Foreground Mining via Contrastive Guidance for Weakly Supervised Object Localization

Wonyoung Lee[1]
lwy8555@yonsei.ac.kr
Minsong Ki[2]
mski1019@lguplus.co.kr
Cheolhyun Mun[1]
cheolhyunmun@yonsei.ac.kr
Sungpil Kho[3]
khosungpil@yonsei.ac.kr
Hyeran Byun[13]
hrbyun@yonsei.ac.kr

[1] Department of Artificial Intelligence
Yonsei University
Seoul, Republic of Korea

[2] AI Imaging Tech. Team
LG Uplus
Seoul, Republic of Korea

[3] Department of Computer Science
Yonsei University
Seoul, Republic of Korea

Abstract

Weakly supervised object localization (WSOL) locates the target object within an image using only image-level labels. Recent methods try to extend the feature activation to cover entire object regions by dropping the most discriminative parts. However, they either overextend the activation into the background or are still limited to covering the most discriminative parts. In this paper, we propose a novel WSOL framework that localizes the entire object to the right extent via contrastive learning. Our framework contains three key components: 1) scheduled region drop, 2) contrastive guidance, and 3) pairwise non-local block. The scheduled region drop progressively erases the most discriminative parts of the original feature at a region-level. The erased feature facilitates the network to discover less discriminative regions in the original feature. Then, our contrastive guidance encourages the foregrounds of the original and erased features to be closer while pushing away from each background. In this manner, the network earns the capacity to differentiate the foregrounds from backgrounds, spreading out the activation within object regions. Last but not least, we utilize the pairwise non-local block, which provides an enhanced attention map to strengthen the spatial correlations between each pixel. In conclusion, our method achieves the state-of-the-art performance on CUB-200-2011 and ImageNet benchmarks regarding *Top-1 Loc*, *GT-Loc* and *MaxBoxAccV2*.

1 Introduction

Fully supervised methods [1, 4, 6, 24] have achieved remarkable performance by training a convolution neural network (CNN) with human-annotated labels (*e.g.*, bounding box for localization, pixel-level mask for segmentation). However, they require expensive annotation

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

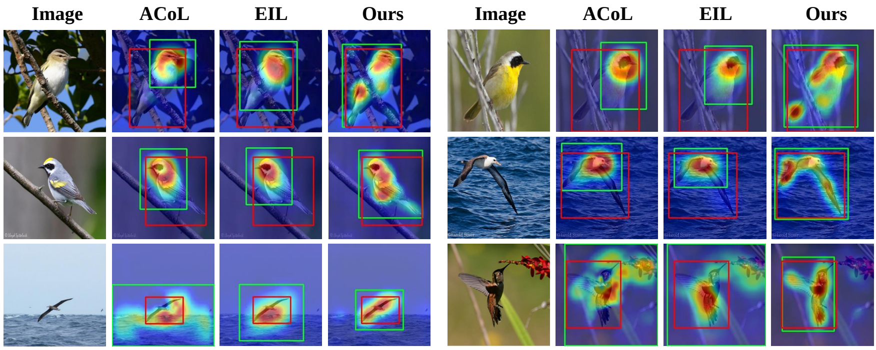


Figure 1: Comparison of localization results with existing WSOL methods on CUB-200-2011 [33] dataset. Both ACoL [38] and EIL [20] attempt to expand activation by discovering complementary regions. However, they rather locate only the most discriminative parts (first, second row) or overextend the activation to the backgrounds (last row). In contrast, ours spread activation on the full extent of the target object without excessively expanding to the background. The ground-truth boxes are in red and predicted boxes are in green.

costs for the target tasks. Therefore, weakly supervised approaches have been actively researched over the various computer vision tasks [2, 17, 26, 28, 30, 39] due to the lower cost to obtain weak supervision. Especially, we focus on weakly supervised object localization (WSOL) task that conducts localization in a given image using only class labels for training.

For example, Zhou *et al.* [41] propose class activation mapping (CAM) that extracts a class-specific localization map with a global average pooling layer (GAP). However, CAM tends to focus on the most discriminative parts of the target object, degrading the localization performance. To relieve this limitation, recent works introduce adversarial erasing (AE) methods [13, 16, 20, 38] to spread out the activation by erasing the most discriminative parts. These methods construct a dual-branch that one activates the most discriminative parts in the original feature map (original branch) while the other mines the complementary regions at the erased feature map (erased branch). However, they still concentrate class-specific local regions or overextend the activation to the backgrounds (Figure 1).

