

# Cross-Modal Scene Representations

Lluís Castrejón  
University of Toronto

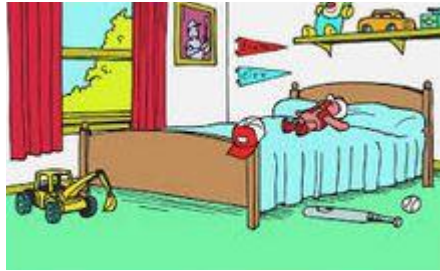
# Motivation



# Motivation



# Motivation



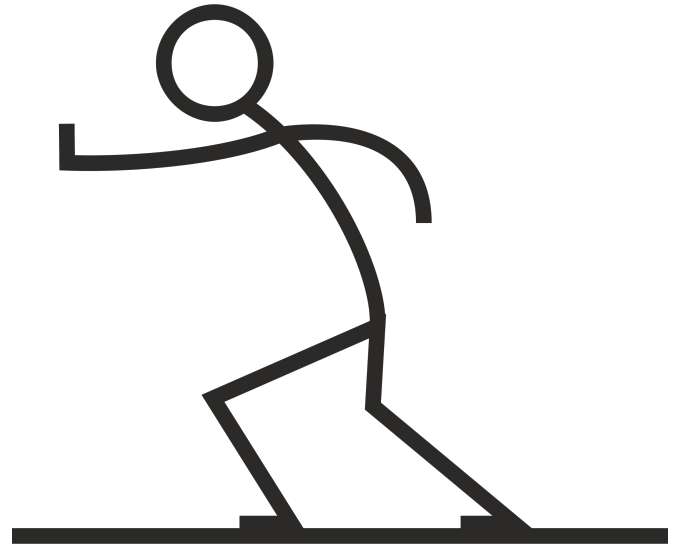
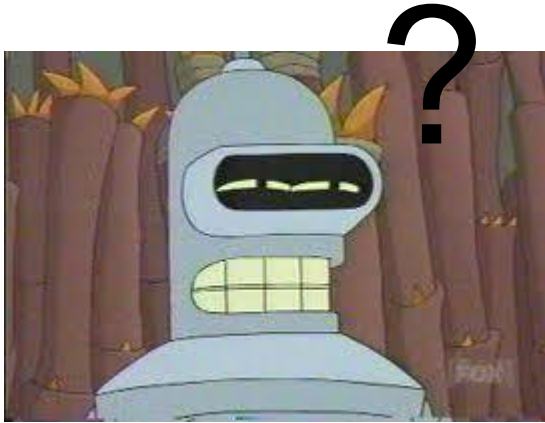
# Motivation



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.





