Domain Adaptation

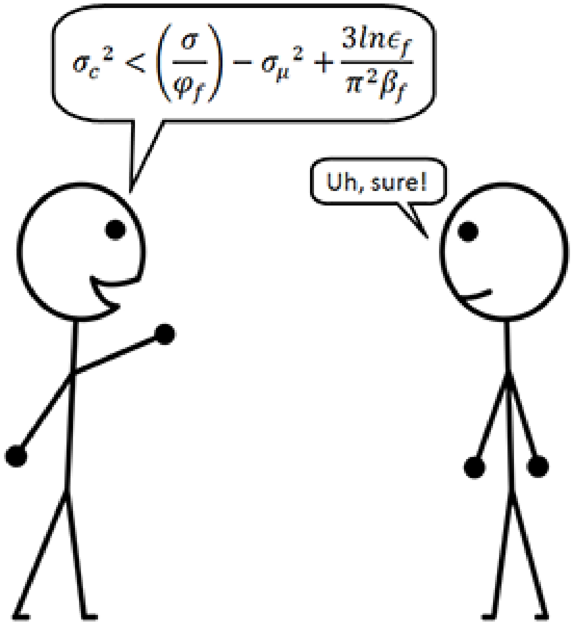
CSC2539 - Visual Recognition with Text
Lluís Castrejón
University of Toronto

What is this?

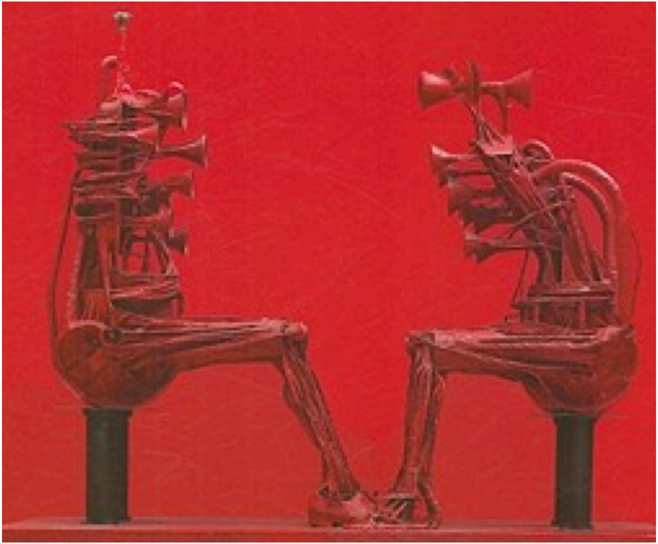
Game: Caption the following images using one short sentence.

What is this?

What is this?


$\sigma_c{}^2 < \left(\frac{\sigma}{\varphi_f}\right) - \sigma_\mu{}^2 + \frac{3ln\epsilon_f}{\pi^2\beta_f}$
Uh, sure!

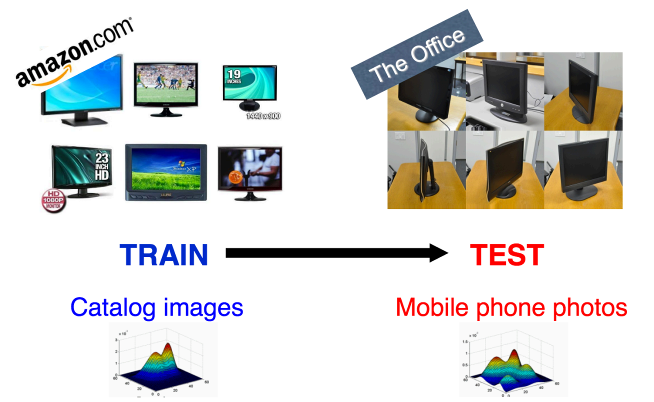
What is this?



