Rough Target Region Extraction with Background Learning

Ryo Nakamura[1,2], Yoshiaki Ueda[1,3], Masaru Tanaka[5], and Jun Fujiki[1,4]

[1] Fukuoka University, Fukuoka Nanakuma 8191, Japan
[2] sd210501@cis.fukuoka-u.ac.jp
[3] uedayos@fukuoka-u.ac.jp
[4] fujiki@fukuoka-u.ac.jp
[5] Prof. Masaru Tanaka deceased in June 2021.

Abstract. Object localization is a fundamental and important task in computer vision, that is used as a pre-processing step for object detection and semantic segmentation. However, fully supervised object localization requires bounding boxes and pixel-level labels, and these annotations are expensive. For this reason, Weakly Supervised Object Localization (WSOL) with image-level (weak) supervision has been the focus of much research in recent years. However, WSOL requires a large dataset to detect the region of an object in images with high performance. When the large dataset is unavailable, it is difficult to localize the image with high performance. This paper proposes a method for extracting target regions using small amounts of target and background images with image-level labels. The proposed method enables the detection of object locations with high performance using relatively less training images by classifying multiple patches cut from the image. This object localization method differs from the typical WSOL method that takes a single image as input and detects the location of an object because it assumes a small patch of area as input. The label of the patch cropped from the image must be labeled with the ground truth. However, the proposed method uses labels attached to images because ground truth labeling is costly. Instead, in the proposed method, the network learns by learning many "background" labeled background patches, and learns to induce the network to classify the mislabeled background patches that resemble ground truth as background. We call this key idea Decision-Boundary Induction(DBI). Moreover, learning many background patches for such a DBI is what we call background learning. In our experiments, we verified that decision boundaries are induced, and accordingly, we could roughly extract the target region. Also, we showed that the Loc. Acc. is higher than that of WSOL.

Keywords: Weakly supervised object localization, Patch-based training, Background learning, Noisy label training

1 Introduction

Object localization is a fundamental and important task in computer vision and is used as preprocessing for object detection [18, 17, 4, 11, 10, 5], semantic segmentation [14, 12, 9, 8], etc. For this, methods of deep learning, such as convolutional

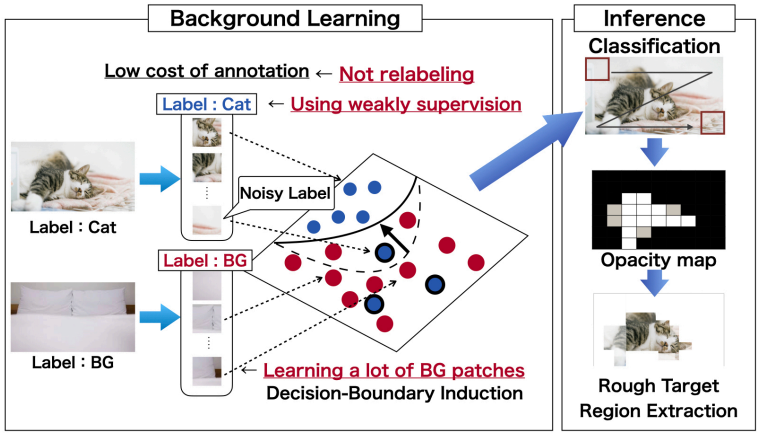

Fig. 1. Overview of proposal method: We consider the task of classifying randomly cropped multiple small patches from an image as whether they are in the target or background (BG). In order to classify patches with high accuracy, it is necessary to attach ground truth to the patches. However, in background learning, we do not provide the ground truth. Instead, the network induces the decision boundary to classify patches of label noise as background by learning many background patches cropped from the background image. Note that label noise is the background patch labeled with the target label. In inference, the background learning network roughly extracts the target region by outputting the backgroundness (sigmoid value) of patches in the target image as a sliding window formula.

neural networks (CNN), are widely used. To achieve high accuracy, these need labeled training datasets: Pixel-level labels and bounding boxes of target objects. When we carry out pixel-level labeling, it often becomes a burden compared to image-level labeling (i.e., attaching labels to images in one-by-one ways).

As a methodology to overcome such problems, Weakly Supervised Object Localization (WSOL) [22, 3, 19, 1, 21, 20, 16, 13] has received recent attention from researchers in the field. This aims for high accuracy object-localization, and we have only to prepare training datasets consisting of image-level labels, and thus we can save time and human resources despite the high accuracy. Typically, WSOL estimates regions in the images, which are recognized to be important for the classification of images. The regions are then used in object localization. For this procedure, we need huge datasets, and thus a large part of the total cost is passed on to the large size of the datasets. This implies that we have to label many images to use WSOL, which would also be a heavy task.

