LOOK AT HERE : UTILIZING SUPERVISION TO ATTEND SUBTLE KEY REGIONS -SUPPLEMENTARY MATERIALS

Anonymous authors
Paper under double-blind review

In this supplementary material, we include detailed description and qualitative results which are not included in the main paper due to the space limit. A code-level description of Cut&Remain is included in Section 1, and qualitative results from original version of region-perturbed augmentation techniques with the same experimental setting in main manuscript are included in Section 2.

1 CUT&REMAIN ALGORITHM

We present the code-level description of Cut&Remain algorithm in **Algorithm 1**. N, W, H, C denote the size of mini-batch, width, height, and channel size of input image. respectively. Using the bounding box annotation with various aspect ratio, relevant regions are simply crop while preserving location information and added to current batch. Note that Cut&Remain is easy to implement with few lines(from line 3 to line 16), so is very practical algorithm giving significant impact on wide range of tasks.

Algorithm 1 Pseudo-code of Cut&Remain with aspect ratio variation.

Require: aspect_ratio r = {1.0, 1.5, 2.0}
1: **for** each iteration **do**
2: input, target, bounding_box=get_minibatch(dataset) ▷ input has $N \times W \times H \times C$ size tensor.
3: **if** mode == training **then** ▷ Cut&Remain starts here.
4: C_x, C_y, w, h = bounding_box
5: **for** $r_w \in r$ **do**
6: **for** $r_h \in r$ **do**
7: x1 = Round(Clip(C_x - $r_w \times w/2$, min=0))
8: x2 = Round(Clip(C_x + $r_w \times w/2$, max=W))
9: y1 = Round(Clip(C_x - $r_h \times h/2$, min=0))
10: y2 = Round(Clip(C_x + $r_h \times h/2$, max=H))
11: aug_img = input[:, x1:x2, y1:y2, :] ▷ Crop the relevant region
12: aug_target = target ▷ aug_target has the same value as target.
13: input $\leftarrow$ input $\cup$ aug_img
14: target $\leftarrow$ target $\cup$ aug_target
15: **end for**
16: **end for** ▷ Cut&Remain ends.
17: **end if**
18: output=model_forward(input)
19: loss=compute_loss(output,target)
20: model_update()
21: **end for**

2 EXPERIMENTS WITH ORIGINAL VERSION OF REGION-PERTURBED AUGMENTATION TECHNIQUES

2.1 BINARY CLASSIFICATION ON CLAVICLE X-RAY DATASET

The experimental result is summarized in Table 1. Compared to the supervision version of other augmentation techniques, Cut&Remain showed the highest AUC-ROC of 98.6 and F1-score of 98.8, and improved the performance of the base network by more than 7.8 and 7.6, respectively. We also observed that Mixup degrades the performance in our dataset. One of the possible reasons is that it provides unnatural artifacts caused by overlap between tissues. The Cut&Remain results supports our claim that mainly considering the lesion is crucial to images with a small lesion.

Method	AUC-ROC	F1-score
ResNet-50	90.8±1.5	91.2±1.8
Mixup	88.2±1.5	88.0±2.0
Cutout	94.0±0.9	93.2±0.8
Cutmix	91.8±0.6	92.6±1.1
Cut&Remain (w/o aspect ratio variation)	98.4±0.4	98.6±0.5
Cut&Remain (w/ aspect ratio variation)	**98.6±0.4**	**98.8±0.4**

Table 1: Results on the clavicle X-ray dataset. The averages of 5-fold cross validation are reported for each setting with standard deviation.

2.2 MULTI-CLASS CLASSIFICATION OF FEMUR FRACTURE ON THE PELVIC X-RAY DATASET

The experimental result is summarized in Table 2. On pelvic X-ray images, applying Cut&Remain to ResNet-50 improved the AUC-ROCs and F1-scores for the normal class compared to the baseline by 6.6, and 6.4, respectively. Lee et al. presented a method to classify femur fractures on X-ray images using deep learning trained with radiology reports. The experiments they conducted achieved an averaged F1 score of 81.7 in the 3-class classification task with whole images, not cut in half vertically. Based on the results, it was not favorable to translate clinical practice. Conversely, the results of the present study indicate that Cut&Remain can significantly improve classification performance with high F1 score and AUCs.