In this paper, we propose a novel AE-based framework using dual-branch features for mining the foregrounds to the right extent of the target object. Our framework consists of three key elements: scheduled region drop (SRD), contrastive guidance (CG), and pairwise non-local block (PNL). The scheduled region drop erases the most discriminative parts progressively on the original feature map at a region-level. It promotes the network to discover less informative regions in an effective way. The contrastive guidance encourages the foreground features of the dual-branch to pull together while pushing away from each background feature. This leads the model to learn the representation of the foregrounds that distinguish from backgrounds, preventing the expansion of activations to the backgrounds. Also, the pairwise non-local block learns the relationship between pixels in the feature map, which accelerates the network to discover other relevant parts of the most distinctive area. We validate that each proposed component plays an important role in improving localization performance. Finally, we verify the effectiveness of our method throughout extensive experiments, considerably outperforming the existing WSOL methods in CUB-200-2011 and ImageNet benchmarks.

2 Related Work

Weakly Supervised Object Localization trains a CNN classifier only using image-level labels and extracts a CAM [41] to highlight discriminative regions. Recent methods [7, 20, 28, 38] propose adversarial erasing (AE) to expand activations from the most discriminative parts to the less discriminative regions. Hide-and-Seek (HaS) [28] divides the input image into patches and randomly hide in training phase. Adversarial Complementary Learning (ACoL) [38] partially drops the most discriminative part to discover non-discriminative parts. Attention-based Dropout Layer (ADL) [7] produces a drop mask for hiding the most discriminative part and an importance map for highlighting informative region. Erasing Integrated Learning (EIL) [20] integrates two branches that one with an erased feature map and one with unerased feature map for both localization and classification. These erasing approaches incompletely eliminate discriminative parts by erasing in pixel-level through simple thresholding. In contrast, our scheduled region drop steps further to erase the discriminative parts at a region-level so that the model to find complementary regions efficiently. Also, we better localize the target object utilizing additional contrastive guidance.

Contrastive Learning has drawn significant attention due to its great achievement in unsupervised representation learning [5, 12, 21, 22, 34]. Their key idea is to maximize the agreement between the positive samples while minimizing that of negative samples. Existing works utilize contrastive learning in various vision tasks [9, 10, 14, 18]. Recently, Ki *et al*. [15] have first introduced contrastive learning in the WSOL task. They construct different views in a single feature map to exploit contrastive loss to cover the sophisticated object region. Different from [15], we design the contrastive samples with dual-branch feature maps. Thus, our proposed contrastive guidance loss optimizes quadruple relation (foreground and background feature maps of the original and erased branch), utilizing complementary discovered regions in the target object. It guides our network to discover the entire object to the right extent.

3 Proposed Method

3.1 Framework Overview

As shown in Figure 2, our WSOL framework utilizes the classification network and trains it with the contrastive guidance loss and classification loss using only class labels. The SRD generates an erased feature map $\bar{\mathbf{X}}$, which becomes an input of the erased branch. This branch shares the weight from the original branch. The network feed-forwards original and erased feature maps ($\mathbf{X}$, $\bar{\mathbf{X}}$) simultaneously and outputs the final feature maps ($\mathbf{F}$, $\bar{\mathbf{F}}$), exploring complementary regions. The pairwise non-local block produces the enhanced feature maps by learning the contextual information between pixel relationships. Then, the enhanced feature maps are served as input to the contrastive guidance to compute our contrastive loss. The contrastive guidance loss $\mathcal{L}_{\mathbf{cg}}$ guides the network to explore the entire object regions without spreading the activation map to the backgrounds. The final objective of our network is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}^{orig} + \mathcal{L}_{cls}^{erased} + \mathcal{L}_{cg} \tag{1}$$