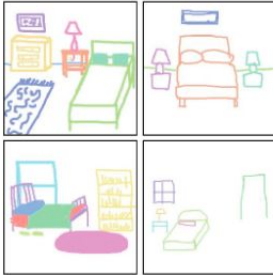





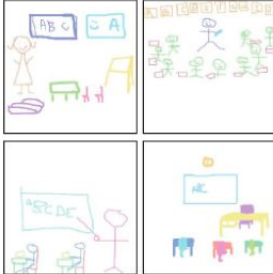


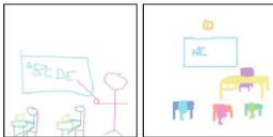
# Motivation



# Motivation



# Cross-Modal Scene Understanding

	Real	Clip art	Sketches	Spatial text	Descriptions																													
Bedroom				<table border="1"> <tr> <td>ceiling</td> <td>ceiling_lamp</td> <td>ceiling</td> <td>wall</td> </tr> <tr> <td>wall</td> <td>wall</td> <td>wall</td> <td>wall</td> </tr> <tr> <td>wall</td> <td>bed</td> <td>pillow</td> <td>headboard</td> </tr> <tr> <td>floor</td> <td>bed</td> <td>carpet_crop</td> <td>bed</td> </tr> </table>	ceiling	ceiling_lamp	ceiling	wall	wall	wall	wall	wall	wall	bed	pillow	headboard	floor	bed	carpet_crop	bed	<p>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.</p>													
	ceiling	ceiling_lamp	ceiling	wall																														
wall	wall	wall	wall																															
wall	bed	pillow	headboard																															
floor	bed	carpet_crop	bed																															
			<table border="1"> <tr> <td>pillow</td> <td>window</td> <td>bed</td> <td>wall</td> <td>wall</td> </tr> <tr> <td>wall</td> <td>bed</td> <td>carpet</td> <td>bed</td> <td>wall</td> </tr> <tr> <td>floor</td> <td>floor</td> <td>carpet</td> <td>floor</td> <td>floor</td> </tr> </table>	pillow	window	bed	wall	wall	wall	bed	carpet	bed	wall	floor	floor	carpet	floor	floor	<p>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.</p>															
pillow	window	bed	wall	wall																														
wall	bed	carpet	bed	wall																														
floor	floor	carpet	floor	floor																														
Kindergarten classroom				<table border="1"> <tr> <td>wall</td> <td>toy</td> <td>toy</td> <td>ceiling</td> </tr> <tr> <td>toy</td> <td>toy</td> <td>toy</td> <td>poster</td> </tr> <tr> <td>floor</td> <td>toy</td> <td>table</td> <td>board</td> </tr> <tr> <td></td> <td></td> <td>table</td> <td>cabinet</td> </tr> <tr> <td></td> <td></td> <td>cabinet_crop</td> <td>cabinet_crop</td> </tr> </table>	wall	toy	toy	ceiling	toy	toy	toy	poster	floor	toy	table	board			table	cabinet			cabinet_crop	cabinet_crop	<p>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.</p>									
	wall	toy	toy	ceiling																														
toy	toy	toy	poster																															
floor	toy	table	board																															
		table	cabinet																															
		cabinet_crop	cabinet_crop																															
			<table border="1"> <tr> <td>poster</td> <td>wall</td> <td>ceiling</td> <td>board</td> <td>wall</td> <td>wall</td> </tr> <tr> <td>wall</td> <td>toy</td> <td>wall</td> <td>wall</td> <td>wall</td> <td>wall</td> </tr> <tr> <td>toy</td> <td>shelves</td> <td>shelves</td> <td>shelves</td> <td>shelves</td> <td>shelves</td> </tr> <tr> <td>shelves</td> <td>shelves</td> <td>table</td> <td>table</td> <td>table</td> <td>table</td> </tr> <tr> <td>floor</td> <td>floor</td> <td>table</td> <td>table</td> <td>table</td> <td>table</td> </tr> </table>	poster	wall	ceiling	board	wall	wall	wall	toy	wall	wall	wall	wall	toy	shelves	shelves	shelves	shelves	shelves	shelves	shelves	table	table	table	table	floor	floor	table	table	table	table	<p>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.</p>
poster	wall	ceiling	board	wall	wall																													
wall	toy	wall	wall	wall	wall																													
toy	shelves	shelves	shelves	shelves	shelves																													
shelves	shelves	table	table	table	table																													
floor	floor	table	table	table	table																													



# 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:

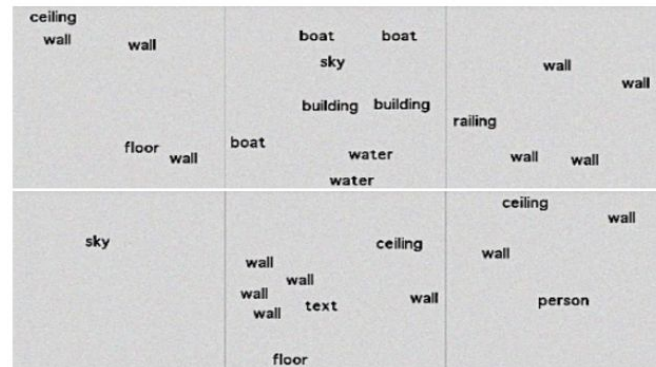
**4,307 training + 2,050 validation examples**

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 ...

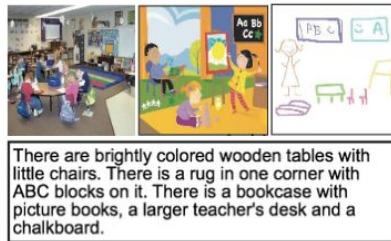
# Strong vs weak alignment

## 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

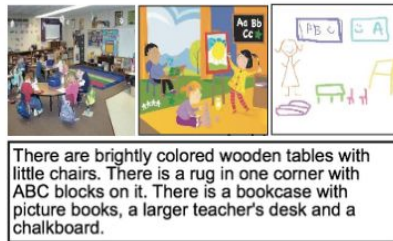
Not scalable!

