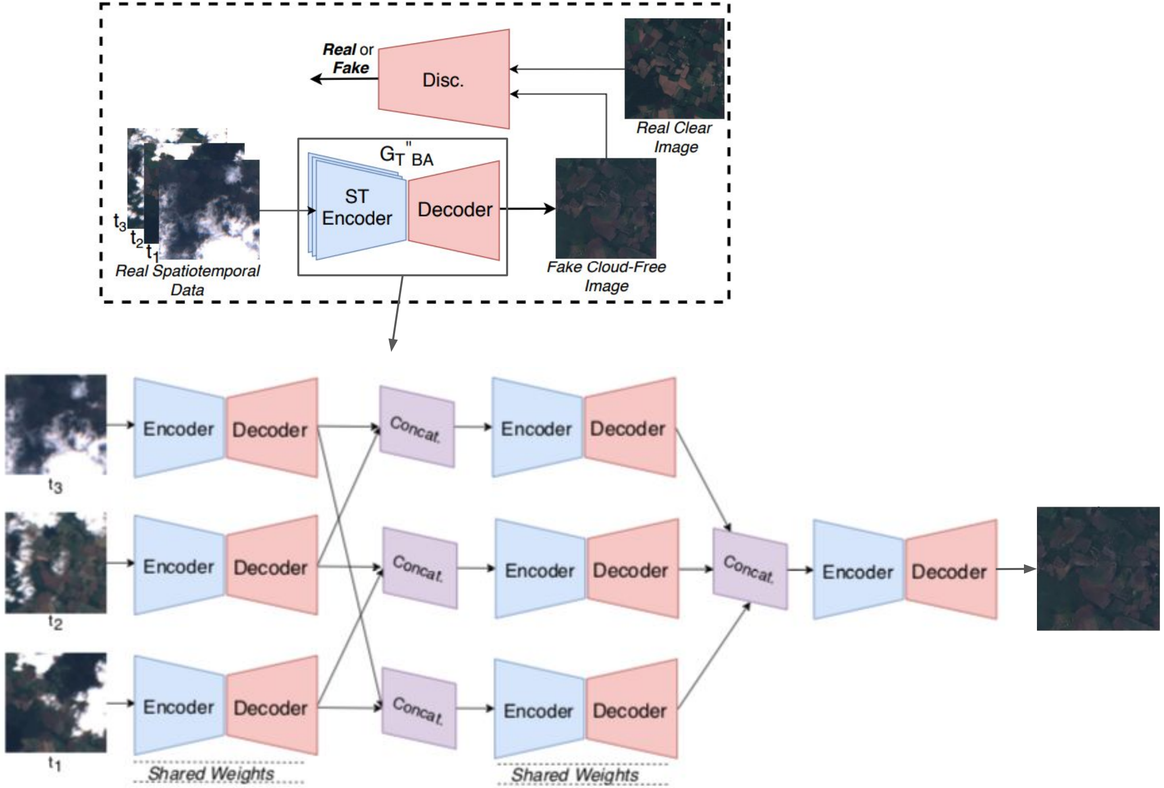
Loss Function :

$$\mathcal{L}_{cGAN}(G^t, D^s) = \mathbb{E}_{x^n,y}[logD^s(x^n, y)] + \mathbb{E}_{x_c^t,z}[log(1 - D^s(x^n, G^t(x_c^t, z))]$$

$$\mathcal{L}_{L1}(G^t) = \mathbb{E}_{x_c^t,y,z}[||y - G^t(x_t^c, z)||_1]$$

$$G^{t*} = arg \min_{D^s} \max_{G^t} \mathcal{L}_{cGAN}(G^t, D^s) + \lambda^t \mathcal{L}_{L1}(G^t)$$

Spatiotemporal GANs - (STGAN-Branched ResNet)



Loss Function :

$$\mathcal{L}_{cGAN}(G^t, D^s) = \mathbb{E}_{x^n,y}[logD^s(x^n, y)] + \mathbb{E}_{x_c^t,z}[log(1 - D^s(x^n, G^t(x_c^t, z))]$$

$$\mathcal{L}_{L1}(G^t) = \mathbb{E}_{x_c^t,y,z}[||y - G^t(x_t^c, z)||_1]$$

$$G^{t*} = arg \min_{D^s} \max_{G^t} \mathcal{L}_{cGAN}(G^t, D^s) + \lambda^t \mathcal{L}_{L1}(G^t)$$

