# Knowledge Guided Disambiguation for Scene Recognition with Multi-Resolution CNNs

Sheng Guo, Limin Wang, Bowen Zhang, Yu Qiao

Shenzhen Institutes of Advanced Technology,  CAS, China

June 26, 2016

# Outline

1. Introduction
2. Knowledge Guided Disambiguation
3. Multi-Resolution CNNs
4. Experiments
5. Conclusions

# Introduction

Figure: Examples of ten scene categories.

# Introduction

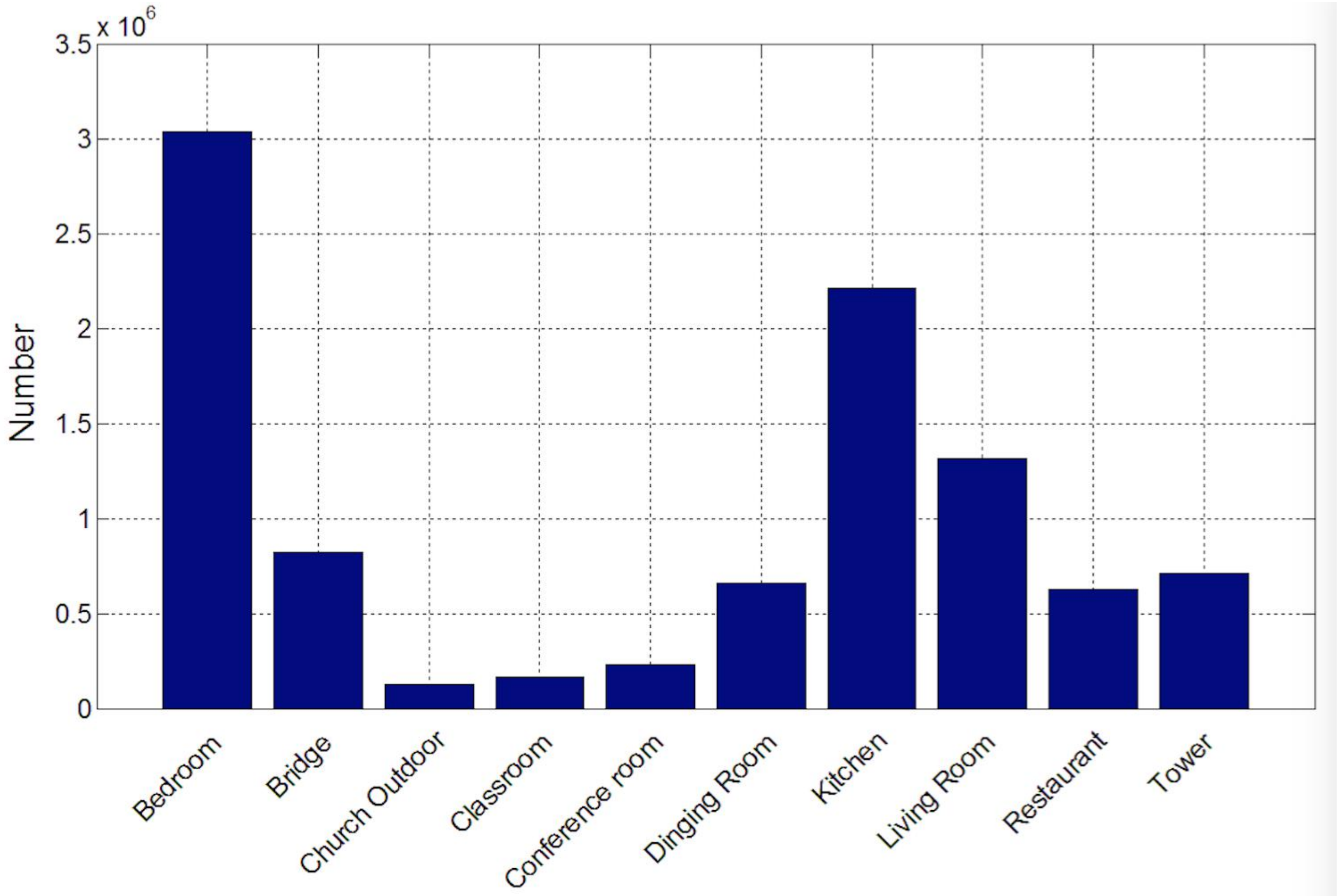
- **LSUN scene classification challenge:**
  - 10 scene categories, each class containing from 126,227 to 3,033,042 images.
  - 10M images for training, 3k images for validation and 10k images for testing.
  - The number of images is much bigger than ImageNet and Places2.
- **Scene classification is more challenging:**
  - The concept of scene is more subjective and high level than object.
  - The number of each class images variations (classes imbalance)
  - Large intra-class variations (visual inconsistency).
  - Small inter-class variations (label ambiguity).

