

# Domain Adaptation

CSC2539 - Visual Recognition with Text

Lluís Castrejón

University of Toronto

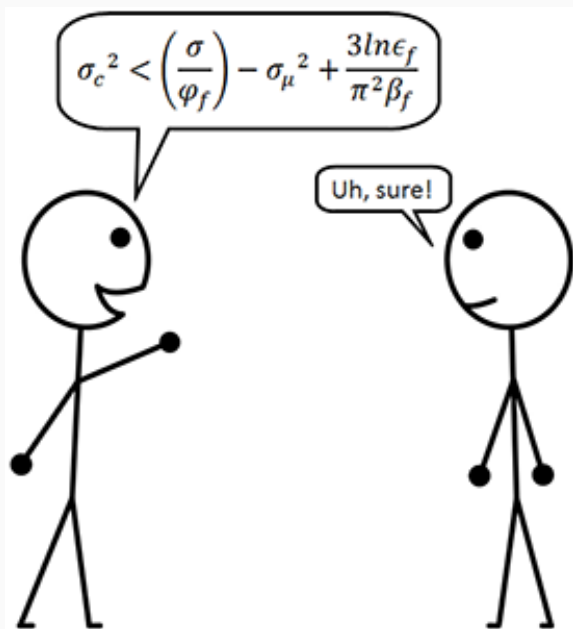
# What is this?

Game: Caption the following images using one short sentence.

What is this?



What is this?





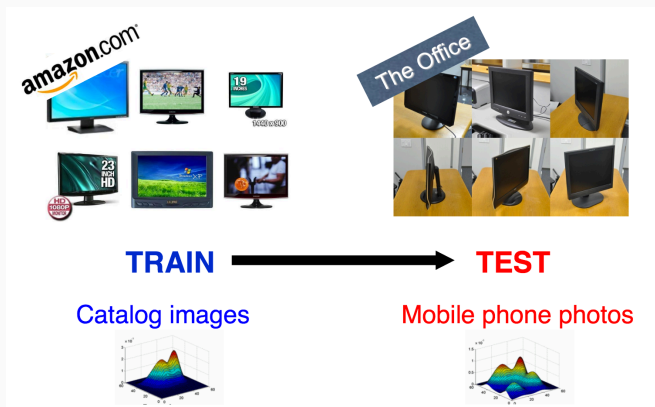
What is this?



# Domain Adaptation

Use the same model with different data distributions in training  
and test

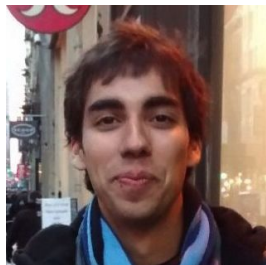
$$P(X) \neq P(X'); P(Y|X) \approx P(Y'|X')$$