Method	AUC-ROC			F1-score		
	Normal	A-type	B-type	Normal	A-type	B-type
ResNet-50	91.2 ±0.8	91.2 ±1.3	87.2 ±1.8	92.2 ±0.9	73.0 ±1.1	47.4 ±2.5
Lee et el.	90.6 ±1.8	91.0 ±1.0	88.4 ±1.5	95.4 ±1.4	88.2 ±1.8	76.4 ±2.9
Mixup	93.6 ±1.9	86.2 ±0.8	90.0 ±0.7	95.6 ±1.0	86.2 ±2.1	53.8 ±3.5
Cutout	97.2 ±1.1	96.2 ±1.3	92.4 ±1.9	96.2 ±0.9	84.6 ±1.1	72.6 ±2.5
CutMix	92.8 ±1.5	84.8 ±2.3	92.8 ±1.3	95.2 ±1.0	82.8 ±2.0	57.4 ±2.5
Cut&Remain (w/o aspect ratio variation)	97.4 ±1.3	**97.4** ±**1.1**	97.0 ±1.1	98.4 ±1.2	**93.8** ±**1.2**	85.8 ±2.7
Cut&Remain (w/ aspect ratio variation)	**97.8** ±**0.8**	97.0 ±1.4	**97.2** ±**1.3**	**98.6** ±**0.8**	93.0 ±1.4	**87.0** ±**2.3**

Table 2: Results on the pelvic X-ray dataset. The averages of 5-fold cross validation are reported for each setting with standard deviation.

2.3 MULTI-LABEL CLASSIFICATION ON THE MS-COCO$_s$ DATASET

As mentioned in the main paper, we adopted a criteria to build the dataset, which is the average area occupied of the object is not more than 2% in the image (Figure 1.). Based on the criteria, we finally selected 66,612 training and 2,805 test images, which include a total of 27 types of objects to make up a training dataset, including car, traffic light, kite, cup, etc., after filtering.

We can see from Table 3 that Cut&Remain provides better results than baseline on the MS-COCO$_s$, with mAP 3.3% higher than baseline network. However, the performance improvement by Cut&Remain is relatively small compared to the above medical image tasks. The reason for this is that, unlike medical images, objects are distributed in various location in natural image domain, making it difficult to utilize location information of the lesion.

Although originally developed for medical image tasks, Cut&Remain improve the classification performance in the natural image domain as well. The result implies that mainly considering object context is not only crucial for X-ray images but is also generally helpful for distinguishing subtle differences in local structures in the natural image domain.

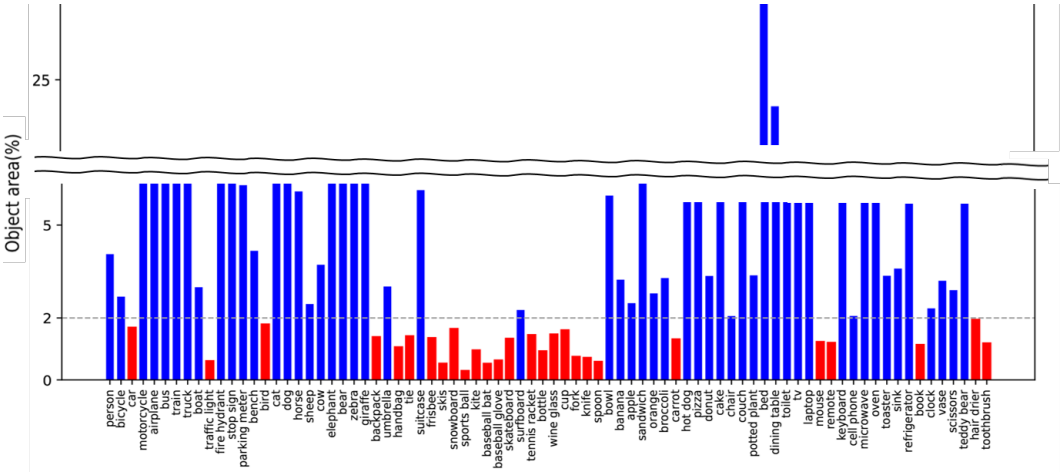

Figure 1: Average area per category

Method	mAP	CF1	OF1
ResNet-50	74.4	69.3	74.0
Mixup	74.1	68.2	72.8
Cutout	75.3	68.7	73.3
Cutmix	75.6	70.1	74.7
Cut&Remain (w/o aspect ratio variation)	75.8	70.3	75.1
Cut&Remain (w/ aspect ratio variation)	**76.7**	**71.8**	75.6

Table 3: Multi-label classification results on the MS-COCO$_s$ dataset. All metrics are in %. Results are reported for input resolution 448.

