Scene recognition with CNNs: objects, scales and dataset bias

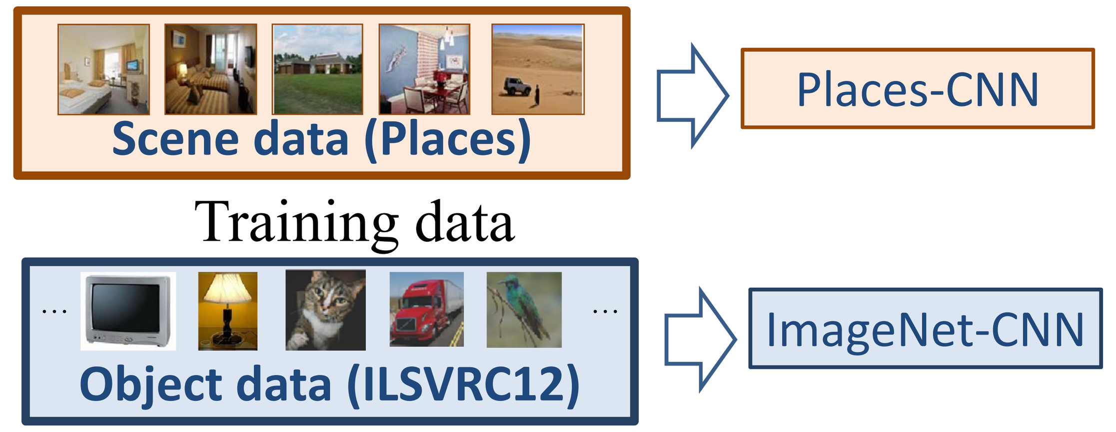
Luis Herranz, Shuqiang Jiang, Xiangyang Li
Institute of Computer Technology, Chinese Academy of Sciences, Beijing, China
{luis.herranz,shuqiang.jiang,xiangyang.li}@vipl.ict.ac.cn

IEEE 2016 Conference on Computer Vision and Pattern Recognition
CVPR 2016

Introduction

We address two **problems**:

1. Effectively implement **multiscale CNN architectures** for scene recognition.
2. Effectively **combine Places and ImageNet**



Motivation

- Scaling (of patches) changes the data distribution.
- This induces a **scale-related bias** if the CNN model is fixed.
- However, this **bias is ignored** in most works.

Contributions

- Study the scale-induced bias
- Multi-scale architecture using **scale-specific CNNs**.

Previous works

1. Scene recognition:

- **Holistic scene-CNNs**[1]. Trained on Places-CNN. *Only **global** scale(s), object-like scales are ignored.
- **Multi-scale local CNN pooling**[2,3,4]. ImageNet-CNNs on multiple scales and aggregated using pooling (e.g. VLAD, FV). *The CNN model is **fixed** black box (scale is ignored).

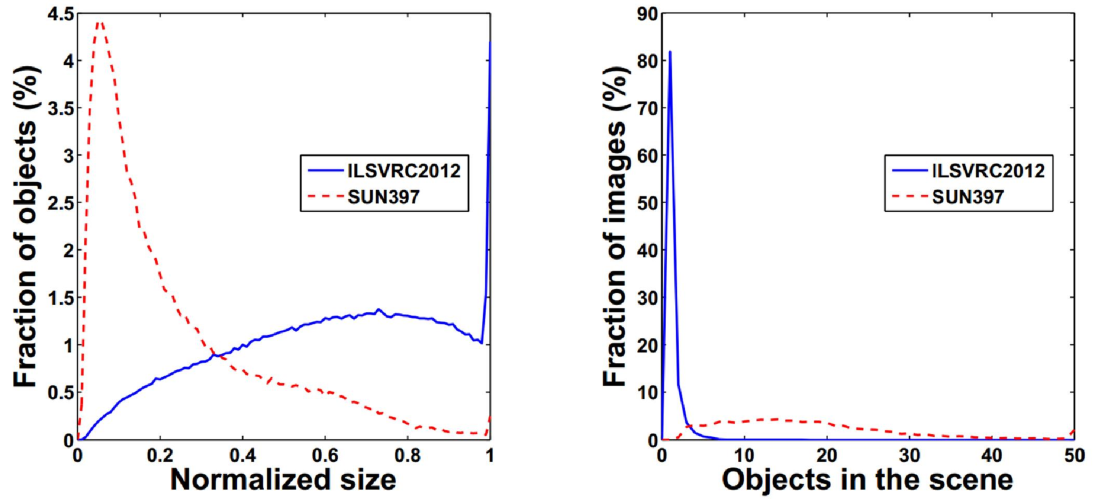
2. Combining object data and scene data in a CNN:

- **Hybrid-CNN**[1]. Trained with all ImageNet and Places data. *Object and scene images are rescaled equally (scale is ignored).