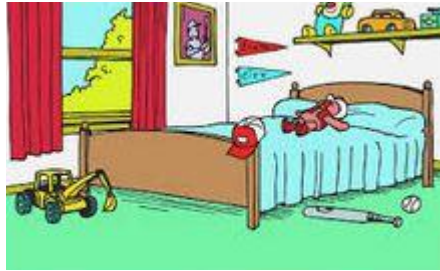
Credit: Kristen Grauman

# Learning Aligned Cross-Modal Representations from Weakly Aligned Data

Lluís Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, Antonio Torralba



# Motivation



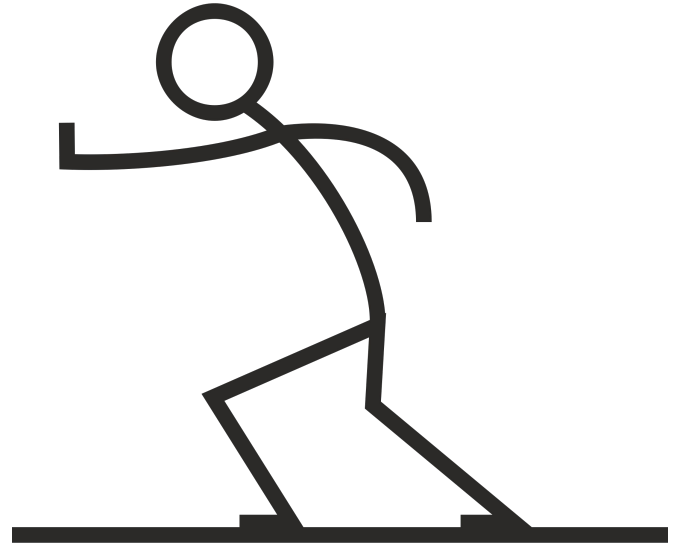
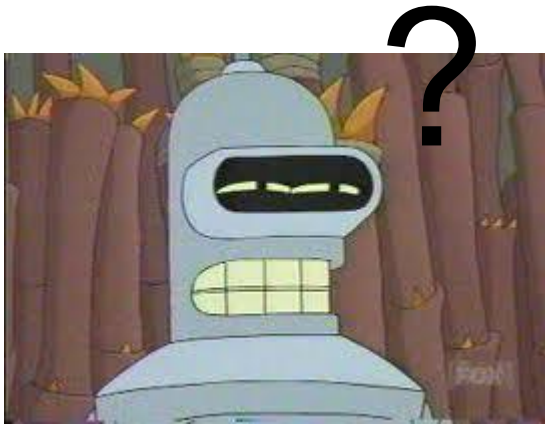
# Motivation



A fire-ship attack on our port. Frigates burned in their berths, honest merchant-men losing their livelihoods. The sacking of the township. Women cut down in their homes. Innocents slaughtered. This must not go unanswered. So I stand in solitude. And I pray. I pray that I am forgiven. I pray that we Dutch are given the year to rebuild our lost vessels and recruit fresh men. That we will right the wrongs done by Charles of England. That renewed, we will take this fight back to England. But this destruction, this murder cannot remain unaddressed. I pray that the sparks of the same fire that burned Schelling are blown across the water to England. That God brings down His fire upon the English and that we Dutch are avenged. That we are spared the necessity of retaliation in the new martial season. I pray that this is done soon, so that God's will is seen. Amen.



# Motivation















# Motivation





# Cross-Modal Scene Understanding

	Real	Clip art	Sketches	Spatial text	Descriptions	
Bedroom				<pre>ceiling wall    wall wall    bed pillow floor</pre>	<pre>ceiling_lamp ceiling wall wall headboard carpet_crop bed</pre>	There is a bed with a striped bedspread. Beside this is a nightstand with a drawer. There is also a tall dresser and a chair with a blue cushion. On the dresser is a jewelry box and a clock.
				<pre>pillow wall floor</pre>	<pre>ceiling wall bed carpet floor</pre>	I am inside a room surrounded by my favorite things. This room is filled with pillows and a comfortable bed. There are stuffed animals everywhere. I have posters on the walls. My jewelry box is on the dresser.
Kindergarten classroom				<pre>wall toy    toy toy    toy floor</pre>	<pre>ceiling poster board table table cabinet cabinet_crop</pre>	There are brightly colored wooden tables with little chairs. There is a rug in one corner with ABC blocks on it. There is a bookcase with picture books, a larger teacher's desk and a chalkboard.
				<pre>poster wall toy shelves shelves floor</pre>	<pre>ceiling board wall wall shelves shelves table</pre>	The young students gather in the room at their tables to color. They learn numbers and letters and play games. At nap time they all pull out mats and go to sleep.



# CMPlaces

## Dataset of 205 scene categories

### Line drawings:

6,644 training + 2,050 validation examples



### Clipart:

11,372 training + 1,954 validation examples



# CMPlaces

## Dataset of 205 scene categories

### Text Descriptions:

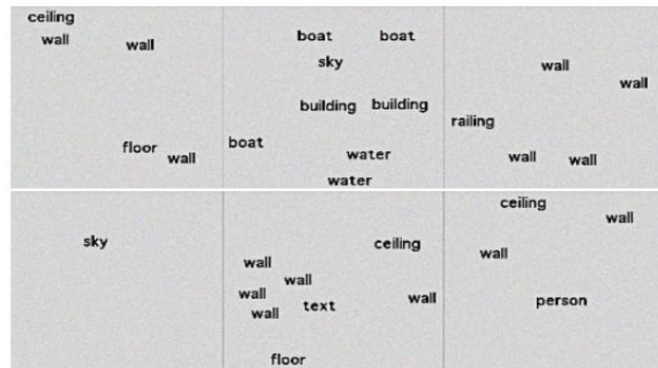
**4,307 training + 2,050 validation examples**

