



# TWO-STAGE POOLING OF DEEP CONVOLUTIONAL FEATURES FOR IMAGE RETRIEVAL



Tiancheng Zhi, Ling-Yu Duan, Yitong Wang, Tiejun Huang  
Institute of Digital Media, School of EE&CS, Peking University

## Background & Motivation

- Convolutional Neural Network (CNN) based image representations have achieved high performance in image retrieval tasks.
- Drawbacks of traditional CNN based global representations include:
  - Provide high-dimensional features
  - Inadequately capture discriminative information in images
- Our goal is:
  - Tackle the limitation of traditional max/mean pooling approach
  - Construct compact and discriminative global feature representations

## Contribution

- We propose an effective and efficient two-stage partial mean pooling (PMP) strategy and embed PMP into an advanced feature extraction framework.
- Experiments show that:
  - The proposed PMP method is superior to max/mean pooling in both intra-patch stage and inter-patch stage.
  - The proposed representation and significantly improves the state-of-the-art retrieval accuracy on benchmark datasets.

## Approach

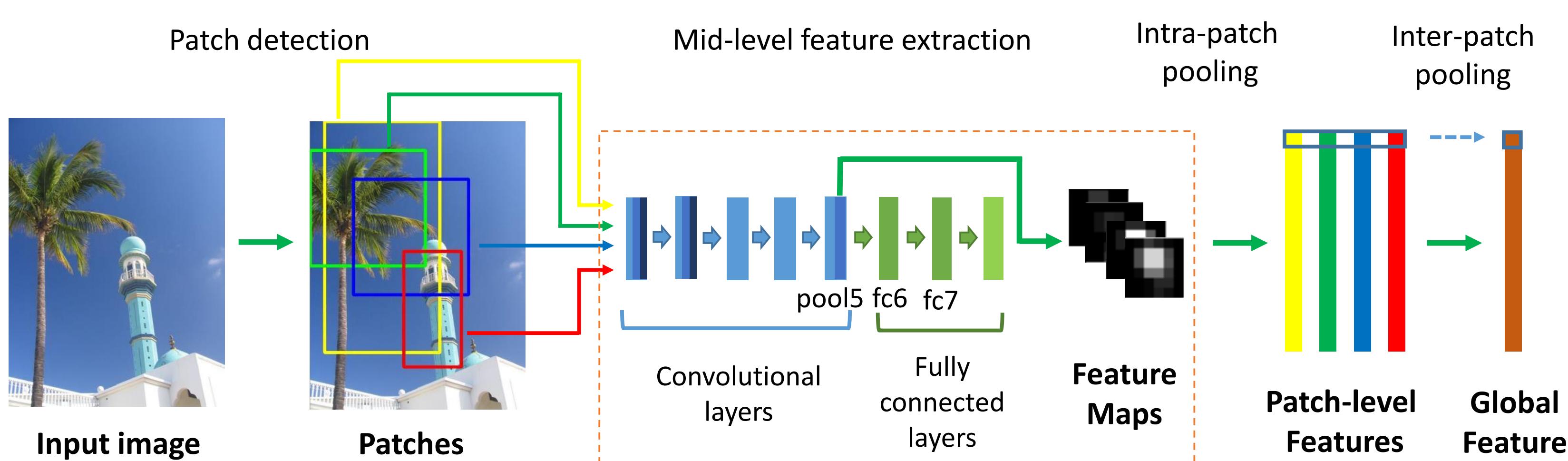


Fig 1. Feature extraction framework

### Step 1: Patch Detection

Use object proposal algorithm to detect  $N$  regions with high objectness  
Handle multiple scales, noisy background and abundant subjects

### Step 2: Mid-level Feature Extraction

Extract the last convolutional/pooling layer feature maps  
Sort the feature within each feature map to avoid hard coding of location