The architecture of Encoder and Decoder

Encoder	***Decoder***
1x(3x3 - 2) - 64C	1x(3x3 - 2) - 128C
1x(3x3 - 2) - 128C	1x(3x3 - 2) - 64C
9x(Residual Layers) - 512C	

Results

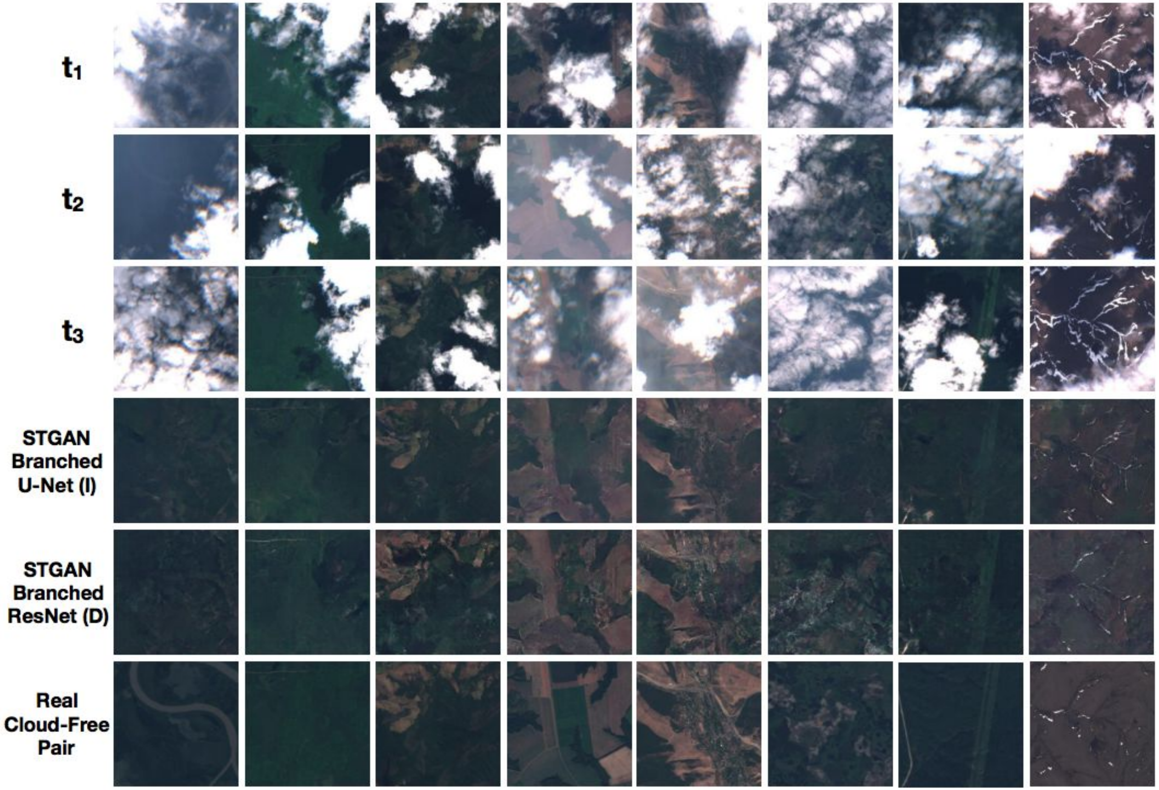
Results on Spatial-only Dataset

	Validation Set		Test Set	
Models	PSNR	SSIM	PSNR	SSIM
Pix2Pix (Real Pairs)	**23.130**	**0.442**	**22.894**	**0.437**
Pix2Pix (Synthetic Pairs)	21.067	0.4342	20.886	0.429
Cloudy Images (Unprocessed)	8.742	0.396	8.778	0.398

Results on Spatiotemporal Dataset

	Validation Set		Test Set	
Models	PSNR	SSIM	PSNR	SSIM
Pix2Pix (Real Pairs)	23.130	0.442	22.894	0.437
Mean Filter	16.962	0.174	16.893	0.173
Median Filter	9.081	0.357	9.674	0.395
STGAN-Stacked U-Net	24.923	0.526	25.163	0.538
STGAN-Stacked ResNet	24.261	0.497	24.771	0.520
STGAN-Branched U-Net (D)	25.879	0.502	26.150	0.533
STGAN-Branched ResNet (D)	25.519	**0.550**	26.000	**0.573**
STGAN-Branched U-Net (I)	25.484	0.534	25.822	0.564
STGAN-Branched ResNet (I)	**26.373**	0.475	**26.940**	0.496
Cloudy Images (Unprocessed)	7.926	0.389	8.289	0.422