There are brightly colored wooden tables with little chairs. There is a rug in one corner with ABC blocks on it. There is a bookcase with picture books, a larger teacher's desk and a chalkboard.

I am inside a room surrounded by my favorite things. This room is filled with pillows and a comfortable bed. There are stuffed animals everywhere. I have posters on the walls. My jewelry box is on the dresser.

### Spatial Text:

**456,300 training + 2,050 validation examples**



# CMPlaces

## Dataset of 205 scene categories

**Natural images (Places dataset):** 2M training + 20,500 validation examples



Scene categories include Art Gallery, Bedroom, Office, Restaurant, River, Airfield, Bar, Canyon ...

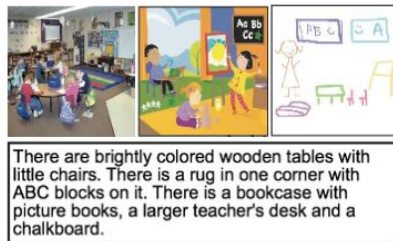
# Strong vs weak alignment

## Strong Alignment (Pairs)



- Cross modal embedding with **pairs**
- CCA, Joint space embeddings, etc.

## Weak Alignment (Category Level)



- Samples are aligned in category level only
- No object level alignment, i.e. **no pairs**

# Strong vs weak alignment

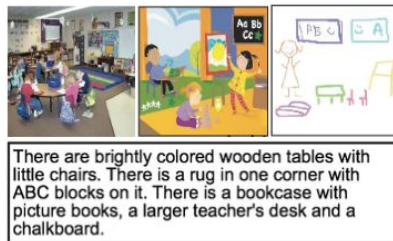
Not scalable!

## Strong Alignment (Pairs)



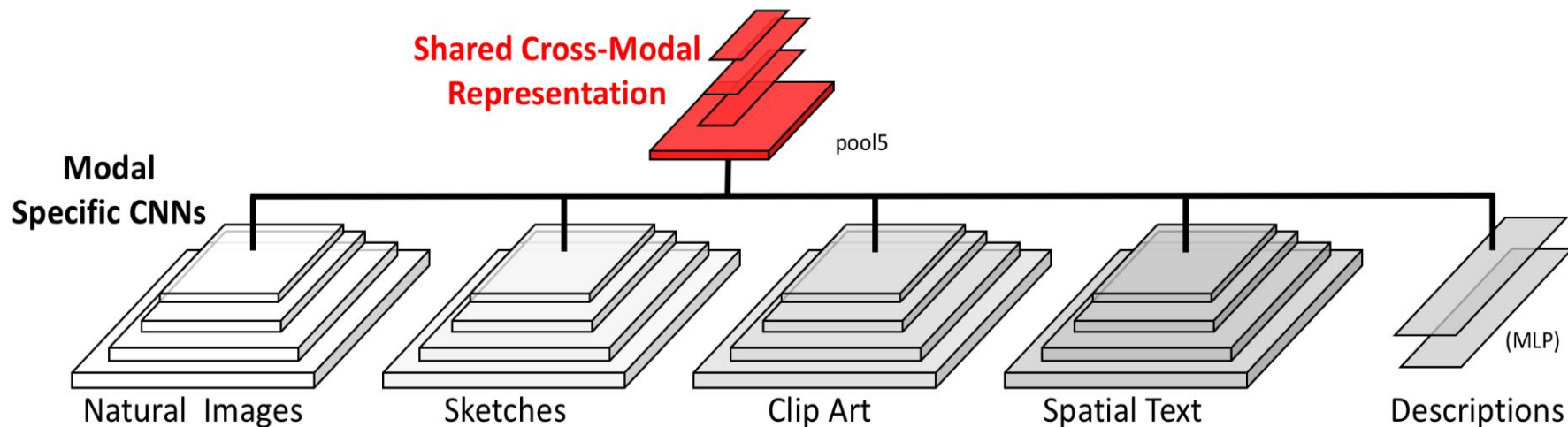
- Cross modal embedding with **pairs**
- CCA, Joint space embeddings, etc.