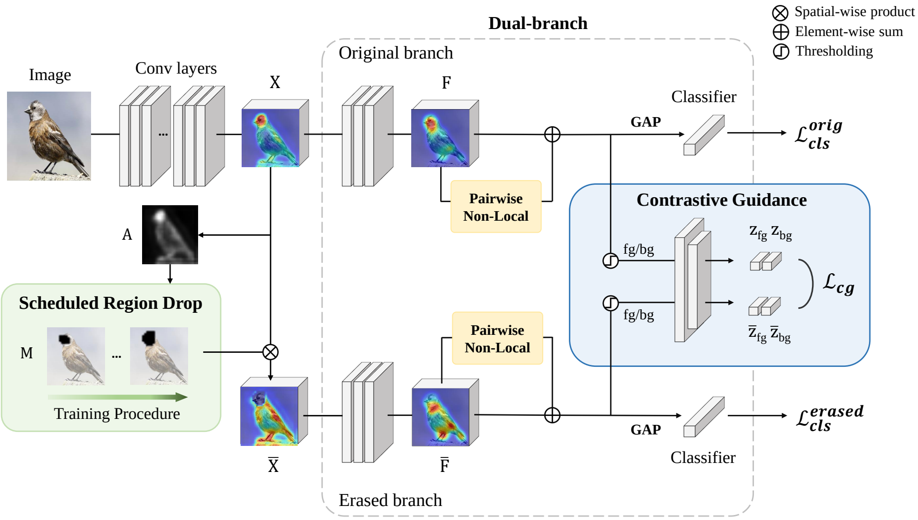


Figure 2: The overview of our framework. The scheduled region drop (SRD) produces the erased feature map $\bar{\mathbf{X}}$ by progressively dropping the most discriminative parts. The pairwise non-local block (PNL) generates an enhanced attention map considering the pixel-wise spatial relationships. Finally, we compute the contrastive guidance loss $\mathcal{L}_{\mathbf{cg}}$ that constructs the foregrounds as positive samples and backgrounds as negative samples.

3.2 Scheduled Region Drop

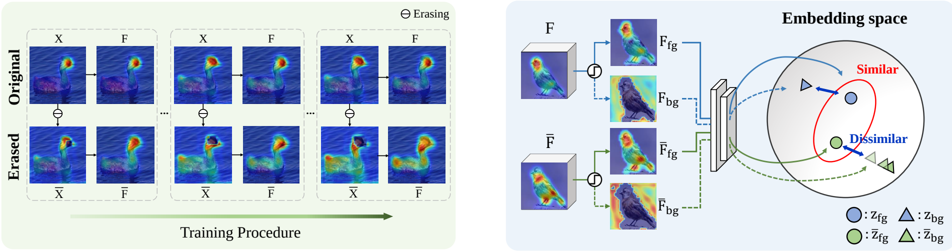
Conventional WSOL methods using adversarial erasing [7, 15, 20, 32, 38] produce erased feature maps by dropping the most discriminative parts at the pixel level. However, it is challenging to remove the adjacent pixels to the most informative parts completely using only pixel-level dropping. These remaining informative pixels hinder the erased branch from discovering complementary regions (*i.e.*, less discriminative parts of the target object).

We propose a region-level erasing strategy to remove the distinctive area more effectively. First, we obtain an attention map $\mathbf{A} \in \mathbb{R}^{1\times H\times W}$ of the original feature map $\mathbf{X}$ by channel-wise pooling. Then, we generate a pixel-level binary mask $\mathbf{M_{pix}} \in \mathbb{R}^{1\times H\times W}$ by:

$$\mathbf{M_{pix}} = \mathbb{1}[\mathbf{A} > \tau_{\text{drop}}], \quad where \quad \tau_{\text{drop}} = max(\mathbf{A}) \times \theta_{\text{drop}} \tag{2}$$

τ_{d} denotes the maximum intensity of $\mathbf{A}$ times pre-defined drop threshold θ_{d}.

We generate region drop mask $\mathbf{M}$ by expanding each pixel in $\mathbf{M_{pix}}$ to the size of $\mathbf{S} \times \mathbf{S}$ squared region. Specifically, we apply max pooling layer with a kernel size of $(\mathbf{S}, \mathbf{S})$ to $\mathbf{M_{pix}}$. At last, the erased feature map $\bar{\mathbf{X}}$ is produced by spatial-wise multiplication between $\mathbf{X}$ and $\mathbf{M}$. Both $\mathbf{X}$ and $\bar{\mathbf{X}}$ are fed into the afterward layers of the network concurrently, which are sharing the weights. In addition, we observe that the fixed drop threshold θ_{d} induces the unstable performance. The erased branch suffers from classifying at the early training phase because of discarding the most discriminative parts in a wide range (*i.e.*, region-level dropping). To remedy this issue, we reduce the discrepancy between a dual-branch at the start of the training by decreasing the drop threshold linearly from 1 to θ_{d}. Overall, our SRD gradually increase the erasing area and successfully expand the activation to less discriminative regions, as in Figure 3-a.