Visual Results



PatchDrop: Dynamic Image Masking using Reinforcement Learning

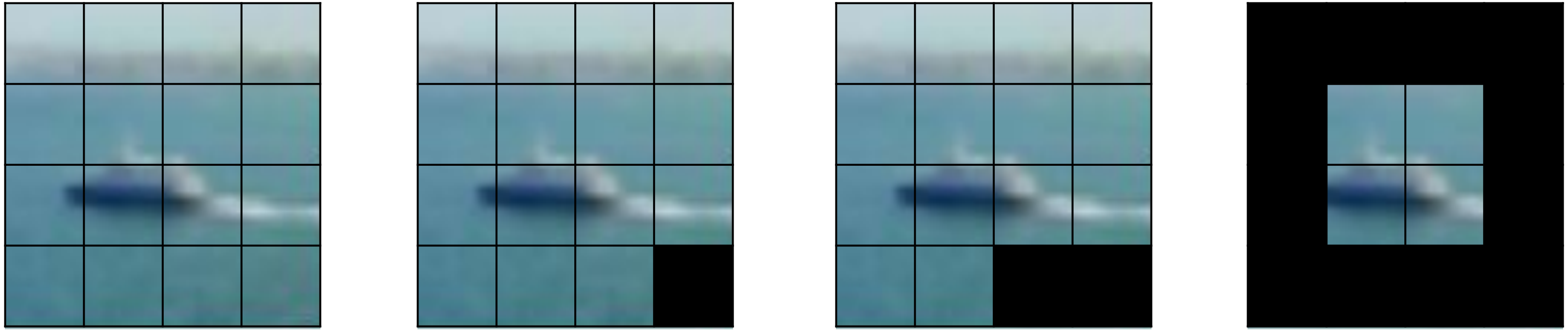
Motivation

Low Resolution Image

High Resolution Image - Patches only Sampled from Semantically Meaningful Points

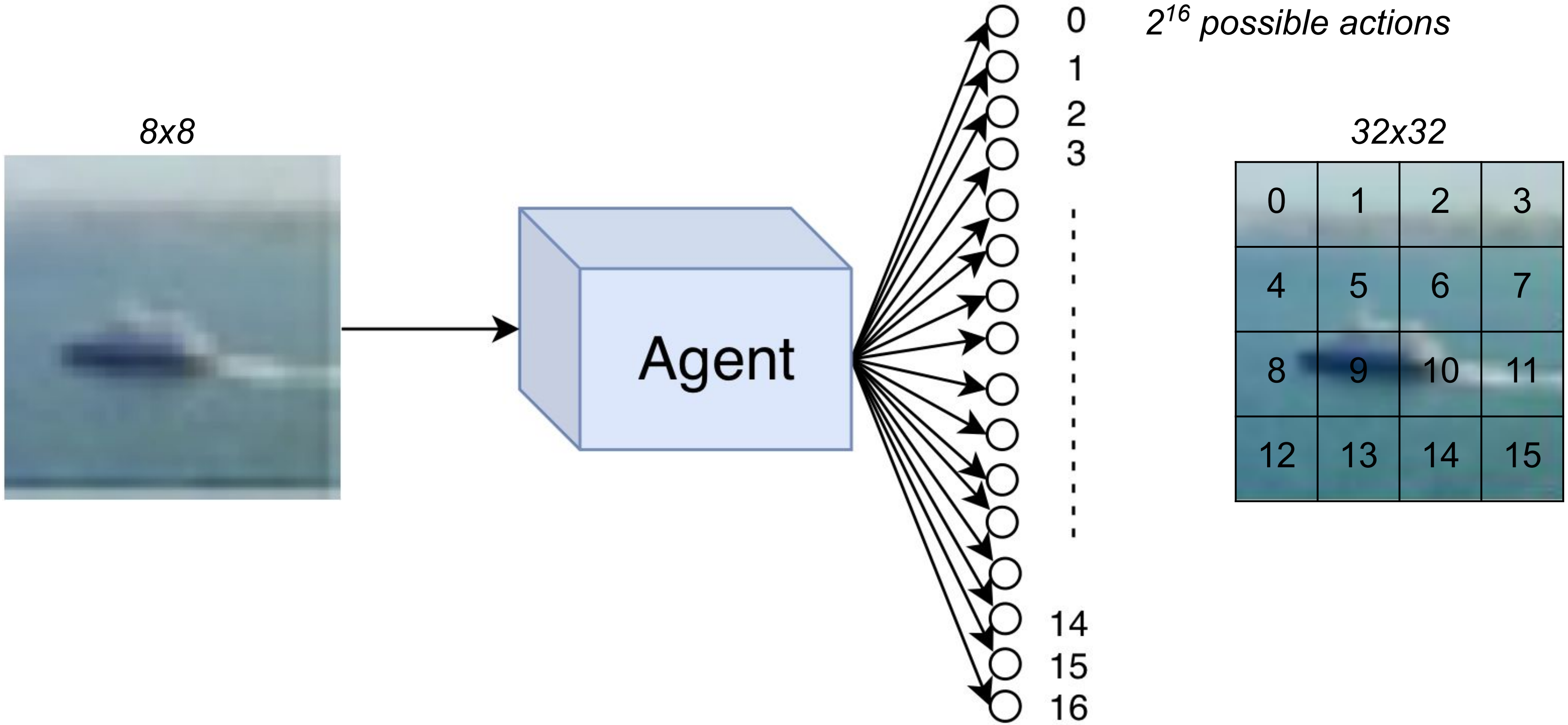
PatchDrop - An Adaptive Patch Sampling Framework