Domain Adaptation

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

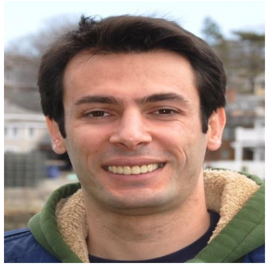
$$P(X) \neq P('X); P(Y|X) \approx P(Y'|X')$$



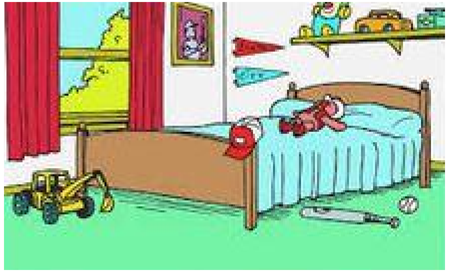
Credit: *Kristen Grauman*

Learning Aligned Cross-Modal Representations from Weakly Aligned Data

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

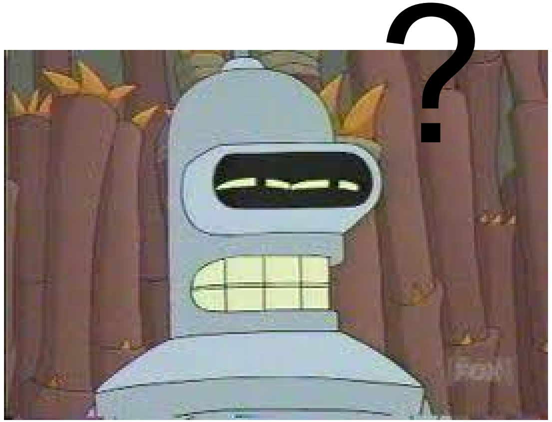

Motivation

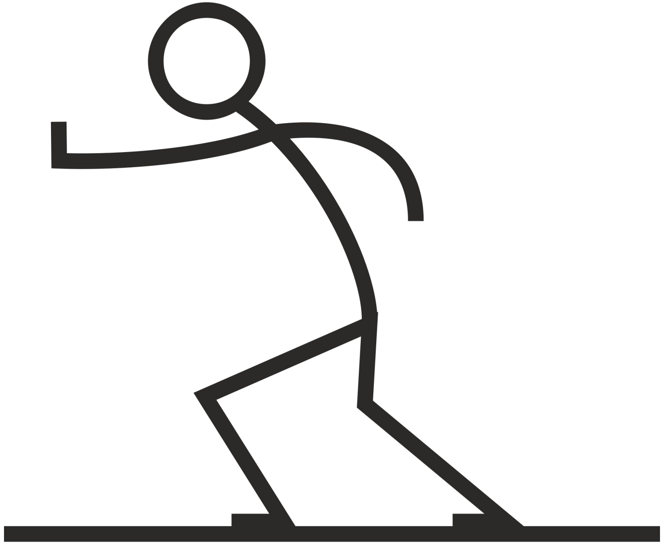


Motivation

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



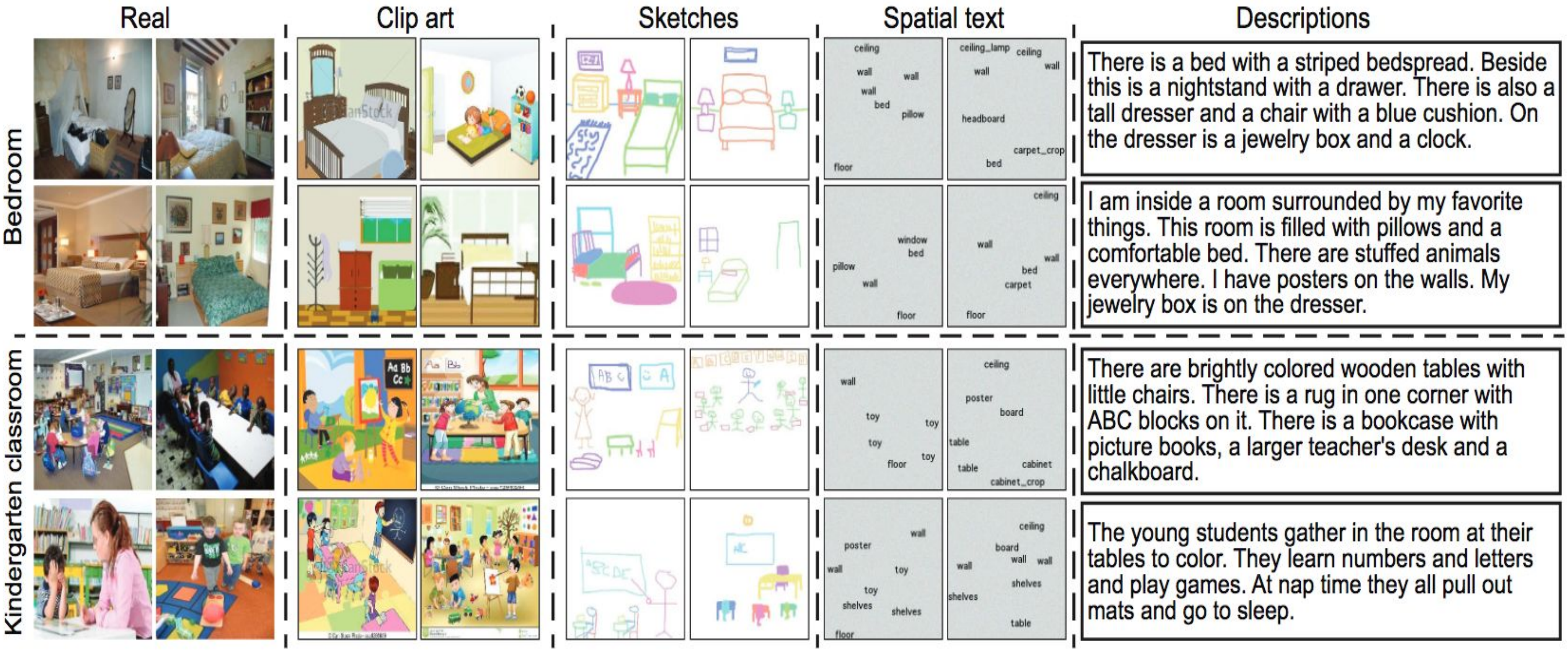


Motivation


Make
More
COFFEE!!


RESTRICTED
ACCESS
AUTHORISED