Scale-induced bias

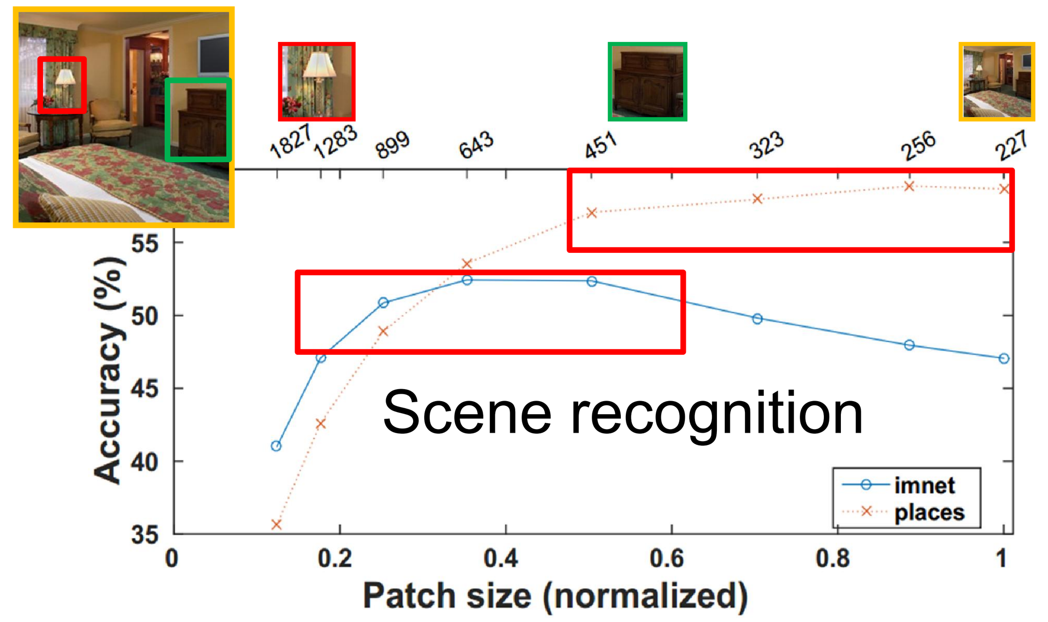
The **distributions of objects** in object datasets and scene datasets are **very different**