Do we need all the patches in an image to infer correct decisions?

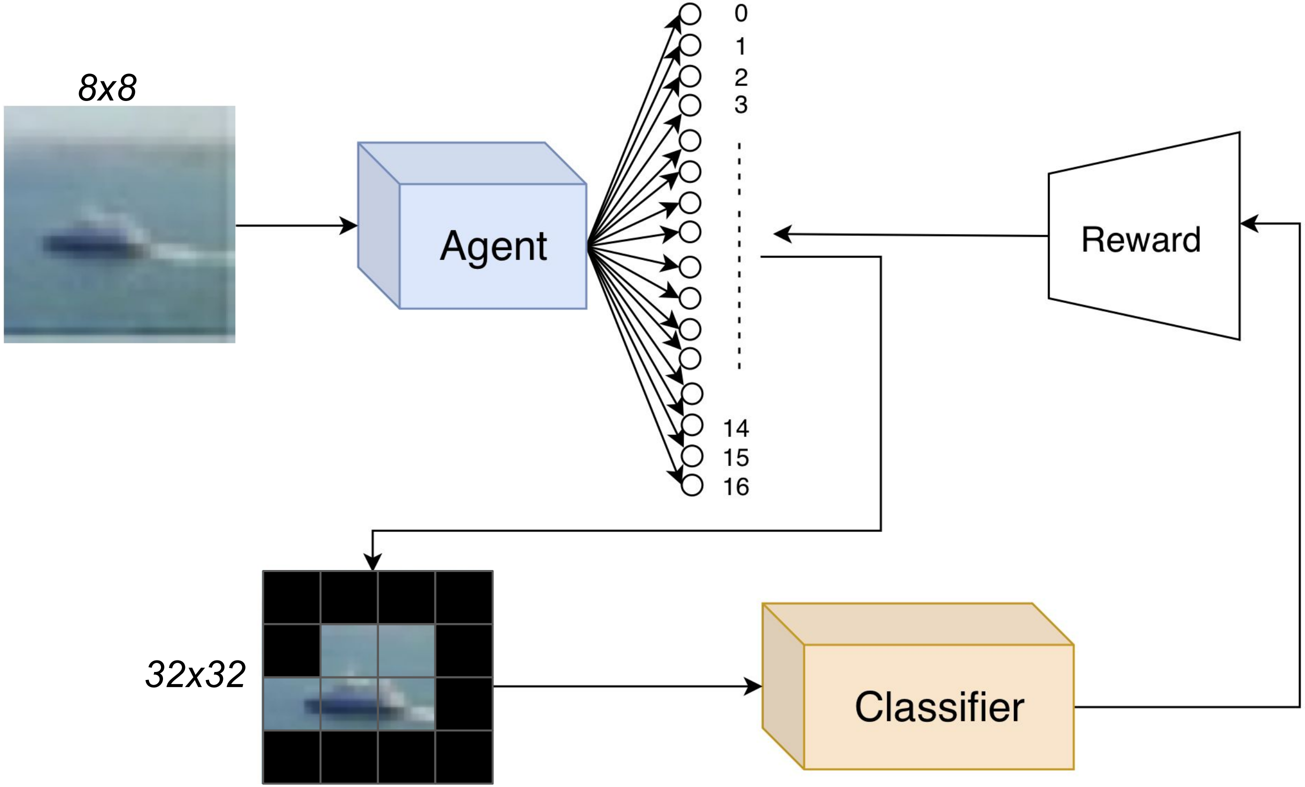


Can we design a conditional patch dropping strategy?