To avoid such difficulty, we propose a novel method to identify the location of the (target) object to be recognized (see Fig. 1). The method consists of two parts, explained below. One is to randomly cut out small parts of images, each of which will be called a patch, and then classify them into two categories with attached labels: 'foreground' (target) and 'background.' More specifically, all patches from an image will be labeled as the common name of an object (e.g., 'cat,' 'dog,' 'horse,' 'owl,' and so on) if the object in the consideration is in the image, and they will be labeled as 'background' otherwise. An image showing a

cat might contain a background area. One should notice that all patches from the image are labeled as 'cat' even if the patch does not contain any part of the cat. In such a case, the false labels of the patches will be called 'label noises.' On the other hand, all patches from an image showing no objects in consideration are assigned with 'background,' in coincidence with the true label. Therefore one can easily collect a large dataset consisting of truly-labeled patches, namely the ones with 'background' labels, since it is easy to prepare images showing no objects.

The other is the method we call 'background learning,' which enables our model to discriminate labels to be attached to new patches by training with the large dataset of patches truly labeled 'background' combined with a relatively small dataset including label-noises. Now, assigning new patches to either true target or true background is what we call object localization and will be realized with the network trained using our background learning. It would be natural to design all patches with label noises, and the ones with background labels should be similar to each other. The true labels of such patches are both 'background.' The main ingredients in the dataset are cut out from background images, which is why we call the method 'background learning'.

All patches in the labeled training dataset for background learning are classified into three classes (see Fig. 2.):

- One consists of patches showing (a part of) objects (thus, these are cut out from images showing objects).
- Another is label-noises, the class consisting of patches cut out from images showing objects (and thus labeled as the target), but the patches themselves show no parts of the object.
- The other consists of patches cut out from background images (images showing no objects), which are truly labeled as 'background.'

In this paper, each state of the classes is said to be positive (i.e., showing objects), false positive (i.e., showing no object but cut out from an image showing objects), and negative (i.e., cut out from background image), respectively. With these terminologies, background learning is a method to determine whether the states of patches showing objects are positive or false positive. Equivalently, it is a method to detect false positives (or extract positives) among states of patches showing objects.

For our experiments, we prepared 240 images showing one of cats, dogs, horses, or owls and 240 background images. We trained networks by WSOL and our background learning separately. As a result, we showed that our background learning successfully extracted target regions (regions lying in positive patches). For the object localization task, we showed that the performance of background learning is superior to WSOL in the sense of Localization Accuracy (Loc. Acc.).

2 Related work

Weakly Supervised Object Localization(WSOL) aims to learn to localize the object using only image-level labels. A popular method for WSOL is Class Acti-

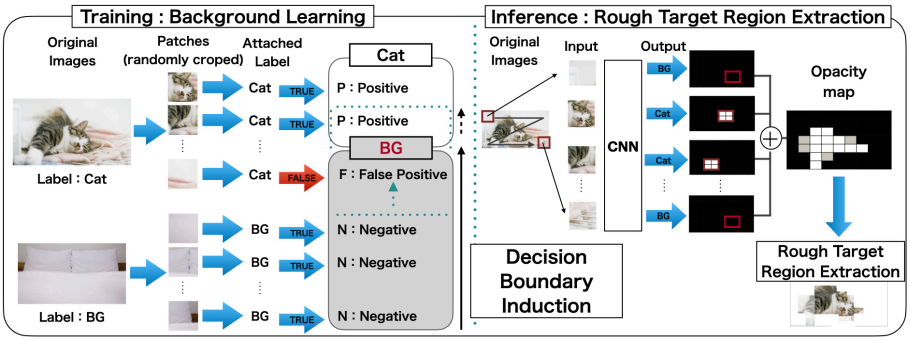

Fig. 2. Explanation of rough target area extraction with background learning. In background learning, a large amount of training data is prepared by randomly cropping multiple small-area patches from the image. Then, annotations to patches are not annotated with the correct labels, instead using image labels, which does not increase the annotation cost. In fully supervised learning, a patch of F is classified as a "cat," but learning a large number of patches of N that resemble F induces it to be classified as a "background (Decision-Boundary Induction). For rough target extraction, CNN with background training is used to output target patches using the sliding window method to generate a targetness map of the corresponding patch region.

vation Map (CAM) [22], which generates localization maps by aggregating deep feature maps using fully connected layers by class. Hwang and Kim [7] simplify the network structure of CAM by removing the last unnecessary fully connected layer. CAM methods are simple and effective, but they can only classify the small classification region of the object. In order to improve the activation map of CAM, HaS [16] and CutMix [20] adopted the region dropout-based strategy from the input image so that the network focuses on the more related region of the object. ADL [3] focuses on the problem that learning while deleting classifiable regions with high performance requires high computational resources and eliminates feature maps corresponding to discriminable regions to localize objects with a lightweight model that is efficient and has low network parameters.