PERSONNEL ONLY

Cross-Modal Scene Understanding



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:

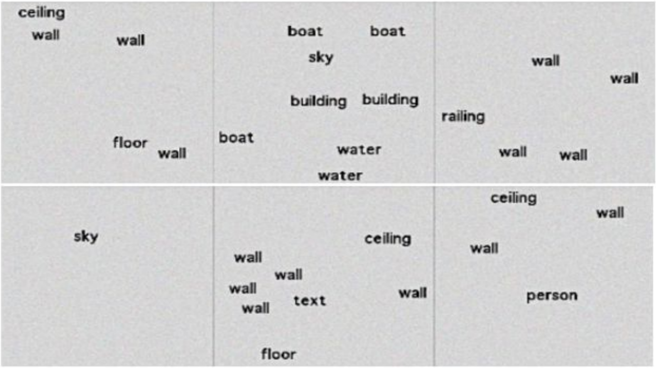
4,307 training + 2,050 validation examples

> 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)

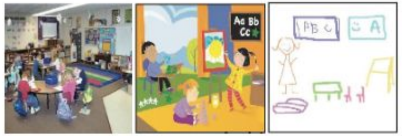
image

text

a man holding a white snow board by some other kids

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

Weak Alignment (Category Level)



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.

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

Strong vs weak alignment

Not scalable!

Strong Alignment (Pairs)