2.4 WHAT DOES MODEL LEARN WITH CUT&REMAIN?

2.4.1 CUT&REMAIN AS A CORRECTION FOR FAILED PREDICTION

We also analyze samples that Cut&Remain helps to reduce false positives. Figure 1. shows the prediction Grad-CAM results of the baseline method (ResNet-50) and Cut&Remain for several samples of COCO dataset. In Figure 2, the baseline method incorrectly predicts that background regions is an object (i.e. false positives). Specifically, the ResNet-50 and other region-perturbed augmentation techniques could not find small features such as 'baseball glove', 'bird', and 'frisbee'. On the other hand, Cut&Remain captures the region and predicts correctly.

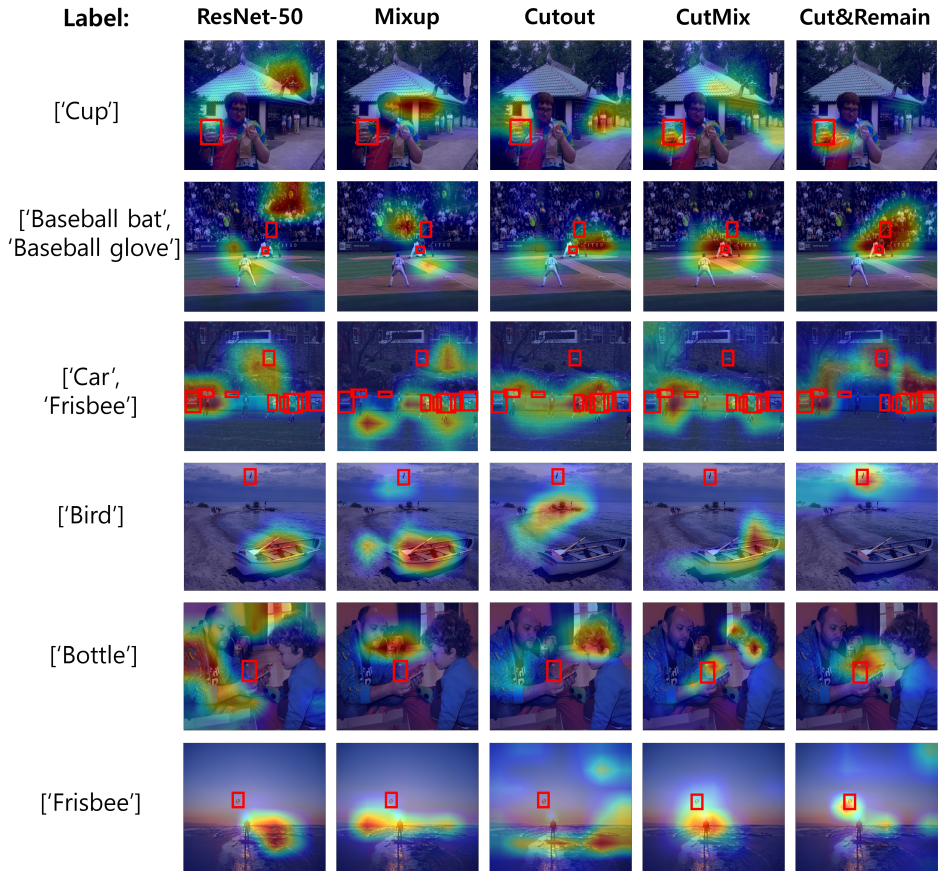

Figure 2: Grad-CAM visualization for the failure samples by baseline method (ResNet-50) on the MS-COCO$_s$ dataset. The red boxes denote ground truth annotations.