3 Rough Target Region Extraction with Background Learning

3.1 Background Learning

Overview. Fig. 2 shows the proposed method. In background learning, we consider the problem of classifying whether the small patches of the image are target or background using a background image. Note that background images are relatively easy to obtain because the images are not required to include the object. When creating the training dataset, multiple patches are randomly cropped from the target and background images, and the patches are annotated with image labels. Then, the patch cropped from the target image has two patches: the target and the background. But we do not relabel the background patches with the target label. (i.e.the incorrect label is annotated). Instead, we use the data

imbalance induced by learning many background patches with the F. We then induce the network to classify the mislabeled background patches as background (We call this induction DBI: Decision-Boundary Induction).

Define target/background. In this paper, "target" and "background" mean the target object to be extracted and the non-target object to be extracted. In a specific example, if the target is the cat, the cat's area in the image represents the target, and the other image areas represent the background. Therefore, if an image contains a cat and a dog, the cat area would be the target, while the other dog and background areas would be the background.

About the type of patches to be cropped. The cropped patch from the image is a small area image of the target or background image. The size of the small area is an optional parameter. Also, the number of patches to be cropped is an optional parameter. We use image labels to label the patches. Therefore, cropped patches have three types of patterns, as shown in Fig. 2. The first is a positive (P) patch that includes the target region cropped from the target image. (P) patches are labeled with the target label. The second is a false positive (F) patch that does not include the target region cropped from the target image. (F) patches are labeled with the target label. Not labeling (F) patches with ground truth are to avoid labeling costs. The third is a negative (N) patch cropped from the background image. (F) patches are labeled with the background label.

Background learning purpose (Decision-Boundary Induction). The purpose of background learning is to induce the "background" classification result of the network when inputting the F patch. Therefore, we induce the classification of the network by training a large amount of N patch on it, causing bias in the learning. In fully supervised learning, the network learns that background F patches labeled as targets are to be classified as backgrounds based on their labels. However, to classify it as "background," we need to relabel F with a background label. But relabeling increases the annotation cost. Therefore we want to train the network to correctly classify P, F, and N patches into target and background patches without relabeling. In this paper, the problem of classifying such P, F, and N patches is called the PFN classification problem. Then, we use the imbalance of the number of training data, i.e., the property of [2] (Decision-Boundary Induction), which induces the class with the largest number of training data to be classified.

Training. For learning patches with label noise such as PFN, we use fully supervised learning, which is widely used in deep learning. In this learning, the error function is optimized to classify the patches so that they correctly answer the label of the image to be cropped, without relabeling. In the paper, binary cross-entropy is used for the error function and Stochastic Gradient Descent (SGD) for optimization.

3.2 Rough Target Region Extraction

The CNN with background training roughly calculates the target region by the following procedure. First, the target image is slided window by a specified step width (in this paper, 8 pixels in height and width), and the backgroundness of

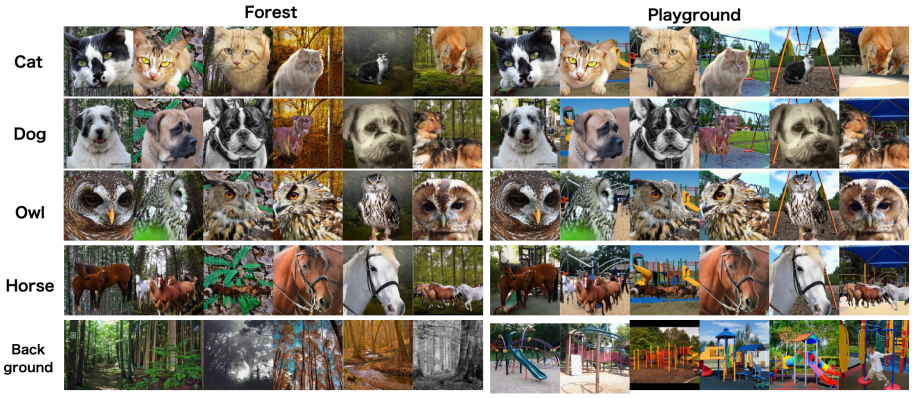

Fig. 3. Sample image of the dataset used in the experiments

each patch is output. Next, for each pixel in the target image, add the final output c, which is the output of the sigmoid function. When this score is calculated for all pixels, the maximum value is M, and the minimum value is m. The "Targetness" is defined as follows for each pixel.

$$\text{Targetness} = \frac{c - m}{M - m} \in [0, 1] \tag{1}$$

In this paper, we consider the targetness to be the probability that each pixel is in the region to be classified. The target region can be visualized as an opacity map by using the targetness as the value of α channel, representing the opacity (see Fig. 2).

4 Experiments

4.1 Experimental setting

Data collection. In order to verify the effectiveness of decision-boundary induction, we prepare the dataset, which includes target (cat, dog, owl, horse) images with forest and playground background regions and background images with a different scene from the target backgrounds. Since there are no publicly available open datasets with such limited backgrounds, we collected 240 targets and 480 backgrounds (240 treated as background areas of targets and 240 treated as background images) from Fricker, manually generated masks of targets and background areas, and used the masks to composite them with targets and background images to construct a data set (see Fig. 3).

