Dual-attention Guided Dropblock Module for Weakly Supervised Object Localization

Junhui Yin, Siqing Zhang, Dongliang Chang, Zhanyu Ma, Jun Guo
Beijing University of Posts and Telecommunications

Abstract. In this paper, we present a dual-attention guided dropblock module, and aim at learning the informative and complementary visual features for weakly supervised object localization (WSOL). The attention mechanism is extended to the task of WSOL, and design two types of attention modules to learn the discriminative features for better feature representations. Based on two types of attention mechanism, we propose a channel attention guided dropout (CAGD) and a spatial attention guided dropblock (SAGD). The CAGD ranks channel attention by a measure of importance and consider the top-\(k\) largest magnitude attentions as important ones. The SAGD can not only completely remove the information by erasing the contiguous regions of feature maps rather than individual pixels, but also simply distinguish the foreground objects and background regions to alleviate the attention misdirection. Extensive experiments demonstrate that the proposed method achieves new state-of-the-art localization accuracy on a challenging dataset.

Keywords: Weakly supervised object localization, spatial attention, channel attention, dropout.

1 Introduction

Weakly supervised object localization (WSOL) requires less detailed annotations to identify the object location in a given image [31] compared to the fully-supervised learning. WSOL is a challenging task since neural networks have access to only image-level labels (“cat” or “no cat”) that confirms the existence of the target object, but not the guidance of the expensive bounding box annotations in an image.

To address the WSOL problem with convolutional neural networks (CNNs), people resort to a general method, e.g., generating Class Activation Mapping (CAM) [31] for performing the object localization. Unfortunately, the CAM often covers only the most important regions of the target object instead of the entire object. The method improves the classification accuracy, but it also leads to localization accuracy degradation [4].

Existing approaches have explored adversarial erasing [29], Hide-and-Seek (HaS) [20], Attention-based Dropout Layer (ADL) [4]. Specifically, the Adversarial Complementary Learning (ACoL) approach [29] can efficiently locate different object regions and learn new and complementary parts belonging to the
same objects by two adversary classifiers. HaS [20] hides patches in an input image randomly, encouraging the network model to seek multiple relevant parts. ADL [4] hides the most discriminative part from feature maps to pursue the full object extent, rewarding the informative part to improve the recognition power of CNNs model. In fact, similar to the pixel-based dropout, these techniques are not really the region-based dropout. However, neighbouring pixels are correlated spatially on the feature map. These adjacent pixels share much of the same information. The pixel-based dropout discards the information, but the information are still passed on from the other adjacent pixels that are still active.

Erasing the most discriminative parts is simple yet powerful method for WSOL. For example, ADL is a lightweight method that uses the attention mechanism to learn the less discriminative regions of the target object. However, the erasing methods abandon all the information on the most discriminative regions. This forces the model to highlight the less discriminative parts and sometimes captures useless information of the background, which leads to the attention misdirection and the biased localization. The bounding box is too large to precisely locate the object, and the classification performance is not as good as before since the focused attention has been changed to other objects.

In this paper, we propose a dual-attention guided dropblock module (DGDM), a lightweight yet powerful method, for WSOL, which is illustrated in Figure 2. It contains two key components, channel attention guided dropout (CAGD) and spatial attention guided dropblock (SAGD), to learn the complementary information in the spatial and channel dimensions, respectively. Specifically, in CAGD, we first aggregate the spatial information of input feature map by GAP, to generate channel attention. We also rank the obtained channel attention according to a measure of importance (e.g., magnitude), and then drop out some elements with low importance. For SAGD, we perform channelwise average pooling to generate a self-attention map. We also generate an importance map using the sigmoid activation to highlight the most discriminative regions of target object and suppress less useful ones. We also threshold the self-attention map to generate a drop mask. It can not only efficiently remove the information by erasing contiguous regions of feature maps and simply sense the foreground objects and background regions to alleviate the attention misdirection.

Deep networks implemented with DGDM incorporate image classification and WSOL. In an end-to-end learning manner, the proposed method captures complementary and discriminative information for precise object localization while achieving good result of image classification.

