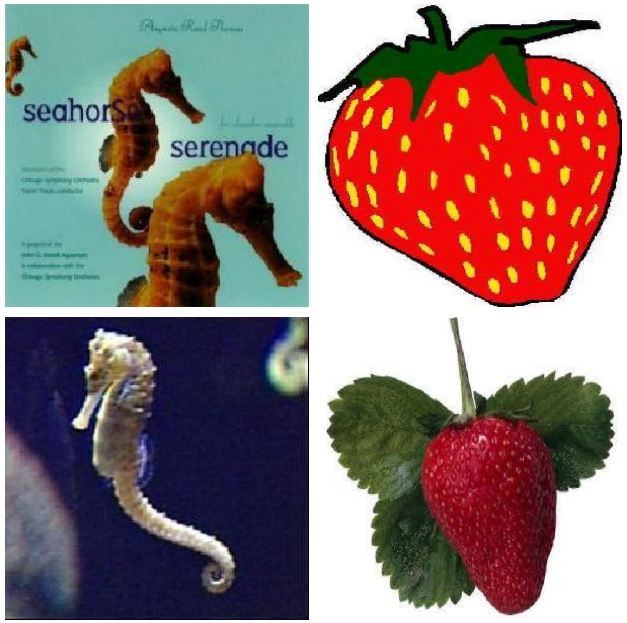
# See More for Scene: Pairwise Consistency Learning for Scene Classification

Gongwei Chen, Xinhang Song, Bohan Wang, Shuqiang Jiang
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, CAS
Beijing, China

# Introduction

## Scene Classification

- Identify an image as a scene concept, such as bedroom, rainforest.
- Demand: “seeing” more comprehensive and informative regions




**Object Images**

**Scene Images**

**Scene Characteristics**

- More semantic concepts
- No clear boundary
- Flexible spatial configuration

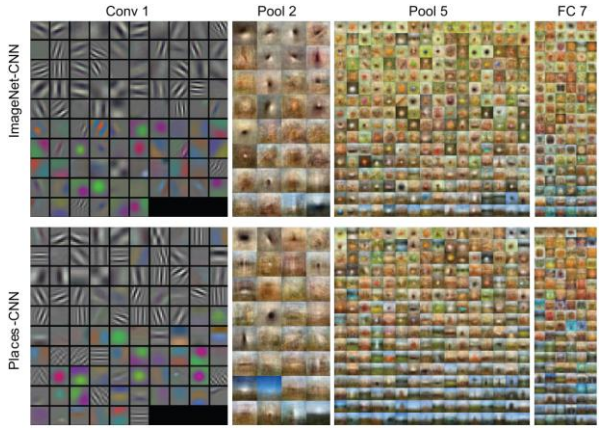
# Introduction

- Current scene classification methods
  - *Region Discovery*: unspecific[1] or specific regions[2]
  - *Region Aggregation*: statistical models[3] or relation modeling[4] methods
- Issues and challenges
  - Incompatibility
  - Inevitable computational consumption
  - Digging into the CNN properties for meeting scene demands?