(a) Scheduled region drop (b) Contrastive guidance

Figure 3: (a) The changes of activation in the feature maps of the original branch ($\mathbf{X}$, $\mathbf{F}$) and the erased branch ($\bar{\mathbf{X}}$, $\bar{\mathbf{F}}$). (b) The foregrounds and backgrounds of the final feature maps ($\mathbf{F}$, $\bar{\mathbf{F}}$) are projected to the embedding space for modeling contrastive guidance loss.

3.3 Contrastive Guidance

Contrastive learning [6, 12, 21, 34] aims to learn a meaningful representation by attracting positive pairs while pushing their negative pairs away. Likewise, we construct the foregrounds as positive pairs and backgrounds as negative pairs for using this concept of contrastive learning (Figure 3-b).

The final feature maps ($\mathbf{F}$, $\bar{\mathbf{F}}$) are encoded from the dual-branch with the original $\mathbf{X}$ and erased feature map $\bar{\mathbf{X}}$, respectively. We generate the foreground and background masks ($\mathbf{M_{fg}}$, $\mathbf{M_{bg}}$) by thresholding the average intensity of channel-wise pooled attention map $\mathbf{A_F}$ as in Section 3.2. Then, we produce foreground and background features ($\mathbf{F_{fg}}$, $\mathbf{F_{bg}}$) multiplied with each mask:

$$\mathbf{M_{fg}} = \mathbb{1}[\mathbf{A}_{\mathrm{F}} > \tau_{\mathrm{fg}}], \quad \mathbf{M_{bg}} = \mathbb{1}[\mathbf{A}_{\mathrm{F}} < \tau_{\mathrm{bg}}], \tag{3}$$

$$\mathbf{F_{fg}} = \mathbf{F} \odot \mathbf{M_{fg}}, \quad \mathbf{F_{bg}} = \mathbf{F} \odot \mathbf{M_{bg}}, \tag{4}$$

where τ_{fg} and τ_{bg} are pre-defined thresholds. Each foreground and background feature is projected to the normalized embedding space with the projection head. It consists of two 1x1 convolution layers with ReLU activation and outputs each 128-dimension of feature vectors ($\mathbf{z}_{\mathrm{fg}}, \mathbf{z}_{\mathrm{bg}}, \bar{\mathbf{z}}_{\mathrm{fg}}, \bar{\mathbf{z}}_{\mathrm{bg}}$). Formally, our contrastive guidance loss is given by:

$$\begin{aligned}\mathcal{L}_{cg} = \Big\{ &\max\left[\|(\mathbf{z}_{\mathrm{fg}} - \bar{\mathbf{z}}_{\mathrm{fg}})\|_2 - \|(\mathbf{z}_{\mathrm{fg}} - \mathbf{z}_{\mathrm{bg}})\|_2 + m,\ 0\right] \\ &+ \max\left[\|(\bar{\mathbf{z}}_{\mathrm{fg}} - \mathbf{z}_{\mathrm{fg}})\|_2 - \|(\bar{\mathbf{z}}_{\mathrm{fg}} - \bar{\mathbf{z}}_{\mathrm{bg}})\|_2 + m,\ 0\right] \Big\},\end{aligned} \tag{5}$$

where m denotes the margin. Our loss function encourages to reduce the distance between the representation of $\mathbf{z}_{\mathrm{fg}}, \bar{\mathbf{z}}_{\mathrm{fg}}$ while enlarging the distance between their own backgrounds. It allows mining diverse complementary foregrounds within the full extent of the target object.