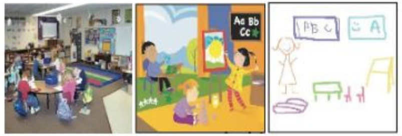
image

text

a man holding a white snow board by some other kids

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

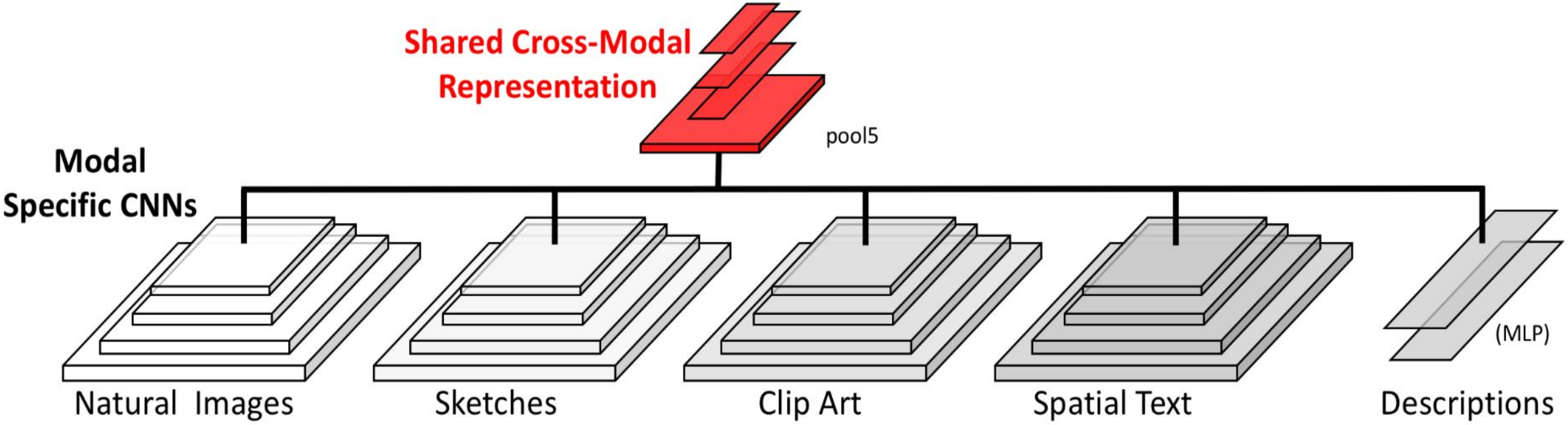
Weak Alignment (Category Level)



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.

- 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

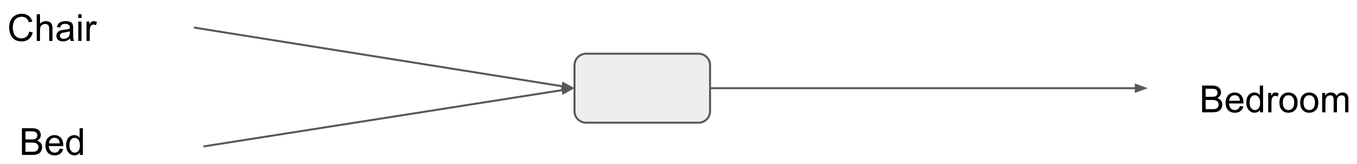
Problem: Parts of the network specialize to certain domains

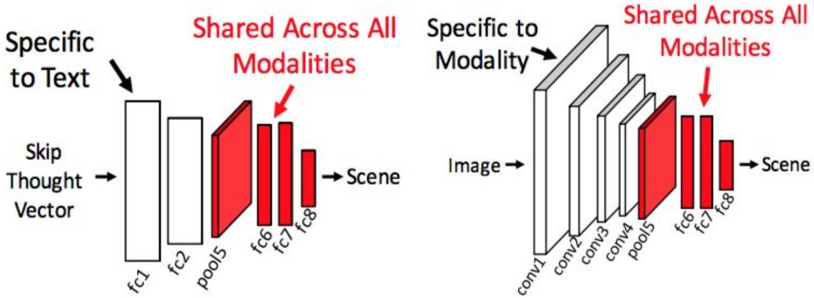
Cross-modal Networks

Solution: Use regularization to enforce alignments

Cross-modal Networks

A) Modality Tuning



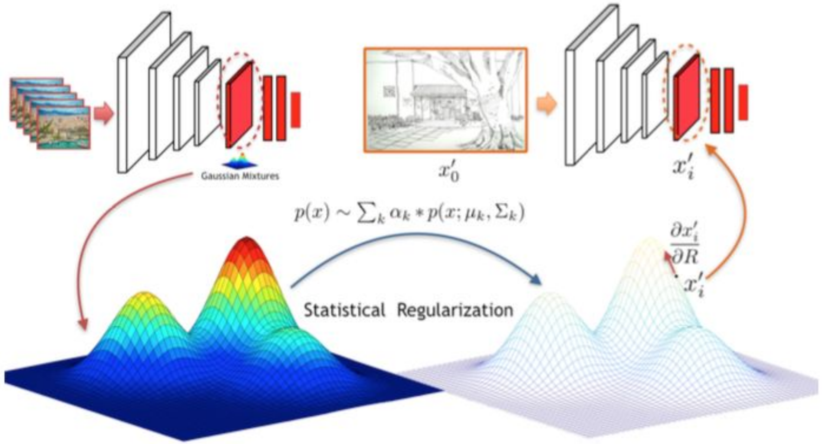


Step 1: Train network with higher-level layers initialized and fixed from Places CNN.