Our main contributions include:

(1) We propose a lightweight and efficient attention module (DGDM) to improve the performance of WSOL. The spatial self-attention mechanism is designed to generate a drop mask and erase the most discriminative part of object. This drop mask can efficiently erase the information by removing contiguous regions of feature maps and simply sense the foreground objects and background regions.
(2) We extend the channel attention mechanism in the task of WSOL to model channel interdependencies. We rank channel attention according to a measure of importance and consider the top-\(k\) largest magnitude attentions as important. We also keep some low-valued elements to increase their value if they become important during training.

(3) The proposed approach can be easily employed to different CNNs classifiers and achieve a new accuracy on CUB-200-2011.

2 Related Work

**Attention mechanism.** Attention mechanism is an artificial data processing method learnt from human perception process [18]. It does not process all the data in equal, but focuses more weights on the most informative parts [18], [1]. Attention mechanisms have demonstrated their utility across various fields, such as scene segmentation [7], image localization and understanding [13], [3], [16], and image inpainting [28], [17]. In particular, the self-attention mechanism [23] is firstly proposed to draw global dependencies of inputs and applies it in machine translation. The squeeze-and-excitation module [12] is introduced to exploit the channel-interdependencies. Convolutional block attention module [26] is proposed to emphasize meaningful features by fusing cross-channel and spatial information together. However, these techniques bear extra training parameters for obtaining the attention map.

**Dropout in convolutional neural networks.** Dropout [11] introduced in 2012 has been proven to be a practical technique to alleviate overfitting in fully-connected neural networks, which drops neuron activations with some fixed prob-
ability during training. All activations are used during the testing phase, but the output is scaled according to the dropout probability. In the years since, many methods inspired by the original dropout technique have been proposed. They include dropconnect [25], variational dropout [14], Monte Carlo dropout [8] and many others. Some complete reviews on dropout methods can be found in a survey [15]. However, applying dropout to the feature map is less powerful. This reduction can largely be attributed to two factors. The first is that fully-connected layers already have much more parameters than convolutional layers, and therefore convolutional layers require less regularization. The second factor is that there is strong correlation between spatially adjacent pixels on the feature map. These pixels have much of the same information. Therefore, the pixel-based dropout abandons the information, but the information still be passed on from the other adjacent pixels that are still active.

Cutout [6] drops out contiguous regions of input images instead of individual pixels in the input layer of CNNs. This method induces the network to better utilize the contextual information of the image, rather than relying on a small set of specific features. Dropblock [9] generalizes Cutout by applying Cutout at every feature map in convolutional networks. Its main difference from regular dropout is that it discards contiguous regions from feature map rather than dropping independent random units. ADL [4] utilizes attention mechanism to find the maximally activated part and then drops them out. However, the method does not discard strongly activated region effectively, but the strongly activated pixels, as discussed in Subsection 3.3.

**Weakly supervised object localization.** WSOL is an alternative cheaper way to identify the object location in a given image by only using image-level supervision, i.e., presence or absence of object categories [5], [31], [2]. The WSOL method decomposes a given image into a “package” of regional proposals and iteratively choosing an instance (a proposal) from each bag (an image with multiple proposals) to minimize the image classification error in step-wised manner [5]. Recent research [31] utilizes CNNs classifier for specifying the spatial distribution of discriminative patterns for different image classes. A way to pursue full object extent is self-paced learning [30]. The self-produced guidance (SPG) approach utilizes a CNNs to incorporate high confident regions, and attention maps are then leveraged to capture the object extent with the auxiliary supervision masks of foreground and background regions. The other way to enhance object localization is about adversarial erasing [29], [20], [4]. These methods first activate the most discriminative regions on the feature map or input image and then drops them so that less discriminative regions can be highlighted during the training phase. Nevertheless, most existing approaches use alternative optimization, or requiring a lot of computing resources to erase the most discriminative regions exactly.
3 Dual-attention guided dropblock module

3.1 Spatial attention module

Let $F \in \mathbb{R}^{H \times W \times C}$ is a convolutional feature map. Note that $C$ denotes the channel number, $H$ and $W$ are height and width of the feature map, respectively. For simplicity, the mini-batch dimension is omitted in this notation. The 2D spatial attention map $M_{self} \in \mathbb{R}^{H \times W}$ is obtained by performing channelwise average pooling (CAP) on feature map $F$. $M_{self}$ is then fed into a sigmoid function to generate our importance map $M_{imp} \in \mathbb{R}^{H \times W}$. The spatial attention focuses on where is a discriminative part. In short, the importance map is computed as:

$$M_{self} = CAP(F),$$

$$M_{imp} = \sigma(M_{self}),$$

where $\sigma$ denotes the sigmoid function.