3.4 Pairwise Non-Local Block

We utilize the pairwise non-local block [36] to strengthen pixel-wise relationships regarding the target object region in the feature maps ($\mathbf{F}$, $\bar{\mathbf{F}}$). It produces the enhanced feature maps, which feed into the contrastive guidance and classifiers. The feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ is projected with three 1x1 convolution layers into $\{\mathbf{q}, \mathbf{k}, \mathbf{v}\} \in \mathbb{R}^{C' \times H \times W}$ which denotes query, key and value, respectively. The weight matrix $\mathbf{W} \in \mathbb{R}^{HW \times HW}$ represents similarities between each pixel that is obtained by whitened dot product operation of $\mathbf{q}, \mathbf{k}$:

$$\mathbf{W} = \sigma\left((\mathbf{q}_i - \mu_{\mathbf{q}})^T (\mathbf{k}_j - \mu_{\mathbf{k}}) \right), \tag{6}$$

where σ is a softmax function and $\mu_{\mathbf{q}}, \mu_{\mathbf{k}}$ are the spatial-wise average values from each pixel i, j in $\mathbf{q}, \mathbf{k}$, respectively. Then, the enhanced feature map $\mathbf{F}'$ is produced as:

$$\mathbf{F}' = \mathbf{F} \oplus h(\mathbf{v} \otimes \mathbf{W}), \tag{7}$$

where $h(\cdot)$ denotes 1x1 convolution layer followed by batch normalization.

The PNL learns where to attend, considering the similarities of the class-specific regions by optimizing the normalized difference between the query and key pixels. Therefore, they provide informative clues to the classifier and contrastive guidance.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate the proposed method on two benchmarks: CUB-200-2011 [33], ImageNet [25], which are given only image-level labels for training. CUB-200-2011 includes 200 species of bird consisting of 5,994 images for the training set and 5,794 images for the test set. ImageNet has 1,000 classes which contains 1.2 million and 50,000 images for training and test sets, respectively. We use CUBV2, ImageNetV2 [23] as a validation set, following [8].

Evaluation metrics. We leverage Top-1 localization (*Top-1 Loc*), GT-known localization (*GT-Loc*), and *MaxBoxAccV2* [8] to evaluate our methods. *Top-1 Loc* indicates the proportion of correctly classified images containing a bounding box intersection over union (IoU) 0.5 with the ground truth. *GT-Loc* measures the ratio, where the predicted box is considered as correct if an IoU greater than 50%. *MaxBoxAccV2* [8] averages the localization performances at three IoU criterions (0.3, 0.5, 0.7) by searching the optimal threshold for generating bounding boxes.

Implementation details. We build our method with three backbone networks: VGG16 [27], InceptionV3 [29], and ResNet50 [11]. All networks start training by loading ImageNet pretrained weights. Our PNL and CG are inserted before the classifier. We set drop threshold θ_{d} as 0.8 for CUB dataset, and 0.9 for ImageNet dataset. Thresholds of foreground τ_{fg} and background τ_{bg} are set to 0.9, 0.8 for VGG16 and others can be found in the supplementary material. For inference, we only utilize the scheduled region drop with its last drop threshold to extract the complementary region, as in [38]. Note that we follow the [3] for generating the class activation map of the target objects.

4.2 Ablation Study

The ablation studies for the proposed components are performed with VGG16 on CUB-200-2011 dataset. Bold texts denote the best performance.

Effects of each proposed component. We propose three components to localize the entire target object. Table 1 shows the effectiveness of individual elements in our framework. Compared to the baseline, our overall framework achieves a large performance gain with 10.12%,

Methods	SRD	CG	PNL	MaxBoxAccV2 (%) 0.3	0.5	0.7	Avg	Top-1 Loc (%)
Baseline* [20]	✗	✗	✗	97.58	78.91	34.64	70.38	59.87
Ours	✓	✓	✓	**99.00**	**88.63**	**53.88**	**80.50**	**65.60**
– SRD	✗	✓	✓	98.65	86.05	46.84	77.18	64.22
– CG	✓	✗	✓	98.29	83.07	41.58	74.31	62.67
– PNL	✓	✓	✗	98.58	86.78	47.26	77.54	63.98

Table 1: The ablation study of the main configurations of our method with VGG16 on CUB dataset in terms of *MaxBoxAccV2* and *Top-1 Loc*. SRD: scheduled region drop, CG: contrastive guidance, PNL: pairwise non-local block. Following [7, 20, 38], we use pixel-level erasing when SRD is not applied. * indicates reproduced results.

