ECCV'20 ONLINE

23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

Rethinking Class Activation Mapping for Weakly Supervised Object Localization

Wonho Bae*

Junhyug Noh*

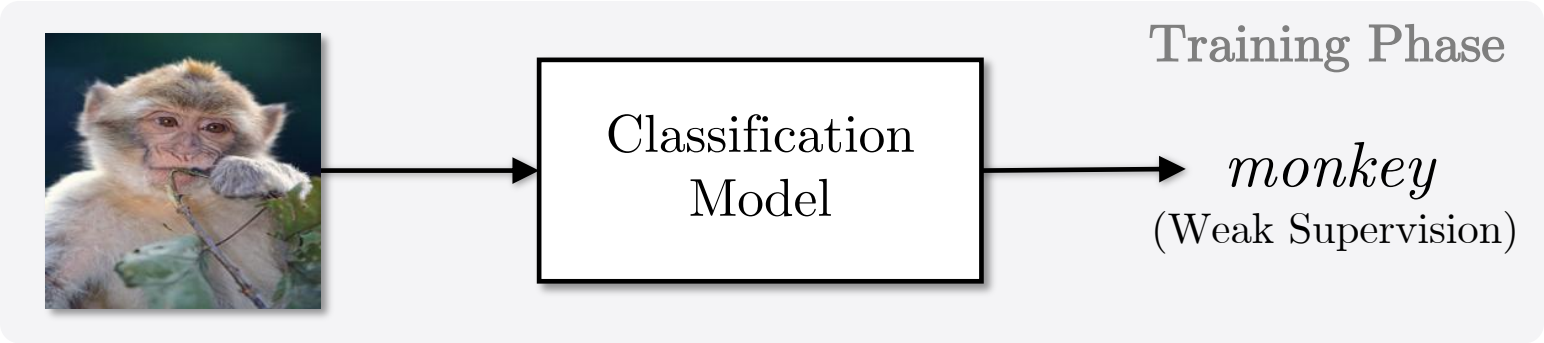
Gunhee Kim

SEOUL NATIONAL UNIV.
VISION & LEARNING

* Equal contributions

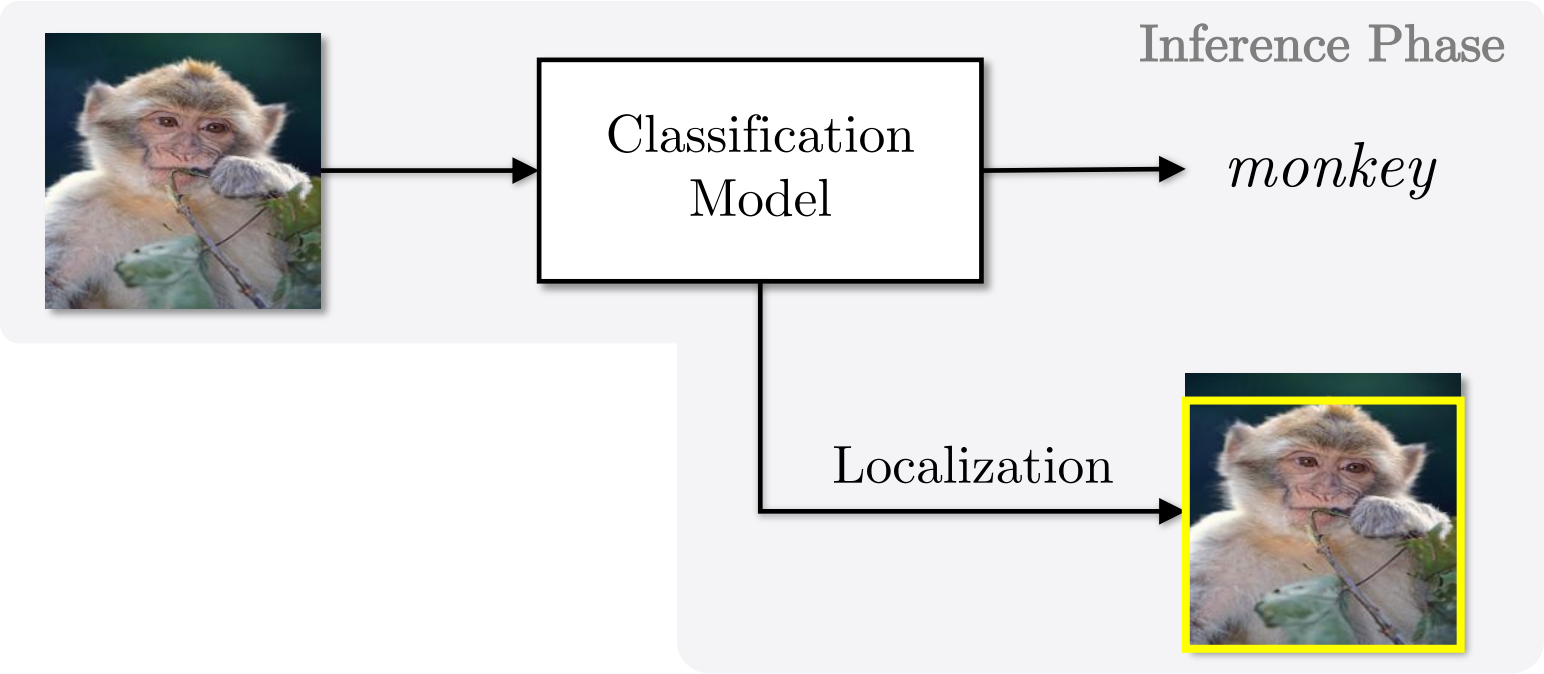
Weakly Supervised Object Localization (WSOL)

- **Goal:** To localize an object only using **image-level class labels** (no annotations for object location is provided).



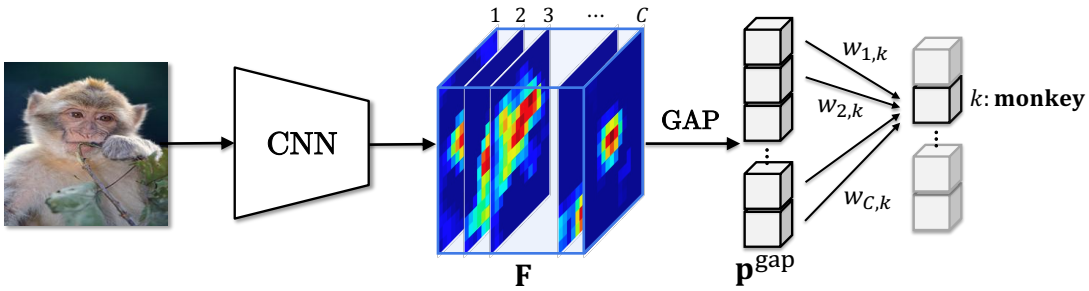
Weakly Supervised Object Localization (WSOL)

- **Goal:** To localize an object only using **image-level class labels** (no annotations for object location is provided).



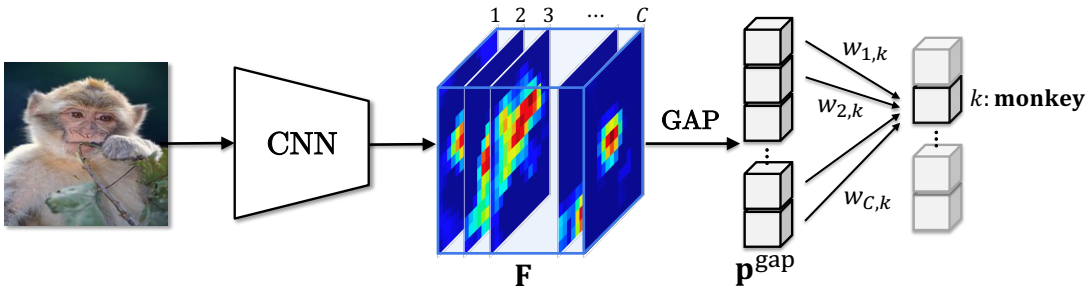
Class Activation Mapping (CAM)

1. Train a classification model.



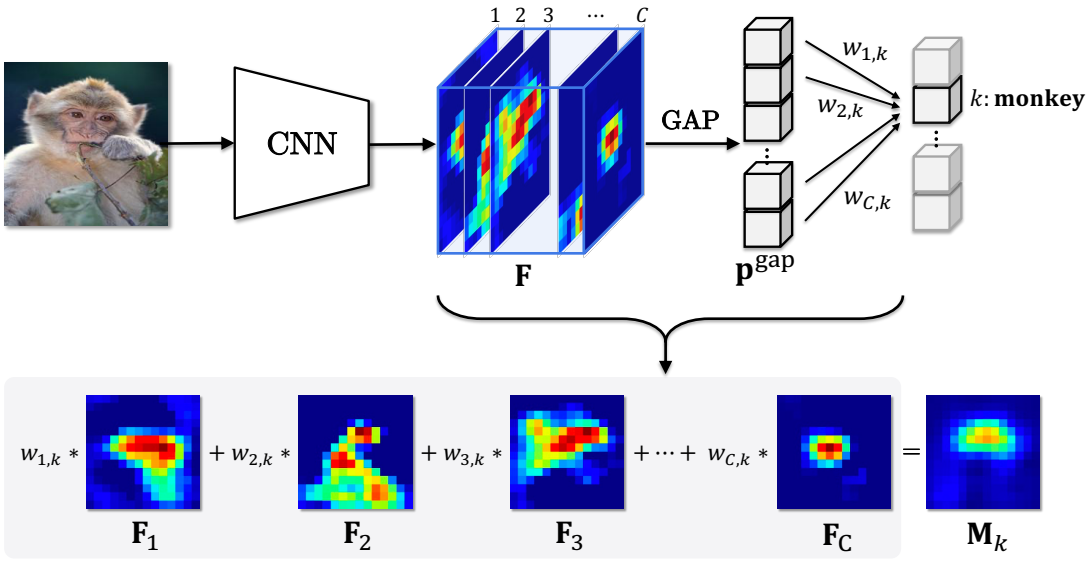
Class Activation Mapping (CAM)

1. Train a classification model.
2. Based on the classification model, apply CAM to localize an object.



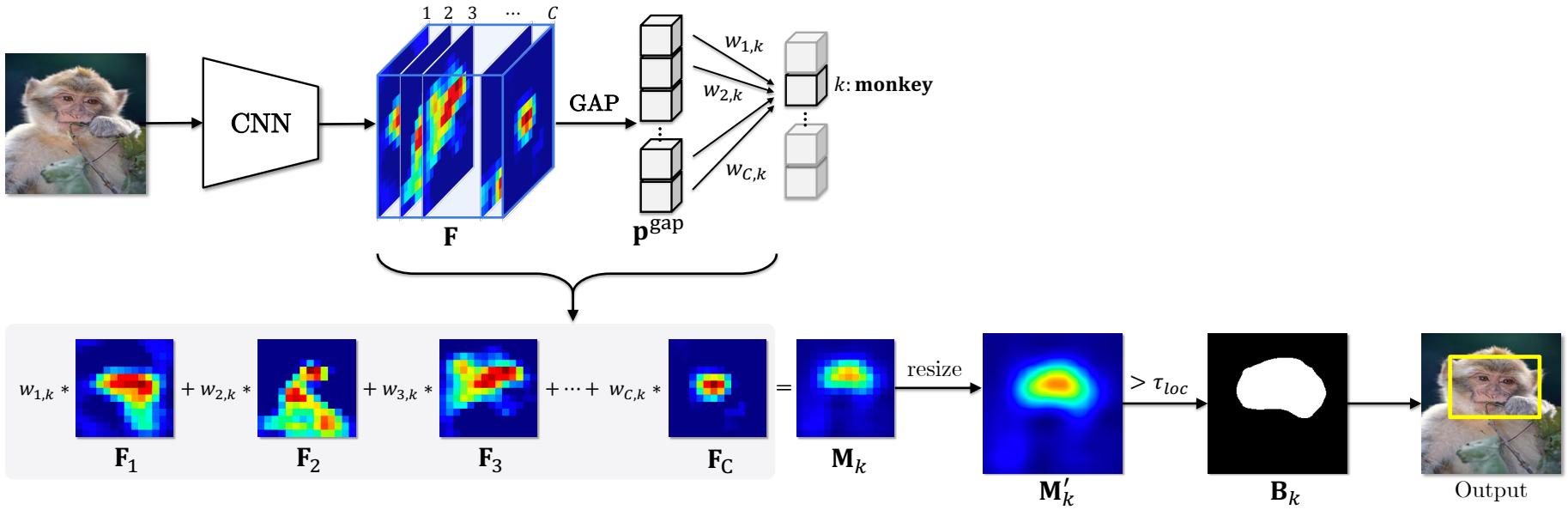
Class Activation Mapping (CAM)

1. Train a classification model.
2. Based on the classification model, apply CAM to localize an object.



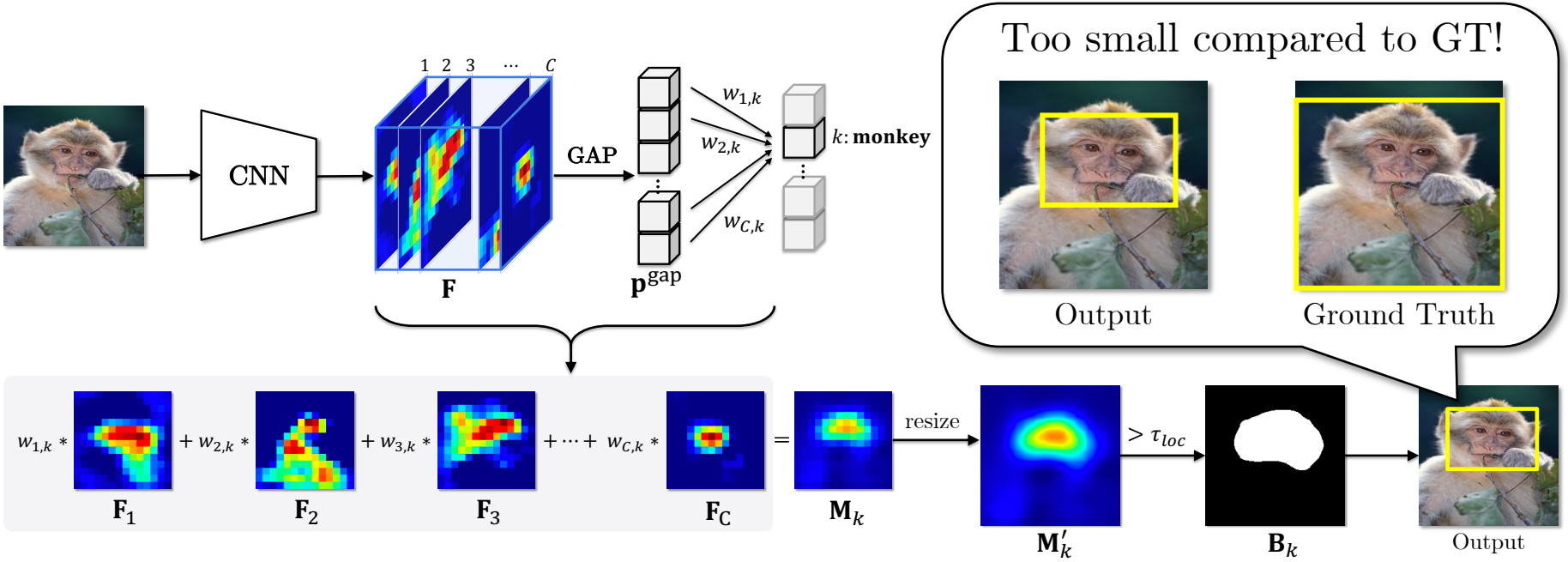
Class Activation Mapping (CAM)

1. Train a classification model.
2. Based on the classification model, apply CAM to localize an object.



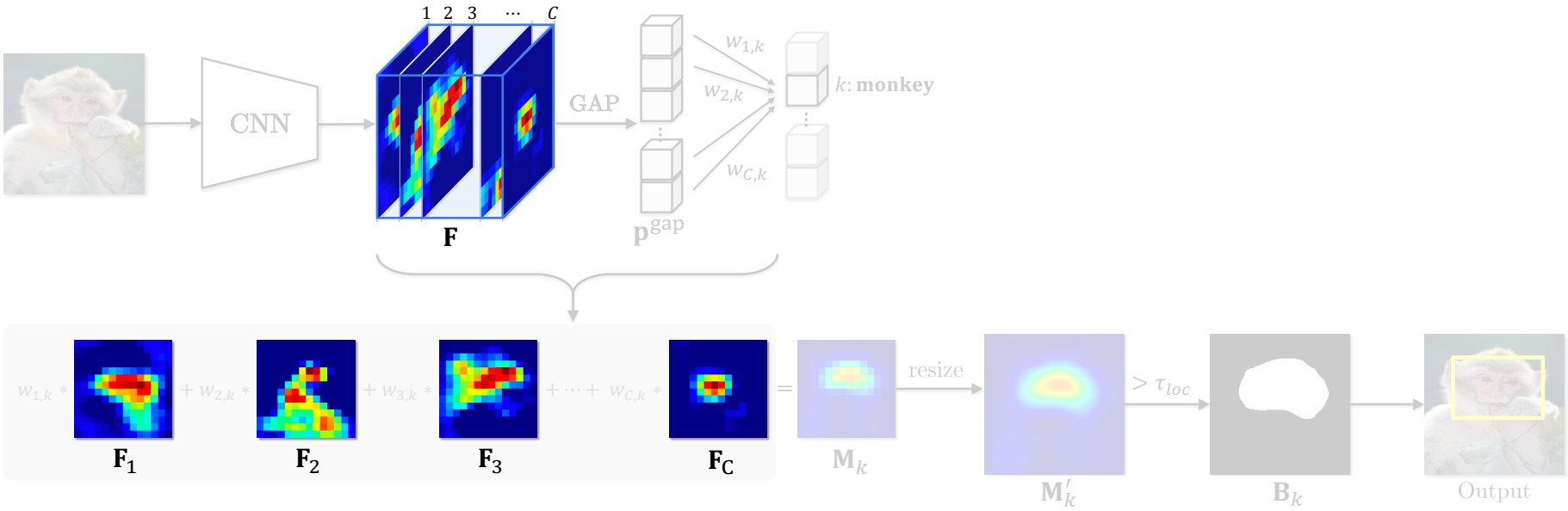
Class Activation Mapping (CAM)

1. Train a classification model.
2. Based on the classification model, apply CAM to localize an object.



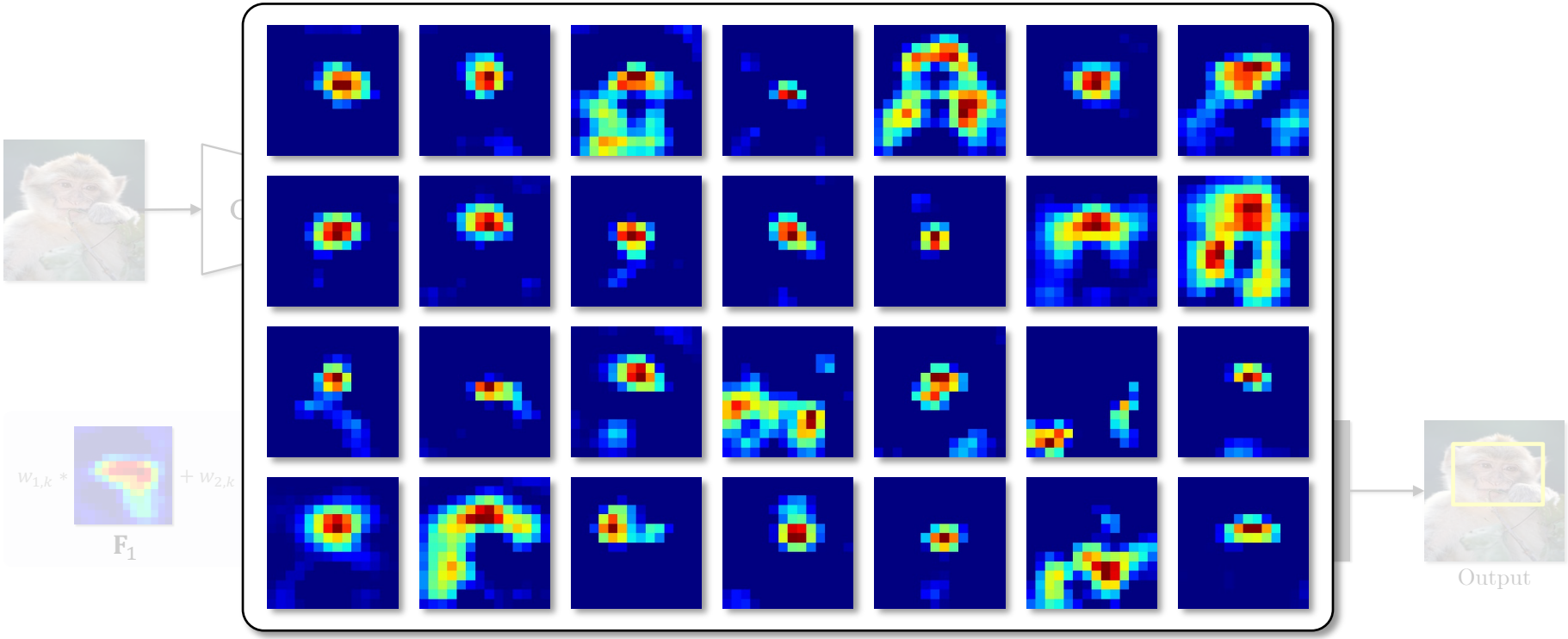
Cause of Limited Localization

- Localization is determined by the activations of a feature map.



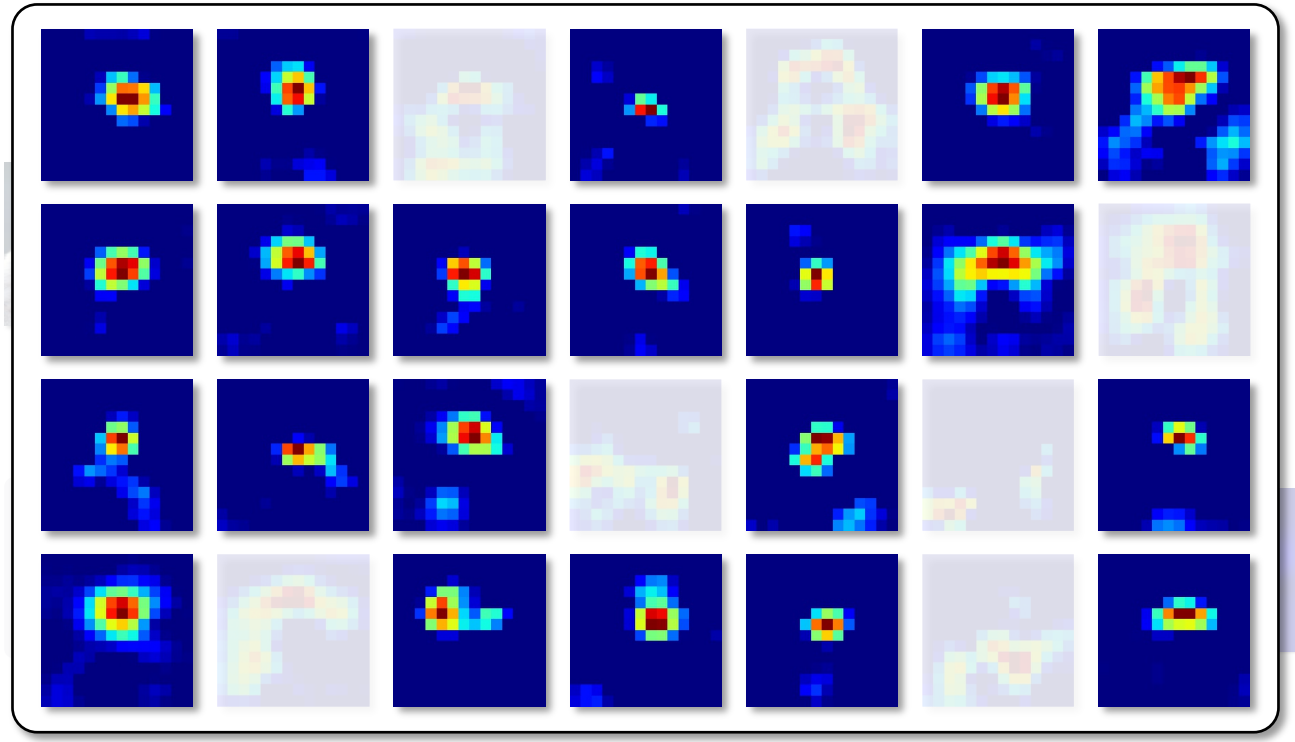
Cause of Limited Localization

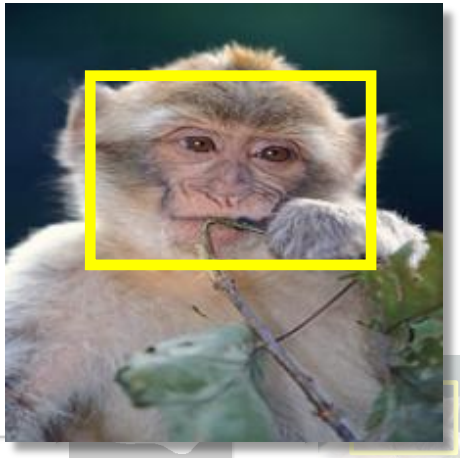
- Localization is determined by the activations of a feature map.



Cause of Limited Localization

- Many of them are highly activated in the **small discriminative region**.
- Looking at the small discriminative region is enough to classify an object.

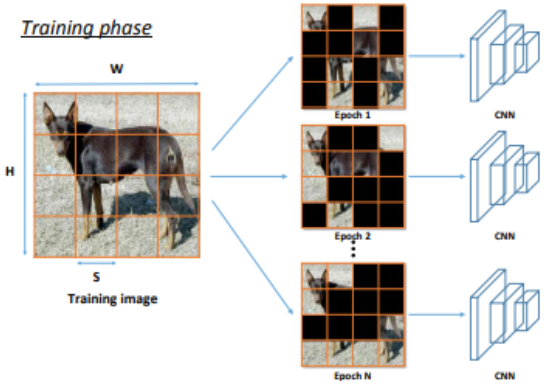




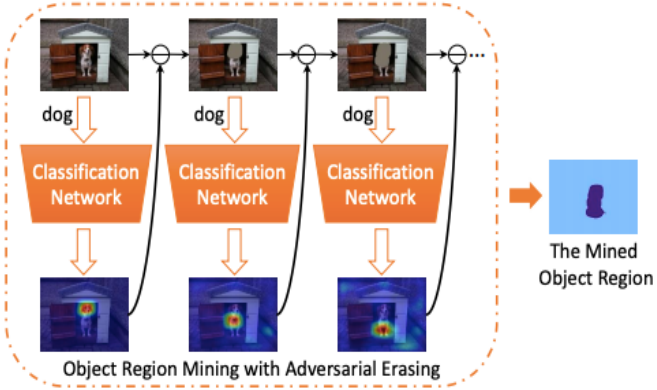
$\mathbf{B}_k$

Output

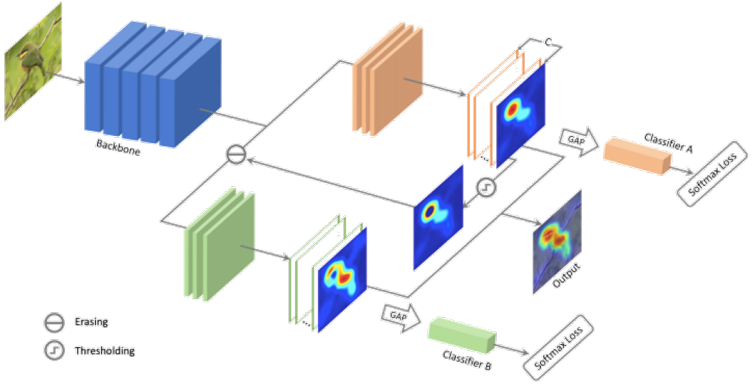
Previous Methods



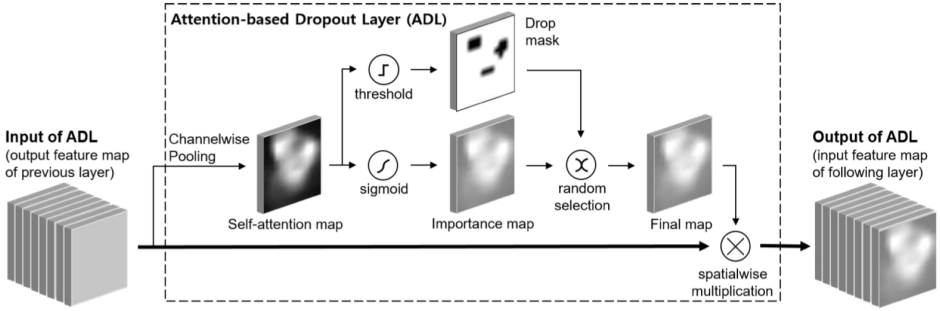
[HaS] Singh, et al. ICCV 2017.



[AE] Wei, et al. CVPR 2017.



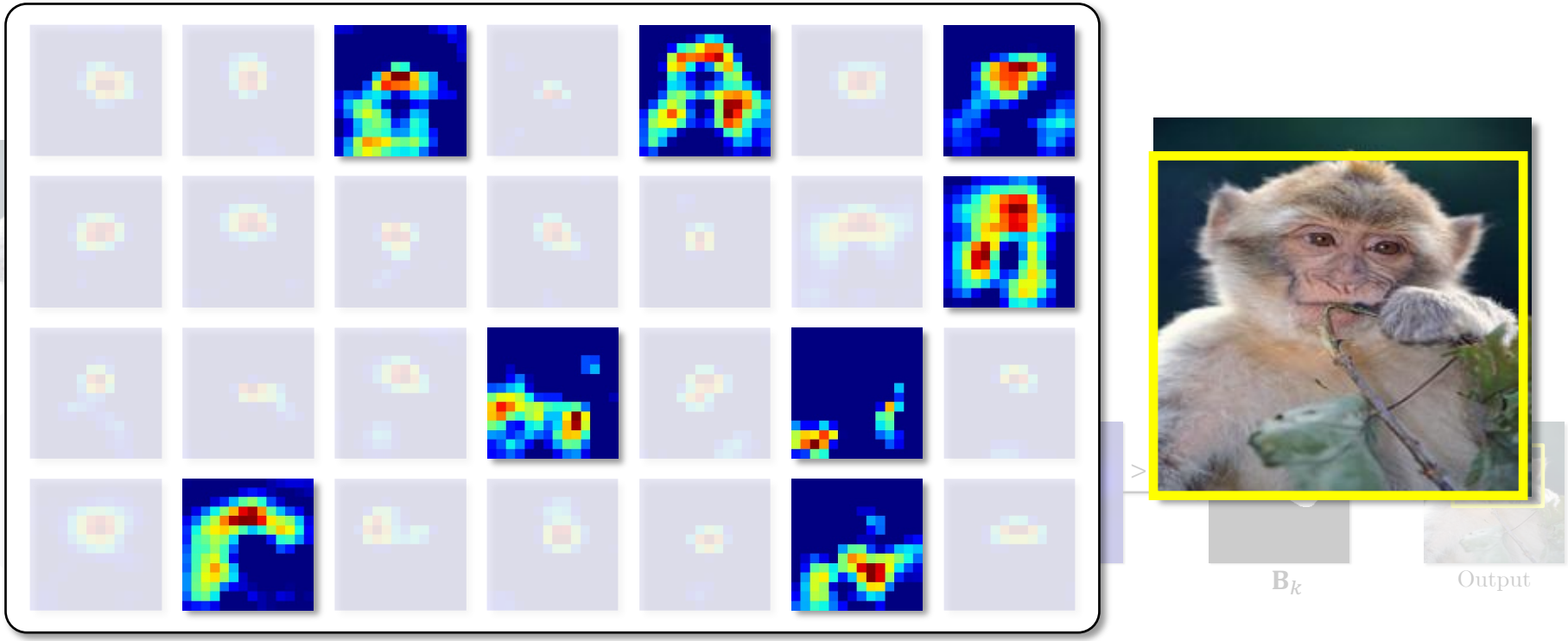
[ACoL] Zhang, et al. CVPR 2018.



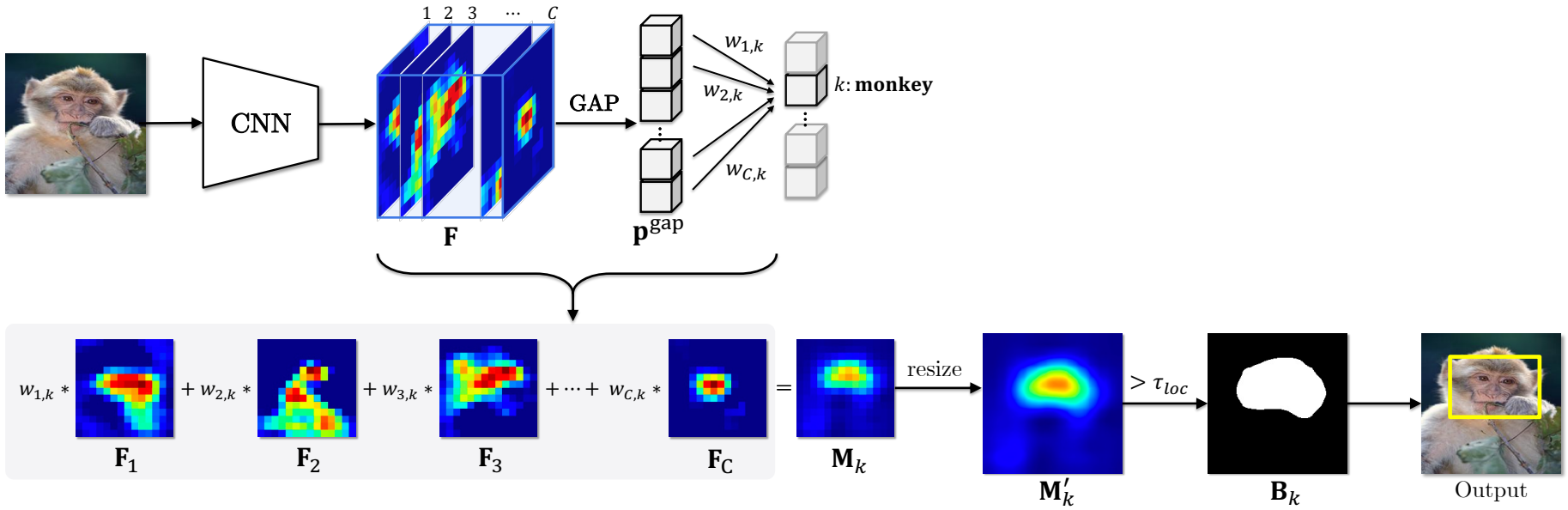
[ADL] Choe, et al. CVPR 2019.

Cause of Limited Localization

- Information that captures the **whole object region** already exists.
- Our goal is to correctly utilize this information.

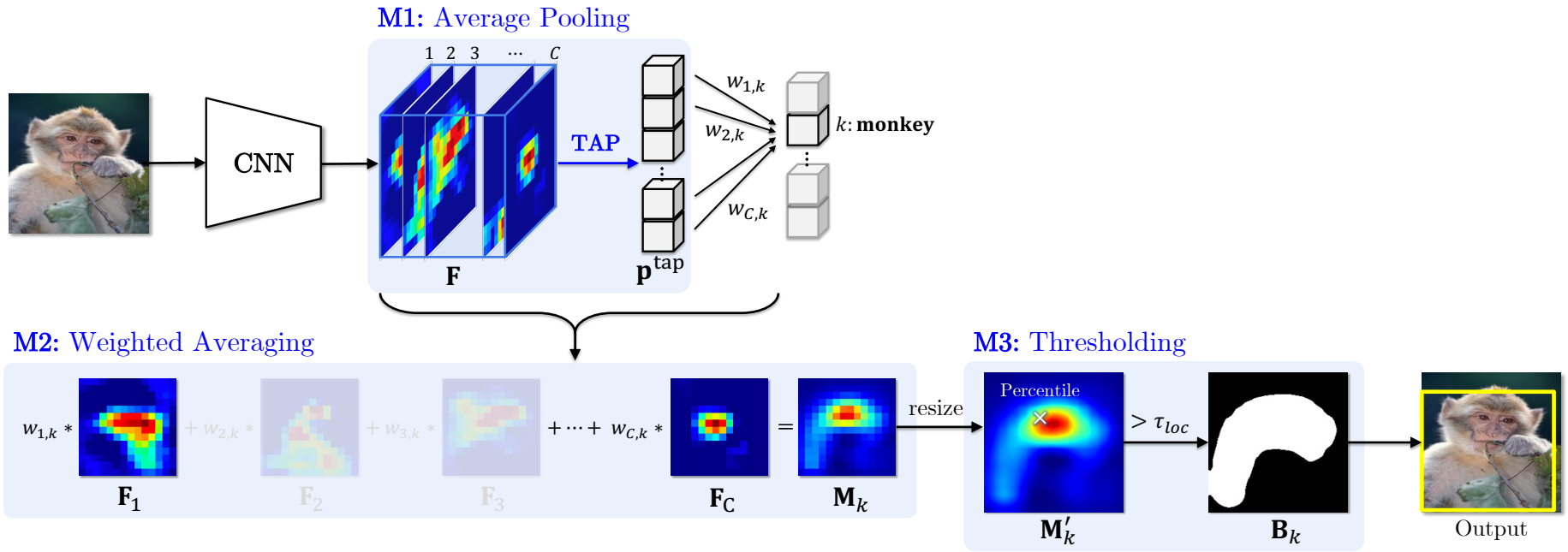


Our Approach: Overall



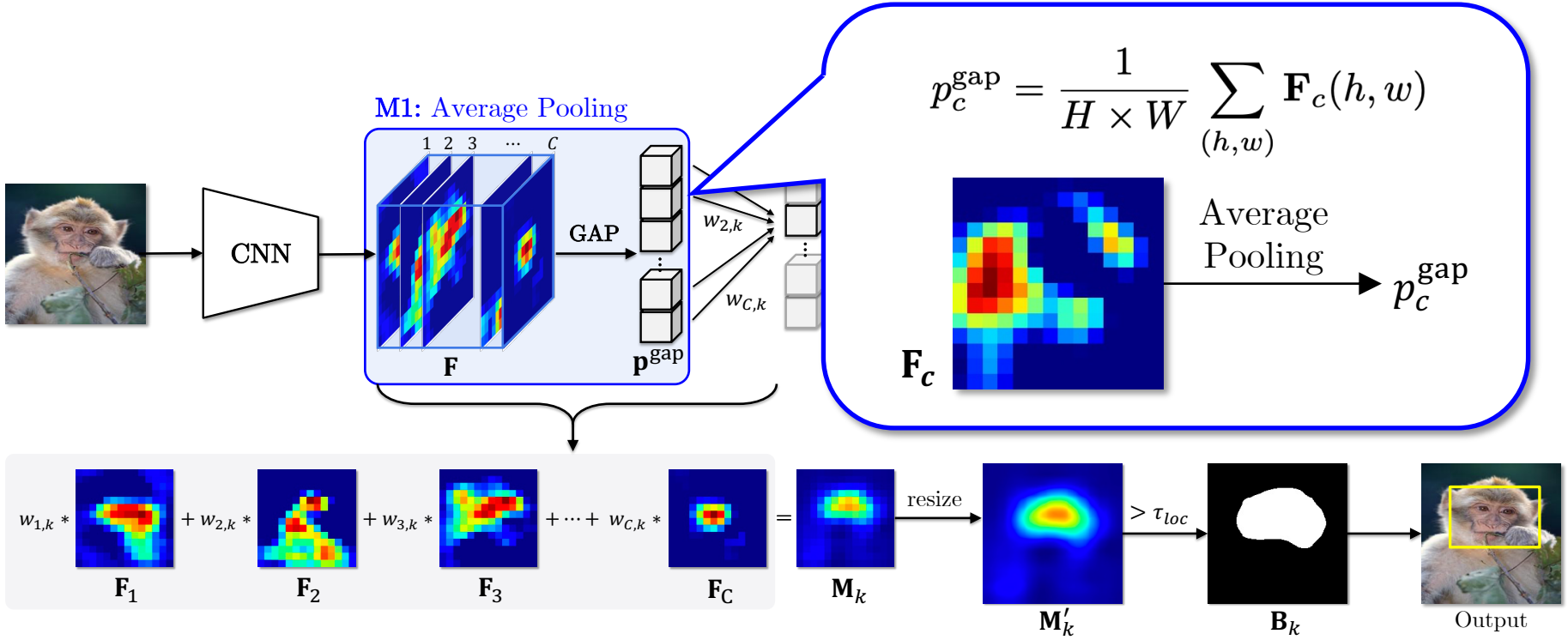
Our Approach: Overall

- Properly utilizes the information by simply modifying three modules.



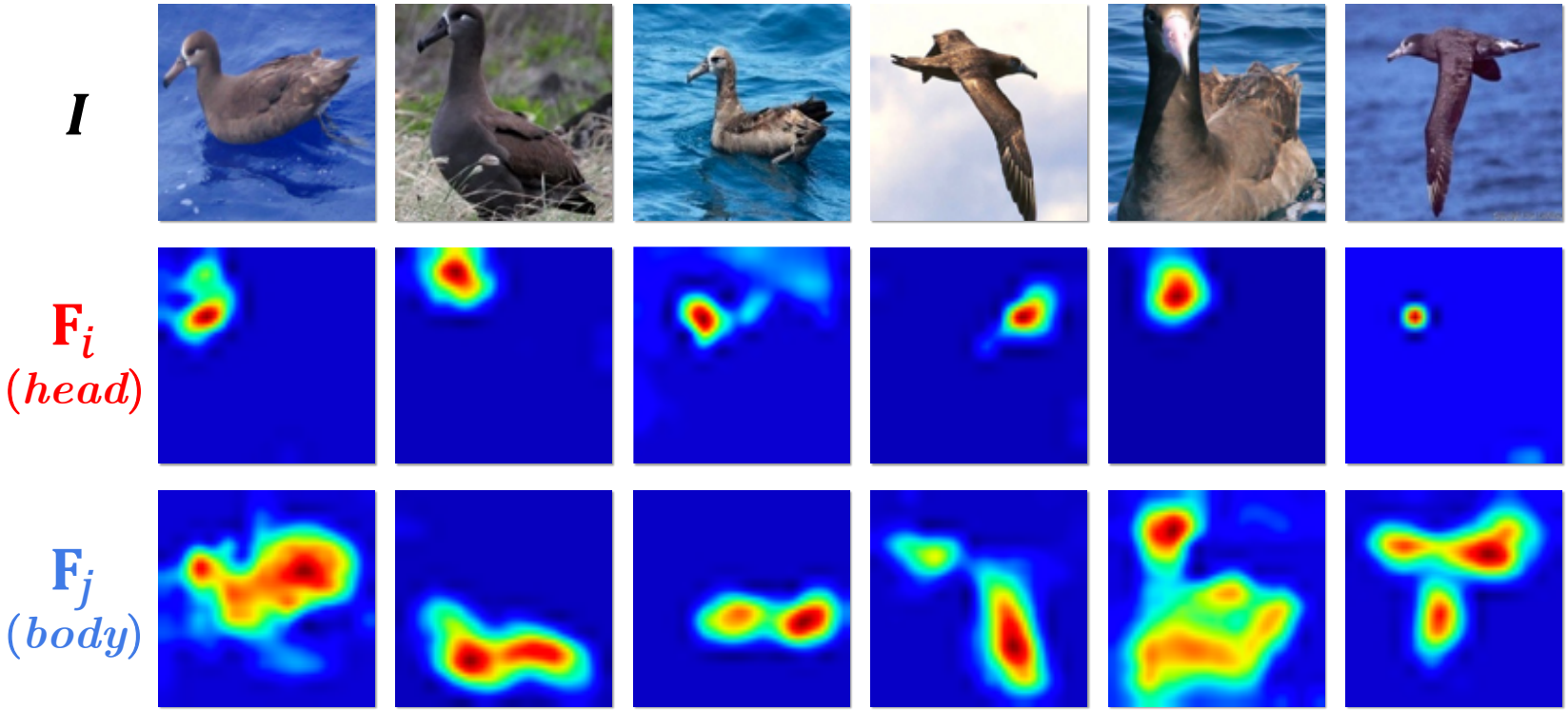
Our Approach (1) Average Pooling

- **Problem:** Global Average Pooling (GAP)



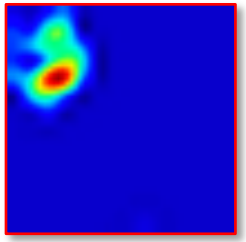
Our Approach (1) Average Pooling

- **Problem:** Global Average Pooling (GAP)

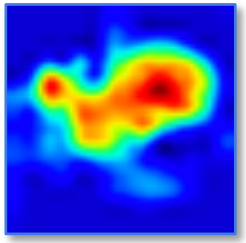


Our Approach (1) Average Pooling

- **Problem:** Global Average Pooling (GAP)



$\mathbf{F}_i$ (max: **64.7**)

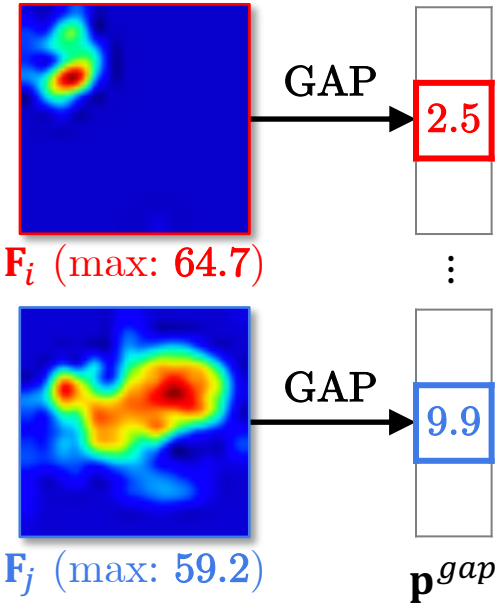


$\mathbf{F}_j$ (max: **59.2**)

Classification phase

Our Approach (1) Average Pooling

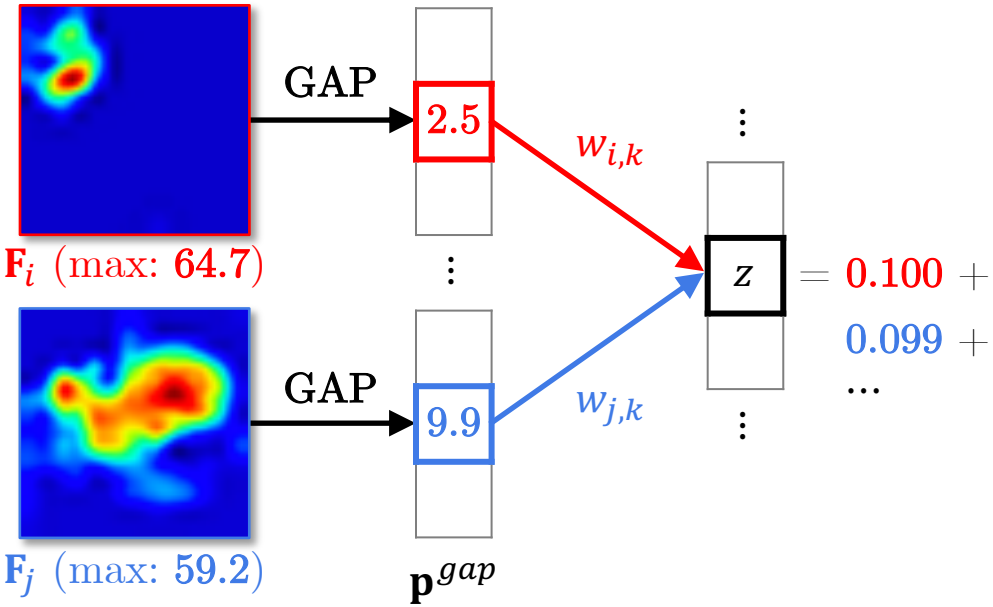
- **Problem:** Global Average Pooling (GAP)



Classification phase

Our Approach (1) Average Pooling

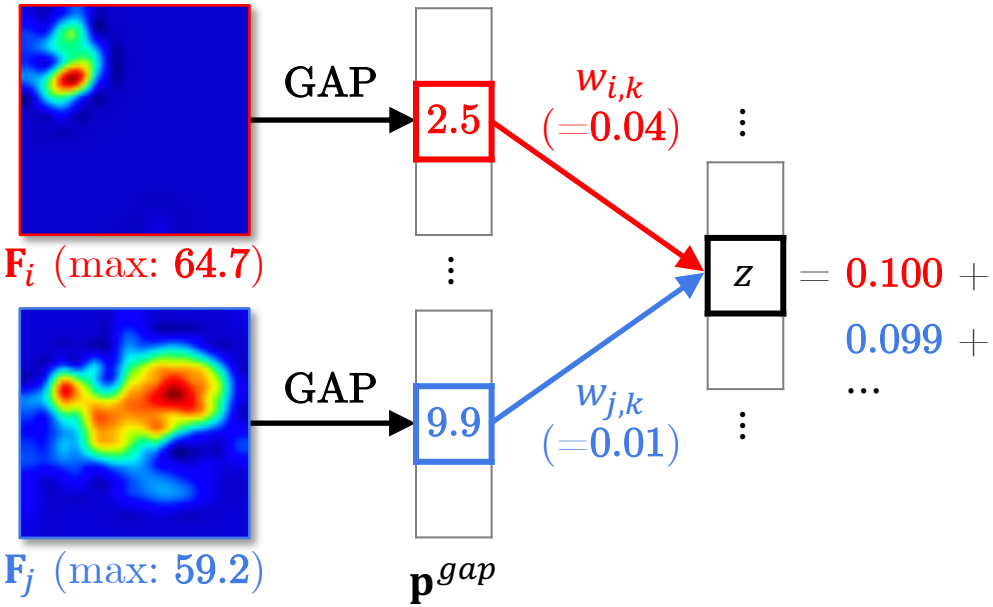
- **Problem:** Global Average Pooling (GAP)



Classification phase

Our Approach (1) Average Pooling

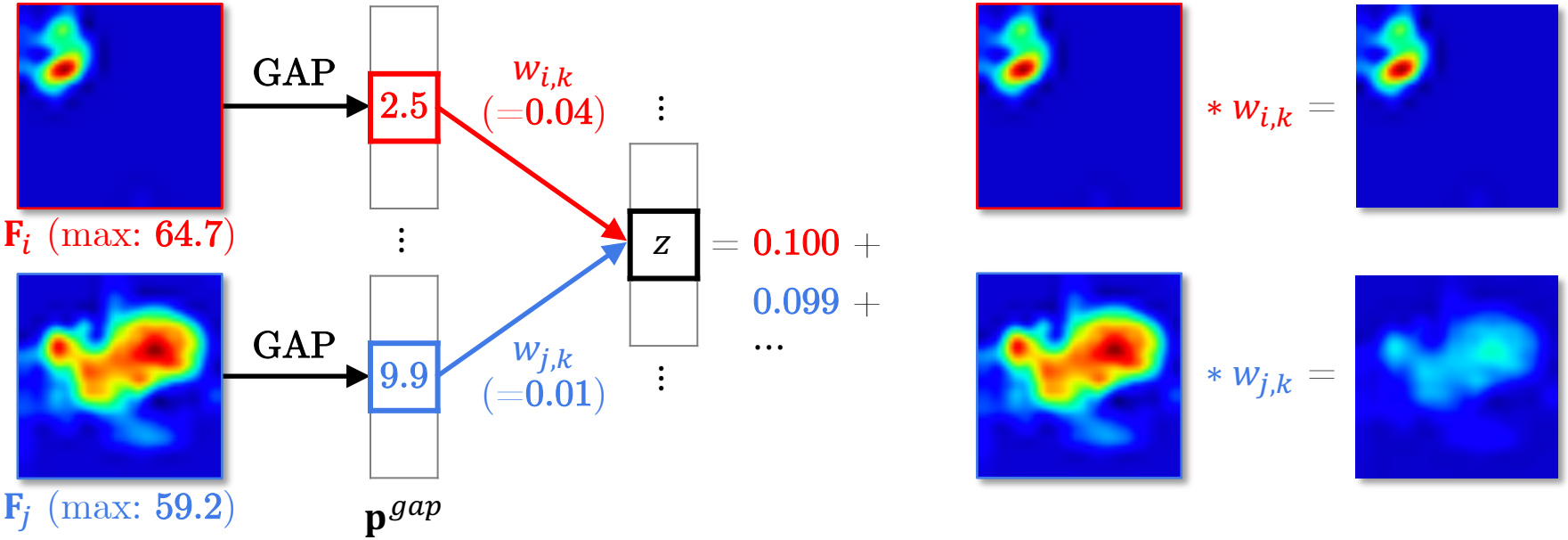
- **Problem:** Global Average Pooling (GAP)



Classification phase

Our Approach (1) Average Pooling

- **Problem:** Global Average Pooling (GAP)

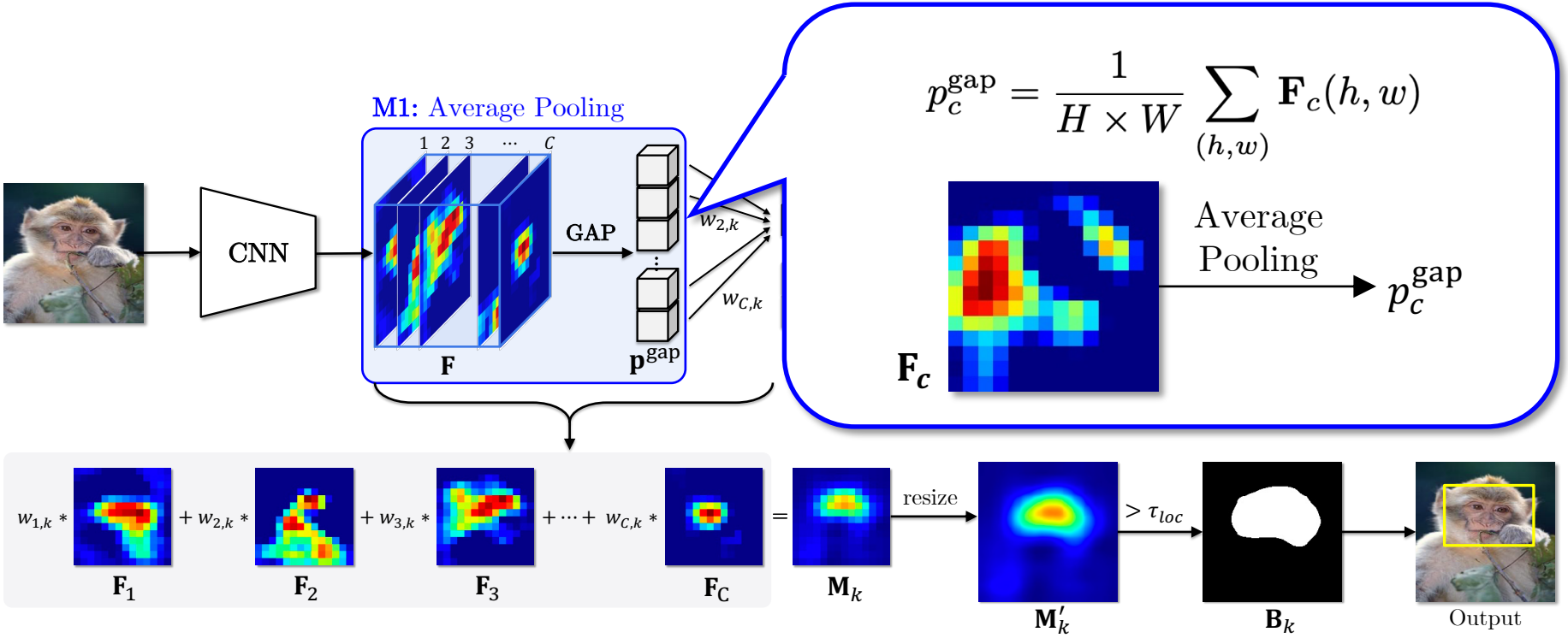


Classification phase

Localization phase

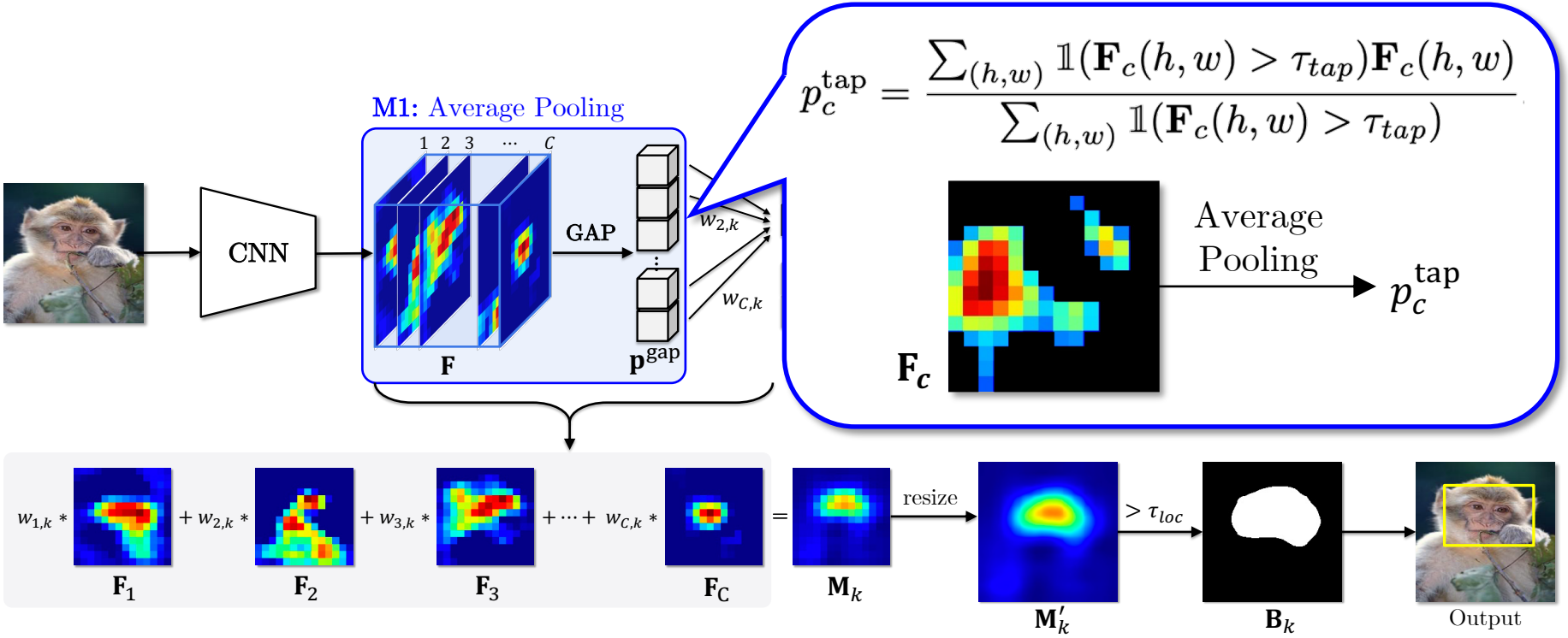
Our Approach (1) Average Pooling

- **Problem:** Global Average Pooling (GAP)



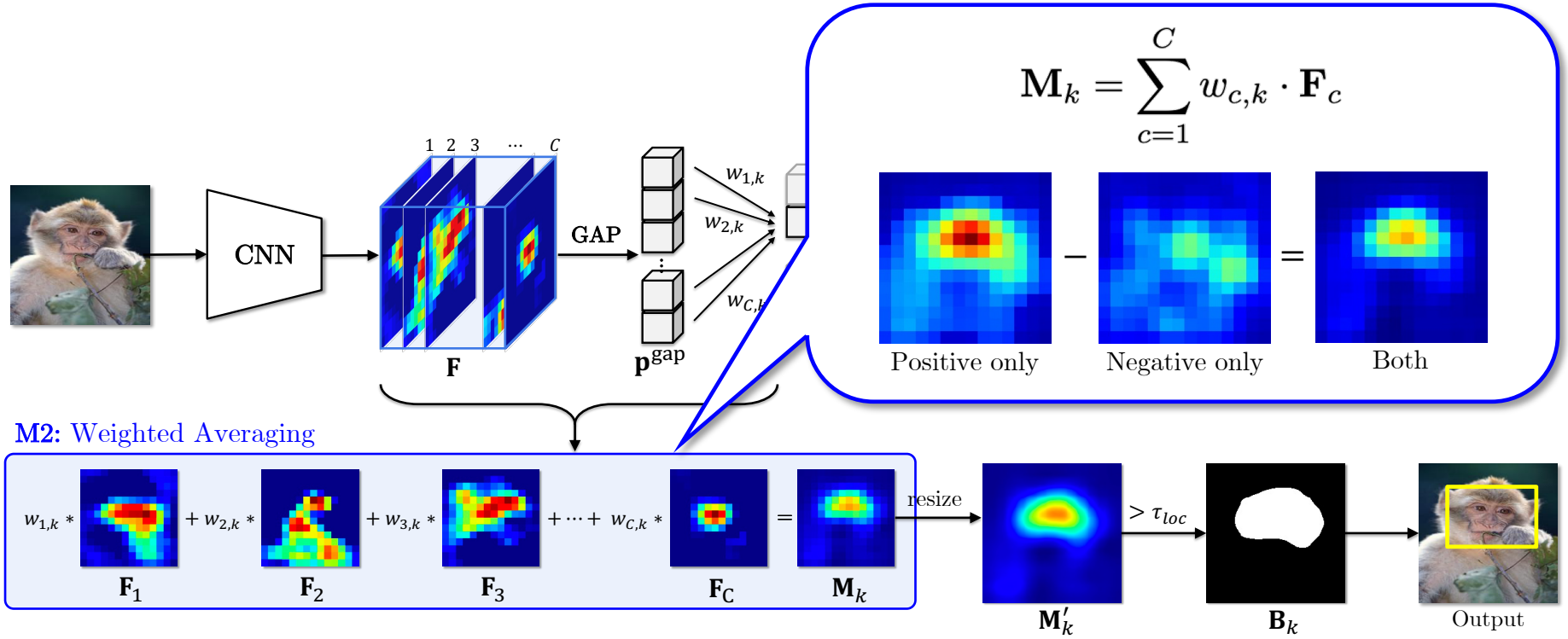
Our Approach (1) Average Pooling

- **Solution:** Thresholded Average Pooling (TAP)



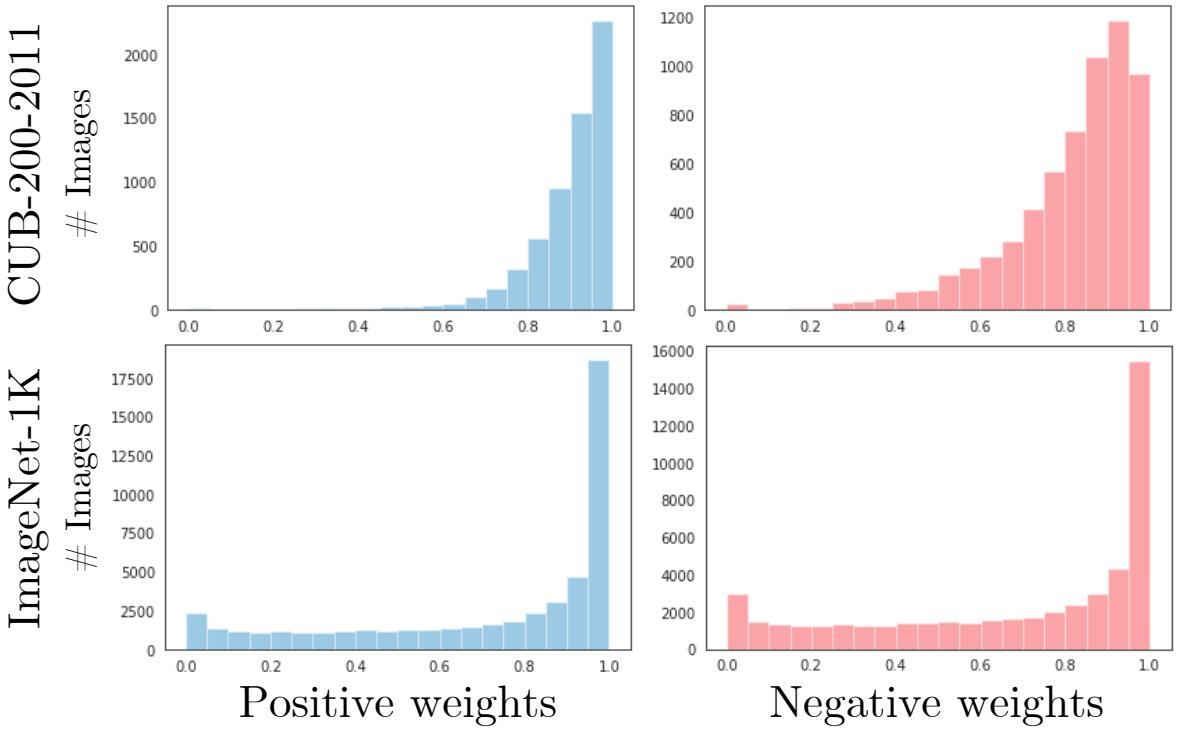
Our Approach (2) Weighted Averaging

- **Problem:** Class Activation Maps



Our Approach (2) Weighted Averaging

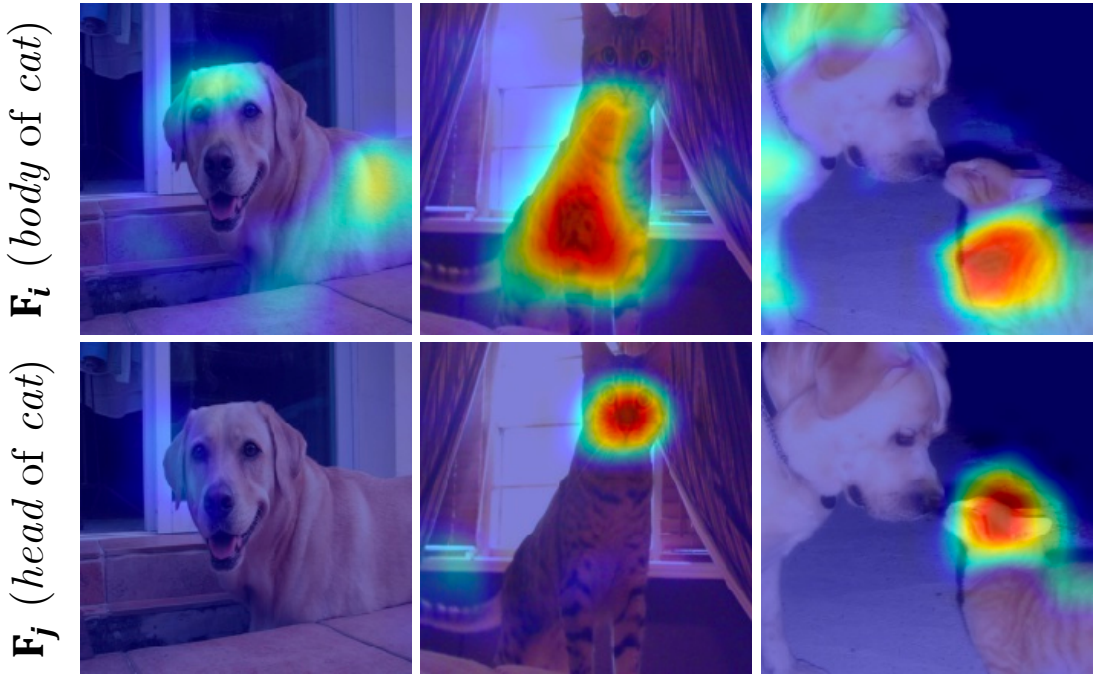
- **Problem:** Class Activation Maps



IoA between the GT boxes and the CAMs

Our Approach (2) Weighted Averaging

- **Problem:** Class Activation Maps

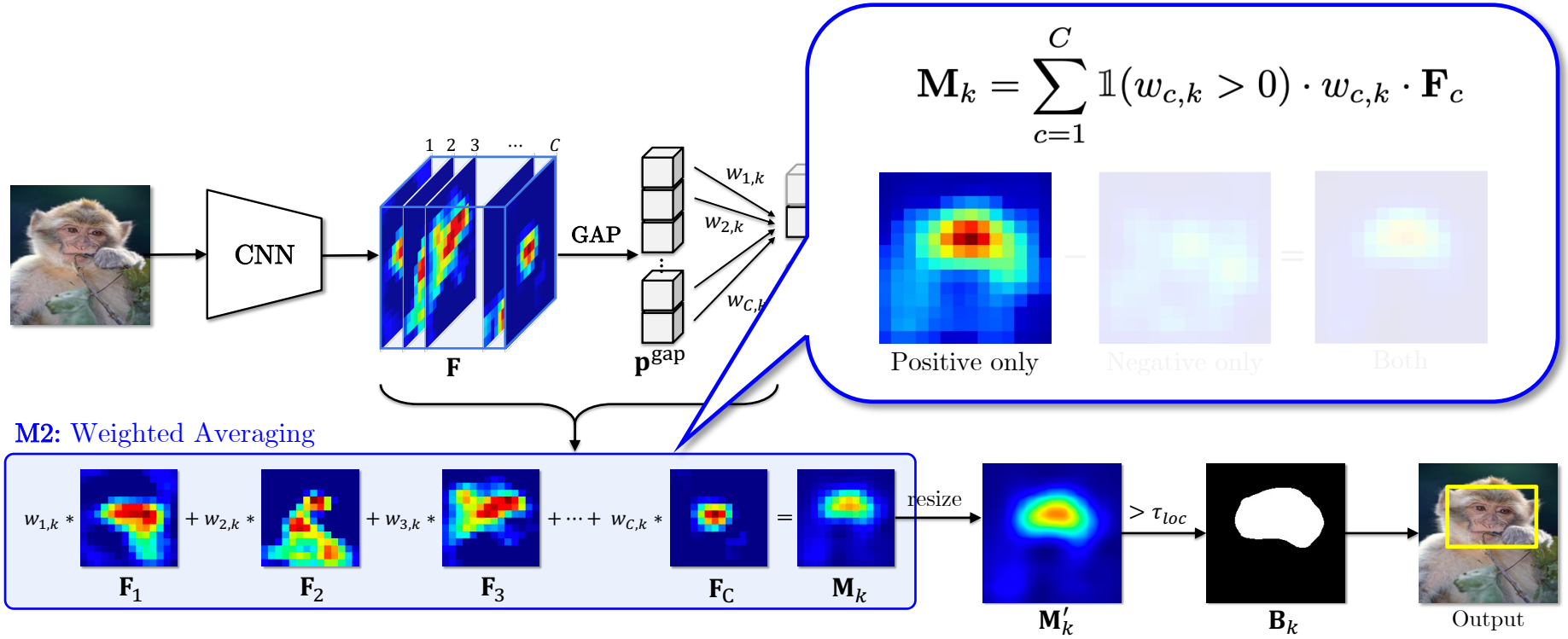


(GT: *dog*, *w*: −) (GT: *cat*, *w*: +) (GT: *dog*, *w*: −)

Activated regions of two features

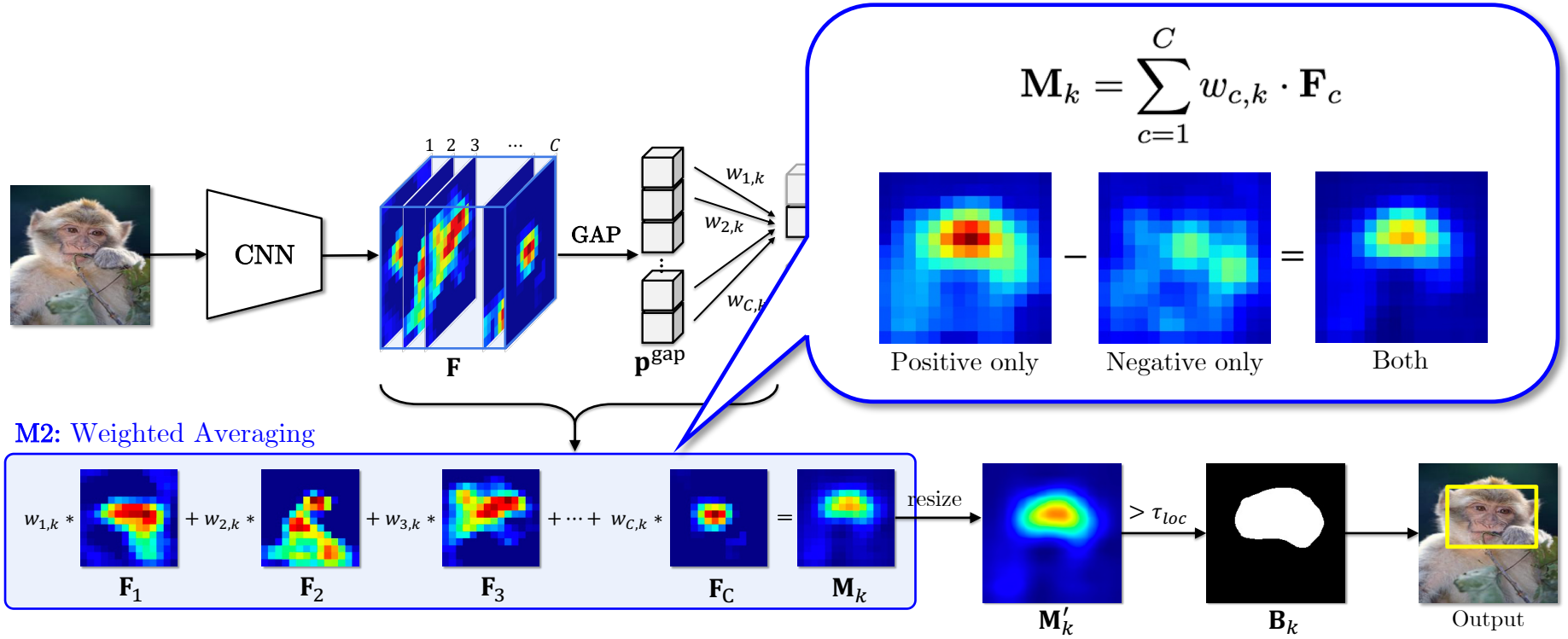
Our Approach (2) Weighted Averaging

- **Solution:** Negative Weight Clamping (NWC)



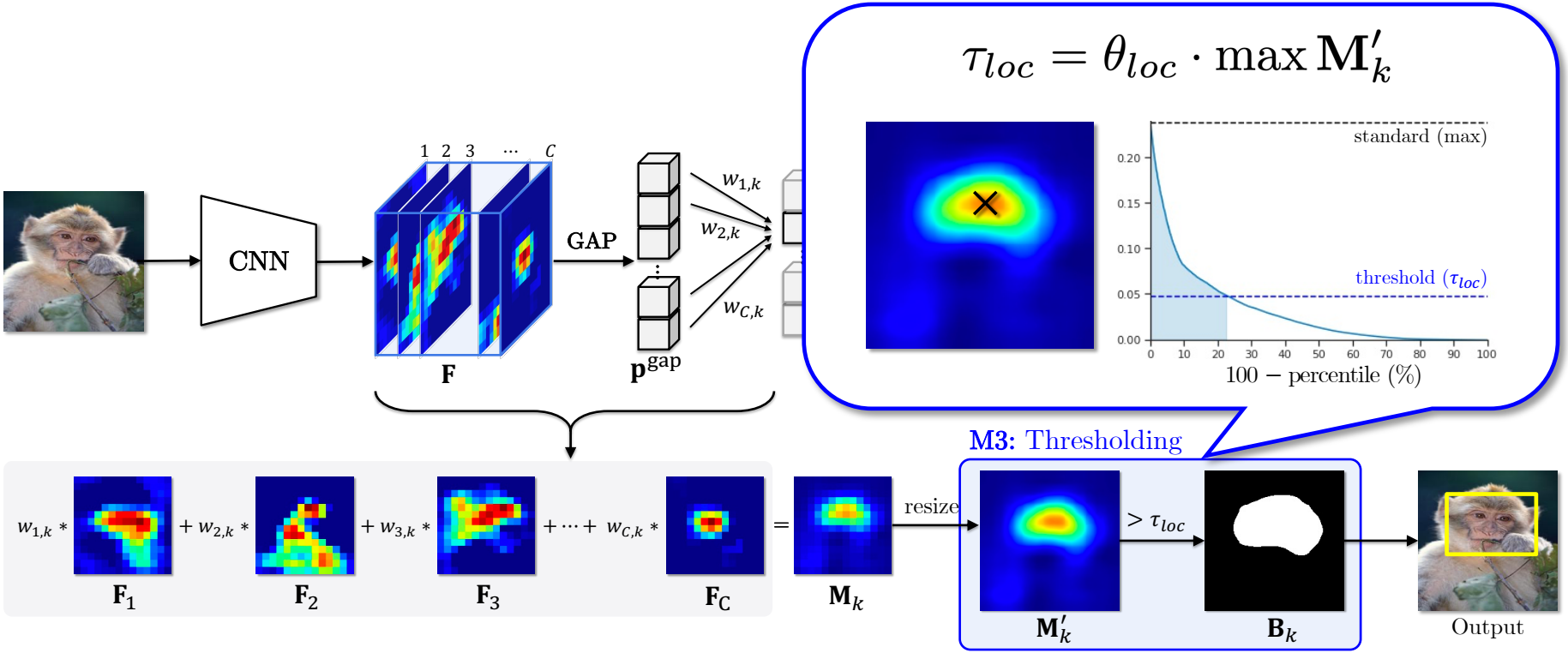
Our Approach (2) Weighted Averaging

- **Problem:** Class Activation Maps



Our Approach (3) Thresholding

- **Problem:** Maximum as a Standard (MaS)



Our Approach (3) Thresholding

- **Problem:** Maximum as a Standard (MaS)

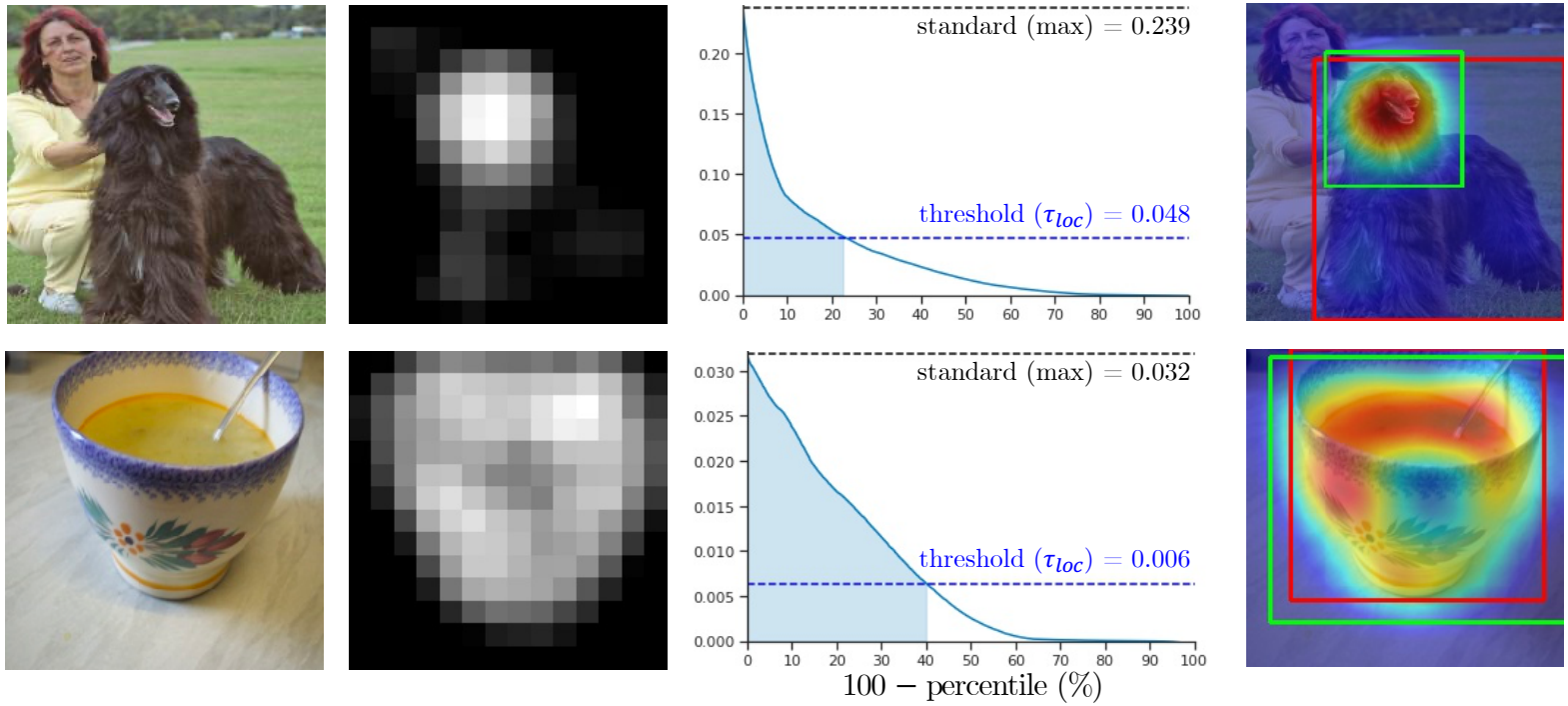


Image | Num of channels (activation $> \tau_{0.8}$) | CAM values (desc. order) + Maximum as a Standard | Result with CAM

Our Approach (3) Thresholding

- **Problem:** Maximum as a Standard (MaS)

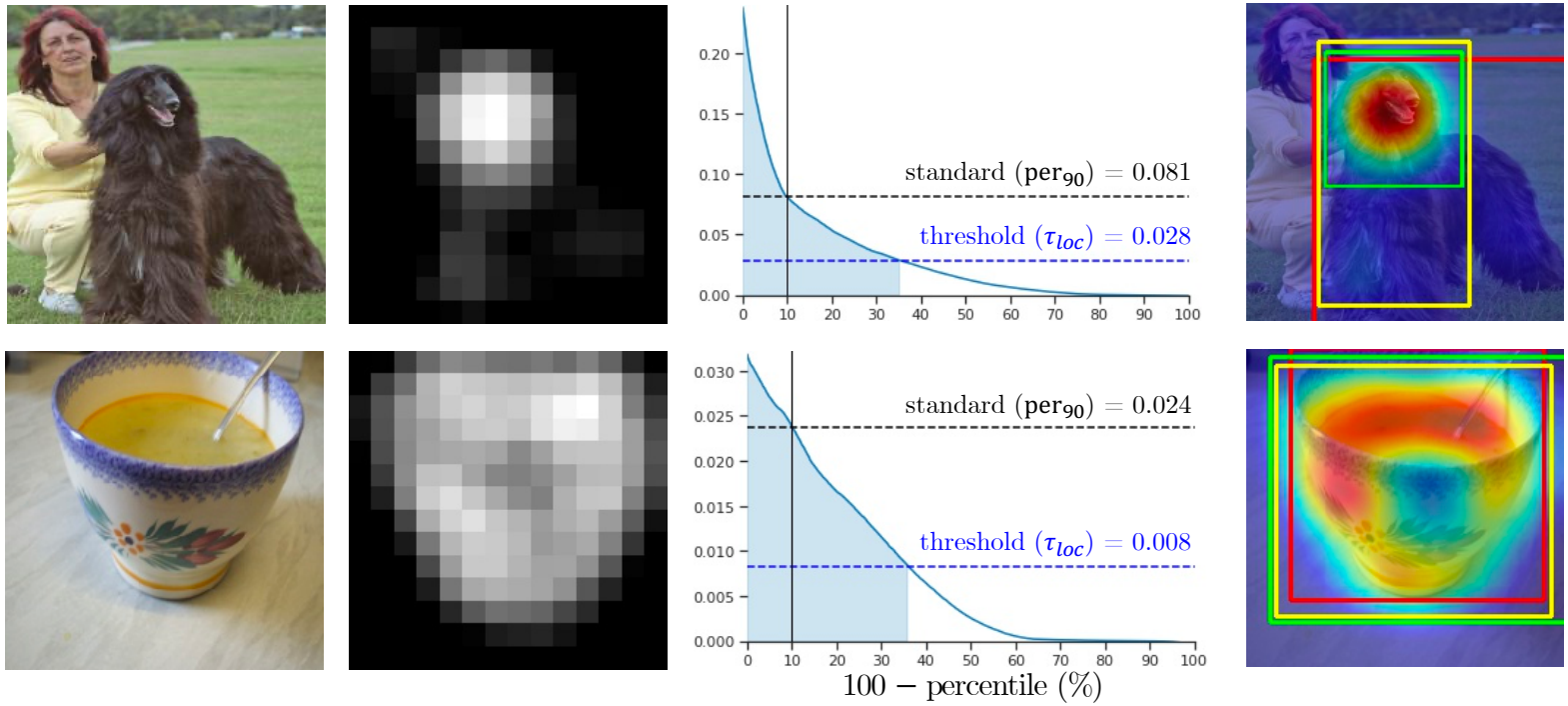
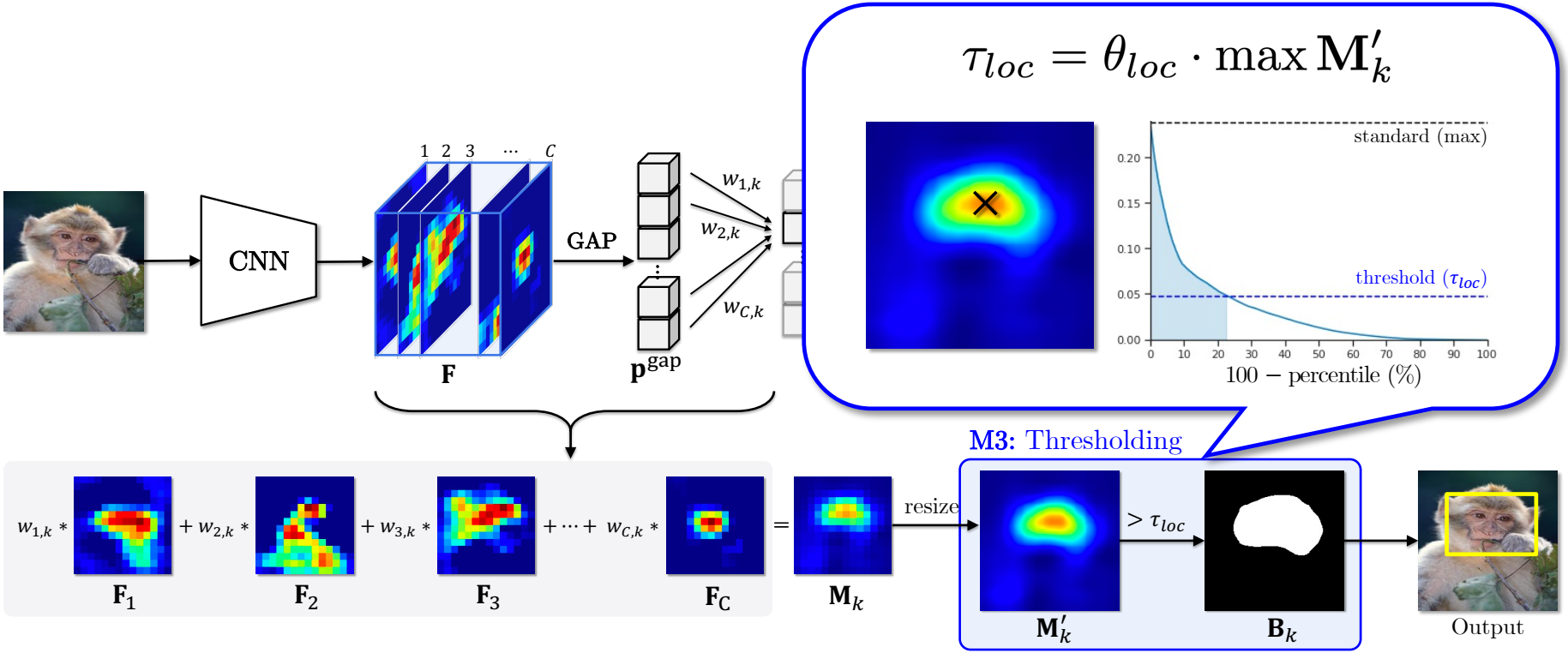


Image | Num of channels (activation $> \tau_{0.8}$) | CAM values (desc. order) + Percentile as a Standard | Result with CAM

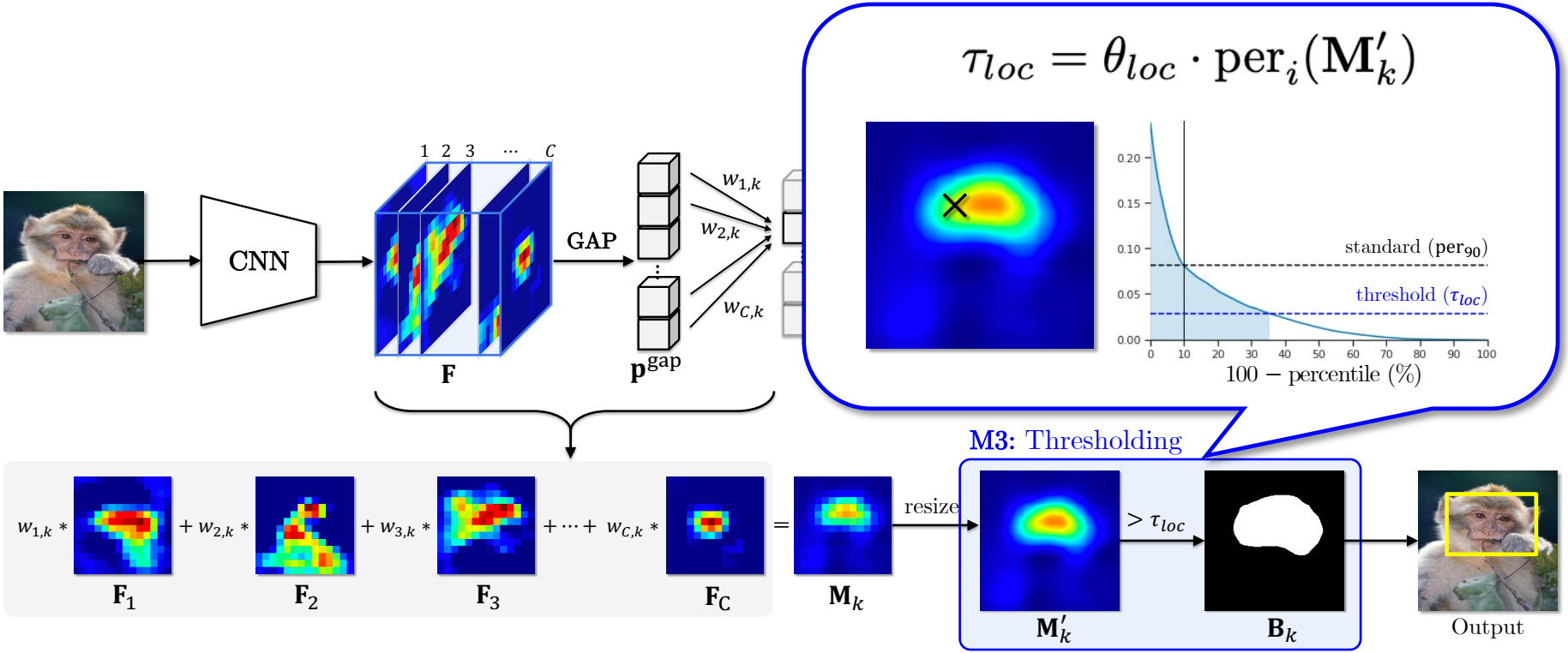
Our Approach (3) Thresholding

- **Problem:** Maximum as a Standard (MaS)



Our Approach (3) Thresholding

- **Solution:** Percentile as a Standard (PaS)



Experimental Setting: Datasets

1. **CUB-200-2011**
   - ~12K images
   - Birds with 200 categories
2. **ImageNet-1K**
   - ~1.35M images
   - General objects with 1000 categories

Experimental Setting: Datasets

1. **CUB-200-2011**
   - ~12K images
   - Birds with 200 categories
2. **ImageNet-1K**
   - ~1.35M images
   - General objects with 1000 categories
3. **OpenImages30K**
   - ~37K images
   - General objects with 100 categories

Experimental Setting: Datasets & Evaluation Metrics

1. **CUB-200-2011**
   - ~12K images
   - Birds with 200 categories
2. **ImageNet-1K**
   - ~1.35M images
   - General objects with 1000 categories

- **Top-1 Cls:** top-1 accuracy of classification
- **GT-known Loc:** localization accuracy with known ground truth class
- **Top-1 Loc:** both classification and localization

3. **OpenImages30K**
   - ~37K images
   - General objects with 100 categories

- **PxAP**: area under a pixel precision and recall curve (independent to the choice of a threshold)

Experimental Setting: Datasets & Evaluation Metrics

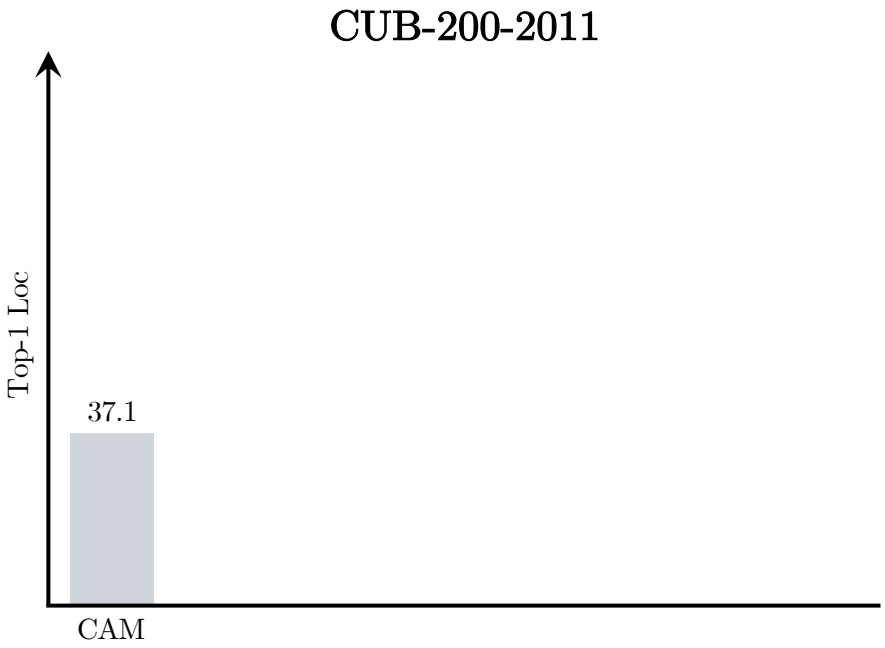
1. **CUB-200-2011**
   - ~12K images
   - Birds with 200 categories
2. **ImageNet-1K**
   - ~1.35M images
   - General objects with 1000 categories

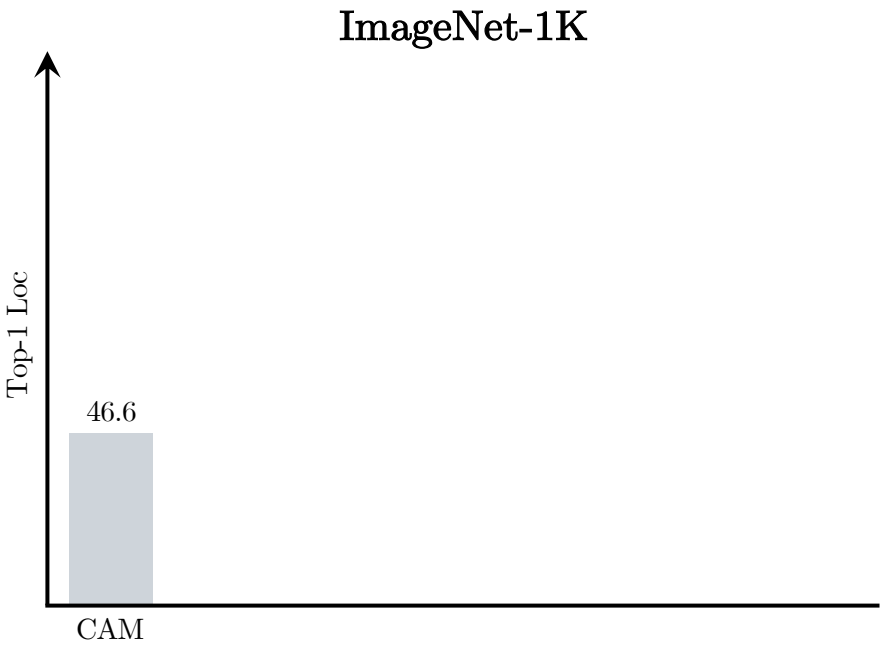
- **Top-1 Cls:** top-1 accuracy of classification
- **GT-known Loc:** localization accuracy with known ground truth class
- **Top-1 Loc:** both classification and localization

3. **OpenImages30K**
   - ~37K images
   - General objects with 100 categories

- **PxAP**: area under a pixel precision and recall curve (independent to the choice of a threshold)

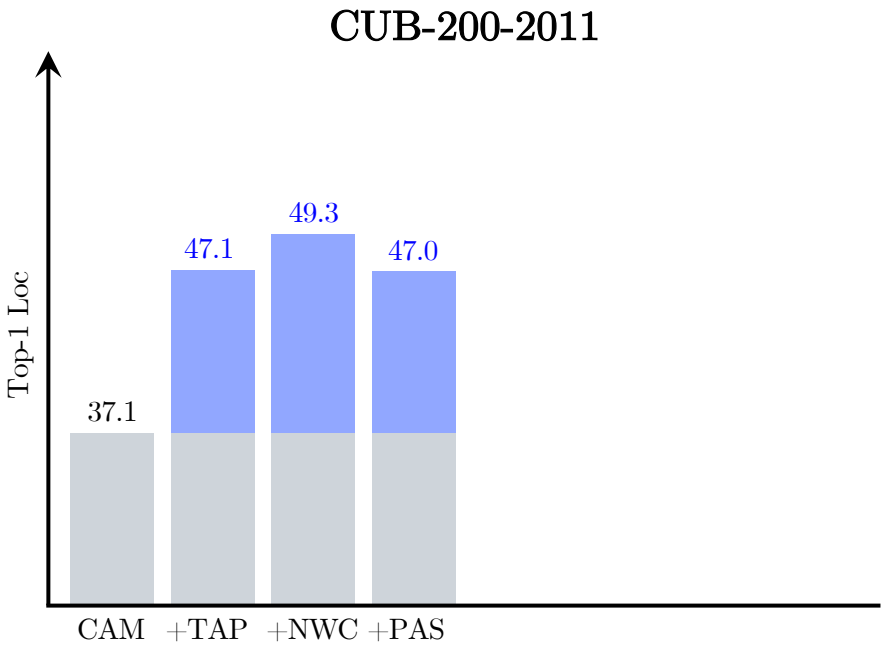
Experiment Results: Different Components

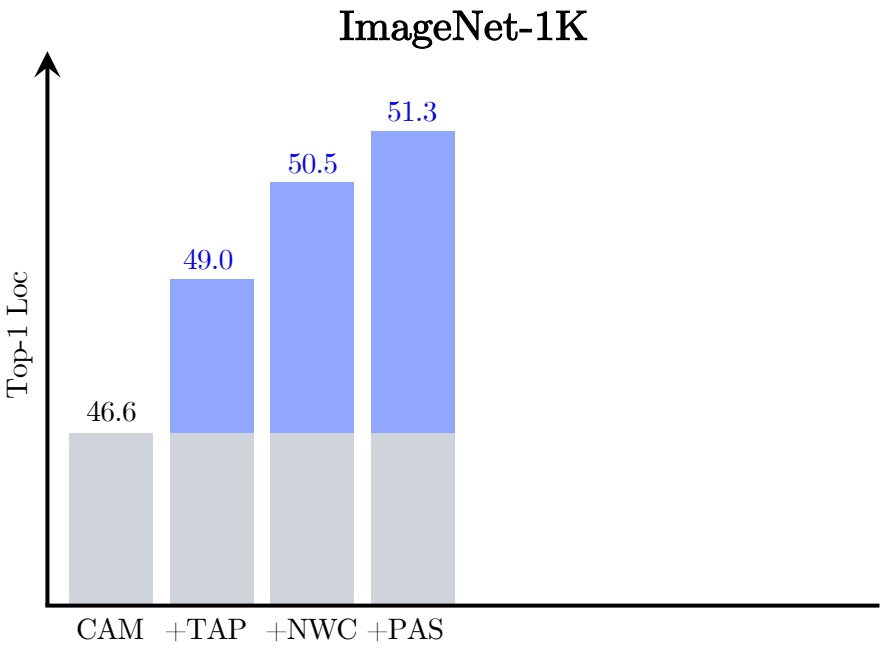




Experiment Results: Different Components

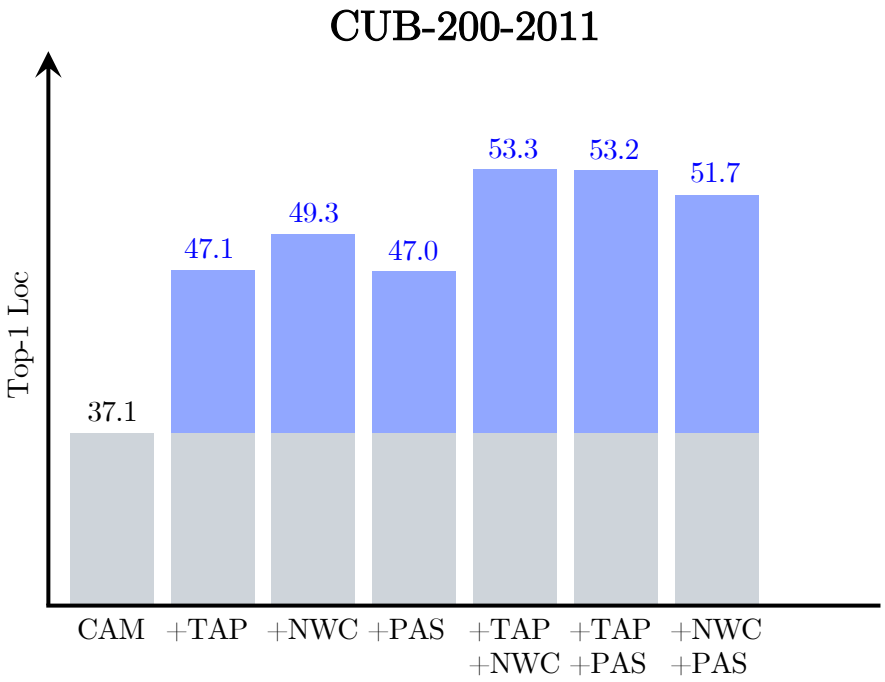
- TAP, NWC and PaS consistently improved the localization performance.

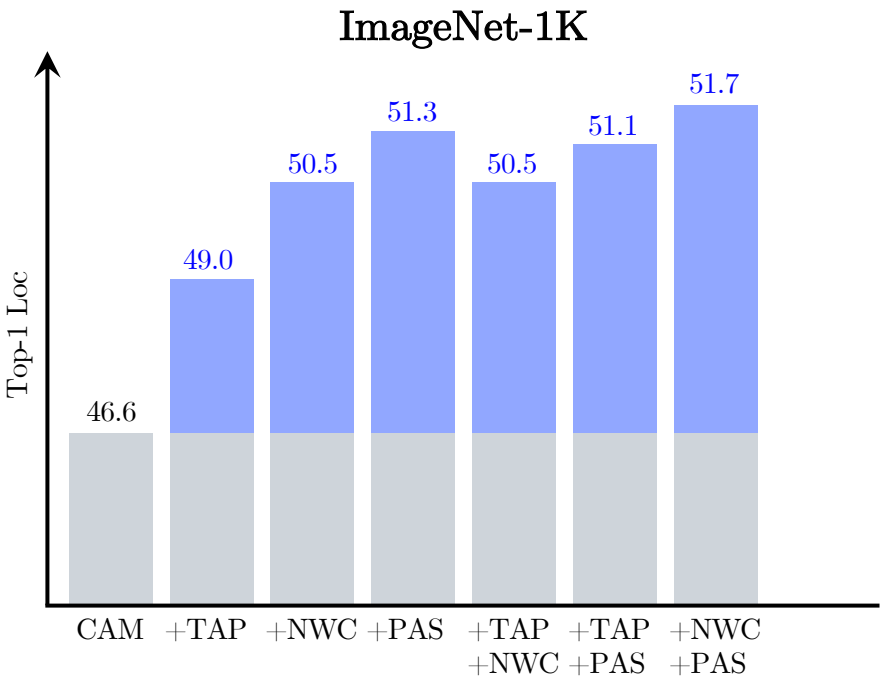




Experiment Results: Different Components

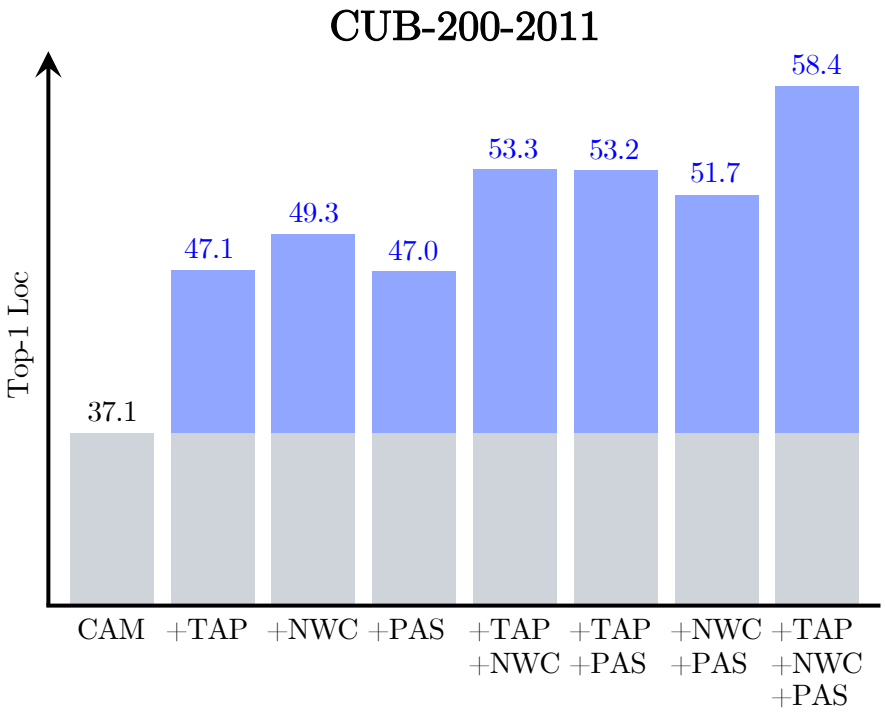
- TAP, NWC and PaS consistently improved the localization performance.

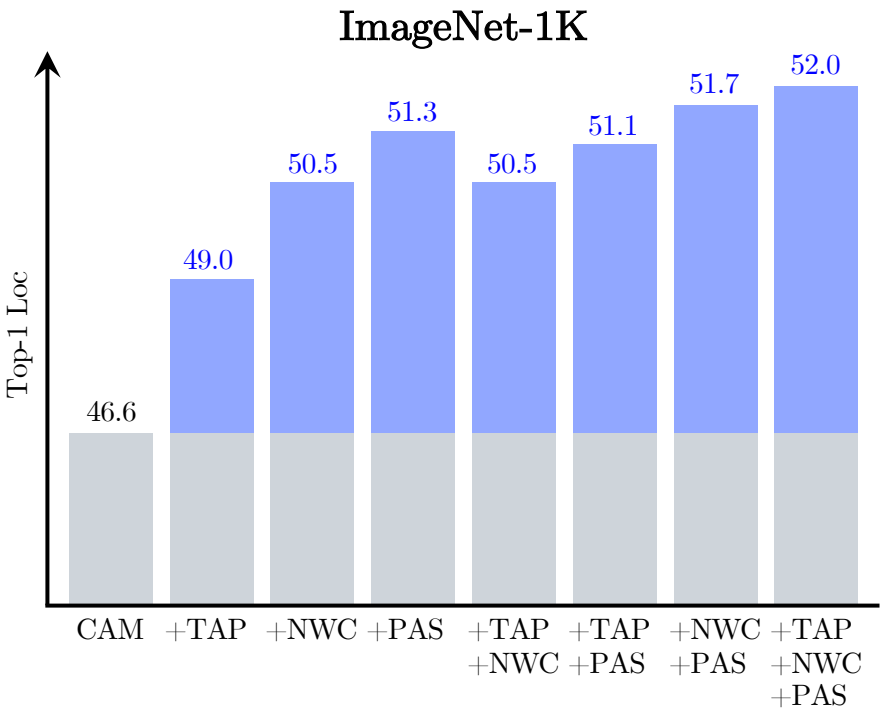




Experiment Results: Different Components

- TAP, NWC and PaS consistently improved the localization performance.
- With all the components applied, the localization performance further improved.





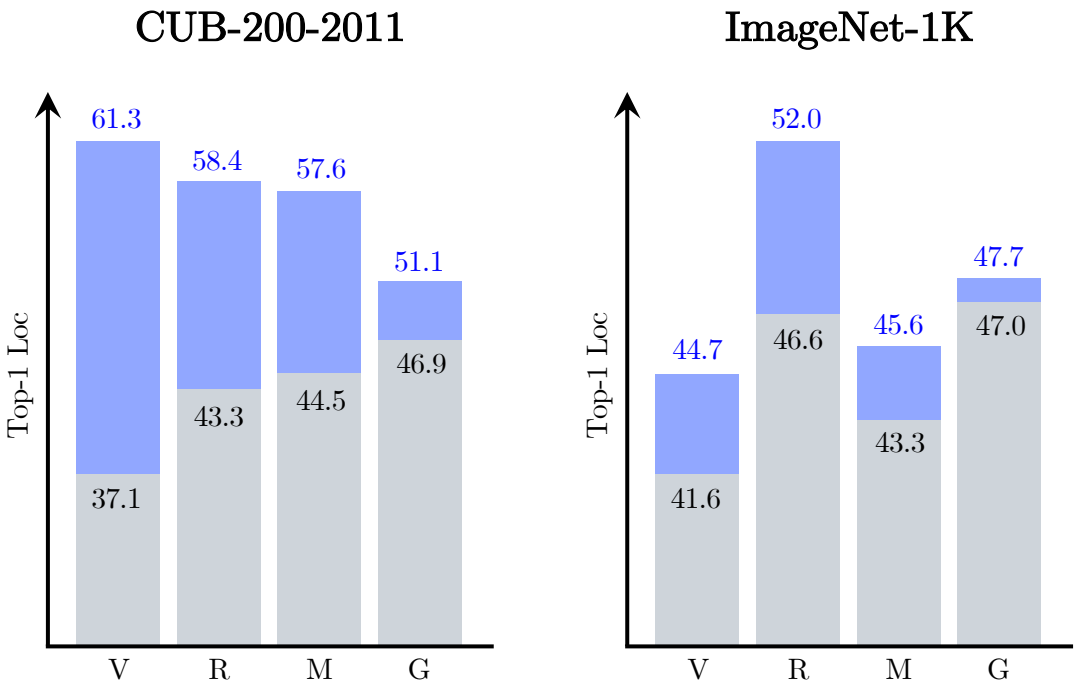
Experiment Results: Different Backbones

- **V**: VGG16
- **R**: ResNet-50
- **M**: MobileNetV1
- **G**: GoogleNet

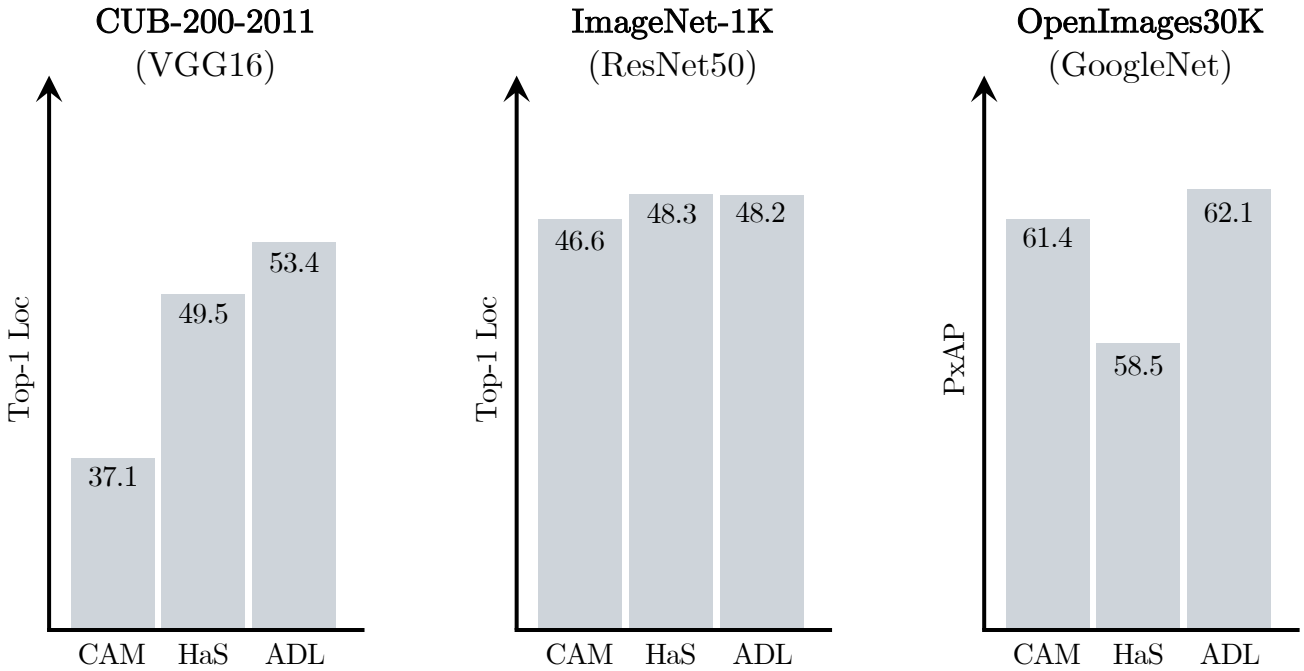
Experiment Results: Different Backbones

- Regardless of a backbone structure, the performance consistently improved. (+ Ours: employ all three components)

- **V**: VGG16
- **R**: ResNet-50
- **M**: MobileNetV1
- **G**: GoogleNet

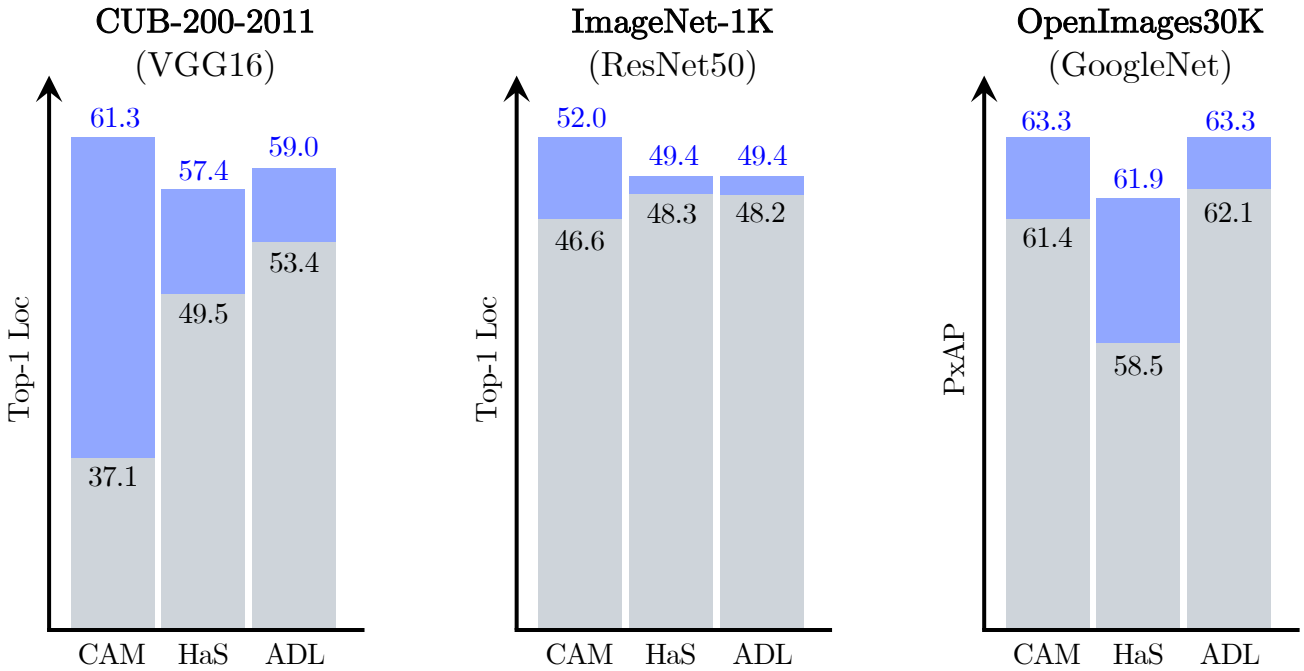


Experiment Results: Different CAM-based Methods



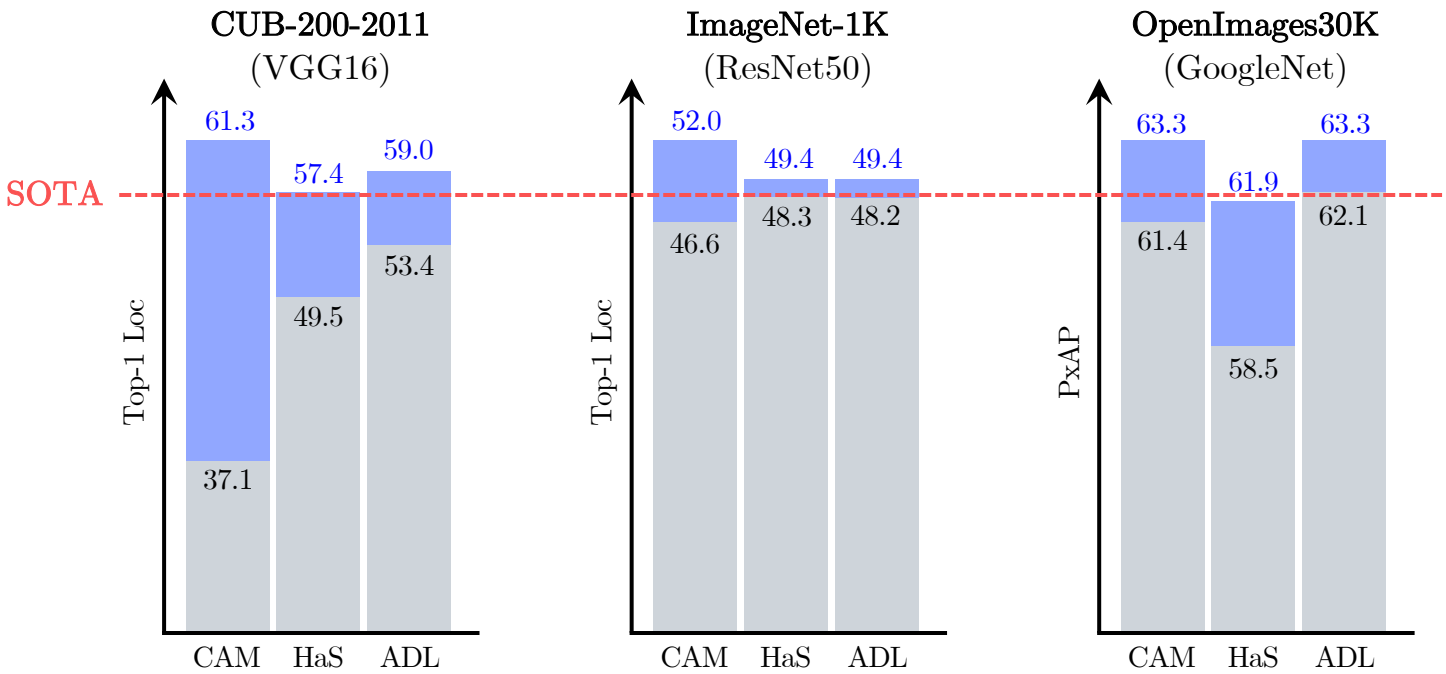
Experiment Results: Different CAM-based Methods

- Our proposed method significantly improved the localization performance.

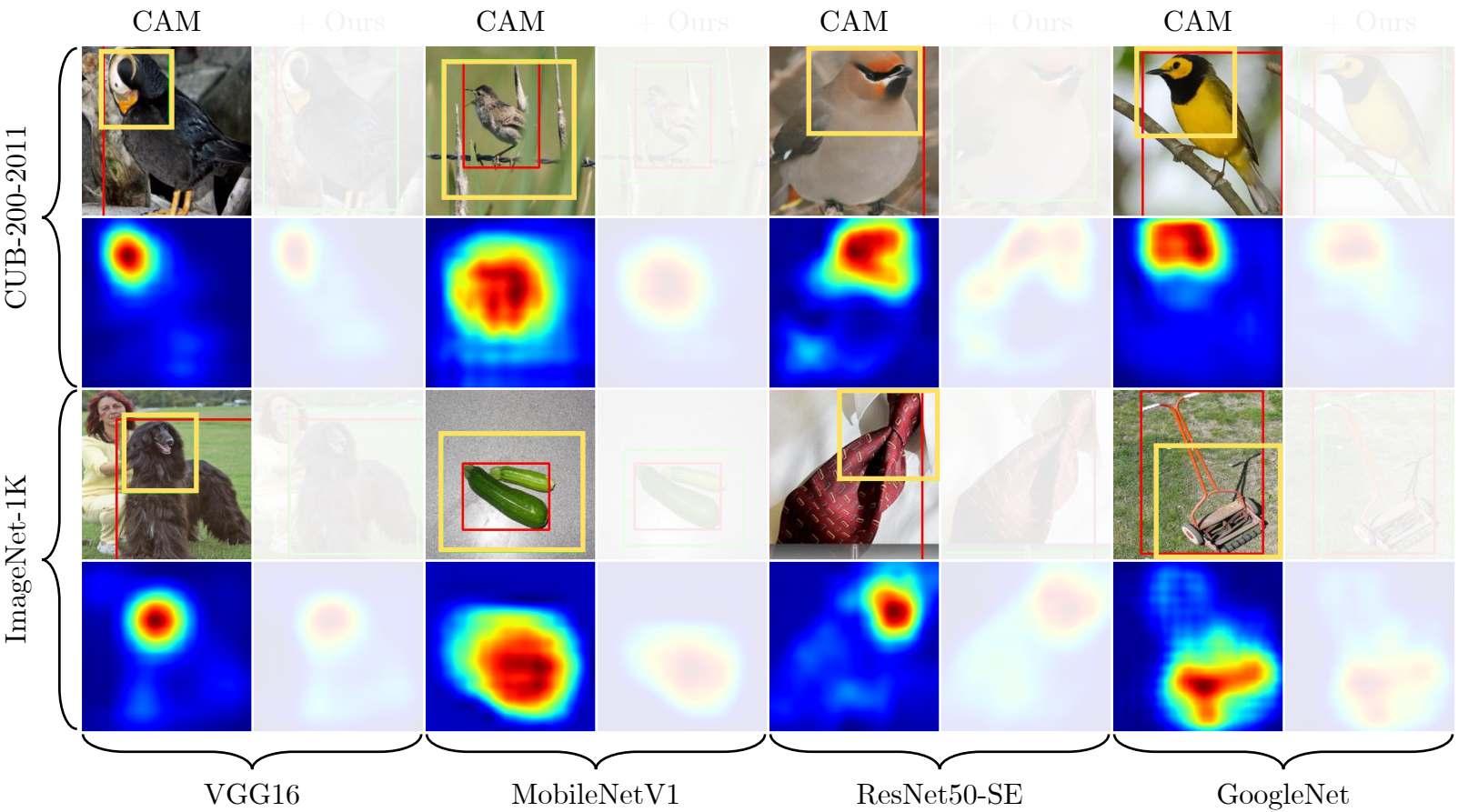


Experiment Results: Different CAM-based Methods

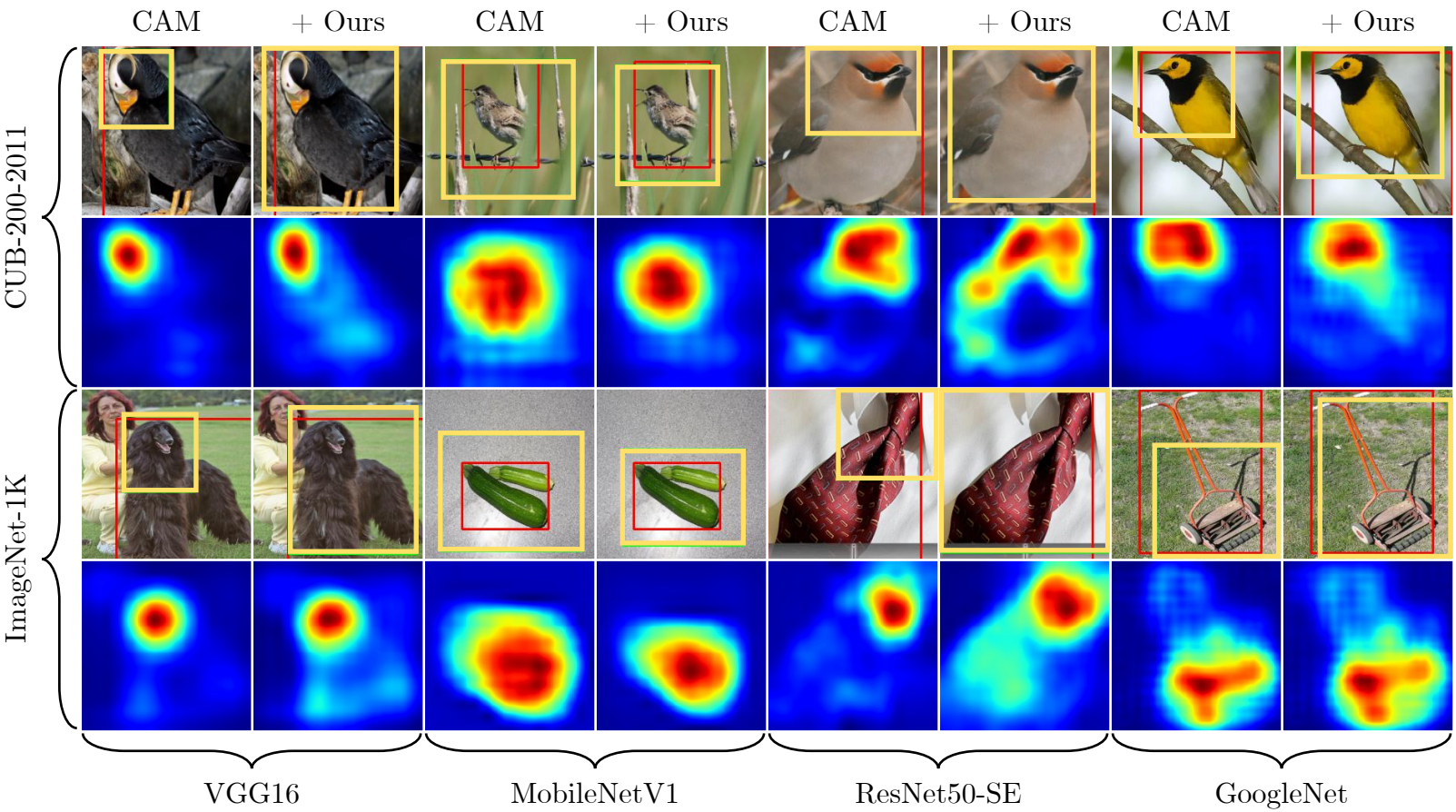
- Our proposed method significantly improved the localization performance.
- As a result, we achieved the new state-of-the-art performance on all datasets.



Qualitative Results



Qualitative Results



Summary

- Demonstrated the underlying issues of CAM, and the mechanism of them making the localization to be limited to a small discriminative region.
- Proposed three solutions that alleviate the issues in each of the corresponding modules of CAM.
- Verified our proposed method consistently improved the localization performance regardless of datasets, backbones, and CAM-based methods, and achieved the new state-of-the-art performance on all three benchmark datasets.

Thank You!

For more details, please check out our project page.

http://vision.snu.ac.kr/projects/rethinking-cam-wsol