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

Cross-modal Networks

B) Statistical Regularization



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

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

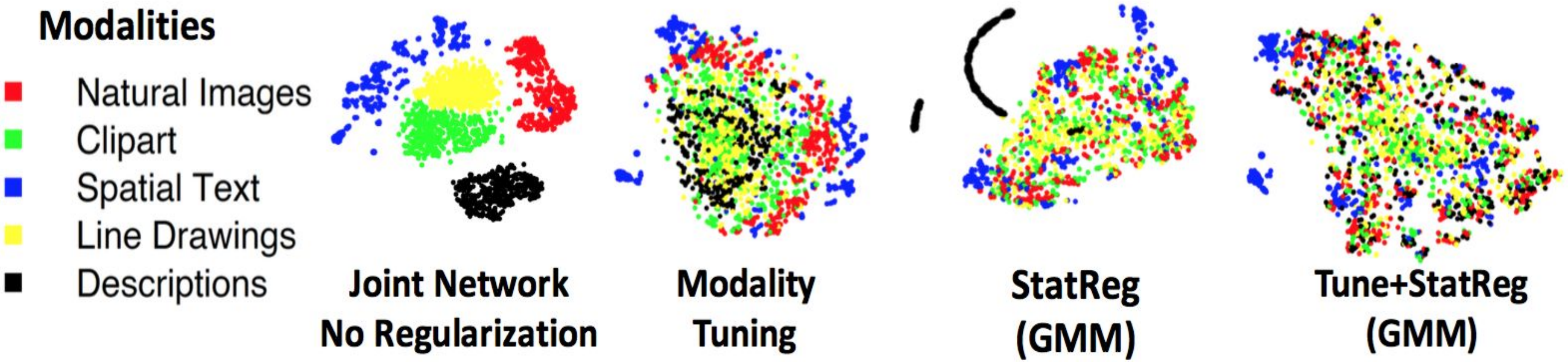
Regularization Term:

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

StatReg with GMM:

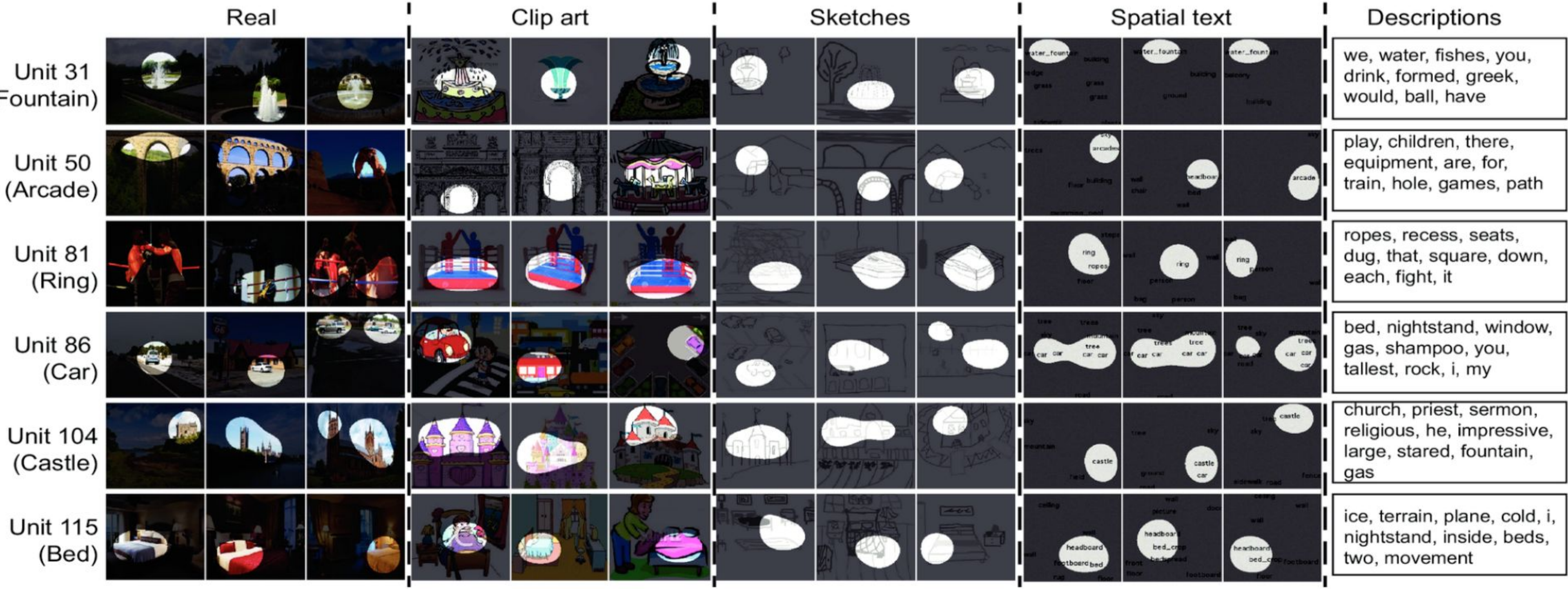
$$\mathcal{R}_i(h; \alpha, \mu, \Sigma) = -\log \sum_{k=1}^{K} \alpha_k \cdot P_k(h; \mu_k, \Sigma_k)$$