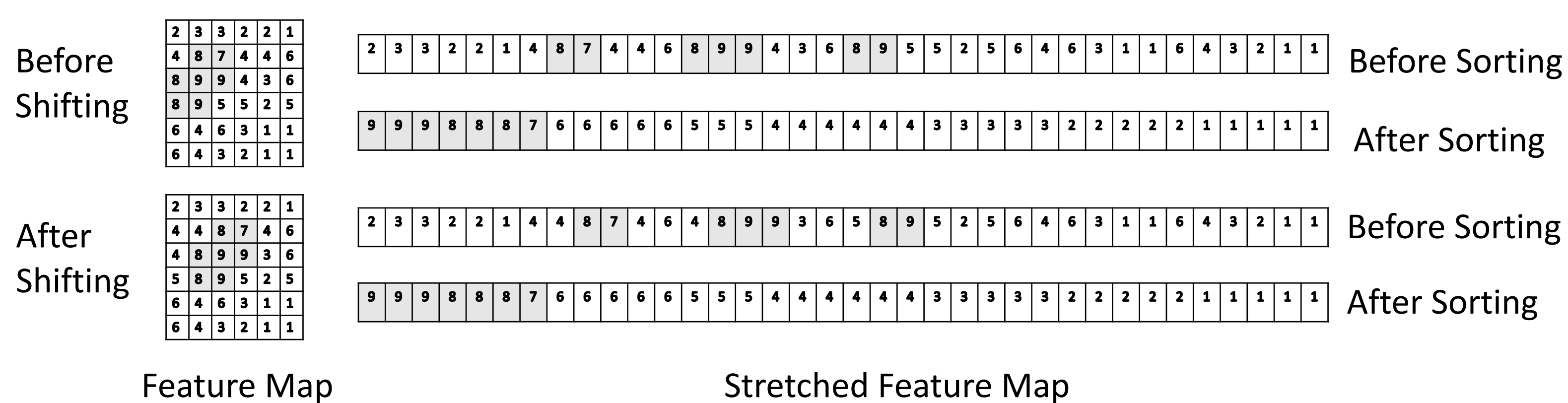


Fig 2. Sorting provides shift-invariance

### Step 3: Two-stage Partial Mean Pooling

#### 1) Intra-patch Partial Mean Pooling

Compute mean value of top  $K_1$  responses for each feature map  
Capture the discriminative responses on the feature maps  
Transform the feature to a low-dimensional representation

