Supplementary Material

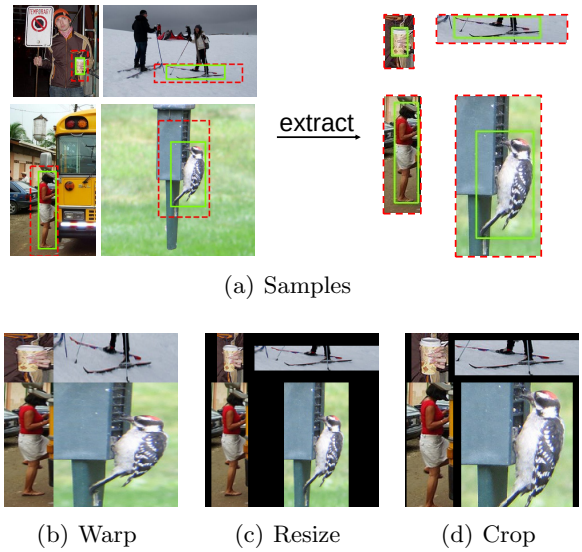
A Details about sample generation

The positive samples will be randomly enlarged to involve more contextual information, which is beneficial for the training process. Specifically, the upper-left and bottom-right corner points of the ground-truth bounding boxes randomly move outwards so that the width/height could be up to 2× of the original width/height. Enlarged bounding boxes beyond original images will be truncated at the edges. Fig. A1 depicts the generation of positive samples. Experimental results indicate that adding context to the positive samples during pre-training will bring 1.6% mAP gain for Faster R-CNN [1] with ResNet-50 backbone.

Fig. A1. The process for generating positive samples. First, we will use ground-truth bounding boxes to locate target regions. Then the regions are randomly enlarged to incorporate context information. Finally the enlarged regions are extracted from the original images as positive samples for pre-training. The green solid lines stand for GT-bounding boxes and blue dash lines for enlarged boxes

B Sample adjustment strategy for Montage assemble process

During the Montage assembled generation process, the samples are adjusted to fit the pre-defined size of in the template. The samples will be randomly cropped or zero-padded, which is conditioned on whether their sizes are smaller or larger than the pre-defined ones. Other possible solutions are to warp or resize the samples. Visualization examples of the different operations are depicted in Fig. A2. From the examples, we can see that resize would result in too many pixels being uninformative and warping will distort the image. Crop is able to retain the shape and preserve more information, which makes it a better choice for size adjustment. Experiment results in Table A1 also indicate that warp and resize lead to suboptimal performances compared with crop.



(a) Samples

(b) Warp (c) Resize (d) Crop

Fig. A2. Visualization of three operations to adjust sample scale. (a) Samples are extracted from the original detection images. Similar to Fig. 2 in the main text, the boxes with green solid lines refer to Ground-Truth bounding boxes, those with red dash lines to the samples, which are randomly enlarged to incorporate more context information. (b) In 'Warp', we change both the aspect ratio and scale of samples. (c) For 'Resize', the aspect ratio is not changed and only the size is changed, then padding is applied to samples whose sizes are smaller than the pre-defined ones. (d) In 'crop', we apply padding or random cropping to samples, conditioned on whether their sizes are smaller or larger than the pre-defined sizes

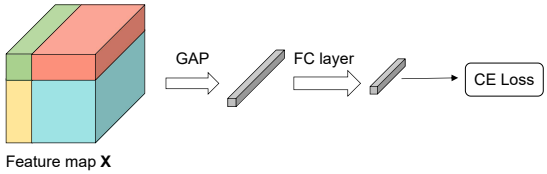
Table A1. Comparison of Warp, Resize and Pad & Crop for scale adjustment during the pre-training process. The network structure is ResNet-50. The three pre-trained models are used for the subsequent detection training of Faster R-CNN [1] and results are evaluated on COCO val2017. The results show that pad & crop is more helpful for obtaining better pre-trained models. For the results of 'Warp', we change both the size and aspect ratio while for those of 'Resize', we only change the size

Method	AP	AP_{50}	AP_{75}
Warp	34.6	54.7	36.8
Resize	34.7	54.7	37.1
Pad&Crop	**35.2**	**55.7**	**37.6**

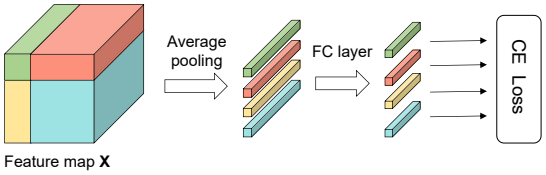
C Visualization and comparison of different classification strategies

This section provides visualization of global-wise and block-wise classification process, respectively, and also shows the comparison of different strategies.

The process of global classification is shown in Fig. A3(b), where global average pooling is exerted on the feature map and we assign the entire image a single global label. The global label is the weighted sum of labels of the four regions according to their region areas. Fig. A3(b) depicts the process of block-wise classification. Different from global classification, the average pooling is independently exerted on the four regions of feature map $\mathbf{X}$ corresponding to samples. Then we will apply classification operation on each region individually according to its label.



(a) Global classification



(b) Block-wise classification

Fig. A3. Visualization on the process of global classification (a) and block-wise classification (b). We use different colors to distinguish regions corresponding to the four samples. Best viewed in color

Table A2 present the experimental results on global-wise, block-wise and ERF-adaptive dense classification, respectively, from which we can see that the performance of the model pre-trained under our ERF-adaptive dense classification strategy is best among three strategies.

D Implementation details

This section provides details of data augmentation during pre-training and training settings of detectors.

Table A2. Comparison of different classification strategies. The backbone CNN is ResNet50 and detection framework is Faster R-CNN[1]. All results are evaluated on COCO val2017. The results show that the ERF-adaptive dense classification strategy clearly outperforms the other two strategies

Strategy	AP	AP_{50}	AP_{75}
Global Cls.	34.3	54.5	36.3
Block-wise Cls.	35.2	55.7	37.6
ERF-adaptive Dense Cls.	**36.3**	**56.5**	**38.9**

Pre-training Augmentation. The samples will first be resized according to a resize ratio chosen randomly from $[0.8, 1.5]$. Both the height and width will be adjusted by the same ratio so that its aspect ratio keeps unchanged. Random horizontal flip with probability 0.5 is also applied on each sample before being assembled into the new image. During Montage assembly, random cropping or zero padding is used to adjust the samples to the pre-defined sizes. After the stitching, the channels of assembled image are normalized with mean $[0.485, 0.456, 0.406]$ and std $[0.229, 0.224, 0.225]$.

Training Details of Detectors. For fair comparisons, we adopt the same training settings on detection for both ImageNet pre-trained models and Montage pre-trained models. We train our models on MS-COCO train2017 split. If not specified, all models are trained for 13 epochs on 8 Tesla V100 GPUs with total batch size 16. We use SGD as the optimizer with momentum 0.9 and weight decay 0.0001. The initial learning rate is 0.02 and decreases by factor 0.1 at epoch 9 and 12. The batch normalization layers are frozen during training. The images are resized to 1333×800 and randomly flipped with probability 0.5 for augmentation.

References

1. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)