## Weak Alignment (Category Level)



- Samples are aligned in category level only
- No object level alignment, i.e. **no pairs**

# Cross-modal Networks



- Inputs from **five modalities** with different low-level statistics
- Represent all modalities in a **high-level shared space**

# Cross-modal Networks

**Problem:** Parts of the network specialize to certain domains

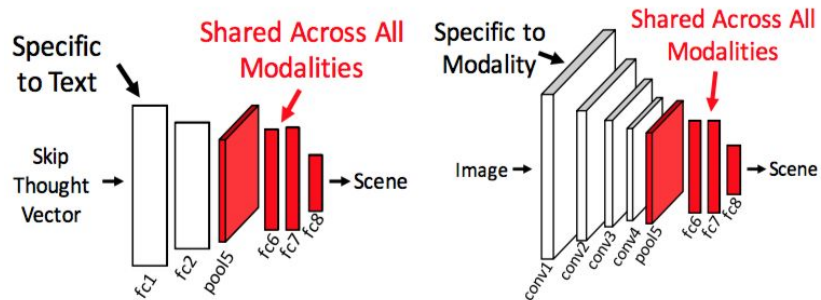
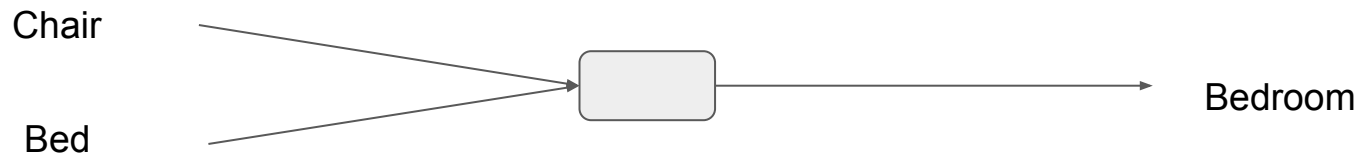
# Cross-modal Networks

**Solution:** Use regularization to enforce alignments



# Cross-modal Networks

## A) Modality Tuning

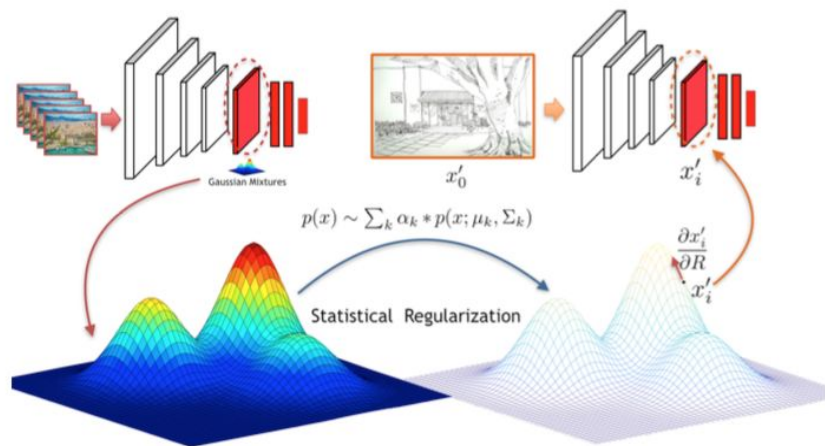


**Step 1:** Train network with higher-level layers initialized and fixed from Places CNN.

**Step 2:** Higher-level layers are released and the model is further fine-tuned end-to-end.

# Cross-modal Networks

## B) Statistical Regularization



$$\min_w \underbrace{\sum_n \mathcal{L}(z(x_n; w), y_n)}_{\text{Softmax Loss for Classification}} + \underbrace{\sum_{n,i} \lambda_i \cdot \mathcal{R}_i(h_i(x_n; w))}_{\text{Statistical Regularization}}$$

Regularize activations in the **shared layers** to follow **similar statistics** across modalities.