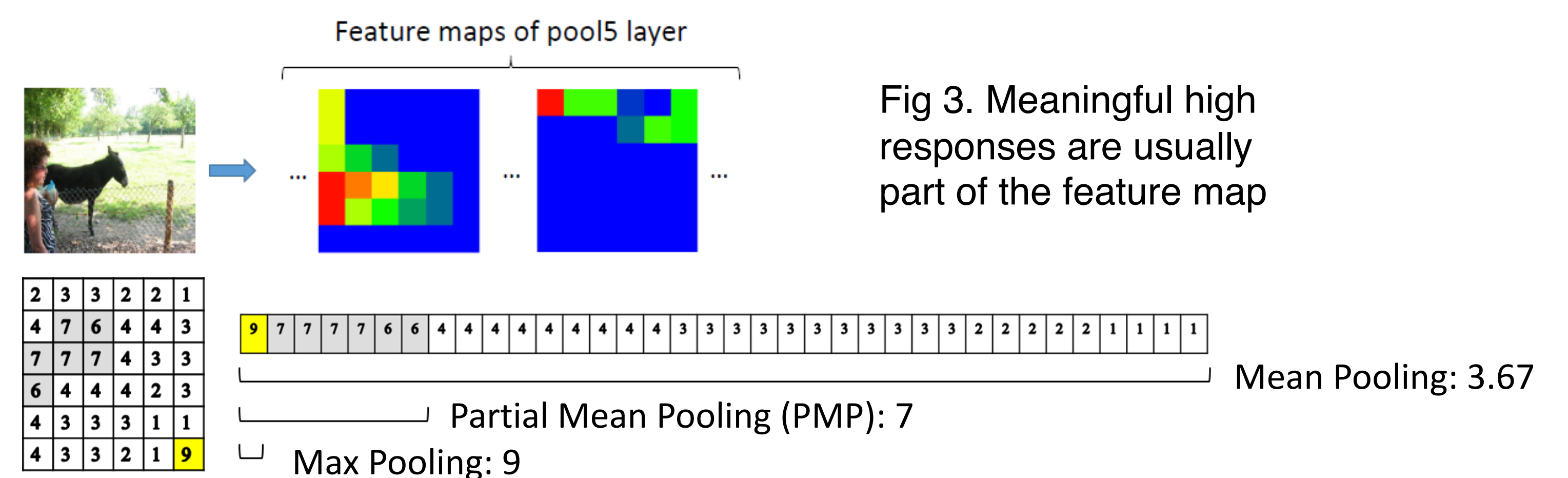


Fig 4. PMP reduces effects of background information and noisy high response

#### 2) Inter-patch Partial Mean Pooling

Compute mean value of top  $K_2$  patch features along each dimension  
Aggregate patch-level features into compact global representation

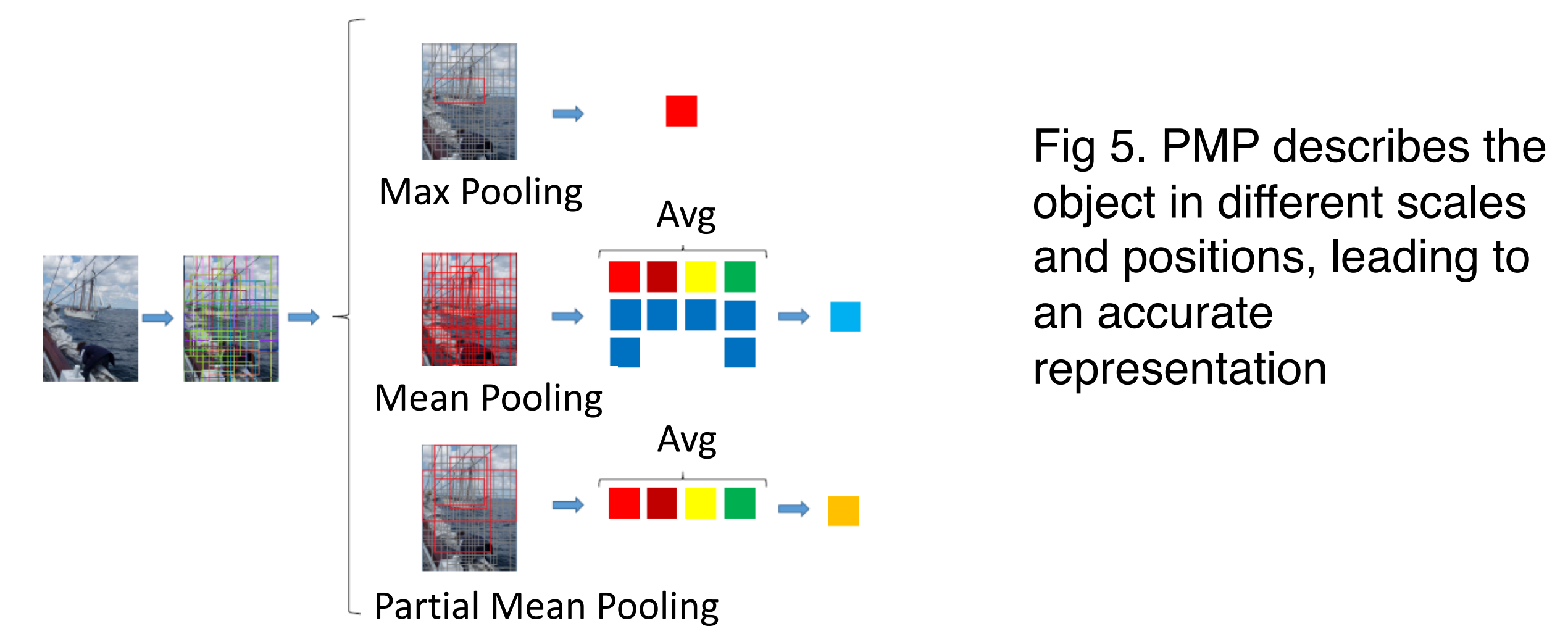


Fig 5. PMP describes the object in different scales and positions, leading to an accurate representation

## Experiment Results

### Impact of Parameters

$N$ : number of detected patches  
 $K_1$ : parameter in intra-patch partial mean pooling  
 $K_2$ : parameter in inter-patch partial mean pooling