3.2 Channel attention guided dropout

We first gather spatial information of a feature map $F$ by performing global average pooling (GAP) operation, generating the global information embedding: $S \in \mathbb{R}^{C}$. Hence, the embedding can be considered as channel attention. According to the relative magnitude of the attention, a binary mask is generated to indicate whether each channel is selected or not. This attention-guided pruning strategies can be treated as a special way to model the interdependencies across the channels. We rank channel attention by a fast and approximate measure of importance (magnitude), and then drop out those elements with low importance. The strategy considers the top-$k$ largest magnitude attentions as important ones.

3.3 Spatial attention guided dropblock

We observe that ADL drops only strongly activated parts according to the drop threshold. In fact, similar to the pixel-based dropout, the ADL is not really the region-based dropout. This is because the drop mask of ADL is generated by the drop threshold $\gamma$. $\gamma$ determines how many activation units to discard. Neighbouring adjacent pixels share much of the same information. Hence, the ADL cannot completely remove the information.

To address above problems, we propose a region-based dropout in a similar fashion to regular dropout, but with an important distinction. The difference is that we drop the contiguous regions of feature maps rather than the individual pixels. Its main difference from Dropblock \cite{9} is that its drop mask $M_{drop2}$ are computed from the self-attention map $M_{self}$ obtained by the spatial attention module. Then, the shared drop mask across different feature channels or each feature channel has the same drop mask. The proposed method has two main
hyperparameters: \textit{block.size} and \( \delta \). The \textit{block.size} presents the size of the block to be discard, and \( \delta \) determines how many activation units to discard. When \textit{block.size} equals to 1 and covers the full feature map, the region-based dropout reduces to the standard dropout and SpatialDropout \cite{22} respectively. This technique can efficiently remove the information on the feature map. Hence, it forces the network to better capture the full context of the feature map, rather than relying on the presence of a small set of the discriminative features.

It is well known that we can divide images into background and foreground regions. Also, the object of interests is usually consisted of the foreground pixels. The work \cite{30} has reported that the attention map stands for the probabilities of corresponding pixel to be background or foreground. The initial background and object can be produced through the values in the self-attention maps. In particular, the regions with very large values are foreground, while the regions with small values are considered as background. Removing discriminative regions forces the CNNs model to capture the less discriminative part, which sometimes leads to the attention misdirection and the biased localization. Base on this, we can simply sense and background parts and the foreground objects by using the drop mask \( M_{\text{drop}} \) according to the self-attention map, which will finally benefit WSOL.

4 Experiment

4.1 Experimental Setup

\textbf{Datasets.} We evaluate the performance of the proposed method in the commonly used CUB-200-2011 \cite{24}. The CUB-200-2011 contains 200 species of birds with 5,994 images for training and 5,794 for testing.

\textbf{Experimental details.} The proposed DGDM is unified with the commonly used CNNs including VGG \cite{19}, ResNet \cite{10}, ResNet-SE \cite{12}, and InceptionV3 \cite{21}. Following the settings of work \cite{4}, we set the drop rate as 75\%, and apply DGDM to higher-level and intermediate layers of CNNs. Three evaluation metrics are used for WSOL evaluation \cite{20}, that is, GT-known Loc, Top-1 Clas, and Top-1 Loc.

4.2 Ablation studies

We conduct some ablation studies on CUB-200-2011 using pre-trained VGG-GAP to investigate the effects of the proposed DGDM.

