Semantic Clustering for Robust Fine-Grained Scene Recognition

Marian George[1], Mandar Dixit[2], Gábor Zogg[1], and Nuno Vasconcelos[2]
1 Department of Computer Science, ETH Zurich, Switzerland **2 Statistical and Visual Computing Lab, UCSD, CA, United States**

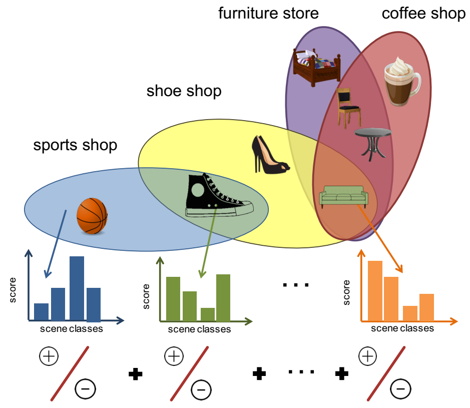
1 Problem

Recognize **fine-grained** scenes in **cross-domain** settings
- Fine-grained scenes share **common** objects
- Varying spatial configurations of objects (cluttered scenes)
  - Especially true in **cross-domain** settings

Example: Store scenes

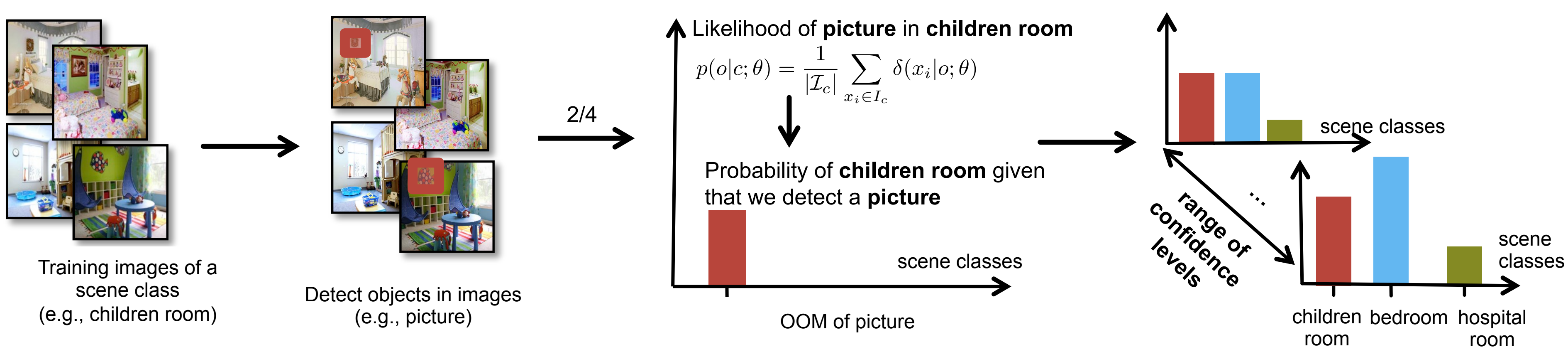
bookstore music store ... office supplies store

2 Semantic Clustering



Exploit semantic structure in fine-grained scenes
- **Semantic scene descriptor**
  - Project scene images to semantic space of object occurrences
  - Convert object occurrences in scenes to scene probabilities
- **Semantic Clustering**
  - Cluster semantic descriptors
  - Learn a discriminative classifier for each discovered topic & combine decisions
  - Better consensus → Better generalization

3 Conditional Scene Probabilities

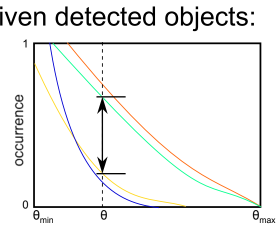


- Represent a scene image as conditional **scene probabilities** given detected objects:

$$p(c|o_i) = \frac{1}{n_i}\sum_k p(c|o_i, \theta = s_k^{(i)})$$

- Filter objects by **discriminative** power:

$$\phi_\theta(o) = \max_{r \in \{1,\ldots,|\mathcal{C}|-1\}} p(\gamma^{-1}(r)|o;\theta) - p(\gamma^{-1}(r+1)|o;\theta)$$