Create train/test dataset. The training dataset is a two-category dataset of target and background images, each with 240 images (a total of 480 images). The test dataset uses target images from the training dataset and Ground truth masks labeled with targets at the pixel level to measure the performance of extracting target regions from the target images in the training dataset. This mask can compare target region extraction performance and extract patches related to PFN from the mask information. Note that the masks containing

Table 1. Performance evaluation of rough target region extraction with background learning with Loc. Acc.: The result that performs well in comparing background patches is assumed to be bold.

		BG : Forest		BG : Playground	
Model	Target	TG:BG=1:1	TG:BG=1:2	TG:BG=1:1	TG:BG=1:2
VGG16	Cat	0.81	**0.98**	0.90	**0.91**
	Dog	0.61	**0.93**	**0.83**	0.79
	Owl	**0.85**	0.81	**0.76**	0.75
	Horse	**0.78**	0.53	**0.75**	0.55
	Avg.	0.76	**0.81**	**0.81**	0.75
ResNet50	Cat	0.88	**0.90**	0.88	**0.90**
	Dog	**0.96**	0.90	**0.86**	0.76
	Owl	**0.83**	0.76	**0.83**	0.80
	Horse	0.53	**0.81**	**0.78**	0.57
	Avg.	0.80	**0.84**	**0.84**	0.76

the target and background merge with the background and evaluate target area extraction performance, but not for training.

Experiment details. We use the VGG16 [15] and ResNet50 [6] models in this experiment. We train the models with the binary cross-entropy loss for 150 epochs using SGD with a learning rate of 0.01. The mini-batch size is 64 for the WSOL methods and 512 for our method. The reason for this is to align the updates of the network parameters. HaS [16] and Cutmix [20] are localized with CAM [22]. In the proposed method, 8 patches are randomly cropped with size 48×48×3ch from a single image (256×256×3ch).

Evaluation. In the evaluation of target area extraction, we evaluate the performance of target area extraction on trained data rather than on the performance of target area extraction on unlearned target images. The number of target images used is 240, the same as the number of training data. We use the localization accuracy metric (Loc. Acc.) to evaluate the roughness of the target region extraction. Loc. Acc. is a metric that calculates the proportion of images with an Intersection over Union (IoU) of 40% or higher. The threshold value of IoU when calculating Loc. Acc. is [0.05, 0.15, ... 0.95], and the best value is used as the experimental result.

4.2 Evaluation of the effectiveness of Decision-Boundary Induction

We show through experiments that Decision-Boudary Induction can be used by adjusting the amount of background patches in the dataset, and that the use of DBI can lead to improved performance in target region extraction. For this experiment, we conducted the following on datasets with one and two times the ratio of background patches to target patches.

- Quantitative evaluation of the extraction performance of the target region by Loc. Acc.
- Qualitative evaluation visualizing relative frequencies of backgroundness of patches of PFN.

– Qualitative evaluation to compare target area extraction results

Evaluation of target area extraction performance with background learning. Table 1 shows the results of Loc. Acc. for each condition when the background patches are trained with the ratio of background patches 1x and 2x compared to the target patches and the target regions are extracted. Cat results showed that the Loc. Acc. was higher in all cases when the background was learned 2x. Dog's results show that for VGG16, the Loc Acc is higher when the background is trained 2x only when the background region is Forest, and 1x is higher for all other cases. In the Owl results, the Loc. Acc. was higher when the background was trained 1x for VGG16, ResNet50, and the background was Forest and Playground. In the Horse results, the Loc. Acc. was higher when the background region was Forest, the Loc. Acc. was higher when the background region was Playground and the Loc. Acc. was higher when the background region was Playground. As shown above, it can be confirmed that adjusting the number of background patches according to the target and the model used contributes to the performance improvement of Loc. Acc.

We consider the bias of the background pattern of the image to be related to the reason that Forest and Playground did better with 2x background patches and 1x background patches, respectively. In the case of Forest, since F contains many similar patterns, such as diverse leaves and trees, it is necessary to learn many N to induce identification. On the other hand, Playground contains many instances of playground equipment, trees, sand, etc., and the patterns are distributed, so the discrimination induction works effectively with a relatively small number of N images.

Comparison by backgroundness relative frequency graph for each patch. Fig.4 is the result of visualizing the sigmoid (backgroundness) of P, F, and N patches as relative frequencies (see Fig.4.) when background learning is performed using Forest as the background on ResNet50, which had high Loc. Acc. in Table 1. In this experiment, for patches, P is defined as if the target is included in 10% or more of the patches cropped from the target image, and F is defined as otherwise. The data for each PFN is 200, for a total of 600. We denote patches extracted from the image containing the region to be identified as X and patches extracted from the background image as Y to clarify the training data structure used for background training in the graph. The CNN trained with X and Y is denoted as Model(X,Y).