T-SNE

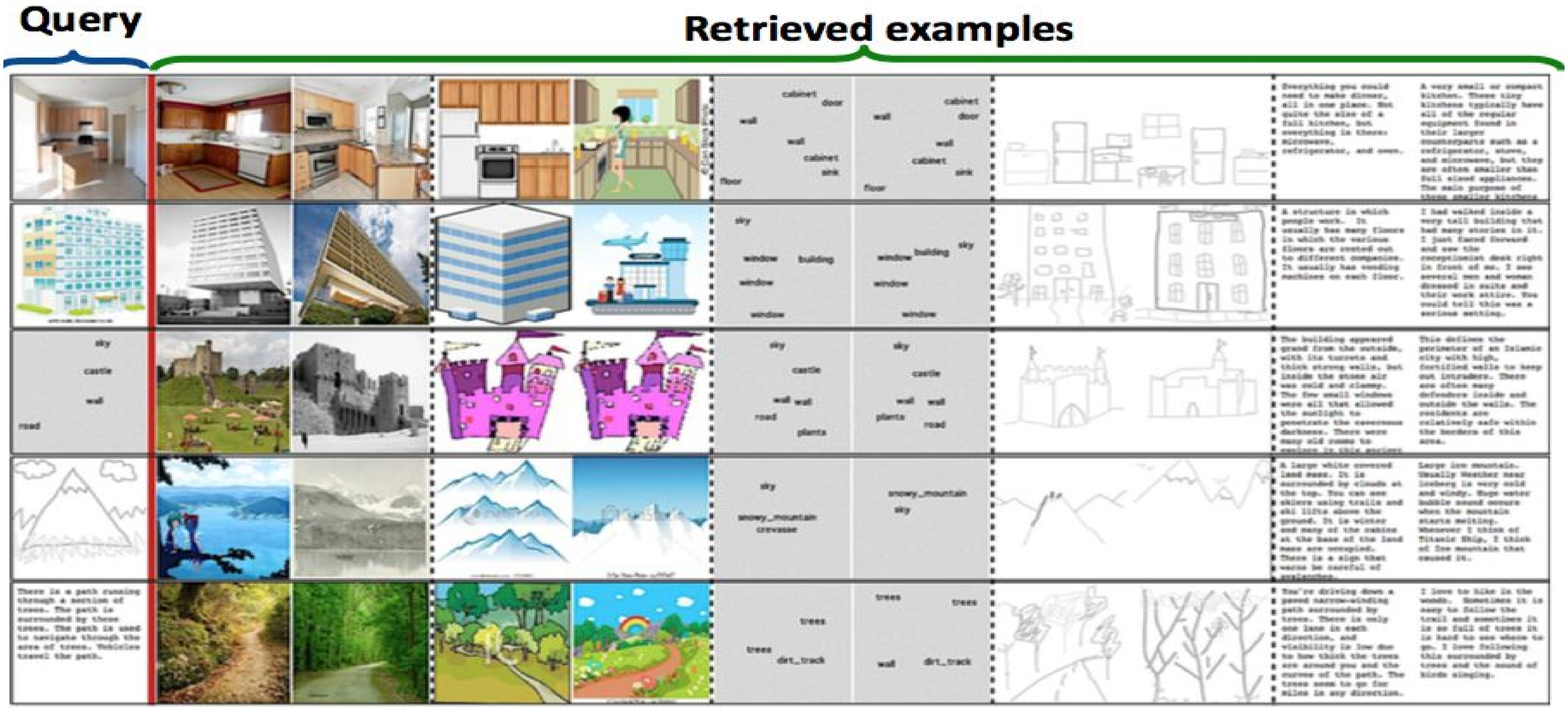


Random samples from all five modalities are embedded onto a 2D space via t-SNE on *fc7* features

Visualizing Activations



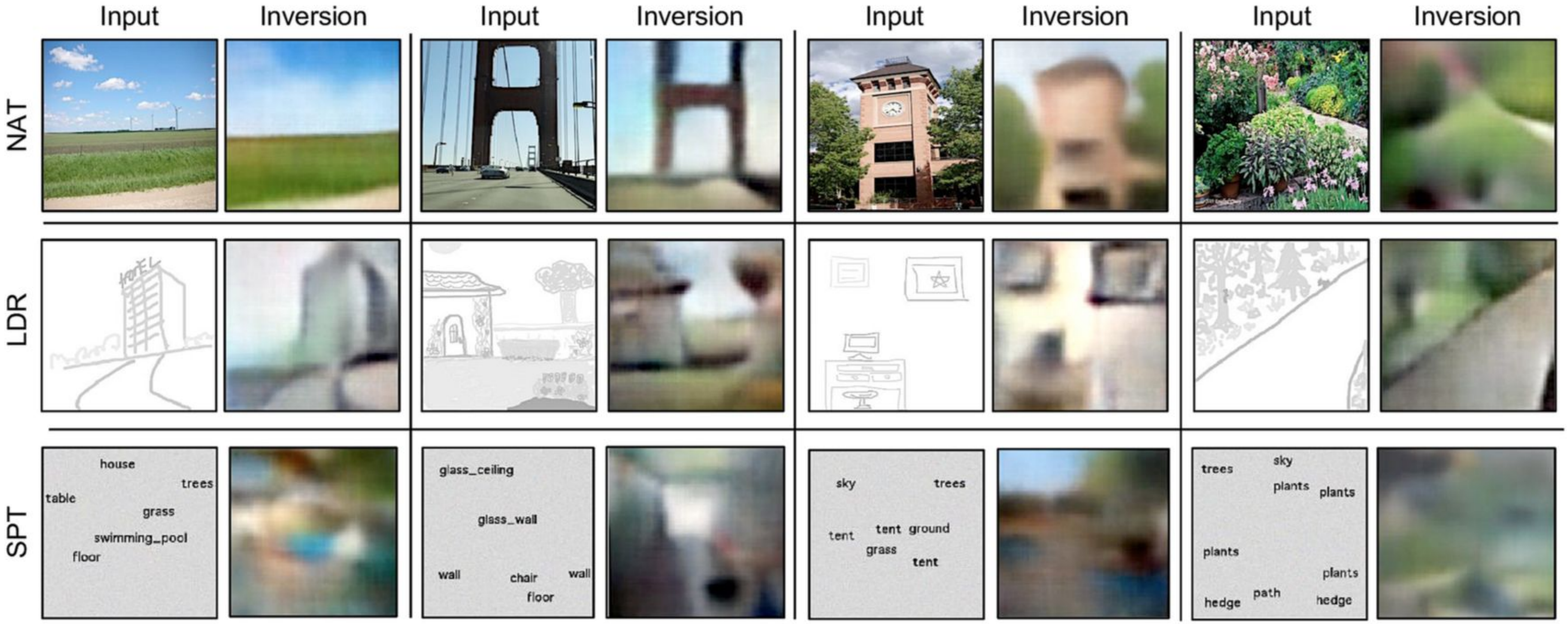
Cross-Modal Retrieval



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	**14.2**

Inverting the representation



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

Thanks!

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