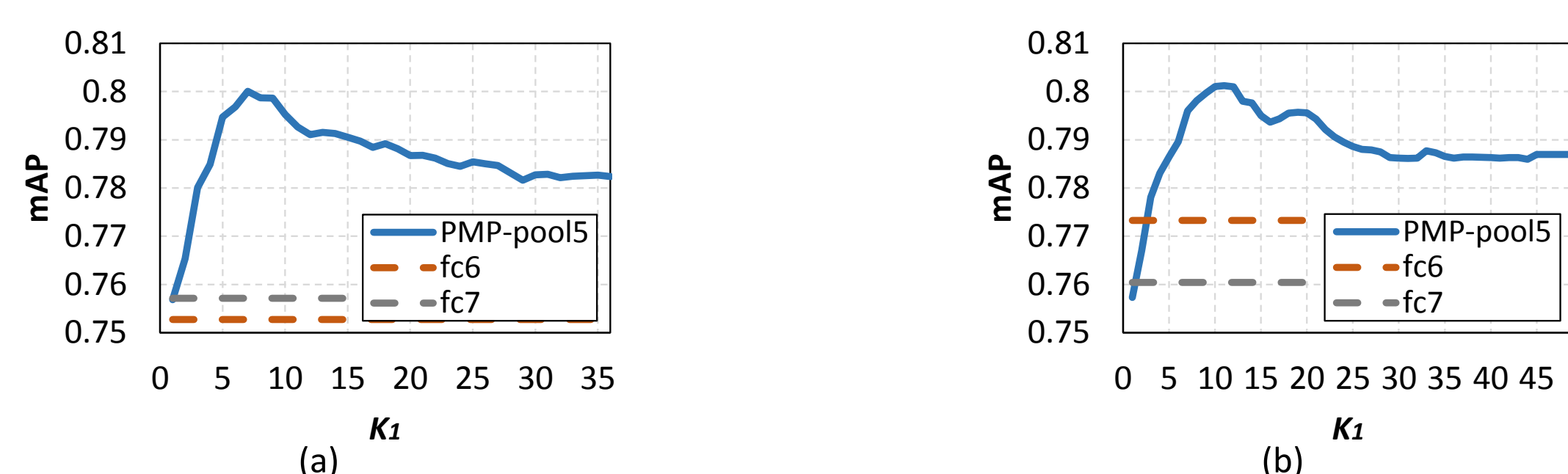


Fig 6. mAP on Holidays dataset (impact of  $K_1$ , (a) CaffeNet, (b) VGGNet)