The Fig.4 shows that the relative frequencies of F and N above 0.8 of the sigmoid are higher when the targets are Cat and Dog by learning the background patches twice. In Table 1, we can verify that Loc. Acc. is also high following the results. On the other hand, when the target is Owl, learning as 2x background patches did not increase the relative frequencies of backgroundness of F and N above 0.8. We consider that one of the reasons why the owl case has not worked is that the features of the forest and the Owl are similar. Owls are mimic animals as they hide and hunt in the forest. Therefore, as a result, we consider that learning to classify the P, F, and N patches became difficult, the learning was unstable, and thus the hypothetical results were not observed.

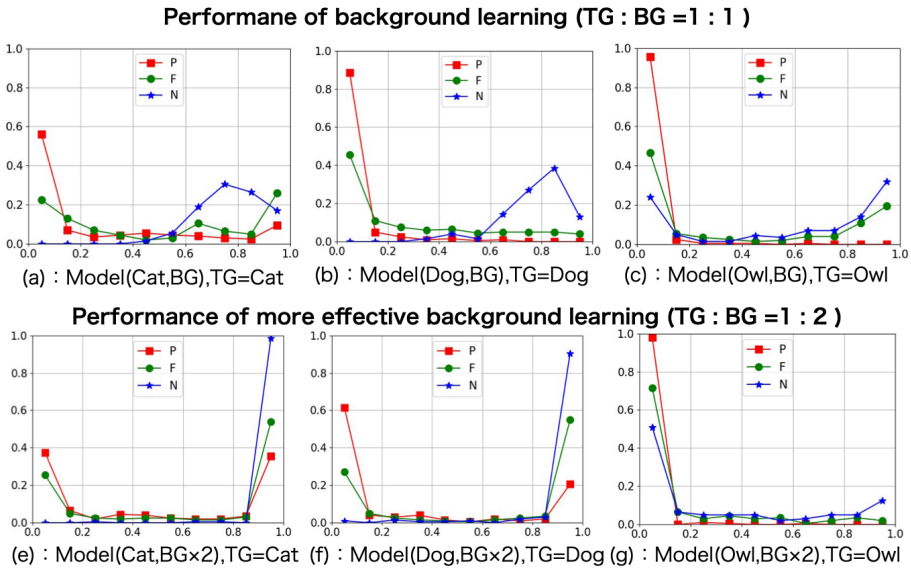

Fig. 4. Relative frequency graph of backgroundness of P, F, and N patches (This experiment uses ResNet50 and Forest background region). The horizontal axis represents the backgroundness (sigmoid value of the background class), and the vertical axis represents the relative frequency. We used forest for the background and ResNet50 for the network.

Comparison by target area extraction map. Fig.5 shows the target region extracted image using the trained network of (a)-(h) in Fig.4. (a)-(d) are the target region extraction images when the number of target patches and background patches are the same. (e)-(h) are the target region extraction images when more background patches are trained. The Cat, Dog, and Horse results show that when the background is trained strongly, the regions are extracted to remove the background region. On the other hand, the Owl results show that when the background is learned too much, the target region is extracted in such a way that the backgroundness body region of the Owl is removed. The effectiveness of can also be verified through qualitative results on the amount of background patches (Fig.4).

4.3 Comparison of target region extraction performance

We compare the proposed method with the WSOL method by Loc. Acc. to show that the proposed method can extract the target region with a smaller amount of images than the WSOL method (The results are shown in Table 2). Avg. in the table is the average of Loc. Acc. for the four targets (Cat, Dog, Owl, and Horse). For each model and target, the highest value of Loc. Acc. is indicated by bold and the second highest by underline.

In the results of Forest in Table 2, our method has the highest Loc. Acc. for all targets in VGG16 and ResNet50, where we adjusted the number of background patches and used DBI effectively. Avg. results show that our method occupies at least the top two positions. It is interesting to note that the results show a higher

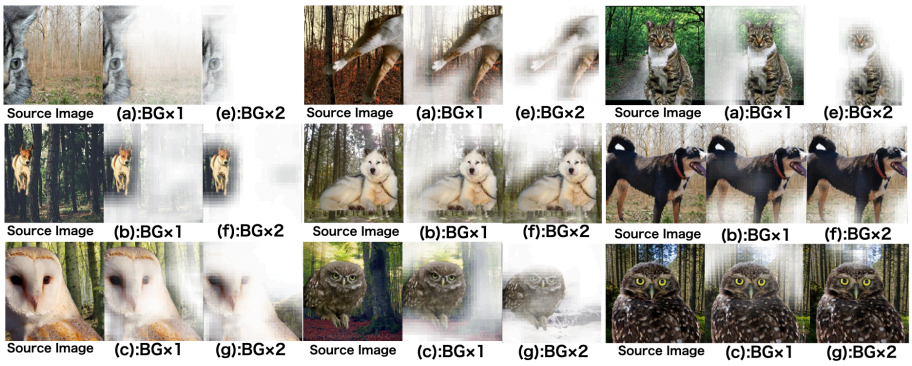

Fig. 5. The results of extracting the target region when background learning: (a)-(h) are the results of the model learned in Fig. 4. It can be shown that background learning induces the classification of the background in the target image. However, if the background learning is too effective, it also induces the classification of target regions that look like the background.