## Strong Alignment (Pairs)



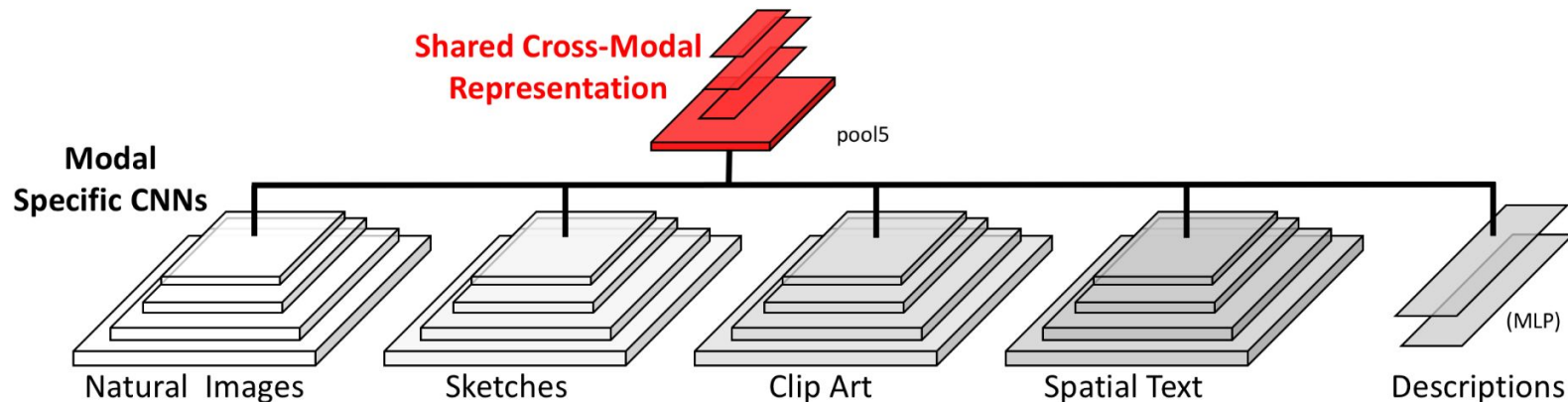
- Cross modal embedding with **pairs**
- CCA, Joint space embeddings, etc.

## 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

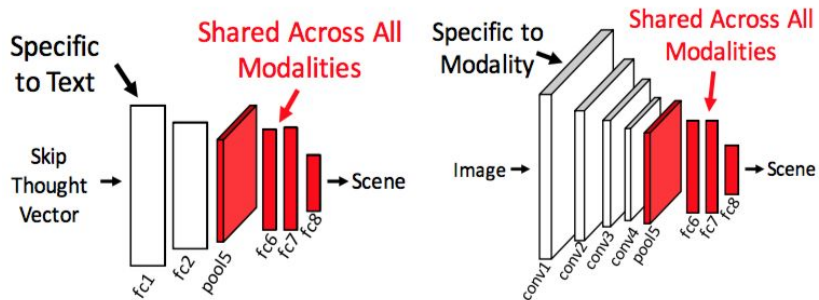
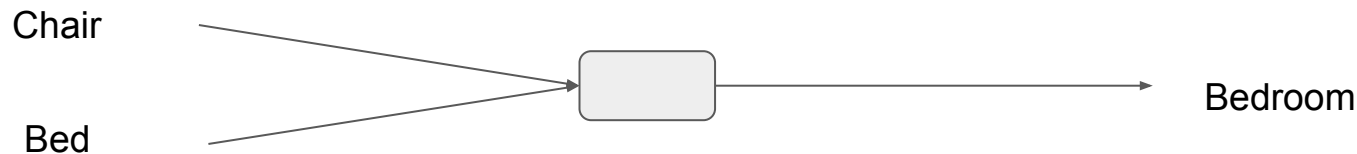
# Cross-modal Networks