Modeling the Agent



PatchDrop



Modeling the Agent and Reward Function

- The agent is trained using the predictions from the classification model.

$$R(u) = \begin{cases} 1 - (\frac{|u|_1}{P})^2 & \text{if } y = y^* \\ -\sigma & \text{Otherwise} \end{cases} \tag{1}$$

$$\pi(u|x,\theta) = \prod_{p=1}^{P} s_p^{u_p}(1 - s_p^{1-u_p}) \tag{2}$$

$$J = E_{u\sim\pi_w}[R(u)] \tag{3}$$

$$\nabla_w J = E[R(u)\nabla_w log\pi_w(u|x)] \tag{4}$$

$$\nabla_w J = E[R(u)\nabla_w log \prod_{p=1}^{P} s_p^{u_p}(1 - s_p^{1-u_p})] \tag{5}$$

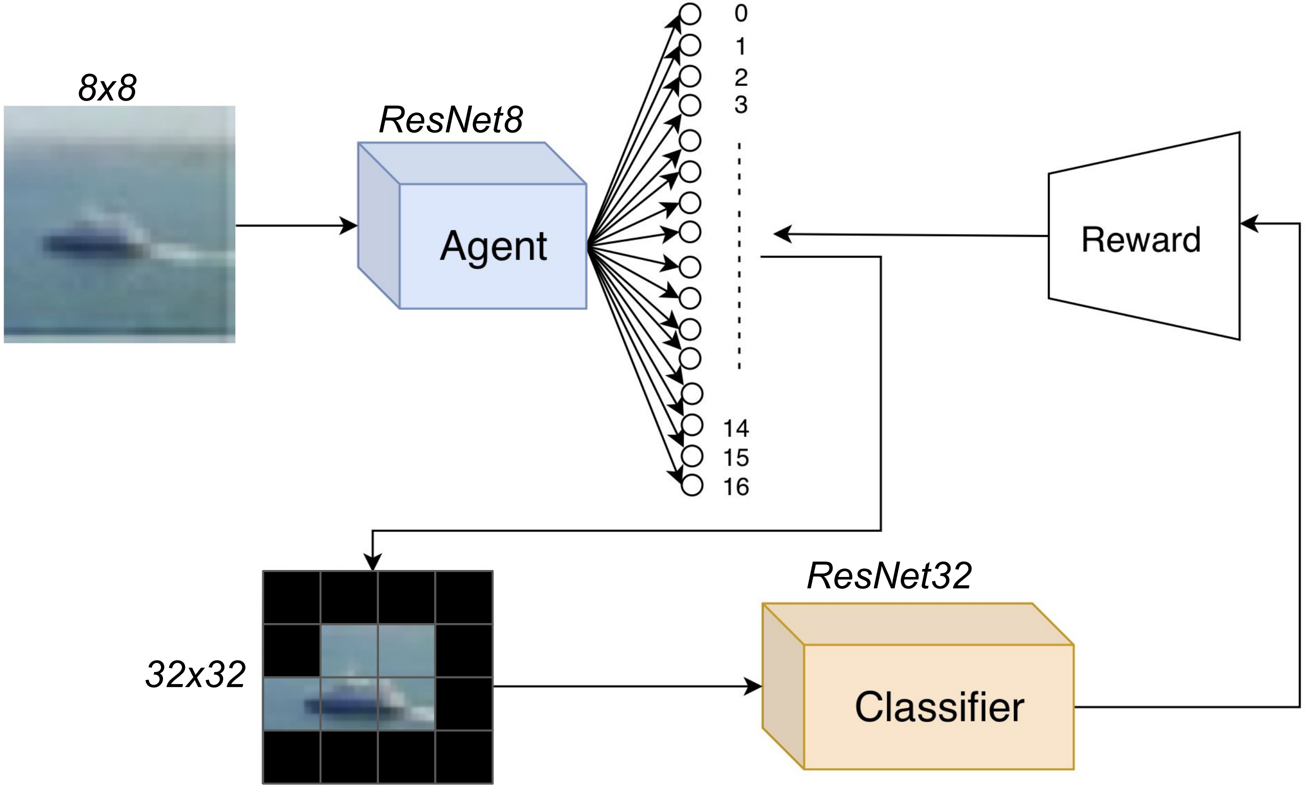
$$\nabla_w J = E[A\nabla_w log \prod_{p=1}^{P} s_p^{u_p}(1 - s_p^{1-u_p})] \tag{6}$$