Shared statistics estimated from a large dataset (Places) and modeled by a parametric distribution. We experimented with:

- Gaussian
- Gaussian Mixture Model

**Regularization Term:**

$$\mathcal{R}_i(h) = -\log P_i(h; \theta_i)$$

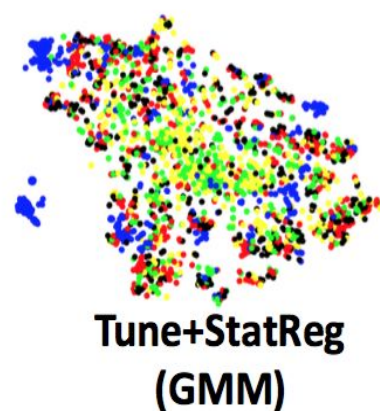
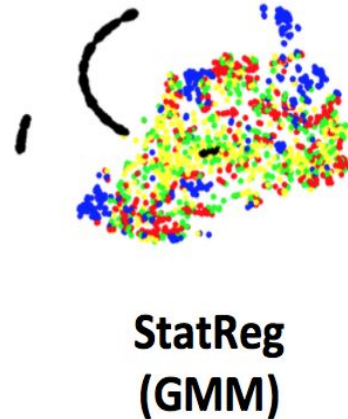
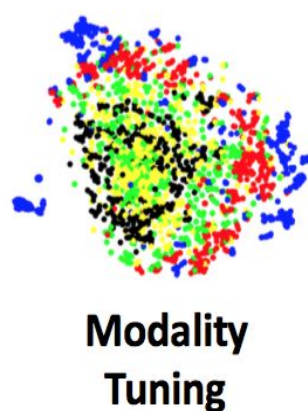
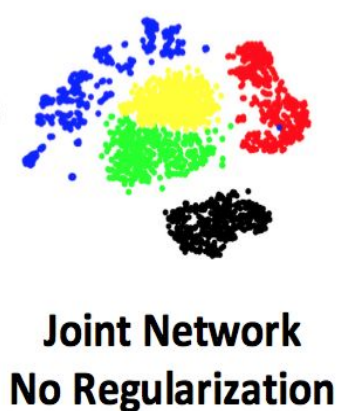
**StatReg with GMM:**

$$\mathcal{R}_i(h; \alpha, \mu, \Sigma) = -\log \sum_{k=1}^K \alpha_k \cdot P_k(h; \mu_k, \Sigma_k)$$

# T-SNE

## Modalities

- Natural Images
- Clipart
- Spatial Text
- Line Drawings
- Descriptions



Random samples from all five modalities are embedded onto a 2D space via t-SNE on *fc7* features

# Visualizing Activations










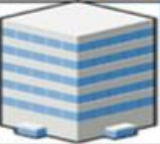




















	Real			Clip art			Sketches			Spatial text			Descriptions
Unit 31 Fountain)													we, water, fishes, you, drink, formed, greek, would, ball, have
Unit 50 (Arcade)													play, children, there, equipment, are, for, train, hole, games, path
Unit 81 (Ring)													ropes, recess, seats, dug, that, square, down, each, fight, it
Unit 86 (Car)													bed, nightstand, window, gas, shampoo, you, tallest, rock, i, my
Unit 104 (Castle)													church, priest, sermon, religious, he, impressive, large, stared, fountain, gas
Unit 115 (Bed)													ice, terrain, plane, cold, i, nightstand, inside, beds, two, movement