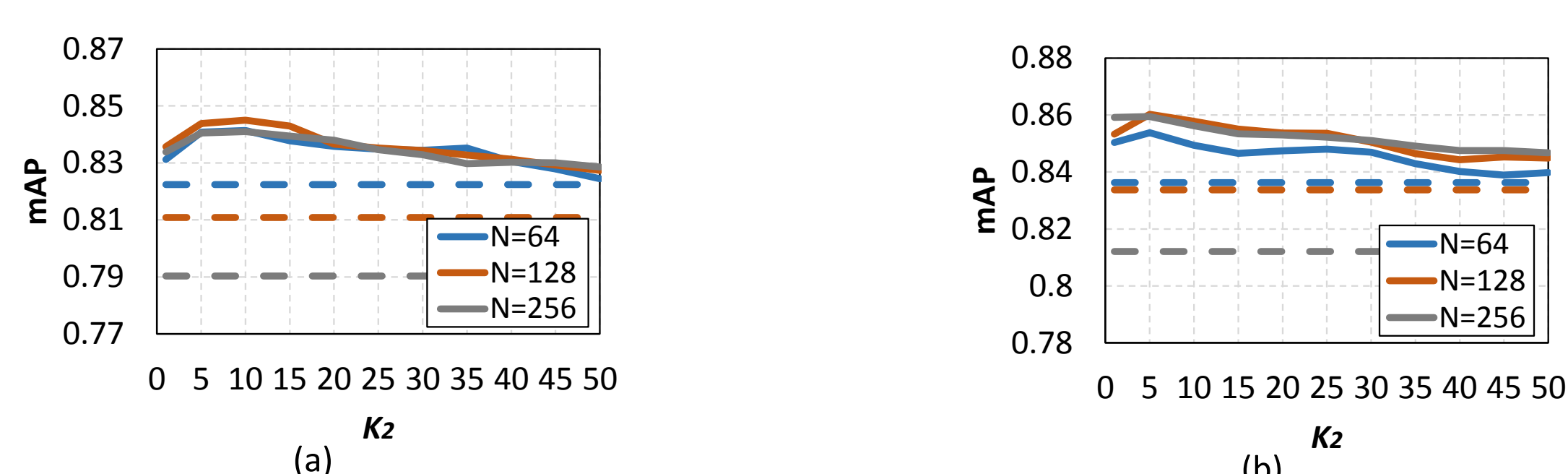


Fig 7. mAP on Holidays dataset (impact of  $N$  and  $K_2$ , (a) CaffeNet, (b) VGGNet)

### Benchmark Results

Method	Dimension	Holidays	UKBench	Oxford5K
SIFT + TE + DA	1024	72.0	3.51	56.0
Neural Codes	4096	79.3	-	-
MOP-CNN + PCA	512	78.3	-	-
CNN + SPoC	256	80.2	3.65	58.9
CNN + DPS + Max Pooling	4096	81.0	3.67	56.0
Two-stage PMP (VGGNet, with PCA-whitening)	512	<b>86.6</b>	<b>3.80</b>	<b>64.0</b>

Fig 8. Performance measured by mean average precision (mAP)

### Example Retrieval Result

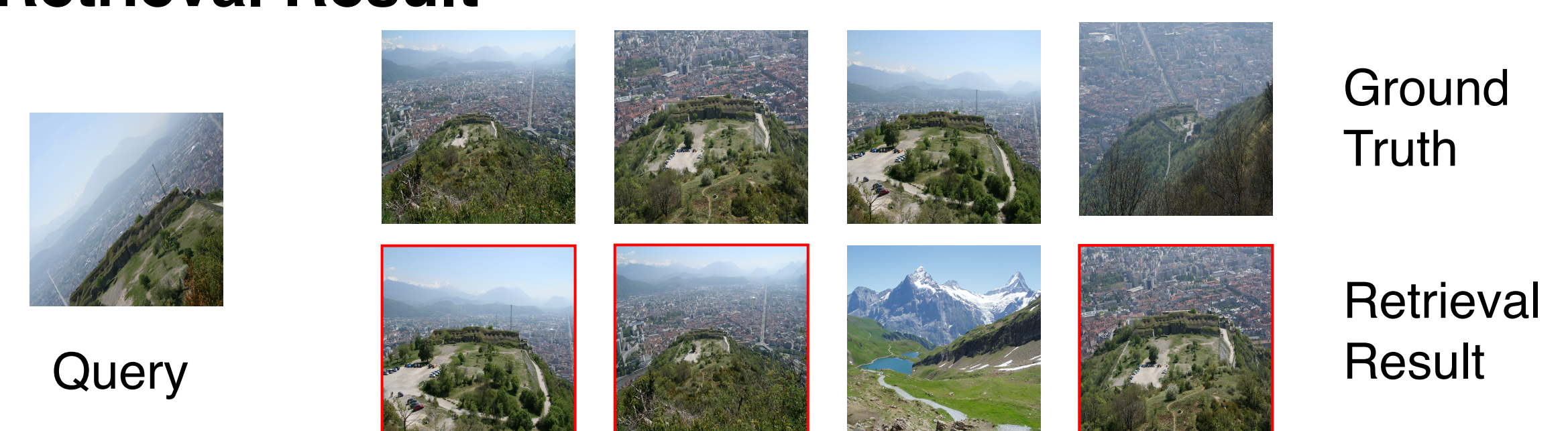


Fig 9. An example from Holidays dataset

## Conclusions & Future Work

- We have proposed a two-stage partial mean pooling strategy towards an advanced CNN feature extraction framework.
- The proposed compact and discriminative image representation outperforms state-of-the-art methods.
- How to incorporate low-level invariant features into this feature extraction framework will be included in our future work.