$$A = R(u) - R(\hat{u}) \tag{7}$$

$$if \quad s_p^{u_p} \geq 0.5 \quad u_p = 1 \tag{8}$$

$$s_p^{u_p} = \alpha s_p^{u_p} + (1-\alpha)(1 - s_p^{u_p}) \tag{9}$$

Pre-training the Agent

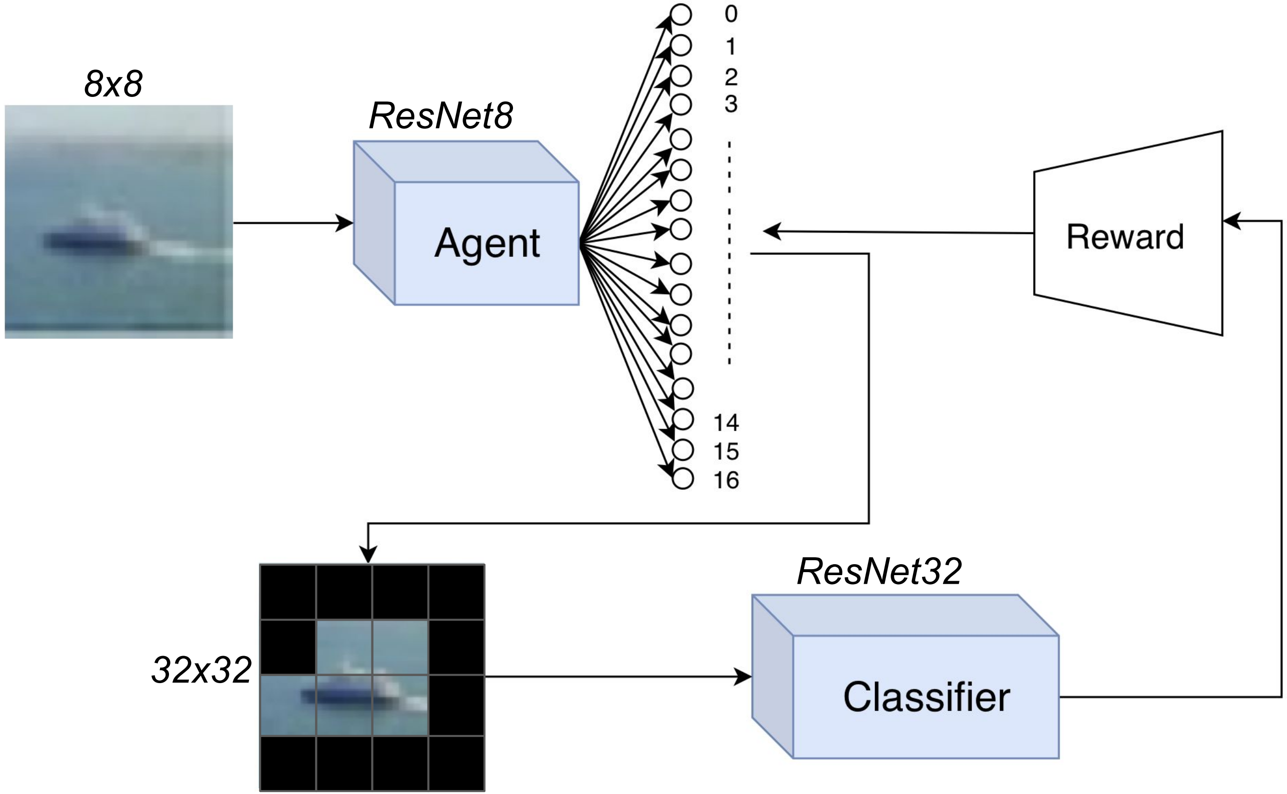


*First, the *classifier* is trained on 32x32 original CIFAR10 images. It achieves 92.3% on test.

*Next, the *agent* is trained on 8x8 low resolution images.