The drop mask map \( M_{\text{drop}} \) can not only remove a small set of the discriminative features to better capture the full context of the feature map, but also identify regions in feature maps as background and foreground, respectively. First, we verify the effectiveness of removing importance parts on accuracy. The upper part of Table 1 presents the experimental results when we use different \textit{block.size}. From these results, it can be seen that we achieve the best localization accuracy when \textit{block.size} is 2. We also present the result when \textit{block.size} is
Table 1. Upper and Middle: Accuracy according to different block size. Lower: Accuracy according to different drop threshold $\beta$. $M_{\text{drop2-stage1}}$: erasing the most discriminative parts of the target object. $M_{\text{drop2-stage1+drop2-stage2}}$: removing the most discriminative parts and alleviating attention misdirection. $M_{\text{drop2-stage1+drop2-stage2+drop1}}$: removing the most discriminative parts, alleviating attention misdirection, and utilizing channel attention guided dropout. Adap_7: adaptive and block size calculated by $[H(W)/7]$.

| Method              | $block\_size$ | GT-known Top-1 Acc(%) | Top-1 Clas(%) | Top-1 Loc (%) |
|---------------------|--------------|-----------------------|--------------|--------------|
| $M_{\text{drop2-stage1}}$ | 1            | 62.70                 | 71.99        | 45.20        |
|                     | 2            | 73.48                 | 68.30        | 52.57        |
|                     | 3            | 69.53                 | 55.56        | 43.10        |
|                     | Adap_7       | 73.52                 | 65.83        | 50.46        |
| $M_{\text{drop2-stage1+drop2-stage2}}$ | 1            | 73.48                 | 69.11        | 52.11        |
|                     | 2            | 73.94                 | 69.68        | 53.27        |
|                     | 3            | 74.23                 | 69.00        | 53.79        |
|                     | 4            | 73.48                 | 69.11        | 52.11        |
|                     | Adap_7       | 72.02                 | 69.21        | 51.42        |
| $M_{\text{drop2-stage1+drop2-stage2+drop1}}$ | 2            | 73.44                 | 69.19        | 52.56        |
|                     | 2.5          | 74.88                 | 69.50        | 54.31        |
|                     | 3            | 75.02                 | 69.89        | 54.34        |
|                     | 3.5          | 72.65                 | 69.68        | 52.67        |

adaptive and calculated by $[H(W)/7]$. It can be observed that the drop masks with $block\_size = 2$ from higher-level layers remove the most discriminative region more accurately than those with other $block\_size$. We observe that the classification accuracy decreases as $block\_size$ increases. This is because the model never captures the most discriminative region. Classification accuracy decreases significantly, which adversely leads to the huge boost in localization performance.

Next, we investigate the effect of removing a small set of background on accuracy. The middle part of Table 1 summarizes the results when we use different $block\_size$. It can be seen that the best object localization accuracy can be established when $block\_size$ is 3. We can also see that the Top-1 Loc increases again (from 52.57% to 53.79%) and the classification accuracy achieve a 0.7% improvement (from 68.30% to 69.00%), which indicates that removing a small set of background boosts the performance of WSOL. The reason lies in that the proposed erasing method doesn’t lead to the attention misdirection when the discriminative parts are erased.

Lastly, we observe the effect of channel attention guided dropout on the accuracy by using four different drop thresholds. The lower part of Table 1 reports the experimental results. Based on this, we can claim that the value of the drop threshold has an important effect on performance of WSOL. It can also
be seen that three evaluation metrics are improved when the drop threshold is 3. This is caused by the classification accuracy increase.

4.3 Comparison with the state-of-the-arts

Our proposed approach is compared with existing WSOL techniques on the CUB-200-2011 test set, and we give the results in Table 2.

Table 2. Quantitative evaluation results on CUB-200-2011 test set with the state-of-the-art results.