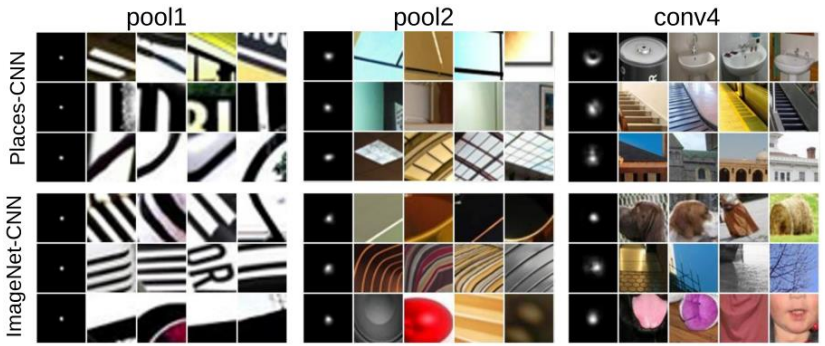

[1] M. D. Dixit and N. Vasconcelos, "Object based Scene Representations using Fisher Scores of Local Subspace Projections," in NeurIPS, 2016, pp. 2811–2819.
[2] Z. Zhao and M. Larson, "From Volcano to Toyshop: Adaptive Discriminative Region Discovery for Scene Recognition," in ACM MM, 2018.
[3] L. Liu et al., "Compositional Model Based Fisher Vector Coding for Image Classification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2335–2348, 2017.
[4] G. Chen, X. Song, H. Zeng, and S. Jiang, "Scene Recognition with Prototype-Agnostic Scene Layout," IEEE Trans. Image Process., vol. 29, pp. 5877–5888, 2020.

# Introduction

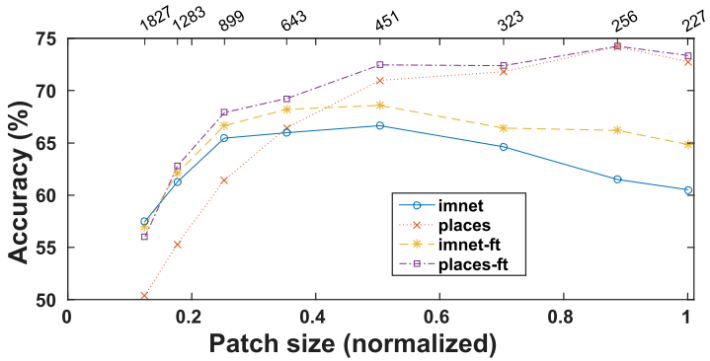
- Comparisons of scene and object classification models




**The mean image method[1]**




**Empirical receptive field[2]**



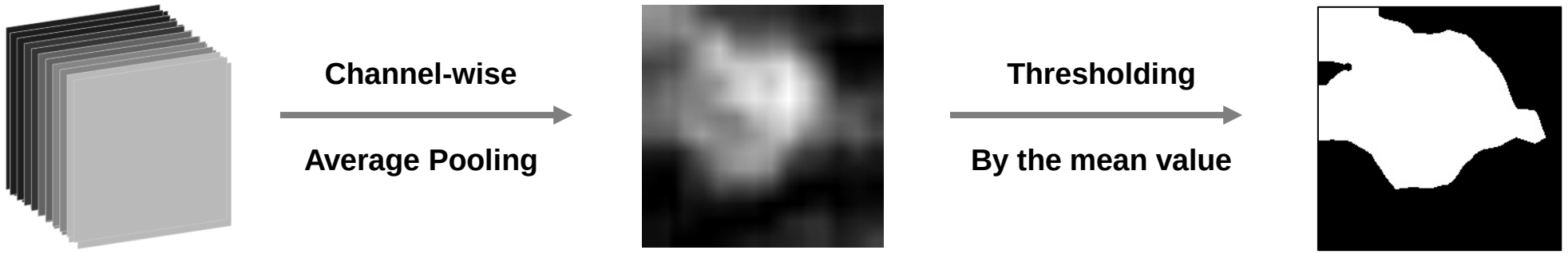

**Transfer results with Scales[3]**


[1] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database," in NIPS, 2014, pp. 487–495.
[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," in ICLR, 2015.
[3] L. Herranz, S. Jiang, and X. Li, "Scene Recognition with CNNs: Objects, Scales and Dataset Bias," in CVPR, 2016, pp. 571–579.