# Cross-Modal Retrieval

## Query

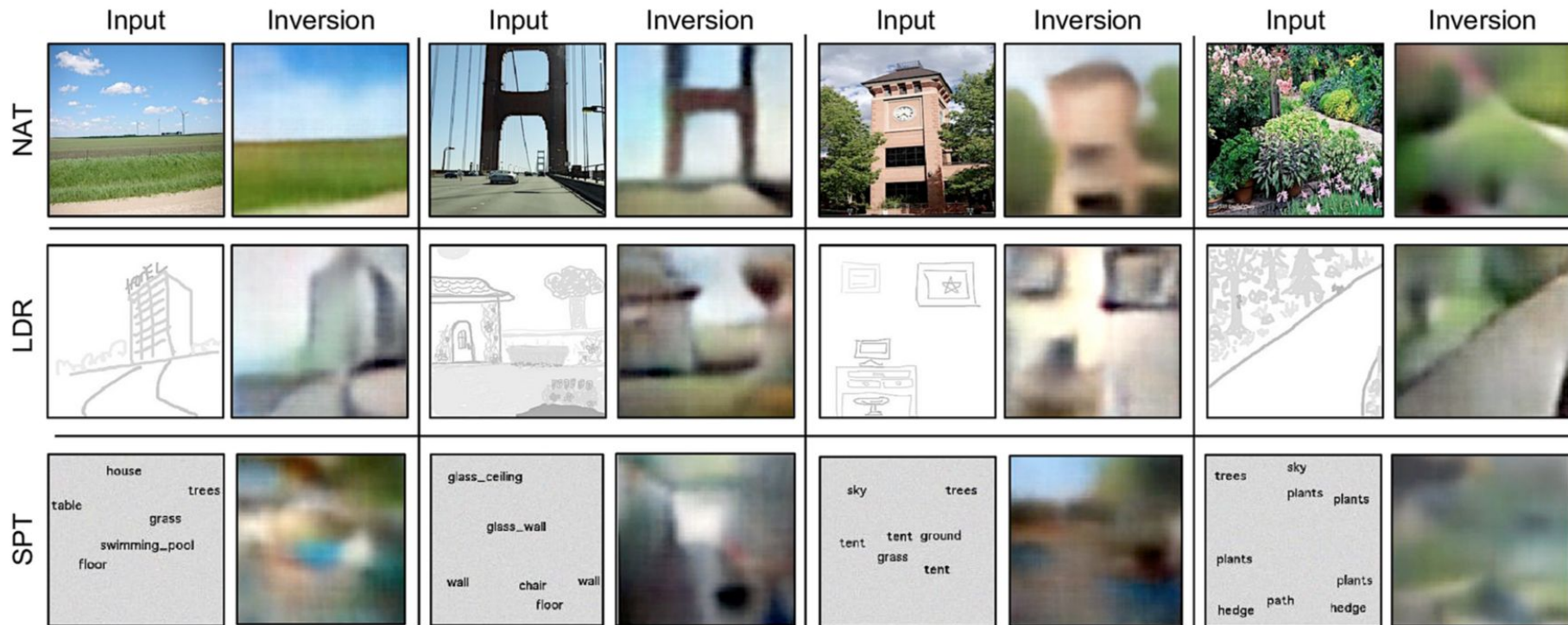
## Retrieved examples

					<p>cabinet door</p> <p>wall</p> <p>wall</p> <p>cabinet sink</p> <p>floor</p>	<p>cabinet door</p> <p>wall</p> <p>wall</p> <p>cabinet sink</p> <p>floor</p>		<p>Everything you could need to make dinner, all in one place. Not quite the size of a full kitchen, but everything is there: microwave, refrigerator, and oven.</p> <p>A very small or compact kitchen. These tiny kitchens typically have all of the regular equipment found in their larger counterparts such as a refrigerator, stove, and microwave, but they are often smaller than full sized appliances. The main purpose of these smaller kitchens</p>
					<p>sky</p> <p>window building</p> <p>window</p> <p>window</p>	<p>window building sky</p> <p>window</p> <p>window</p>		<p>A structure in which people work. It usually has many floors in which the various floors are rented out to different companies. It usually has vending machines on each floor.</p> <p>I had walked inside a very tall building that had many stories in it. I just leaned forward and saw the receptionist desk right in front of me. I saw several men and women dressed in suits and their work attire. This would tell this was a serious setting.</p>
					<p>sky</p> <p>castle</p> <p>wall wall</p> <p>road plants</p>	<p>sky</p> <p>castle</p> <p>wall wall</p> <p>plants road</p>		<p>The building appeared grand from the outside, with its towers and thick stone walls, but inside the stone air was cold and damp. The few small windows were all that allowed the sunlight to penetrate the enormous darkness. There were many old rooms to explore in this ancient</p> <p>This defines the perimeter of an Islamic city with high, fortified walls to keep out invaders. There are often many defensive towers and outside the walls. The residents are relatively safe within the borders of this area.</p>