**Curriculum learning* is applied to stabilize training.

Joint Fine-tuning

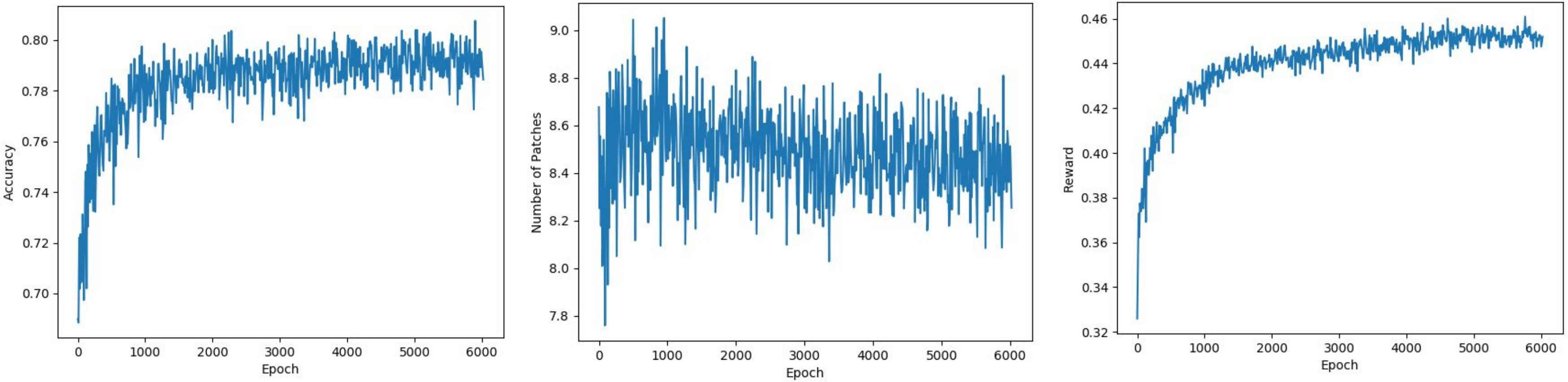


*The *pre-trained agent* is used to drop patches from the original image.

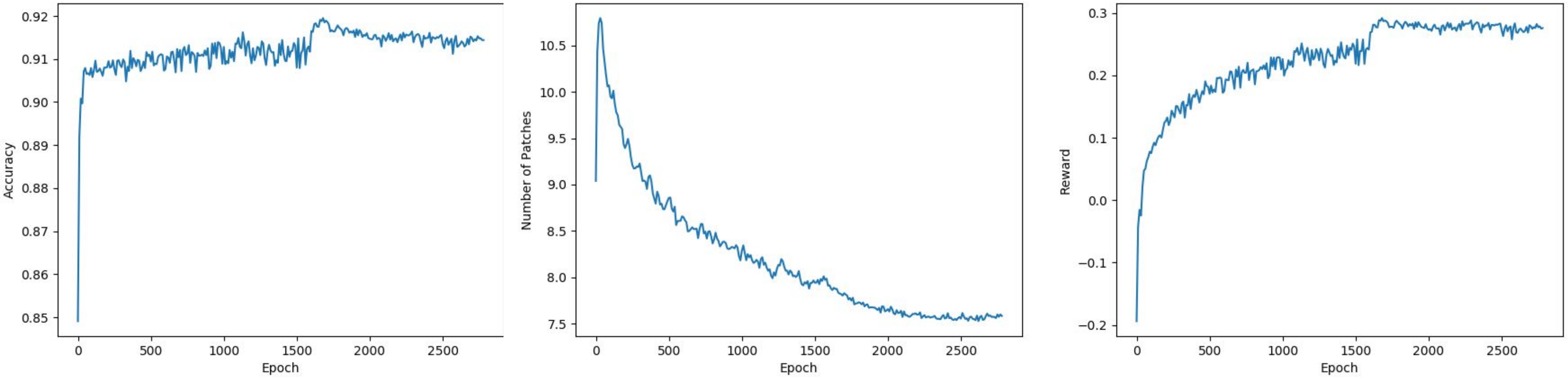
*The *classifier* is then trained *jointly* with the *agent*.