Location	MaxBoxAccV2 (%)	Top-1 Loc (%)
conv4_3	**80.50**	**65.60**
pool3	79.84	64.91
pool2	78.91	64.89

Table 2: Localization performance with VGG16 on CUB dataset regarding the location of scheduled region drop. MaxBoxAccV2 averages the performance at three IoU criterions.

		S 1	3	5	7
	0.8	77.5 / 64.4	**80.5 / 65.6**	77.3 / 64.1	68.2 / 55.3
θ_d	0.6	78.3 / 64.7	80.1 / 64.3	76.9 / 60.1	71.8 / 52.8
	0.4	79.3 / 64.9	78.9 / 62.3	69.8 / 52.2	56.6 / 38.8

Table 3: *MaxBoxAccV2* (%) / *Top-1 Loc* (%) performance with various combination of drop threshold (θ_d) and kernel size (**S**) in scheduled region drop with VGG16 on CUB dataset.

5.73% regarding *MaxBoxAccV2*, *Top-1 Loc*, respectively. Ours without the CG achieves 6.19% lower performance in terms of *MaxBoxAccV2* than the full setting, and especially degrades 12.30% at IoU 0.7. It is necessary to provide guidance on the foreground and background area of complementary feature maps in a given image to the network to localize the entire object. SRD also improves the performance by 3.32%. Except for the PNL in our framework, the performance decreases by 2.96%, and the degradation is the smallest compared to the two elements. As a result, we show the best performance when all components are employed.

Location and size of our SRD. First, we analyze the impact of the erasing location on the performance. As in Table 2, we achieve the best performance when SRD is inserted after *conv4_3* layer. However, in the case of SRD located at early layers (*pool2*, *pool3*), the performance slightly decreases. As discussed in previous works [7, 20], we note that this is because the earlier layers extract general features, activate locally distinctive parts (*e.g.*, edges, corners) in the feature map. In addition, we investigate the performance according to different drop threshold (θ_d) and kernel sizes (**S**) of the erased region in Table 3. We show the best performance by setting the θ_d to 0.8 and **S** to 3. The selection of smaller θ_d and larger **S** results poor localization performance since it erases excessive information in the original feature map. Although our SRD gradually increases the erasing area, we believe that the erased branch suffers in optimizing contrastive guidance loss and classification loss without sufficient clues of the target object.

Comparison with existing contrastive loss and our CG loss. Table 4 shows the results when CG loss is replaced with conventional contrastive loss (*i.e.*, *InfoNCE* loss [5, 22]). According to experimental results, we observe that our method still surpasses the existing state-of-the-art WSOL performances in a large margin of 7.7%, even though using *InfoNCE* loss. However, it is significantly inferior to ours w/ CG (last row) at IoU 0.7. Also, the per-

Methods	MaxBoxAccV2 (%)				Top-1 Loc (%)
	0.3	0.5	0.7	Avg	
Ours (w/o CG)	98.29	83.07	41.58	74.31	62.67
Ours (w InfoNCE)	98.44	86.38	48.88	77.90	63.46
Ours†	98.79	87.50	50.19	78.89	64.21
Ours	**99.00**	**88.63**	**53.88**	**80.50**	**65.60**

Table 4: Ablation study of contrastive guidance (CG) loss with VGG16 on CUB dataset. Ours† indicates that we only use the background of the original feature map as a negative sample.

Methods	CUB-200-2011				ImageNet			
	VGG	Inc	Res	Avg	VGG	Inc	Res	Avg
CAM [41]	63.7	56.7	63.0	61.1	60.0	63.4	63.7	62.4
HaS [28]	63.7	53.4	64.7	60.6	60.6	63.7	63.4	62.6
ACoL [38]	57.4	56.2	66.5	60.0	57.4	63.7	62.3	61.2
SPG [39]	56.3	55.9	60.4	57.5	59.9	63.3	63.3	62.2
ADL [7]	66.3	58.8	58.4	61.1	59.8	61.4	63.7	61.7
CutMix [37]	62.3	57.5	62.8	60.8	59.4	63.9	63.3	62.2
InCA [15]	66.7	60.3	63.2	63.4	61.3	62.8	65.1	63.1
MinMaxCAM [31]	70.2	-	68.0	-	62.2	-	**65.7**	-
Ours	**80.5**	**75.8**	**73.3**	**76.5**	**65.3**	**64.8**	65.5	**64.7**