Our methods are pre-trained on imagenet, place and place2.

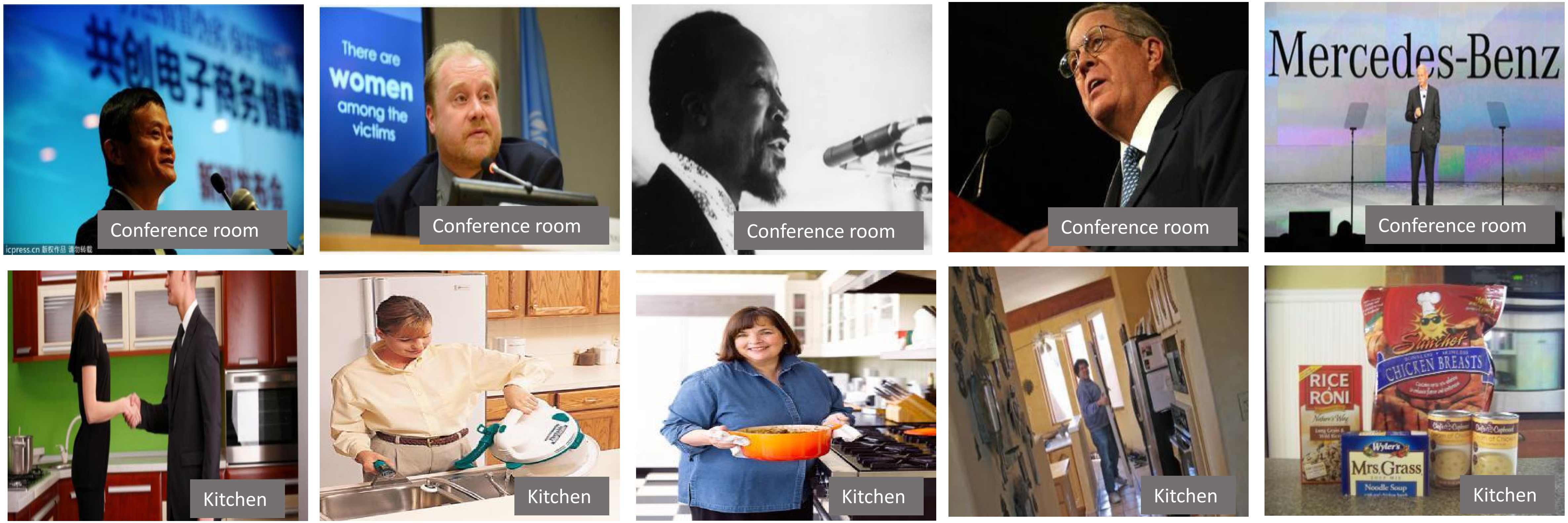
B. Zhou, A. Khosla, A. Lapedriza, A. Torralba and A. Oliva ,
*Places2: A Large-Scale Database for Scene Understanding*, in *Arxiv, 2015*

# Introduction

Classes Imbalance



## Visual Inconsistency



## Label Ambiguity (Example 1)

Church Outdoor


Tower

# Introduction

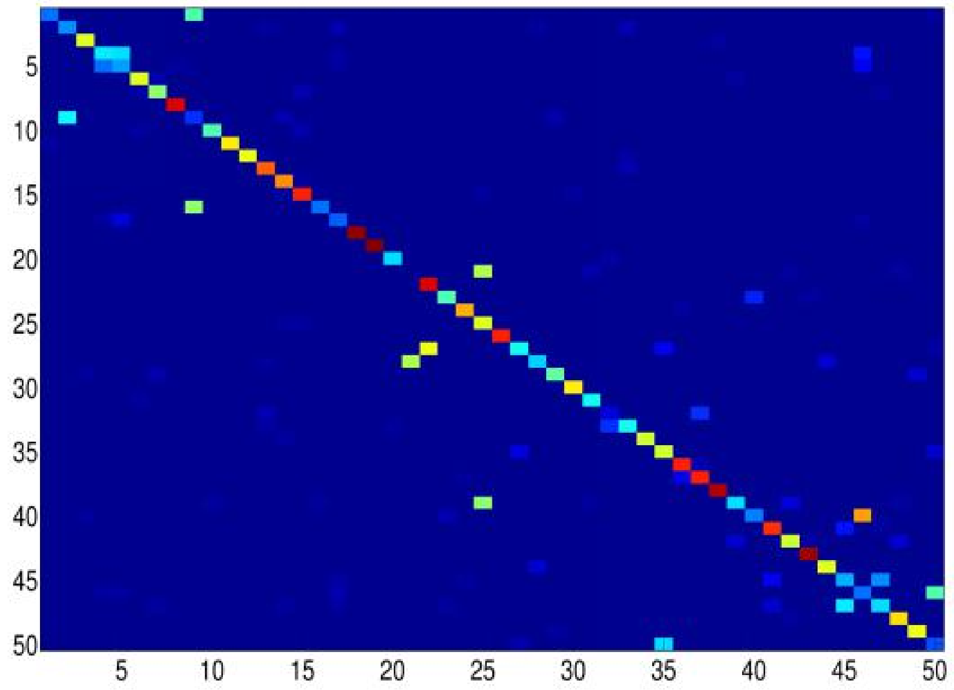
## Label Ambiguity (Example 2)

Conference Room

Classroom

1. Introduction
2. Knowledge Guided Disambiguation
3. Multi-Resolution CNNs
4. Experiments
5. Conclusions

# Knowledge Guided Disambiguation



cubicle/office
office cubicles → Super category1

athletic outdoor
track outdoor → Super category2

baseball field
stadium baseball → Super category3

……

waterfall plunge
waterfall block
waterfall fan → Super category n

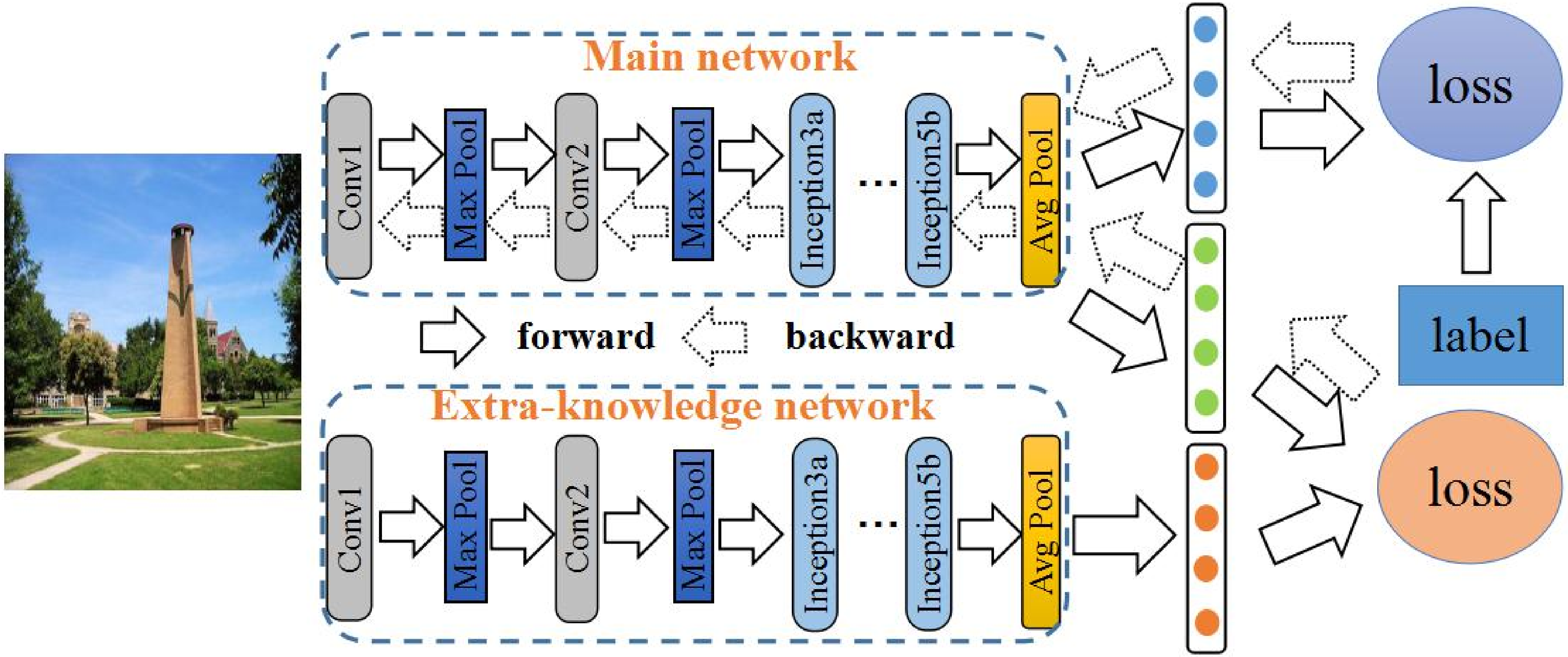
**401 original categories** → **372 super categories**

**Knowledge from confusion matrix**

# Knowledge Guided Disambiguation

- We propose a hierarchical strategy to merge similar categories into a super category, according to the confusion matrix on the validation data.
- The images of different scene categories, that belong to the same super category, will be given the same label.
- Totally, we reduce the number of scene categories from the Places2 dataset into 372 super categories.

# Knowledge Guided Disambiguation




**Knowledge from networks trained on other datasets**

# Knowledge Guided Disambiguation

- In previous scenario, all the images belonging to the same super category are constrained to have the same label, without considering the difference between images.
- We propose to automatically assign a soft code to each image, which is able to better encode the visual information of natural images.
- In the soft code space, the images from easily confused categories are equipped with similar codes.
- Finally, we design a multi-task framework to predict both hard code and soft code.

1 Introduction

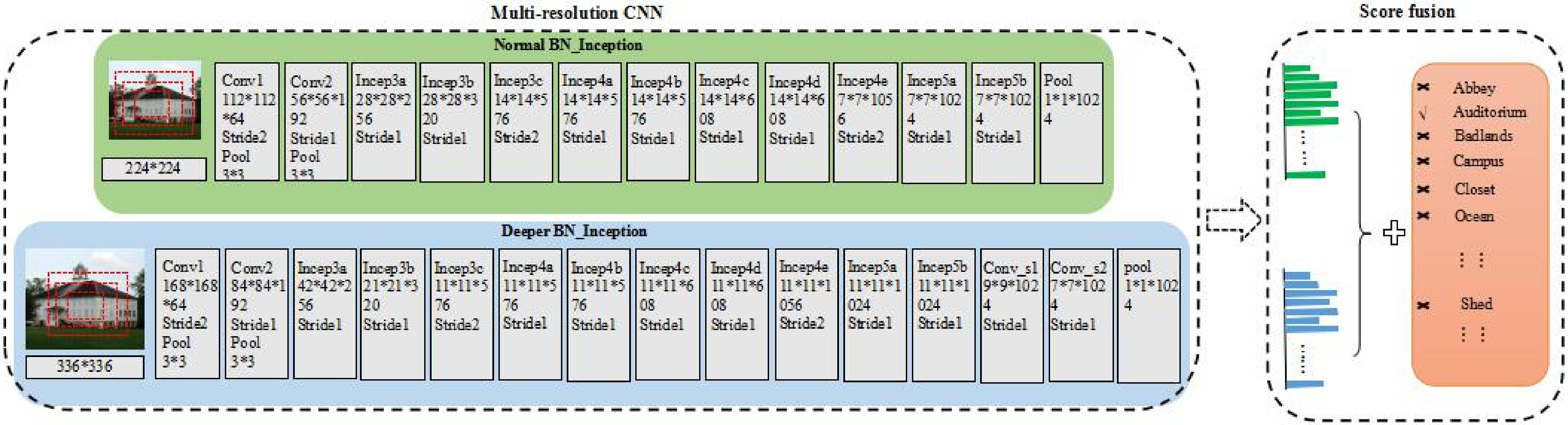
2 Knowledge Guided Disambiguation

3 **Multi-Resolution CNNs**

4 Experiments

5 Conclusions

# Multi-Resolution CNNs



# Multi-Resolution CNNs

Implementation details

- **Architectures:**
  - Low resolution: image (256*256), crop(224*224), inception2 network [2]
  - High resolution, image (384*384), crop(336*336), inception2+2 convs
- **Knowledge networks:**
  - Object nets: inception2 trained with ImageNet
  - Currently, knowledge disambiguation only for low resolution CNNs
- **Implementation details:**
  - Resample images to balance the class distribution
  - Data augmentation: fixed crop, scale jittering, horizontal flipping [1,6]