# Motivation

## The focus area

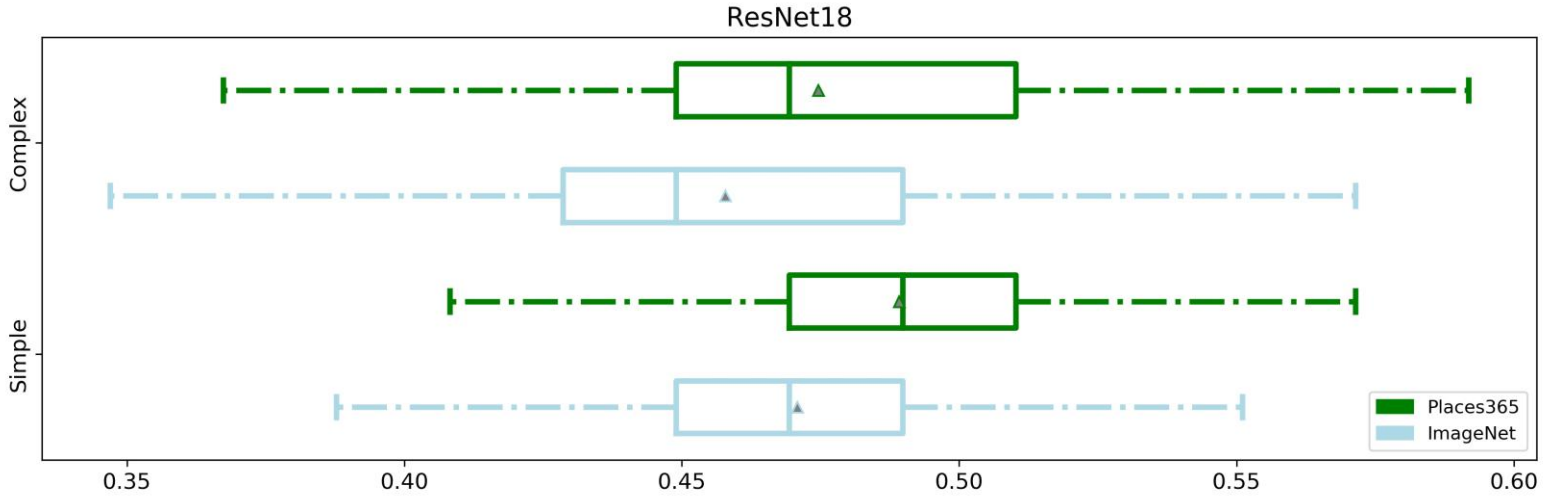
Regions that consist of pixels with larger aggregated activation values than the mean value.



# Motivation

## Object versus Scene Networks

### The distribution of coverage ratio of the focus area




Center line: median
triangle: mean

# large focus area

Scene → More
Object → Less

Simple Train → More
Complex Train → Less

### The classification results

| | ResNet18 | | ResNet50 | |
|---|---|---|---|---|
| | Simple | Complex | Simple | Complex |
| ImageNet | 68.80 | 69.93 | 74.72 | 75.78 |
| Places365 | 54.36 | 54.30 | 55.53 | 55.75 |