Table 5: *MaxBoxAccV2* [8] comparison with the WSOL state-of-the-art methods. InCA [15], MinMaxCAM [31] values are taken from their respective papers and the others are from [8].

formance of ours w/o CG loss seriously degrades at IoU 0.7. It indicates that our CG loss provides adequate guidance to the network rather than the existing contrastive loss to cover the entire object well. Moreover, we also validate the effectiveness of dual-branch in contrastive learning (third row). Similar to [15], Ours† only uses the background of the original feature map as a negative sample. It shows the performance drops when the background of the erased feature map is discarded. Consequentially, the background of the erased feature map plays an important role to find out less discriminative parts by extending the activation within the boundary of the target object. The detailed objective function can be found in the supplementary material.

4.3 Comparison with State-of-the-art Methods

We compare our method with WSOL state-of-the-art methods on CUB-200-2011 and ImageNet datasets in terms of *MaxBoxAccV2* [8], *GT-known Loc*, and *Top-1 Loc*.

MaxBoxAccV2 [8]. In Table 5, our method outperforms all the others on CUB and ImageNet datasets in terms of the *MaxBoxAccV2* for three backbones. We achieve remarkable improvement on CUB (+13.1%), and on ImageNet (+1.6%). In particular, our method gains 15.5% over InCA [15] on CUB-InceptionV3 and 3.1% over MinMaxCAM [31] on ImageNet-VGG16. The detailed performance at three IoU criterions can be found in the supplementary material.

GT-known Loc and Top-1 Loc. Table 6 shows quantitative results using conventional metrics. On both CUB and ImageNet datasets, our method achieves the state-of-the-art performance regarding *GT-Loc*, *Top-1 Loc*.

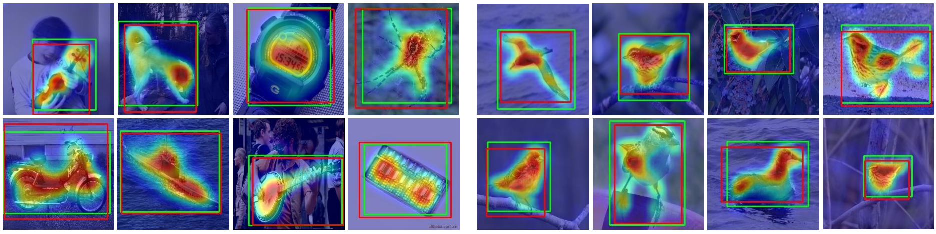
Methods	Backbone	CUB-200-2011		ImageNet	
		GT-Loc	Top-1 Loc	GT-Loc	Top-1 Loc
CAM [41]	VGG16	56.00	44.15	57.72	42.80
ACoL [38]	VGG16	54.10	45.92	62.96	45.83
ADL [7]	VGG16	75.41	52.36	-	44.92
MEIL [20]	VGG16	-	57.46	-	46.81
RCAM [3]	VGG16	80.72	61.30	61.69	44.69
GCNet [19]	VGG16	81.10	63.24	-	-
I2C [40]	VGG16	-	-	63.90	47.41
Ours	VGG16	**88.54**	**65.60**	**65.04**	**48.01**
CAM [41]	InceptionV3	55.10	43.70	62.68	46.30
SPG [39]	InceptionV3	-	46.64	64.69	48.60
DANet [35]	InceptionV3	67.70	52.52	-	47.53
RCAM [3]	GoogLeNet	65.10	51.05	62.76	47.70
GCNet [19]	InceptionV3	75.30	58.58	-	49.10
I2C [40]	InceptionV3	-	55.99	**68.50**	**53.11**
Ours	InceptionV3	**87.95**	**64.72**	66.86	50.63
CAM [41]	ResNet50	-	49.41	51.86	38.99
CutMix [37]	ResNet50	-	54.80	-	47.30
ADL [7]	ResNet50-SE	-	62.29	-	48.53
RCAM [3]	ResNet50-SE	74.51	58.39	64.40	51.96
I2C [40]	ResNet50	-	-	**68.50**	**54.83**
Ours	ResNet50	**85.17**	**69.71**	66.46	52.59

Table 6: Localization performance comparison with the state-of-the-art methods.