2.4.2 FEATURE REPRESENTATION PROPERTY

To verify that Cut&Remain helps to generate a background-independent feature vector (i.e. whether the model mainly focus on the lesion), we analyze the vector similarity between from original and augmented images. For comparison, we use ResNet-50 trained without any augmentation techniques and trained by Mixup, Cutout, Cutmix, and our proposed Cut&Mix strategy. We conducted this study in the training set of clavicle X-rays using the same experimental setting in Section 4.1. We visualize the feature vectors embedded in 2D space by the t-Stochastic Neighbor Embedding(t-SNE) for models trained with each augmentation techniques.

The experimental results are shown in Figure 3. We observed that Cut&Remain makes similar representation vector for the original and augmented sample. On the other hand, conventional augmentation techniques, which randomly remove, mix, or replace regions in images, resulted in increased dissimilarity because informative features might be lost.

2.4.3 PERFORMANCE ALONG THE AMOUNT OF ANNOTATIONS

We conducted ablation study on the MS-COCO$_s$ dataset using the same experimental setting in the main paper. We evaluated Cut&Remain with different amounts of the training dataset which the augmentation was applied. The ratio γ of the training data used for Cut&Remain augmentation was $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

The performance of Cut&Remain with different γ is given in Figure 4. Cut&Remain achieved best performance when the data augmentation is adapted to almost of training dataset (i.e. $\gamma > 0.8$). Specifically, our simple data augmentation achieved highest mAP (76.7) at γ = 1.0, CF1 (72.0) at

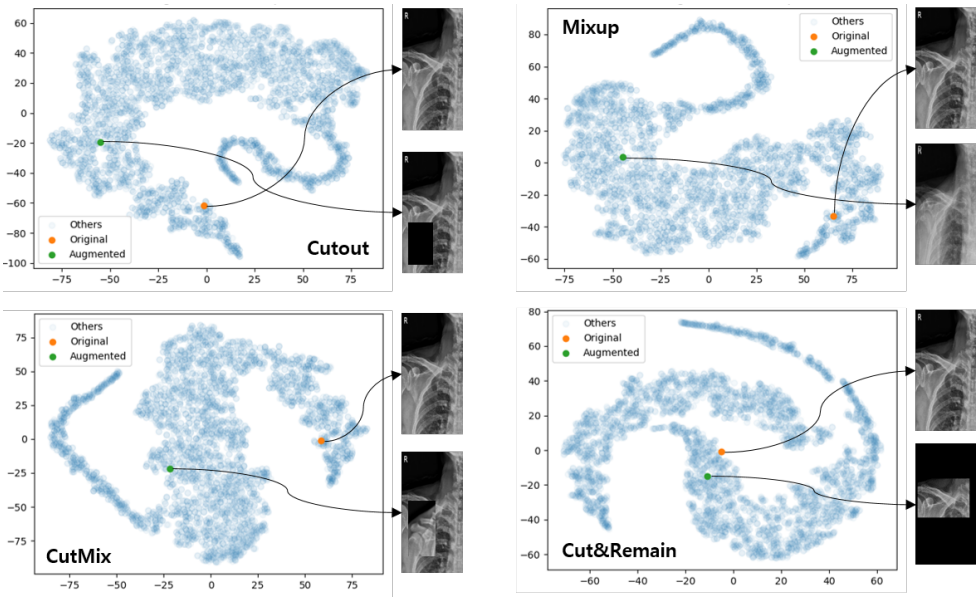

Figure 3: t-SNE visualization of feature vectors of training samples including original and augmented samples by Cutout, Mixup, CutMix, and Cut&Remain.

$\gamma = 0.8$, and OF1 (75.6) at $\gamma = 1.0$. Consequently, we confirmed that performance improvement is guaranteed even with a limited amount of annotations.

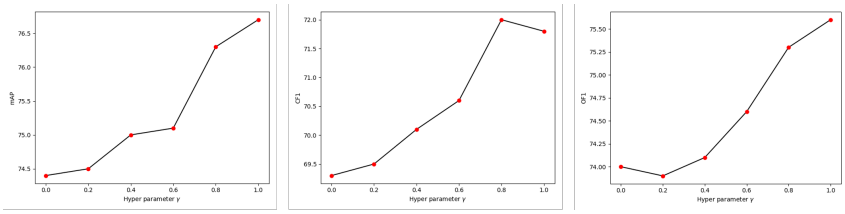

Figure 4: Impact of the amount of Cut&Remain-augmented images in training dataset