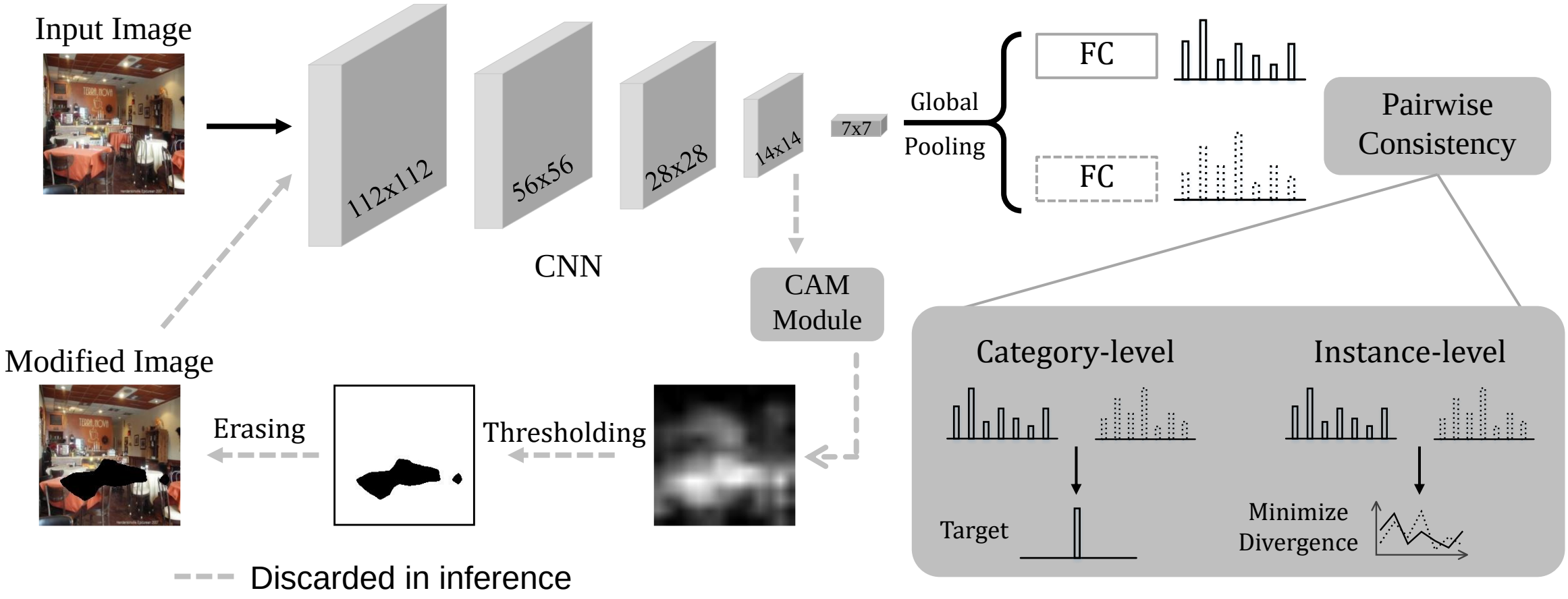
→ Inferior Results
→ Similar Results

**The optimal training method considering scene characteristics**

?

# Method

- The overview of pairwise consistency learning method



# Method

- **The final loss function**

$$L = l_{main} + l_{side} + \alpha l_{mod} + \beta l_{kl}$$
$$= \sum_{k=1}^{K} -g_k \log(p_k) + \sum_{k=1}^{K} -g_k \log(r_k)$$
$$+ \alpha \sum_{k=1}^{K} -g_k \log(q_k) + \beta \sum_{k=1}^{K} -q_k \log\left(\frac{q_k}{p_k}\right)$$

Category-level Pairwise Consistency, CPC:

$$l_{main} + l_{side} + \alpha l_{mod}$$

Instance-level Pairwise Consistency, IPC:

$$l_{main} + l_{side} + \beta l_{kl}$$

# Experiments

- Datasets
  - *Places365*:
    - 365 scene categories with 1.8 million training images, and 36500 validation images.
  - *Places365-small*:
    - a subset of Places365, 365,000 training images.
  - *ImageNet*:
    - 1000 object categories with 1.3 million training images, and 50,000 validation images.

**Places365 Examples**

**ImageNet Examples**

# Experiments

## Main Results

| Dataset | Model | Param. | RF | Baseline | | Our Method | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Top-1 | Top-5 | Top-1 | Top-5 | |
| Places365 -small | ShuffleNetV2 | 2.85M | 527 | 47.67 | 78.68 | 48.91 | 79.76 | |
| | ResNet18 | 11.36M | 435 | 48.21 | 79.22 | 49.70 | 80.62 | |
| | ResNet34 | 21.47M | 899 | 48.52 | 79.58 | 50.72 | 81.31 | ↑ 2.20% |
| | ResNet50 | 24.26M | 427 | 49.66 | 80.67 | 50.92 | 81.70 | |
| | DenseNet121 | 7.33M | 2071 | 49.48 | 80.87 | 51.55 | 82.20 | ↑ 2.07% |
| Places365 | ShuffleNetV2 | 2.85M | 527 | 54.01 | 84.05 | 54.95 | 84.56 | |
| | ResNet18 | 11.36M | 435 | 54.36 | 84.50 | 55.12 | 84.94 | |
| | ResNet34 | 21.47M | 899 | 54.84 | 84.72 | 56.23 | 85.79 | ↑ 1.39% |
| | ResNet50 | 24.26M | 427 | 55.53 | 85.85 | 56.61 | 86.15 | |
| | DenseNet121 | 7.33M | 2071 | 55.40 | 85.50 | 56.84 | 86.44 | ↑ 1.44% |