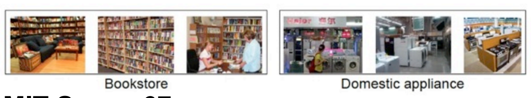


- **Model across a range of confidence levels**
  - Flexible objects arrangements in scenes across domains
- **High-level quantization**
  - Imparts invariance on representation
  - Generalizes better than lower-level features
- **No spatial encoding of objects**

Experimental Evaluation

Datasets

SnapStore
- 18 fine-grained store scenes
- Training: web & testing: real stores



MIT Scene 67
- 67 indoor scenes
- Coarse-grained & same domain

SnapStore, SUN & Places
- 9 store scene classes
- Cross-dataset performance

Dataset Bias

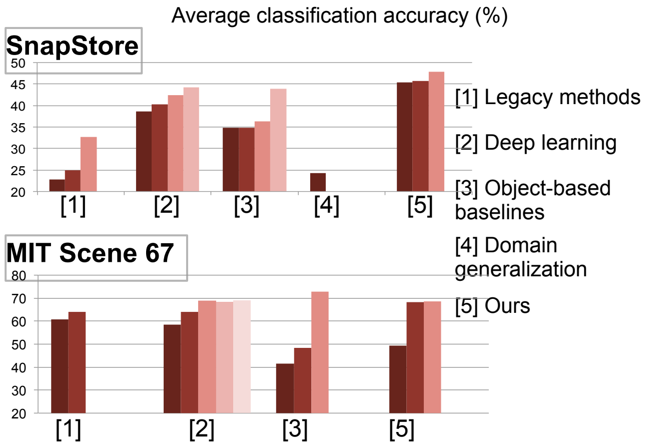
Training/Test	SUN (Web datasets)	SnapStore Web (Web datasets)	Places (Web datasets)	SnapStore phone (Phone dataset)
SUN	68.7 *	57.1	65.7	56.5
SnapStore Web	62.7	71.9 *	60.9	58.2
Places	64.2	59.2	67.6 *	53.8

Average classification accuracy (%)

* Same-dataset recognition accuracy (ground truth)
- performance drop > 12% when testing on **phone** images
- **SUN** and **Places** have very similar distributions → not suitable for domain generalization (only ~3% drop)

Discovered Clusters

Comparison with State-of-the-Art



Cross-Dataset Recognition

Train	Test	DeCaF	DeCaF-C	U-B	DICA	OB	OB-SC	**OOM**	**OOM-SC**
SnW	SnP	58.2	56.3	N/A	42.1	30.0	37.4	61.1	**62.0**
SUN	SnP	56.5	53.9	N/A	45.5	39.2	35.9	54.4	**56.9**
Pla	SnP	53.8	49.1	N/A	37.7	27.6	28.3	**54.8**	54.6
SnW,SnP	Pla,SUN	59.1	59.9	52.3	49.2	22.7	25.7	57.3	**60.6**
SnW,SUN	SnP,Pla	60.6	58.5	50.3	52.2	37.4	37.7	61.0	**63.2**
SUN,Pla,SnW	SnP	59.7	57.2	47.8	53.5	36.3	39.1	61.6	**62.5**
SUN,SnP,SnW	Pla	**63.8**	62.2	33.8	50.8	27.4	30.2	59.8	63.3
Average		58.8	56.7	46.0	47.2	32.9	33.4	58.5	**60.4**

- Semantic clustering outperforms other methods
- Clustering DeCaF performs worse than baseline DeCaF → low-level spatial maps vs. high-level semantic features
- Similarity between SUN and Places benefits DeCaF

Scene Likelihoods (OOM)