					<p>sky</p> <p>snowy_mountain</p> <p>crisscross</p>	<p>snowy_mountain sky</p>		<p>A large white covered land mass. It is surrounded by clouds at the top. You can see skiers using trails and ski lifts above the ground. It is winter and many of the cabins at the base of the land mass are occupied. There is a sign that warns be careful of avalanches.</p> <p>Large low mountains. Usually weather near iceberg is very cold and windy. Huge water bubble round corners when the mountain starts melting. Whenever I think of Titanic Ship, I think of low mountain that saved it.</p>
					<p>trees</p> <p>trees dirt_track</p>	<p>trees trees</p> <p>wall dirt_track</p>		<p>You're driving down a paved narrow-winding path surrounded by trees. There is only one lane in each direction, and visibility is low due to how thick the trees are around you and the curve of the path. The trees seem to go far along in any direction.</p> <p>I love to hike in the woods. Sometimes it is easy to follow the trail and sometimes it is so full of trees it is hard to see where to go. I love following this surrounded by trees and the sound of birds singing.</p>

# Cross-Modal Retrieval

Cross Modal Retrieval	Query	NAT				CLP				SPT				LDR				DSC				Mean
	Target	CLP	SPT	LDR	DSC	NAT	SPT	LDR	DSC	NAT	CLP	LDR	DSC	NAT	CLP	SPT	DSC	NAT	CLP	SPT	LDR	mAP
BL-Ind		17.8	15.5	10.1	0.8	11.4	13.1	9.0	0.8	9.0	10.1	5.6	0.8	4.9	7.6	6.8	0.8	0.6	0.9	0.9	0.9	6.4
BL-ShFinal		10.3	13.5	4.0	12.7	7.2	8.7	2.8	8.2	8.1	5.7	2.2	9.3	2.4	2.5	3.1	3.2	3.3	3.4	8.5	2.4	6.1
BL-ShAll		15.9	14.2	9.1	0.8	8.9	10.9	7.0	0.8	8.4	7.4	4.2	0.8	4.3	5.6	5.7	0.8	0.6	0.9	0.9	0.9	5.4
A: Tune		12.9	23.5	5.8	19.6	9.7	15.5	4.0	13.7	19.0	13.5	5.6	24.0	4.1	3.8	5.8	5.9	6.4	4.5	9.5	2.5	10.5
A: Tune (Free)		14.0	29.8	6.2	18.4	9.2	17.6	3.7	12.9	21.8	15.9	6.2	27.7	3.7	3.1	6.6	5.4	5.2	3.5	10.5	2.1	11.2
B: StatReg (Gaussian)		18.6	20.2	10.2	0.8	11.1	15.4	8.5	0.8	13.3	15.1	7.7	0.8	4.7	6.6	6.9	0.9	0.6	0.9	0.8	0.9	7.2
B: StatReg (GMM)		17.8	23.7	9.5	5.6	13.4	18.1	8.9	4.6	16.7	16.2	8.8	5.3	6.2	8.1	9.4	3.3	3.0	4.1	4.6	2.8	9.5
C: Tune + StatReg (GMM)		14.3	32.1	5.4	22.1	10.0	19.1	3.8	14.4	24.4	17.5	5.8	32.7	3.3	3.4	6.0	4.9	15.1	12.5	32.6	4.6	<b>14.2</b>

# Inverting the representation



We used up-convolutional networks for inversion [Dosovitskiy & Brox]

# Thanks!

<http://cmplaces.csail.mit.edu/>