Table 2. Loc. Acc. comparison on a dataset where the background region.

		Forest					Playground				
Model	Method	Cat	Dog	Owl	Horse	Avg.	Cat	Dog	Owl	Horse	Avg.
VGG16	CAM [22]	0.83	0.58	0.67	0.42	0.62	0.72	0.57	0.68	0.51	0.62
	ADL [3]	0.73	0.75	0.67	0.38	0.63	0.74	0.60	0.69	0.46	0.62
	HaS [16]	0.70	0.65	0.78	0.55	0.67	0.86	0.64	0.68	0.46	0.66
	Cutmix [20]	0.89	0.65	0.73	0.44	0.68	0.88	0.76	0.69	0.46	0.70
	Ours(BG×1)	0.81	0.61	**0.85**	**0.78**	0.76	0.90	**0.83**	**0.76**	**0.75**	**0.81**
	Ours(BG×2)	**0.98**	**0.93**	0.81	0.53	**0.81**	**0.91**	0.79	0.75	0.55	0.75
ResNet50	CAM [22]	0.66	0.66	0.75	0.44	0.63	0.85	0.72	0.74	0.50	0.70
	ADL [3]	0.85	0.73	0.71	0.55	0.71	0.89	0.81	0.78	0.62	0.78
	HaS [16]	0.70	0.59	0.68	0.52	0.62	0.72	0.66	0.69	0.54	0.65
	Cutmix [20]	0.70	0.50	0.66	0.47	0.58	0.86	0.66	0.74	0.53	0.70
	Ours(BG×1)	0.88	**0.96**	**0.83**	0.53	0.80	0.88	**0.86**	0.83	**0.78**	**0.84**
	Ours(BG×2)	**0.90**	0.90	0.76	**0.81**	**0.84**	**0.90**	0.76	0.80	0.57	0.76

Avg. than WSOL, even though the parameters for the amount of background patch are roughly chosen.

The Playground results in Table 2 also show that when DBI is used effectively, our method has the highest Loc. Acc. for all targets in VGG16 and ResNet50. The difference from Forest results was that ResNet50 had the second-best performance with ADL of Avg. The reason for this is the difference in the features of the background since there are more variations of objects, such as playground equipment, the ground, and trees, in the Playground than in the Forest. This can be considered to increase the learning of the target object.

Fig. 6 shows the results of the visualization images where each target is extracted using WSOL and our method. The images extracted by CAM and ADL, among the conventional WSOL methods compared, tended to be extracted for frequently appearing parts, such as cat's whiskers, dog's face, owl's face, and horse's feet. Region dropout-based methods that mask part of the image, such

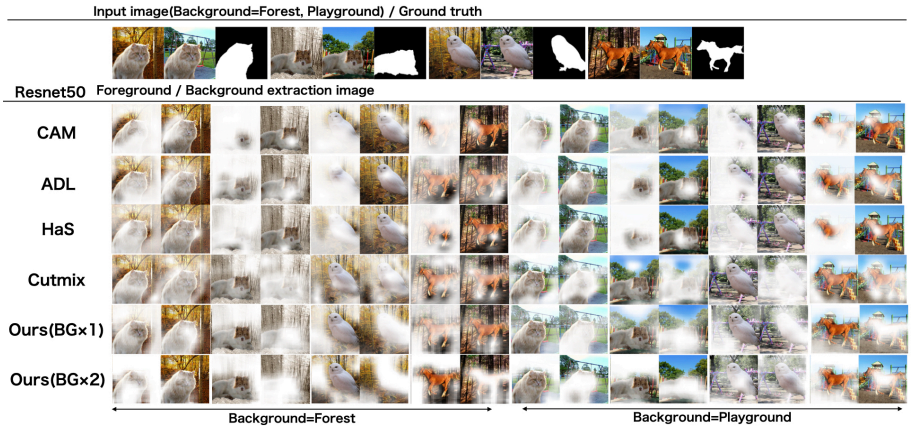

Fig. 6. Visualization results of target region extraction for existing WSOL methods and our method. We use ResNet50 for the network.

Table 3. The relationship between patch size and the performance of target area extraction with background learning. We study the relationship between the patch size {32, 48, 64, 96, 128}pixels.