**Solution:** Use regularization to enforce alignments



# Cross-modal Networks

## A) Modality Tuning

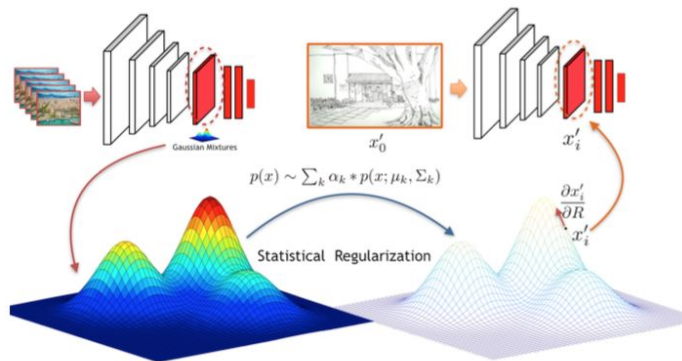


**Step 1:** Train network with **higher-level layers** initialized and fixed from **Places CNN**.

**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

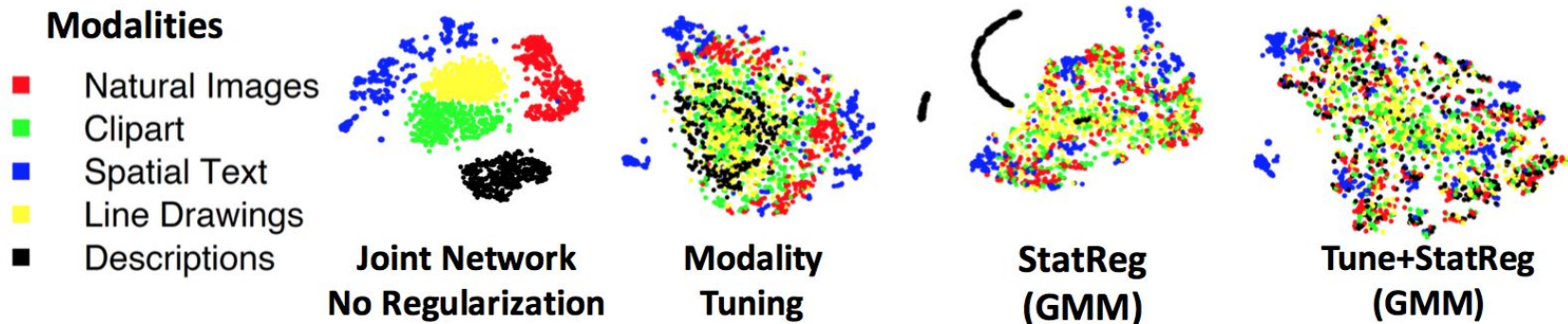
**Regularization Term:**

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

**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



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

Units emerge in our *pool5* representation that fire on concepts independently of the modality

# Cross-Modal Retrieval

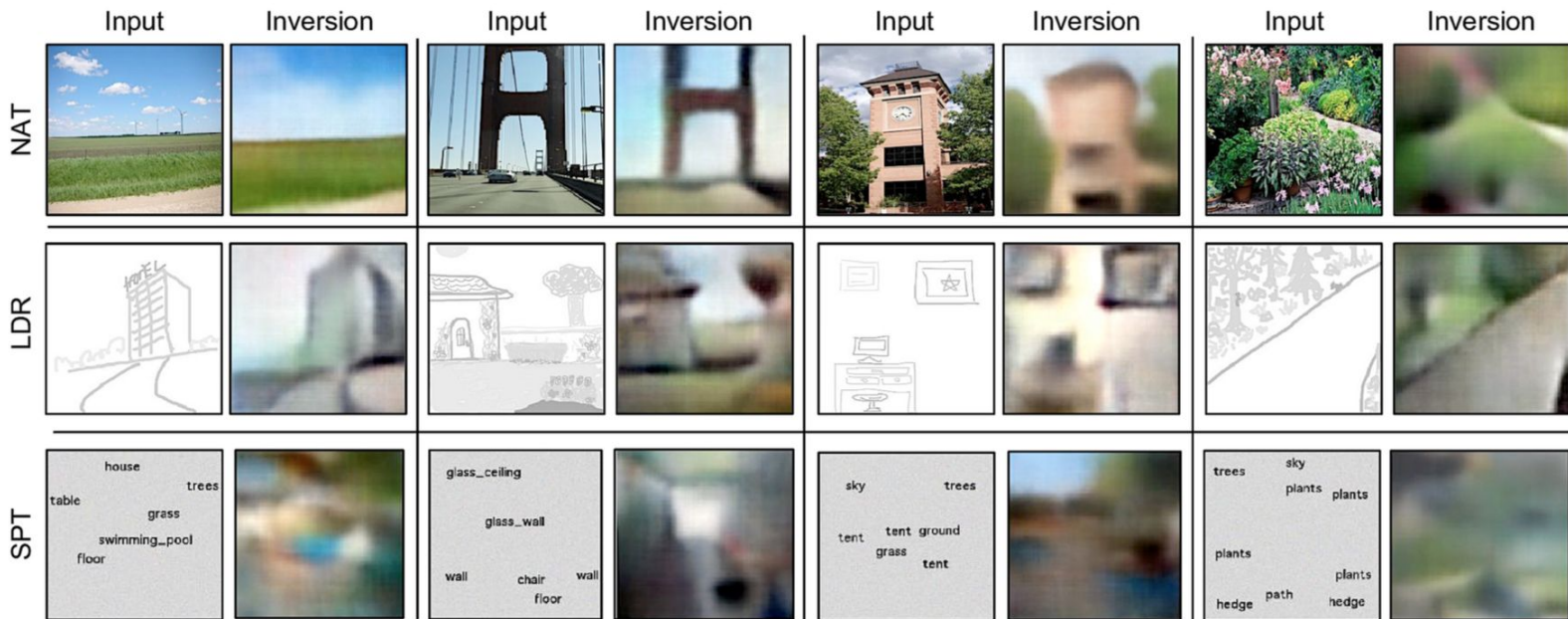
Given a query in one modality, we retrieve nearest neighbors in other modalities