1. Introduction
2. Knowledge Guided Disambiguation
3. Multi-Resolution CNNs
4. Experiments
5. Conclusions

# Experiments (pretrained models)

| Method | Imagenet(top1/top5) | Places(top1/top5) | Places2(top1/top5) |
| --- | --- | --- | --- |
| AlexNet | 40.7%/18.2% | 50.0%/- | 57.0%/- |
| VGGNet | 27.0%/8.8% | 39.4%/11.5% | 52.4%/- |
| Normal BN-Inception | 24.7%/7.2% | 38.1%/11.3% | 48.8%/17.4% |
| Deeper BN-Inception | 23.7%/6.6% | 37.8%/10.7% | 48.0%/16.7% |
| Multi-resolution CNN | 21.8%/6.0% | 36.4%/10.4% | 47.4%/16.3% |

Table1. Performance of our pretrained models with Multi-Resolution CNNs on the validation data from the datasets of ImageNet, Places and Places2.

# Experiments (pretrained models)

| Method | Places2 (top5) |
|---|---|
| Normal BN-Inception (256 × 256) | 17.4% |
| Normal BN-Inception + object networks | 17.4% |
| Normal BN-Inception + scene networks | 16.7% |
| Normal BN-Inception + confusion matrix | 17.3% |
| Deeper BN-Inception (384 × 384) | 16.7% |
| Deeper BN-Inception + object networks | 16.3% |
| Deeper BN-Inception + scene networks | **16.1%** |

Table2. Performance of our pretrained models with different knowledge guided disambiguation techniques on the dataset of Places2.

# Scene Classification Results of Imagenet2015

| Rank | Team | Top5 | Rank | Team | Top5 |
|---|---|---|---|---|---|
| 1 | WM | 16.9 | 5 | NTU_Rose | 19.3 |
| 2 | Our (best) | 17.4 | 6 | Mitsubishi Electric | 19.4 |
| 3 | Qualcomm | 17.6 | 7 | HiVision | 20.0 |
| 4 | Trimps-Soushen | 18.0 | 8 | DeepSEU | 20.0 |

*Our SIAT_MMLAB team (Limin Wang, Sheng Guo, Weilin Huang and Yu Qiao) secures the 2nd place for scene recognition at ILSVRC 2015.*

# Experiments (LSUN)

- The challenge dataset contains 10 scene classes (7 indoor scene closses and 3 outdoor scene classes).
- The image numbers of different scene categories are very different, to balance data and calculation, we use 100,000 images from each category for training.
- As we can not access the label of evaluation data, we mainly train our models on the development data and report the results on the validation data.
- We finetune our challenge results on pretrain models that from Imagenet, Places and Places2.

# Experiments（LSUN)

| | Method | LSUN (top1) |
|---|---|---|
| A0 | inception_256(pretrain on Imagenet) | 89.73% |
| A1 | inception_384(pretrain on Imagenet) | 90.73% |
| | A0+A1 | 90.83% |
| A2 | inception_256(pretrain on Places2) | 89.90% |
| A3 | inception_384(pretrain on Places2) | 90.53% |
| | A2+A3 | 90.95% |
| A4 | incpetion_256(pretrain on Places) | 90.47% |
| A5 | inception_384(pretrain on Places) | 90.70% |
| | A4+A5 | 91.10% |
| A6 | inception_384_kd_object(pretrain on Places2) | 90.69% |
| | Fusion all | 91.77% |

Table3. Performance of our methods on the validation data.

# Experiments（LSUN)

## Misclassified Examples

dining_room
kitchen

dining_room
restaurant

bedroom
living_room


classroom
confer_room

# Experiments（LSUN)

## Experiments Results (Test Data)

| Rank | Team | Top1 |
|---|---|---|
| 1 | SIAT_MMLAB(our) | 91.61% |
| 2 | SJTU-ReadSense | 90.43% |
| 3 | TEG Rangers | 88.70% |
| 4 | ds-cube | 83.02% |
| | Google（last year winner) | 91.20% |

# Conclusions

- Large scale scene datasets with many categories come along with increased ambiguity between the class labels (e.g. bedroom vs.living room).
  - Knowledge guided disambiguation aims to regularize CNN training with extra knowledge and improve the generalization capacity.
- Scene or Places, defined by  containing objects, spatial layout, human events, and global contexts, are more high-level concepts.
  - Multi-Resolution CNNs take images of different sizes as input and capture visual information from different levels.

# Thank you!