	Forest				Playground			
	Cat		Dog		Cat		Dog	
Patch size	VGG16	ResNet50	VGG16	ResNet50	VGG16	ResNet50	VGG16	ResNet50
32	0.954	**0.979**	**0.914**	**0.902**	0.786	0.942	**0.967**	0.7
48	**0.958**	0.975	0.858	0.793	**0.962**	**0.979**	0.931	**0.943**
64	0.967	0.975	0.846	0.846	0.958	0.975	0.894	0.923
96	0.913	0.86	0.7	0.663	0.942	0.876	0.801	0.7
128	0.786	0.777	0.627	0.570	0.847	0.802	0.570	0.542

as HaS and Cutmix, increased the target area to be extracted, but Cutmix also tends to extract more background areas as foreground, and HaS tended to extract some parts of the body as targets, but still tended to extract parts that appeared frequently.

4.4 Ablation Study

Relationship Study with Patch Size (Table 3). We studied the effect of patch size on the extraction of the target area. Specifically, we changed the size of patches extracted from images in {32, 64, 96, 128} and studied the relationship between Loc. Acc. We used VGG16 and ResNet50 as the networks. We used VGG16 and ResNet50 as the network and cat and dog as the target images. Then, we used the forest and playground as the background. The background patch is cropped twice as many times as the target patch to create the dataset. The results are shown in Table 3. The experimental results show that the patch sizes with the highest Loc. Acc. were 32 for the forest and 48 for the playground.

We consider the reason that the accuracy increases as the patch size decreases is because the proportion of patches that contain the target decreases.

Table 4. Study of Background Patch Ratio and Performance of Target Area Extraction with Background Learning.

	Forest				Playground			
	Cat		Dog		Cat		Dog	
Patch size	VGG16	ResNet50	VGG16	ResNet50	VGG16	ResNet50	VGG16	ResNet50
BG×1.0	0.93	0.9	0.923	0.846	0.934	**0.979**	0.902	**0.947**
BG×2.0	0.95	0.954	0.914	0.902	**0.962**	**0.979**	0.931	0.943
BG×4.0	**0.983**	**0.975**	**0.971**	**0.951**	0.925	0.917	**0.939**	0.906

Background learning aims to DBI a background patch F with a target label by learning many N true background patches. Therefore, the smaller the F patches are, the more likely they produce DBI. Therefore, the smaller the patch size, the higher Loc. Acc. However, if the patch size is too small, DBI will also occur in the body of the target image, as shown by the owl in Fig. 5, so a moderately small patch is important for target area extraction.

Also, patch size differs depending on the background because of the background's complexity. The forest has a bias toward trees, leaves, and other patterns that appear in the background. Small patches will cause pattern bias in the clipped patches when there are many background patterns. Since it is difficult to perform DBI on a background with few patterns, we believe that increasing the size of the patch increases the number of patterns in the image and reduces pattern bias, which is effective in improving accuracy.

Study of Relationships with Background Patches Ratio (Table 4). We studied the relationship between the ratio of background patches to target patches and the performance of target region extraction. Changing the ratio of each patch is equivalent to adjusting the strength of the DBI effect. Specifically, we changed the proportion of background patches by {1.0, 2.0, 4.0} and studied the relationship between Loc. Acc. We used VGG16 and ResNet50 as the network and cat and dog as the target images are cat and dog. Then, we used the forest and playground as the background. The patch size is 32×32. The results are shown in Table 4. The results for forest show that increasing the ratio of background patches results in a better Loc. Acc. In the playground, no consistent trend was found comparing the models.

In the forest results, the DBI works correctly and improves accuracy. On the other hand, the playground results showed some variation depending on the model and target. We consider this because the patch size of 32×32 is relatively small, and the image does not contain a variety of background patterns. If there is a large piece of playground equipment in the background image, a large percentage of the image will contain the equipment if the patches are cut out randomly. If the patch size is large, a sandbox and playground equipment appear as patterns in the patch, which can be classified as background because it contains equipment frequently appearing as background. However, if the patch size is small, the number of patches containing only playground equipment patterns increases, making it difficult to distinguish other patterns from the background, in our opinion.

4.5 Limitation

Our proposed method can extract the target region with higher performance than the conventional method in a situation where only a small number of images are available, but it has some challenges.

First, it is not easy to classify the target body and background in the patch input. Fig. 4 shows the relative frequency of the PFN backgroundness when the synthetic data is background trained. Comparing background learning and the effect of background learning enhancement, the backgroundness of most F improved, but the backgroundness of some P also increased. A possible reason for the increased backgroundness of P is that P has patches of the target body, which are not discernable from the background patches, so the DBI is also working on the patches of the body. This problem represents the limitation of learning only patches. To deal with this problem, it is necessary to incorporate a mechanism that can determine the target's body from the target's structural information based on the relationship of the positions of the cropped patches.

Second, there is a need to establish clear indicators for use in selecting background images for background studies. In the background learning, F and P classification results are decided by the background image prepared as the dataset. One approach to effectively DBI F is to use a large background image, but this approach is likely to cause an unbalance in the number of images in the target and background images, and thus DBI for P as well. Currently, we perform background learning by preparing random background images, but to perform DBI effectively, it is important to incorporate a mechanism to select an effective N for DBI. Due to the lack of a clear metric for background images, we consider it difficult to apply the method to large datasets with target images from diverse backgrounds in the current situation.