**Larger receptive field → Better performance improvement**

# Experiments

- Comparison with object classification

| Dataset | Places365-small | Places365 | ImageNet |
|---|---|---|---|
| Baseline | 47.80 | 54.30 | 69.93 |
| Ours | 48.68 | 54.78 | 69.57 |

Superior Inferior

- More adaptive to Scene classification

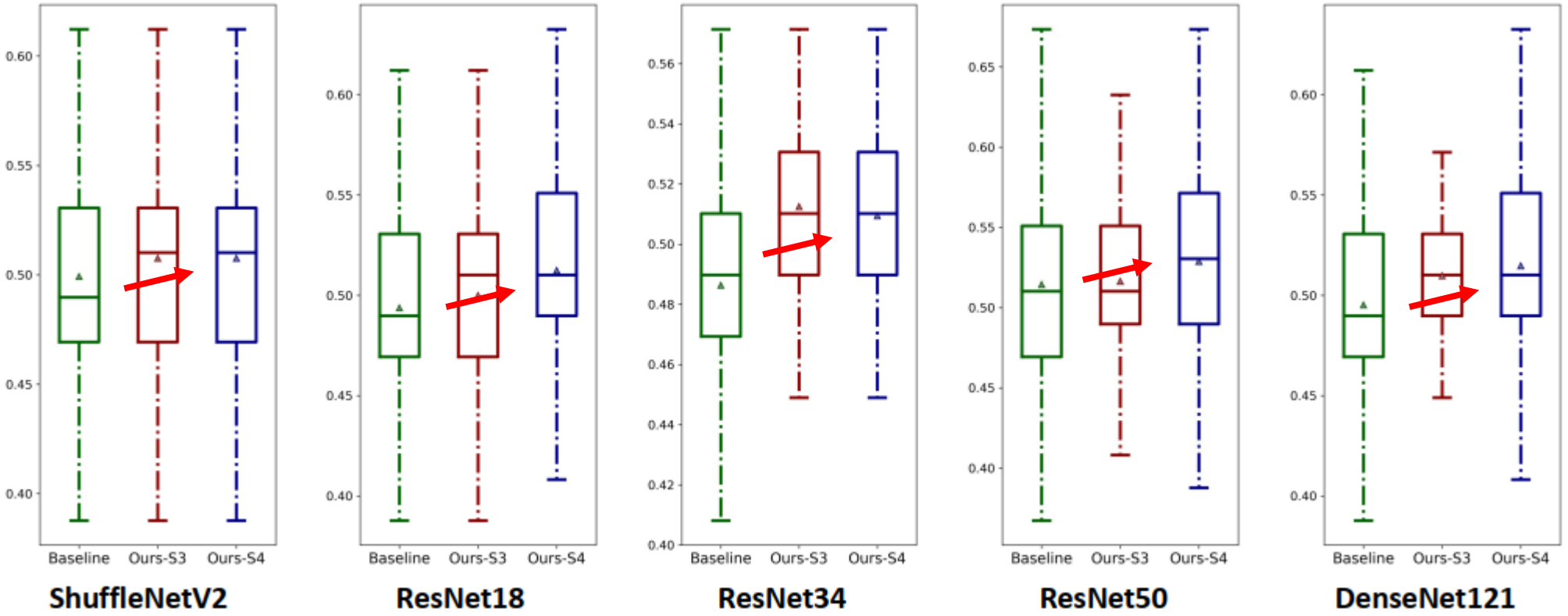
- Comparison with other erasing methods

| Methods | HaS | ADL | ACoL | Ours |
|---|---|---|---|---|
| ResNet18 | 49.25 | 48.70 | 49.03 | 49.70 |
| ResNet34 | 49.91 | 48.68 | 49.55 | 50.72 |