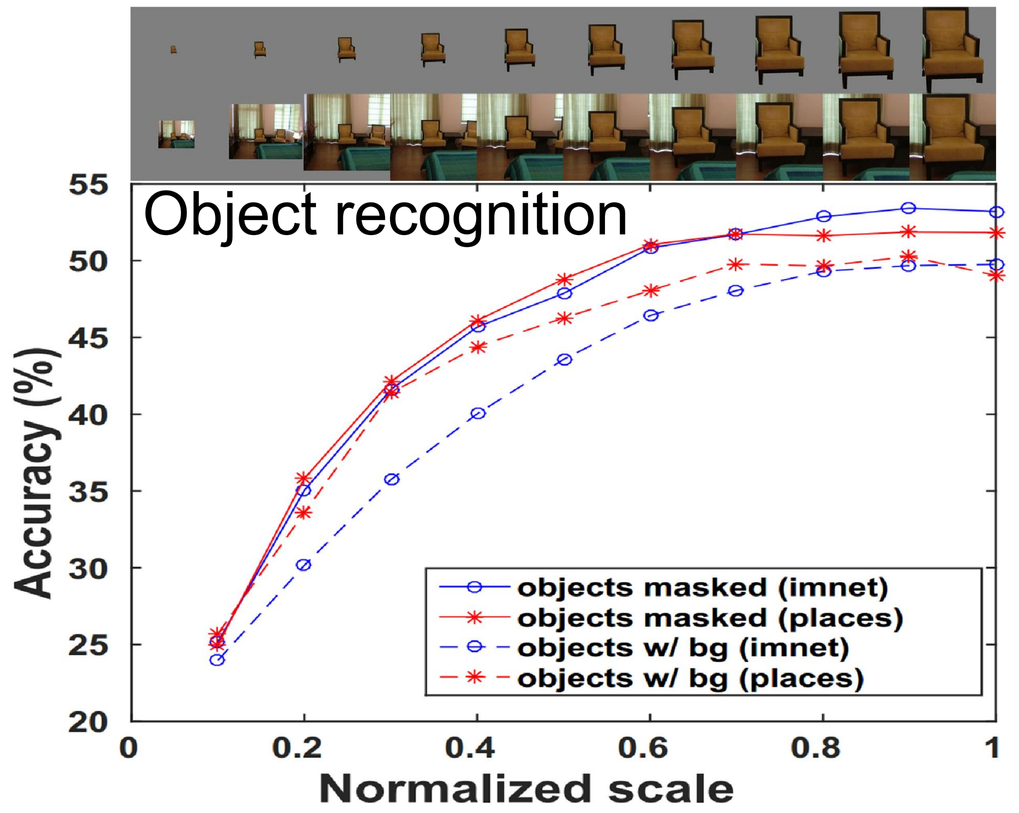
- **Scale** is one of the main factors



Effects of scaling:

- Changes the distribution of visual features
- Content in patches shifts from scenes to objects



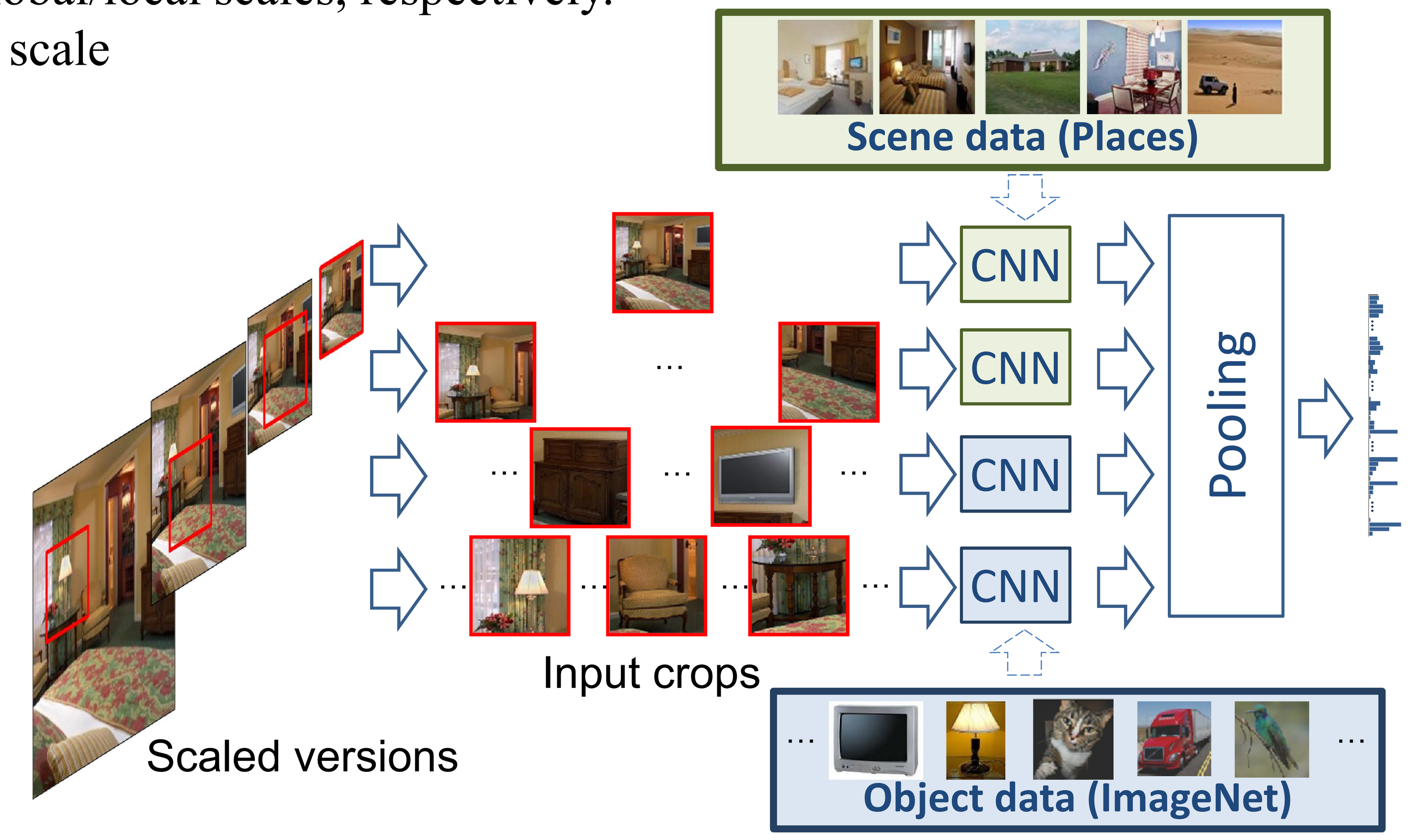


Multi-scale architecture with scale-specific CNNs

How to **correct scale-induced bias?**

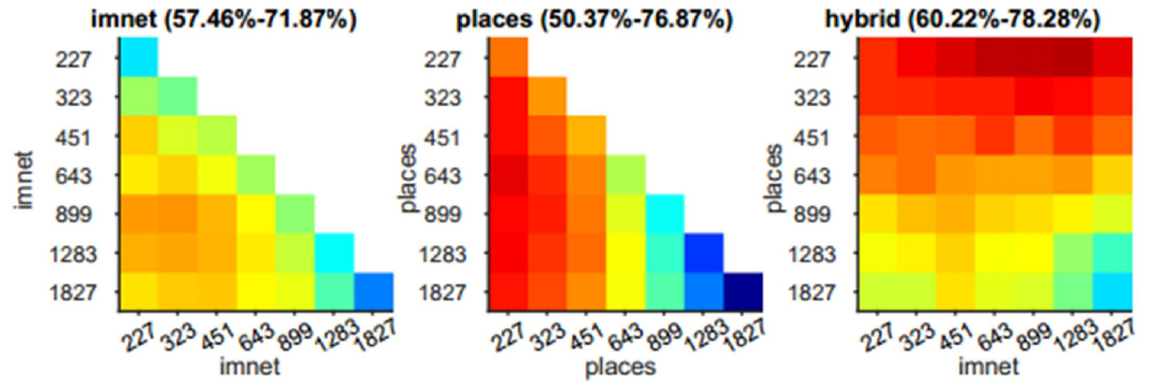
Scale-specific CNNs (instead of a fixed one) adapted to the patches at each scale. We study two ways:

- **Switch** Places-CNNs/ImageNet-CNNs, for global/local scales, respectively.
- **Fine tune** with patches extracted at the target scale

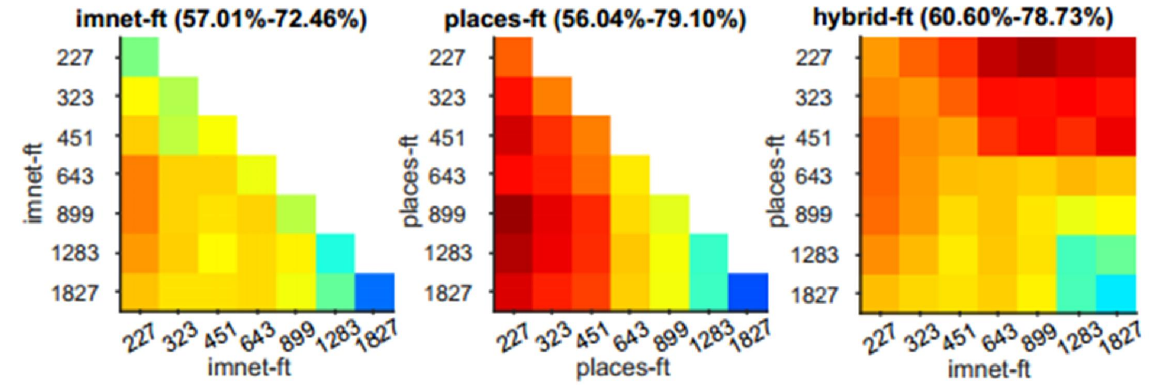


Experimental results

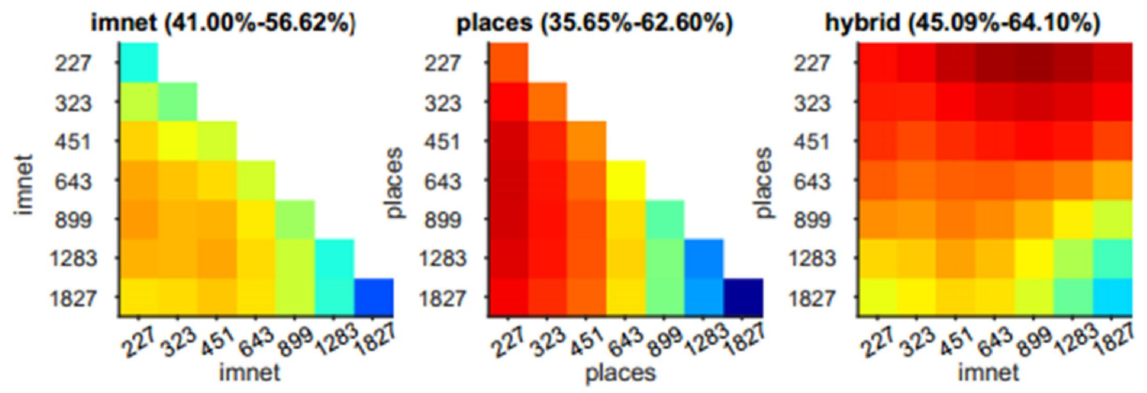
MIT Indoor 67 (dual architecture)



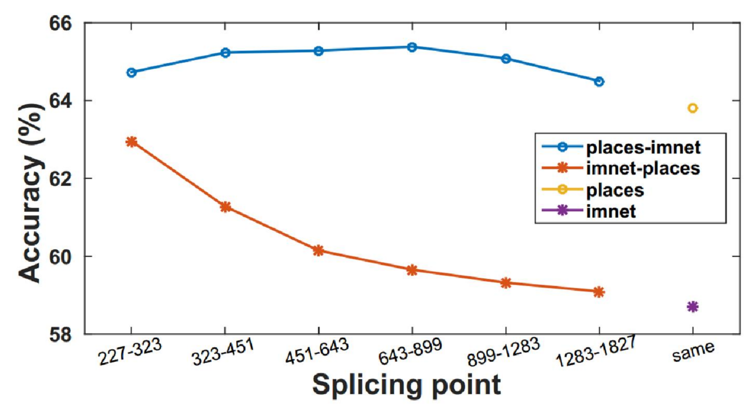
(dual architecture, fine tuned)



SUN397 (dual architecture)



(spliced architecture)



Architecture	Data	MIT Indoor 67 Alex	MIT Indoor 67 VGG	SUN 397 Alex	SUN 397 VGG[1]
Best single	IN	66.64	76.42	52.42	59.71
	PL	72.76	80.90	58.88	66.23
Dual	IN	71.87	79.04	56.62	61.07
	PL	76.87	83.43	62.60	68.49
Dual hybrid	IN/PL	78.28	85.59	64.10	69.20
Three	IN/PL	78.28	**86.04**	63.03	**70.17**
Full (7 scales)	IN	74.33	70.22	58.71	55.18
	PL	78.21	77.81	63.81	58.80
Full hybrid (spliced)	IN/PL	**80.97**	80.22	65.38	63.19
Double full hybrid	IN/PL	**80.97**	80.7	**66.26**	62.01
Human (good)[4]		-		68.5%	
Human (expert)[4]		-		70.6%	

[1] Six scales (1827x1827 was not included).

References

1. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, Learning deep features for scene recognition using Places database, In NIPS, 2014.
2. Y. Gong, L. Wang, R. Guo, and S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, In ECCV, 2014
3. M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene classification with semantic fisher vectors, In CVPR, 2015.
4. D. Yoo, S. Park, J.-Y. Lee, and I. So Kweon. Multiscale pyramid pooling for deep convolutional representation, In CVPR Workshops, 2015