# (Supplementary Material) 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

This supplementary material contains four parts:

- Section A describes the implementation detail of our contrastive guidance replaced by existing contrastive loss.
- Section B provides hyperparameters of each dataset and backbone.
- Section C shows the quantitative results of our method upon all the backbones and datasets in terms of MaxBoxAccV2 [2].
- Section D illustrates more qualitative results of our method.

## A Implementation details of *InfoNCE* loss

In Table 4 of the main paper, we report the performance when our contrastive guidance loss $\mathcal{L}_{\text{cg}}$ is replaced by two different cases. First, we simply replace Eq.(5) with *InfoNCE* loss [1, 5] as:

$$\begin{aligned}\mathcal{L}_{info} = &-log(\frac{exp(sim(\mathbf{z}_{\text{fg}}, \bar{\mathbf{z}}_{\text{fg}})/\tau)}{exp(sim(\mathbf{z}_{\text{fg}}, \bar{\mathbf{z}}_{\text{fg}})/\tau) + exp(sim(\mathbf{z}_{\text{fg}}, \mathbf{z}_{\text{bg}})/\tau)}) \\ &-log(\frac{exp(sim(\bar{\mathbf{z}}_{\text{fg}}, \mathbf{z}_{\text{fg}})/\tau)}{exp(sim(\bar{\mathbf{z}}_{\text{fg}}, \mathbf{z}_{\text{fg}})/\tau) + exp(sim(\bar{\mathbf{z}}_{\text{fg}}, \bar{\mathbf{z}}_{\text{bg}})/\tau)}),\end{aligned} \tag{1}$$

where *sim* denotes cosine similarity between normalized embedded features, $\tau$ denotes temperature parameter.

| | ImageNet | | | CUB | | |
|---|---|---|---|---|---|---|
| | VGG | Inc | Res | VGG | Inc | Res |
| $\tau_{\text{fg}}$ | 0.7 | 1.0 | 0.9 | 1.1 | 0.9 | 1.3 |
| $\tau_{\text{bg}}$ | 0.6 | 0.8 | 0.8 | 1.0 | 0.7 | 0.9 |

Table 1: Hyperparameters ($\tau_{\text{fg}}$, $\tau_{\text{bg}}$) for contrsative guidance. VGG: VGG16 [6], Inc: InceptionV3 [7], Res: ResNet50 [3].

Additionally, we replace our $\mathcal{L}_{\text{cg}}$ with using only one negative sample(*i.e.*, background of original feature map $z_{\text{bg}}$) as:

$$\mathcal{L}_{cg}^{\dagger} = \left\{ \max \left[ \|(\mathbf{z}_{\text{fg}} - \bar{\mathbf{z}}_{\text{fg}})\|_2 - \|(\mathbf{z}_{\text{fg}} - \mathbf{z}_{\text{bg}})\|_2 + m,\ 0 \right] \right\}. \tag{2}$$

## B Hyperparameter setting

For scheduled region drop in section 3.2 of the main paper, we set the square size $\mathbf{S}$ to three for all the experiments. Also, for generating foreground and background masks ($\mathbf{M_{fg}}$, $\mathbf{M_{bg}}$) in contrastive gudiance (Eq.(4)), we set $\tau_{\text{fg}}$ and $\tau_{\text{bg}}$ as in Table 1. Both thresholds are multiplied with the average intensity of channel-wise pooled attention map $\mathbf{A_F}$.

## C Quantitative results at three IoU criterions in terms of MaxBoxAccV2

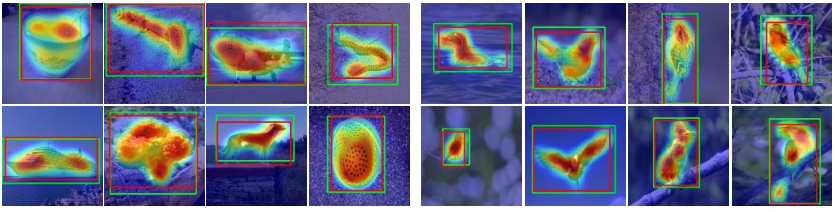
We further provide the performance upon all the backbones and datasets in terms of *MaxBoxAccV2* on the three IoU criterions, as in Table 2.

| | Backbone | Method | MaxBoxAccV2 (%) | | | | Top-1 Cls (%) |
|---|---|---|---|---|---|---|---|
| | | | 0.3 | 0.5 | 0.7 | Avg | |
| CUB-200-2011 | VGG16 | InCA [9] | 96.20 | 77.20 | 26.75 | 66.72 | 73.35 |
| | | Ours | 99.00 | 88.63 | 53.88 | **80.50** | 73.47 |
| | InceptionV3 | InCA [9] | 95.89 | 67.93 | 17.20 | 60.34 | 64.01 |
| | | Ours | 99.45 | 87.95 | 39.92 | **75.77** | 72.49 |
| | ResNet50 | ACoL [9] | 96.96 | 77.29 | 25.03 | 66.43 | 71.07 |
| | | Ours | 99.36 | 85.23 | 35.36 | **73.32** | 81.10 |
| ImageNet | VGG16 | InCA [9] | 81.45 | 63.20 | 39.20 | 61.33 | 69.21 |
| | | Ours | 84.12 | 66.89 | 45.03 | **65.35** | 68.89 |
| | InceptionV3 | CutMix [5] | 84.05 | 66.51 | 41.02 | 63.86 | 69.16 |
| | | Ours | 84.64 | 67.45 | 42.38 | **64.83** | 70.99 |
| | ResNet50 | InCA [9] | 84.26 | 67.62 | 43.58 | 65.15 | 76.54 |
| | | Ours | 84.54 | 67.43 | 44.61 | **65.53** | 74.75 |

Table 2: *MaxBoxAccV2* comparison with the state-of-the-art methods on each dataset and backbone. We also report *Top-1 Classification* for the reference.

## D Additional visualization results of our method

Figure 1 shows more qualitative results on ImageNet and CUB-200-2011 datasets. Our method covers the area of the target object accurately and also suppresses activations of the background.



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

Figure 1: Qualitative results of our method. The ground-truth boxes are in red and predicted boxes are in green.

# References


[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.