- Superiority from Pairwise Consistency

# Experiments

- **The analyses of the focus area**
  - Our method: Large focus area on **more** images




Center line: median  triangle: mean

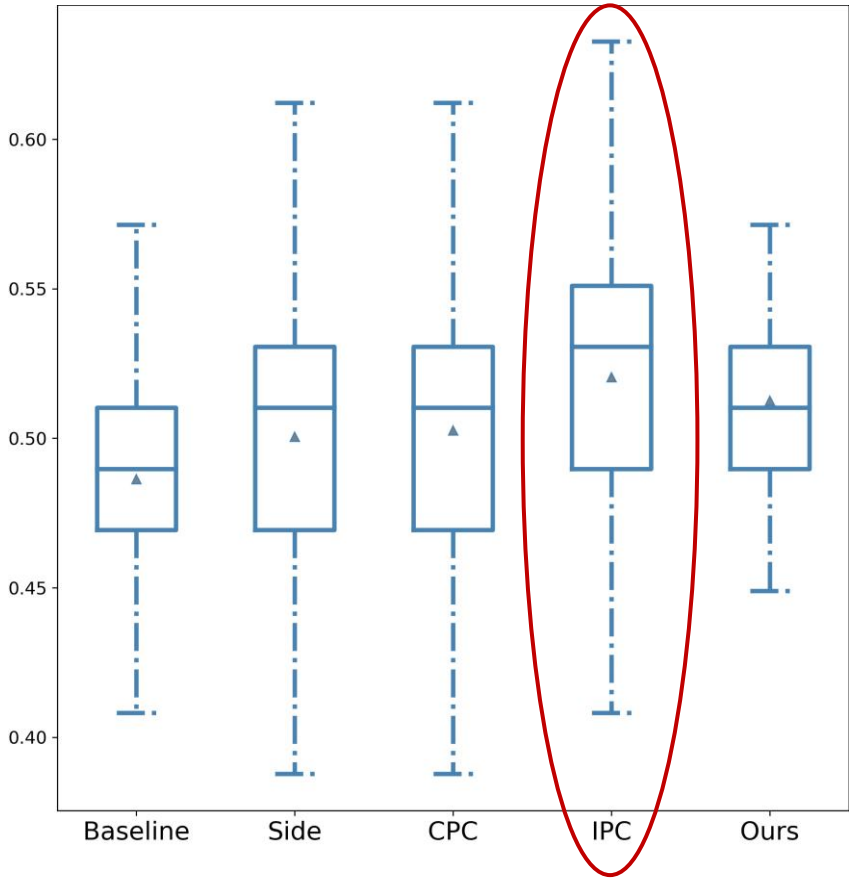
# Experiments

## Comparisons of different loss items

**Classification Results on Places365-small**

| Side | CPC | IPC | ResNet | | |
|---|---|---|---|---|---|
| | | | 18 | 34 | 50 |
| ✓ | - | - | 48.75 | 48.76 | 49.52 |
| ✓ | ✓ | - | 49.12 | 49.92 | 50.43 |
| ✓ | - | ✓ | 49.65 | 50.36 | 50.85 |
| ✓ | ✓ | ✓ | 49.70 | 50.72 | 50.92 |
| Baselines | | | 48.21 | 48.52 | 49.66 |

- Combing CPC and IPC yields a slightly better and robust model
- IPC: Superior ability of expanding the focus area.



# Conclusions

- We investigated the CNN classification models in terms of the *focus area* and show the difference between scene and object networks.
- We proposed a new learning framework with a tailored loss to force CNN to expand the focus area for improving scene classification.
- Experiments on Places365 and ImageNet verify the effectiveness of our approach, and also indicate that it is specifically designed for scenes by capturing their unique attributes.

# Thank you!


中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences