# Attention-based Selection Strategy for Weakly Supervised Object Localization

Zhenfei Zhang and Tien D.Bui

# Introduction

Weakly Supervised Object Localization aims to locate objects by only using image-level labels.

# Limitations

- Class Activation Map (CAM) can only highlight the most discriminative parts.

- Current hiding method is very blind.

# Method



First selection

Second selection

Last selection

**Algorithm 1** Selection strategy

**Input:** $x \subseteq R^{H \times W \times C}$: feature map; $a$: the number of the most discriminative pixels in one featuere map; $b$: the total pixel number of one feature map; $v$: the most discriminative pixel dispersion; $d$: total pixels of target
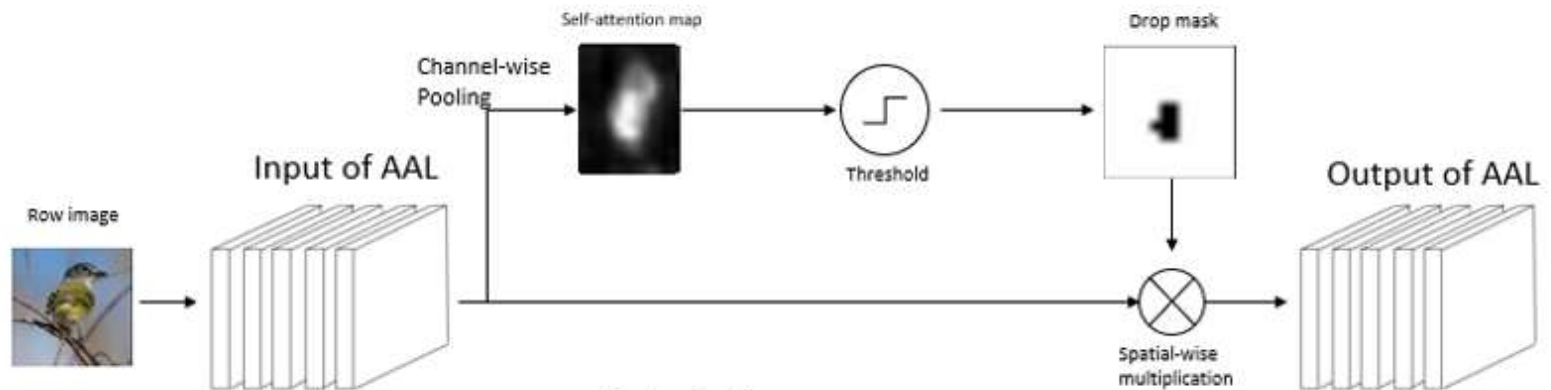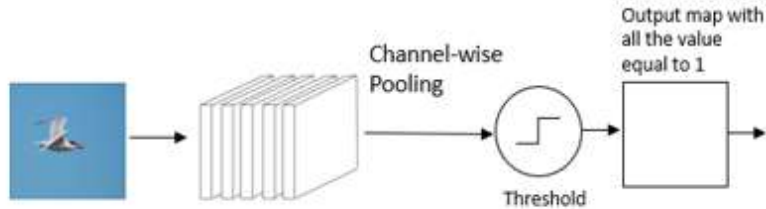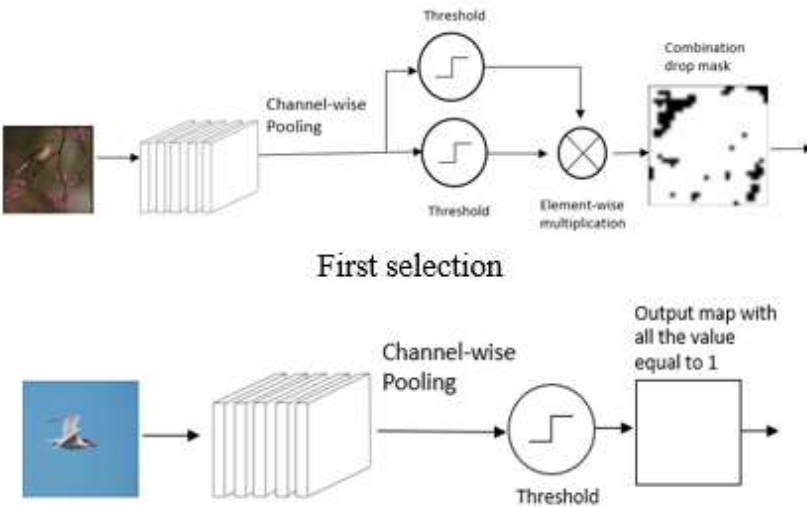
**Output:** $y$ : dropped feature maps

1: **for** $i = 0 \rightarrow batch\_size$ **do**
2:     **if** $v >=$ dispersion threshold **and** $a >= 0.05b$ **then**
3:         selected_map[i] = a drop mask that can remove unhelpful information
4:     **else if** $d/b <$ small area threshold **or** $a/d >$ drop mask proportion threshold **then**
5:         selected_map[i] = a feature map that all the values are 1
6:     **else**
7:         selected_map[i] = a drop mask to remove the most discriminative part
8:     **end if**
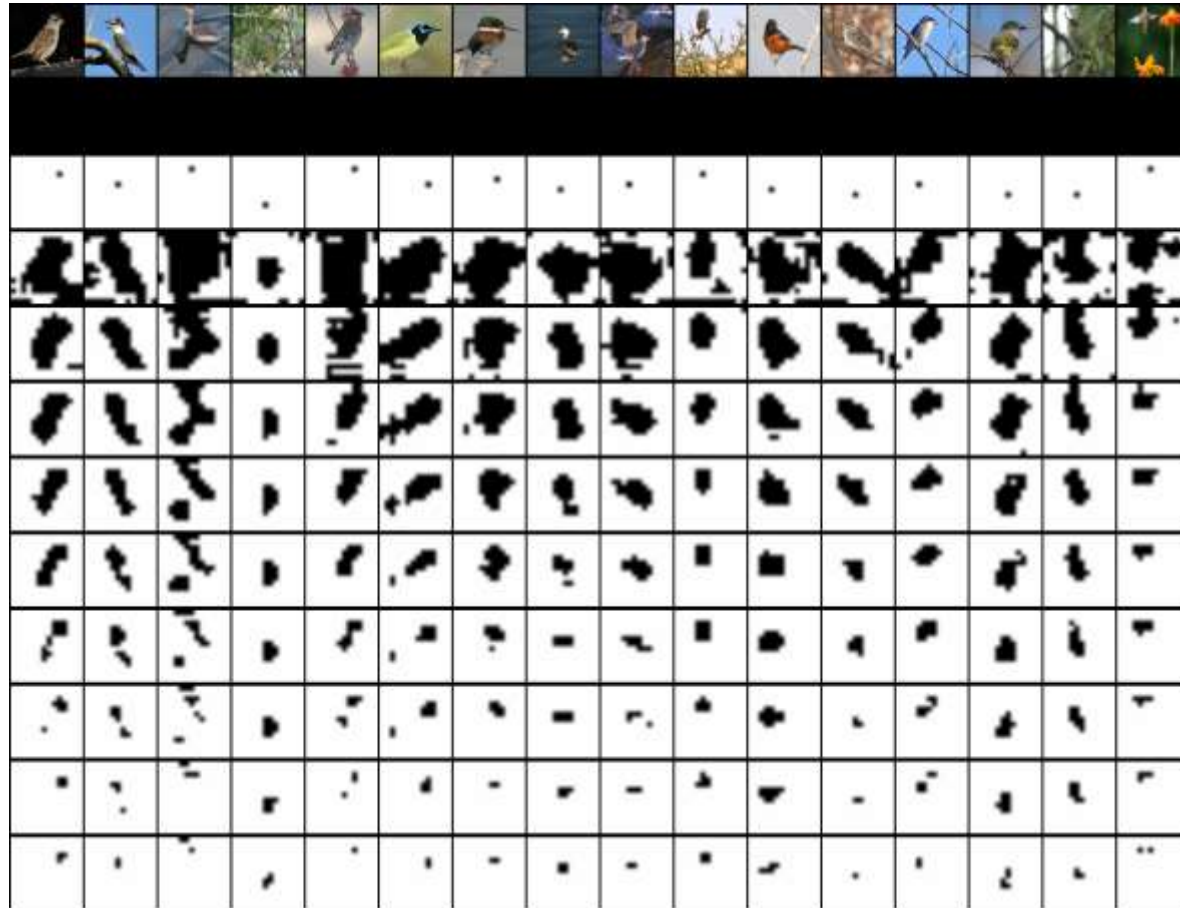9: **end for**
10: **return** y = x · selected_map

The drop masks using different threshold

# Results
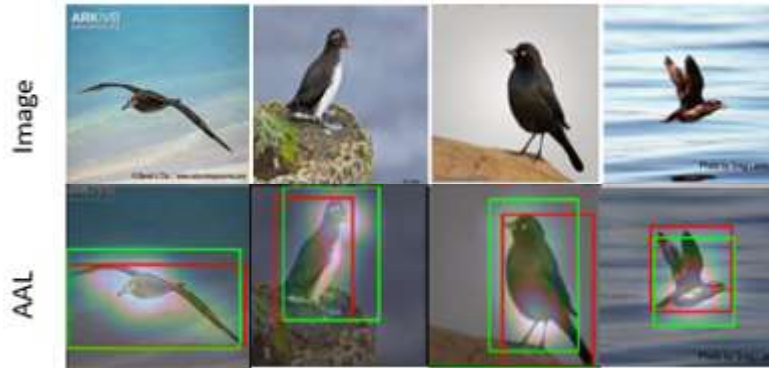


| Method | Top-1 cls.err | Top-1 loc.err |
|---|---|---|
| VGGnet-CAM [8] | **23.86** | 66.05 |
| VGGnet-ACoL [12] | 28.10 | 54.08 |
| VGGnet-SPG [11] | 24.50 | 50.00 |
| VGGnet-ADL [9] | 34.73 | 47.64 |
| VGGnet-DANet [10] | 24.60 | 47.48 |
| VGGnet-Cutmix [13] | - | 47.47 |
| VGGnet-ASS(Ours) | 25.51 | **45.45** |

| Method | Top-1 cls.err | Top-1 loc.err |
|---|---|---|
| VGGnet-CAM [8] | 33.40 | 57.20 |
| VGGnet-ACoL [12] | 32.50 | 54.17 |
| VGGnet-ADL [9] | **30.52** | 55.08 |
| VGGnet-Cutmix [13] | - | 56.55 |
| VGGnet-ASS(Ours) | 30.59 | **53.76** |

# Summary

- we propose a simple but effective method for Weakly Supervised Object Localization task.

- Our method dynamically generates the drop masks according to different input feature map.

# Thank you for listening