Training on CIFAR10

Pre-training

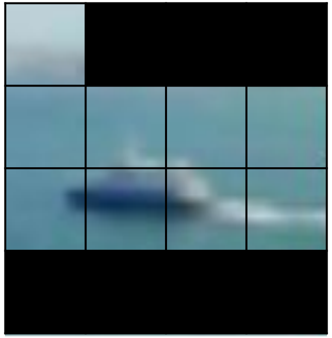


Joint Fine-tuning

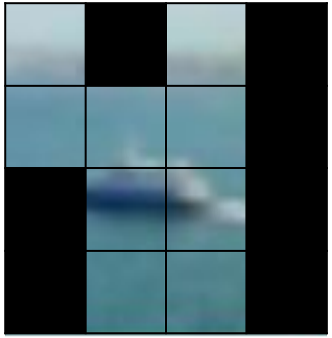


Baseline Models - Fixed Policy

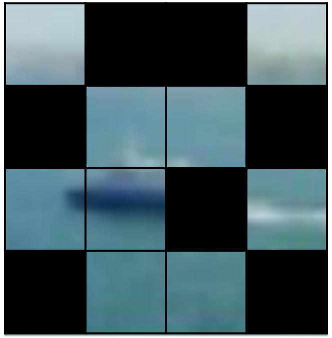
Central P-I = 9



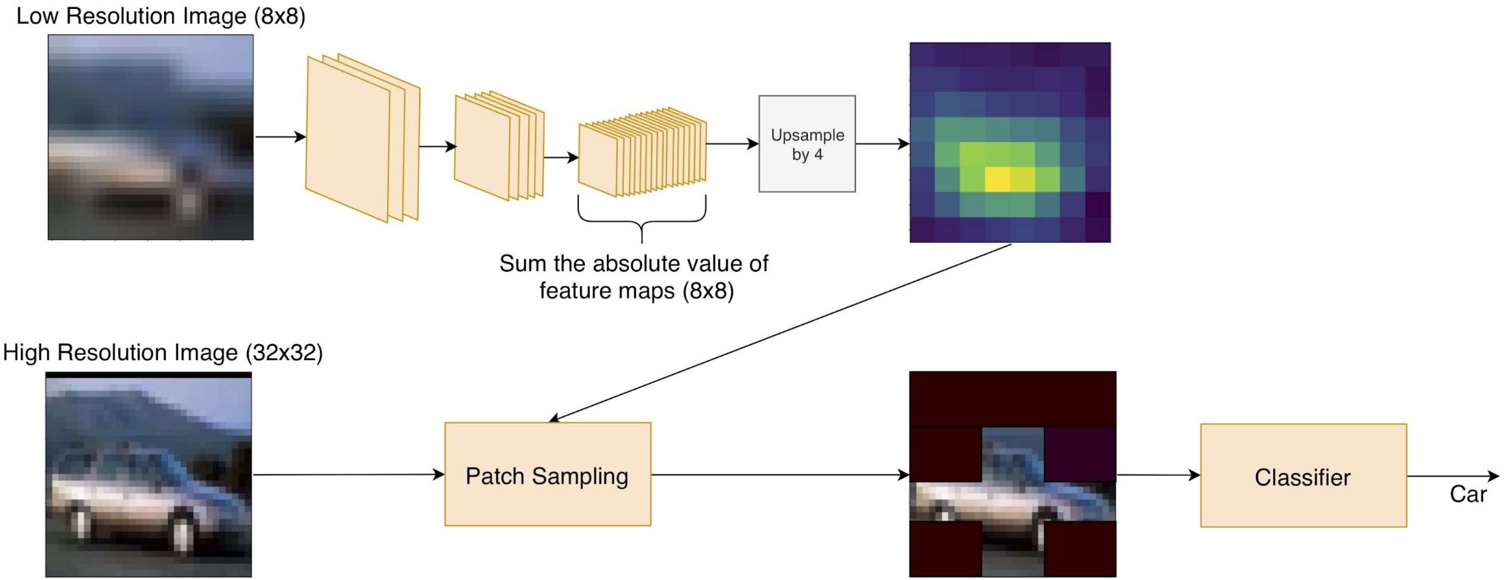
Central P-II = 9



Random P = 9



Baseline Models - Activation Maps



Results on CIFAR10

	Accuracy (%) ***(Pre-training)***	**P**	**Accuracy (%)** ***(Joint Fine-tuning)***	**P**
Central P-I	71.2	9	88.8	9
Central P-II	64.7	9	88.4	9
Random P	40.6 ∓ 1.2	9	88.1∓0.4	9
Activation Map	68.6	9	85.2	9
Ours	**80.6**	**8.5**	**92.0**	**7.8**
NoDrop	N/A	N/A	92.3	16

PatchDrop - Visualizing Agent's Output