4.4 Qualitative results

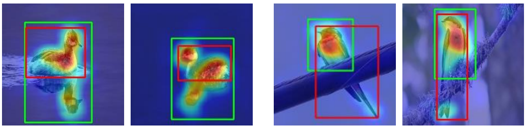
Figure 4 illustrates activation maps and estimated bounding boxes. Our method localizes on the full object correctly and outputs tight bounding boxes compared with ground truth. We constrain the background region using SRD and CG loss at the training phase. Therefore, our method not only spreads out to the less discriminative parts but also suppresses the activations on backgrounds.

Unfortunately, some challenging case exists, as in Figure 5. The reflection on the water surface and occlusion of the target object generates either larger or smaller bounding boxes.



(a) ImageNet (b) CUB-200-2011

Figure 4: Qualitative results of our method on ImageNet and CUB-200-2011 dataset. The ground-truth boxes are in red and predicted boxes are in green.



(a) Reflection (b) Occlusion

Figure 5: Qualitative results of failure cases with our method on CUB dataset. The ground-truth boxes are in red and predicted boxes are in green.

5 Conclusion

In this paper, we propose a novel WSOL framework using adversarial erasing strategy in a dual-branch. The scheduled region drop gradually erases discriminative parts of the original feature map using region-level dropping to capture complementary parts of the target object. The contrastive guidance leverages foreground and background features in dual-branch to encourage their foregrounds to be similar and penalize each corresponding background. Also, the pairwise non-local block learns the pixel correlation of feature maps which provide enhanced feature maps. In this way, our method allows the model to cover the right extent of the target object. Finally, we achieve the state-of-the-art performance on CUB-200-2011 and ImageNet datasets.

Acknowledgements. This work was supported by the National Research Foundation of Korea grant funded by Korean government (No. NRF-2019R1A2C2003760) and Artificial Intelligence Graduate School Program (YONSEI UNIVERSITY) under Grant 2020-0-01361.

References

[1] Manoj Acharya, Tyler L Hayes, and Christopher Kanan. Rodeo: Replay for online object detection. In *The British Machine Vision Virtual Conference (BMVC)*, 2020.

[2] Sadbhavana Babar and Sukhendu Das. Where to look?: Mining complementary image regions for weakly supervised object localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1010–1019, 2021.

[3] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *European Conference on Computer Vision*, pages 618–634. Springer, 2020.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020.

[6] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2061–2069, 2019.

[7] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

[8] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020.

[9] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, J. Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *ECCV*, 2020.

[10] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. *arXiv preprint arXiv:2008.01065*, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[13] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *arXiv preprint arXiv:1810.09821*, 2018.

[14] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X. Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. *ArXiv*, abs/2105.00957, 2021.

[15] Minsong Ki, Youngjung Uh, Wonyoung Lee, and Hyeran Byun. In-sample contrastive learning and consistent attention for weakly supervised object localization. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[16] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017.

[17] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11320–11327, 2020.

[18] Dichao Liu, Yu Wang, Jien Kato, and Kenji Mase. Contrastively-reinforced attention convolutional neural network for fine-grained image recognition. In *BMVC*, 2020.

[19] Weizeng Lu, Xi Jia, Weicheng Xie, Linlin Shen, Yicong Zhou, and Jinming Duan. Geometry constrained weakly supervised object localization. *arXiv preprint arXiv:2007.09727*, 2020.

[20] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8766–8775, 2020.

[21] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[24] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7406–7415, 2020.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[26] Tong Shen, Guosheng Lin, Lingqiao Liu, Chunhua Shen, and Ian Reid. Weakly supervised semantic segmentation based on co-segmentation. In *BMVC*, 2017.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Krishna Kumar Singh and Y. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553, 2017.

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[30] Eu Wern Teh, Mrigank Rochan, and Yang Wang. Attention networks for weakly supervised object localization. In *The British Machine Vision Virtual Conference (BMVC)*, pages 1–11, 2016.

[31] Kaili Wang, Jose Oramas, and Tinne Tuytelaars. Minmaxcam: Improving object coverage for cam-basedweakly supervised object localization. *arXiv preprint arXiv:2104.14375*, 2021.

[32] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[33] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[34] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[35] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6589–6598, 2019.

[36] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020.

[37] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[38] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018.

[39] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 597–613, 2018.

[40] Xiaolin Zhang, Yunchao Wei, and Yi Yang. Inter-image communication for weakly supervised localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 271–287. Springer, 2020.

[41] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.