| Approach | Backbone | # of | Overheads | CUB-200-2011 |
|----------|----------|------|-----------|--------------|
|          |          |      |           | Params parameter(%) | computation(%) | Top-1 Loc (%) | Top-1 Clas (%) |
| CAM      | VGG-GAP  | 29.08| 0         | 0            | 34.41         | 67.55         |
| ACoL     | VGG-GAP  | 43.44| 71.51     | 131.27       | 45.92         | 71.90         |
| ADL      | VGG-GAP  | 29.08| 0         | 0            | 52.36         | 65.27         |
| DANet    | VGG-GAP  | 48.56| 66.99     | 32.58        | 52.52         | 75.40         |
| Ours     | VGG-GAP  | 29.08| 0         | 0            | 54.34         | 69.85         |
| ADL      | ResNet50 | 23.92| 0         | 0            | 46.29         | 79.72         |
| DANet    | ResNet50 | 28.53| 19.27     | 38.21        | 51.10         | 81.60         |
| Ours     | ResNet50 | 23.92| 0         | 0            | 59.40         | 76.20         |
| CAM      | InceptionV3 | 25.69| 0         | 0            | 43.67         | -             |
| SPG      | InceptionV3 | 37.63| 46.48     | 561.10       | 46.64         | -             |
| ADL      | InceptionV3 | 25.69| 0         | 0            | 53.04         | 74.55         |
| DANet    | InceptionV3 | 29.73| 15.72     | 46.07        | 49.45         | 71.20         |
| Ours     | InceptionV3 | 25.69| 0         | 0            | 52.62         | 72.23         |

CUB-200-2011. Table 2 summaries the quantitative evaluation results on the CUB-200-2011 test set. With a VGG-GAP backbone, our method reports 1.98% higher Top-1 Loc and 4.58% higher Top-1 Clas compared with the ADL approach [4]. With a ResNet50 backbone, it reports 8.30% performance gain over the DANet approach [27] at the cost of little classification performance. This is the outcome of the trade-off relationship between classification and localization accuracy discussed in Subsection 3.4. This method also obtains a new state-of-the-art performance (59.40%). When InceptionV3 is used as a backbone, this method also has comparable accuracy.

In addition to obtaining a better performance of WSOL, this method has high efficiency. Table 2 presents the number of parameters and both parameter and computation overheads as well as Top-1 Clas and Top-1 Loc. Similar to ADL [4], the proposed method has no additional parameter for training, and there are no computation overheads upon the backbone network.
5 Conclusions

In this paper, we presented a simple yet powerful dual-attention guided drop-block module (DGDM) for WSOL. We designed two key components of DGDM, CAGD and SAGD, and integrated them with the deep learning framework. The proposed method hides the most discriminative part of the object and then encourages the CNNs classifier to learn the less discriminative part. We defined a pruning strategy so that CAGD can be treated as a special way to model the interdependencies across the channels. In addition, SAGD can not only completely remove the information by erasing contiguous regions of feature maps rather than independent individual pixels, but also simply sense the target objects and background regions to alleviate the attention misdirection. We have achieved a new state-of-the-art performance for WSOL on CUB-200-2011.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Bency, A.J., Kwon, H., Lee, H., Karthikeyan, S., Manjunath, B.: Weakly supervised localization using deep feature maps. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 714–731. Springer (2016)
3. Cao, C., Liu, X., Yang, Y., Yū, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., Xu, W., et al.: Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2956–2964 (2015)
4. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2219–2228 (2019)
5. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(1), 189–203 (2016)
6. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
7. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3146–3154 (2019)
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning (ICML). pp. 1050–1059 (2016)
9. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 10727–10737 (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
11. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141 (2018)
13. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 2017–2025 (2015)
14. Kingma, D.P., Salimans, T., Welling, M.: Variational dropout and the local reparameterization trick. In: Advances in Neural Information Processing Systems (NIPS). pp. 2575–2583 (2015)
15. Labach, A., Salehinejad, H., Valaee, S.: Survey of dropout methods for deep neural networks. arXiv preprint arXiv:1904.13310 (2019)
16. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9215–9223 (2018)
17. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 85–100 (2018)
18. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: Advances in Neural Information Processing Systems (NIPS). pp. 2204–2212 (2014)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3544–3553. IEEE (2017)
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
22. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 648–656 (2015)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS). pp. 5998–6008 (2017)
24. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
25. Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R.: Regularization of neural networks using dropconnect. In: International Conference on Machine Learning (ICML). pp. 1058–1066 (2013)
26. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
27. Xue, H., Liu, C., Wan, F., Jiao, J., Ji, X., Ye, Q.: Danet: Divergent activation for weakly supervised object localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 6589–6598 (2019)
28. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5505–5514 (2018)
29. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1325–1334 (2018)
30. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 597–613 (2018)

31. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016)