5 Conclusions

In this paper, we proposed the patch-based rough target region extraction method to extract target object regions with a relatively small number of images and a small annotation cost. The proposed method learns the network to robustly classify whether a small patch in the image is the target or background by using the method we call background learning. Also, the trained network is used to localize objects by determining whether the small areas of the image are targets or backgrounds. There are two important aspects to classifying target and background patches with fewer training images and less annotation cost. First, we do not use a single image to train the network but rather assume small patches randomly cropped from multiple images as input. This assumption is data efficient and enables the neural network to be trained by cropping many patches from several training images. Second, even if background images have noisy labels, background learning can improve classification robustness. Background learning is a method of classifying target and background patches by learning many background patches of ground truth cropped from background images, even when the labeling of patches cropped from the target image is done roughly. Using this learning, we provide a more robust classification of patches without using the

cost of relabeling. For patch-based rough target area extraction, we calculate the backgroundness of the patch by inputting a small region of the image in a sliding window to the network. The result is calculated as the map of backgroundness, and the maps are merged by averaging the maps of backgroundness calculated for all regions. In our experiments, we verified that DBI works and improves the performance of target region extraction on an ideal dataset with similar background images and target object backgrounds. In addition, we verified that our method could extract target regions with higher Loc Acc than the existing WSOL method, although limited.

Acknowledgements I am grateful to Associate Professor Takafumi Amaha of Fukuoka University for advice on writing the paper. I would like to take this opportunity to thank him.

References

1. Bae, W., Noh, J., Kim, G.: Rethinking class activation mapping for weakly supervised object localization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 618–634. Springer International Publishing, Cham (2020)
2. Chawla, N., Japkowicz, N., Kołcz, A.: Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explorations **6**, 1–6 (06 2004). https://doi.org/10.1145/1007730.1007733
3. Choe, J., Lee, S., Shim, H.: Attention-based dropout layer for weakly supervised single object localization and semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12), 4256–4271 (2021). https://doi.org/10.1109/TPAMI.2020.2999099
4. Cinbis, R., Verbeek, J., Schmid, C.: Multi-fold mil training for weakly supervised object localization. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2409–2416. IEEE Computer Society, Los Alamitos, CA, USA (jun 2014). https://doi.org/10.1109/CVPR.2014.309, https://doi.ieeecomputersociety.org/10.1109/CVPR.2014.309
5. Eu Wern Teh, M.R., Wang, Y.: Attention networks for weakly supervised object localization. In: Richard C. Wilson, E.R.H., Smith, W.A.P. (eds.) Proceedings of the British Machine Vision Conference (BMVC). pp. 52.1–52.11. BMVA Press (September 2016). https://doi.org/10.5244/C.30.52, https://dx.doi.org/10.5244/C.30.52
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2016), in CVPR
7. Hwang, S., Kim, H.E.: Self-transfer learning for weakly supervised lesion localization. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. pp. 239–246. Springer International Publishing, Cham (2016)
8. Khoreva, A., Benenson, R., Omran, M., Hein, M., Schiele, B.: Weakly supervised object boundaries (2016), in CVPR, pages 183–192
9. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation (2016), in ECCV, pages 695–711
10. Liang, X., Liu, S., Wei, Y., Liu, L., Lin, L., Yan, S.: Towards computational baby learning: A weakly-supervised approach for object detection. In: 2015 IEEE

International Conference on Computer Vision (ICCV). pp. 999–1007 (2015). https://doi.org/10.1109/ICCV.2015.120
11. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free? - weakly-supervised learning with convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 685–694 (2015). https://doi.org/10.1109/CVPR.2015.7298668
12. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation (2015), in ICCV, pages 1796–1804
13. Rahimi, A., Shaban, A., Ajanthan, T., Hartley, R., Boots, B.: Pairwise similarity knowledge transfer for weakly supervised object localization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 395–412. Springer International Publishing, Cham (2020)
14. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR **abs/1312.6034** (2013)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015), iCLR
16. 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 (2017). https://doi.org/10.1109/ICCV.2017.381
17. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. p. II–1611–II–1619. ICML'14, JMLR.org (2014)
18. Wang, C., Ren, W., Huang, K., Tan, T.: Weakly supervised object localization with latent category learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 431–445. Springer International Publishing, Cham (2014)
19. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6488–6496 (2017). https://doi.org/10.1109/CVPR.2017.687
20. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6022–6031 (2019). https://doi.org/10.1109/ICCV.2019.00612
21. Zhang, C.L., Cao, Y.H., Wu, J.: Rethinking the route towards weakly supervised object localization. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13457–13466 (2020). https://doi.org/10.1109/CVPR42600.2020.01347
22. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319