Query			Retrieved examples			
			<ul style="list-style-type: none"> <li>cabinet</li> <li>door</li> <li>wall</li> <li>wall</li> <li>cabinet</li> <li>sink</li> <li>floor</li> </ul>	<ul style="list-style-type: none"> <li>wall</li> <li>door</li> <li>wall</li> <li>cabinet</li> <li>sink</li> <li>floor</li> </ul>		<ul style="list-style-type: none"> <li>Everything you could need to make dinner. These tiny kitchens typically have all of the regular equipment found in their larger counterparts such as an refrigerator, stove, and microwave, but they are often smaller than full sized appliances. The main purpose of these spaces is to</li> <li>I had noticed inside a very tall building that had many stories in it. I just faced forward and saw the rectangular ones right in front of me. I saw several men and women dressed in suits and their work attire. You could tell this was a serious setting.</li> </ul>
			<ul style="list-style-type: none"> <li>sky</li> <li>window</li> <li>building</li> <li>window</li> <li>window</li> <li>window</li> </ul>	<ul style="list-style-type: none"> <li>window</li> <li>building</li> <li>sky</li> <li>window</li> <li>window</li> <li>window</li> </ul>		<ul style="list-style-type: none"> <li>I had noticed inside a very tall building that had many stories in it. I just faced forward and saw the rectangular ones right in front of me. I saw several men and women dressed in suits and their work attire. You could tell this was a serious setting.</li> </ul>
			<ul style="list-style-type: none"> <li>sky</li> <li>castle</li> <li>wall</li> <li>wall</li> <li>road</li> <li>plants</li> </ul>	<ul style="list-style-type: none"> <li>sky</li> <li>castle</li> <li>wall</li> <li>wall</li> <li>road</li> <li>plants</li> </ul>		<ul style="list-style-type: none"> <li>The building appeared grand from the outside, with the towers and thick stone walls. Not inside the stone did was mild and slumpy. The few small windows were all that allowed the sunlight to penetrate the concrete darkness. There were many and some to</li> <li>This defines the perimeter of an island city with high, fortified walls to keep out invaders. There are often many different lands and outside the walls. The residents are relatively safe within the borders of this area.</li> </ul>
			<ul style="list-style-type: none"> <li>sky</li> <li>snowy_mountain</li> <li>crusade</li> </ul>	<ul style="list-style-type: none"> <li>sky</li> <li>snowy_mountain</li> </ul>		<ul style="list-style-type: none"> <li>I large white covered and ended. It is surrounded by clouds at the top. The sun and shines using trails and like the snow on the ground. It is winter at the base of the land have are snowed. There is a sign that says be careful of</li> <li>Large ice mountain. Thickly forested deep layers is very cold and windy. Hope select bubble sound occurs when the mountain starts melting. Whenever I visit of Titanic Ship, I think of the mountain that caused it.</li> </ul>
			<ul style="list-style-type: none"> <li>trees</li> <li>tree</li> <li>tree</li> <li>tree</li> <li>tree</li> <li>tree</li> </ul>	<ul style="list-style-type: none"> <li>trees</li> <li>tree</li> <li>tree</li> <li>tree</li> <li>tree</li> <li>tree</li> </ul>		<ul style="list-style-type: none"> <li>I see're driving down a road 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 corners of the path. The trees seem to go in both in any direction.</li> <li>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 the sound of the water and the sound of birds singing.</li> </ul>

# Cross-Modal Retrieval

Cross Modal Retrieval	Query	NAT				CLP				SPT				LDR				DSC				Mean mAP
	Target	CLP	SPT	LDR	DSC	NAT	SPT	LDR	DSC	NAT	CLP	LDR	DSC	NAT	CLP	SPT	DSC	NAT	CLP	SPT	LDR	
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/>

