This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Erasing Integrated Learning : A Simple yet Effective Approach for Weakly Supervised Object Localization

Jinjie Mai Meng Yang* Wenfeng Luo
School of Data and Computer Science, Sun Yat-sen University, *Corresponding author
The Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Sun Yat-sen University
waynemaibutterfly@gmail.com yangm6@mail.sysu.edu.cn luowf5@mail2.sysu.edu.cn

Abstract

Weakly supervised object localization (WSOL) aims to localize object with only weak supervision like image-level labels. However, a long-standing problem for available techniques based on the classification network is that they often result in highlighting the most discriminative parts rather than the entire extent of object. Nevertheless, trying to explore the integral extent of the object could degrade the performance of image classification on the contrary. To remedy this, we propose a simple yet powerful approach by introducing a novel adversarial erasing technique, erasing integrated learning (EIL). By integrating discriminative region mining and adversarial erasing in a single forward-backward propagation in a vanilla CNN, the proposed EIL explores the high response class-specific area and the less discriminative region simultaneously, thus could maintain high performance in classification and jointly discover the full extent of the object. Furthermore, we apply multiple EIL (MEIL) modules at different levels of the network in a sequential manner, which for the first time integrates semantic features of multiple levels and multiple scales through adversarial erasing learning. In particular, the proposed EIL and advanced MEIL both achieve a new state-of-the-art performance in CUB-200-2011 and ILSVRC 2016 benchmark, making significant improvement in localization while advancing high performance in image classification.

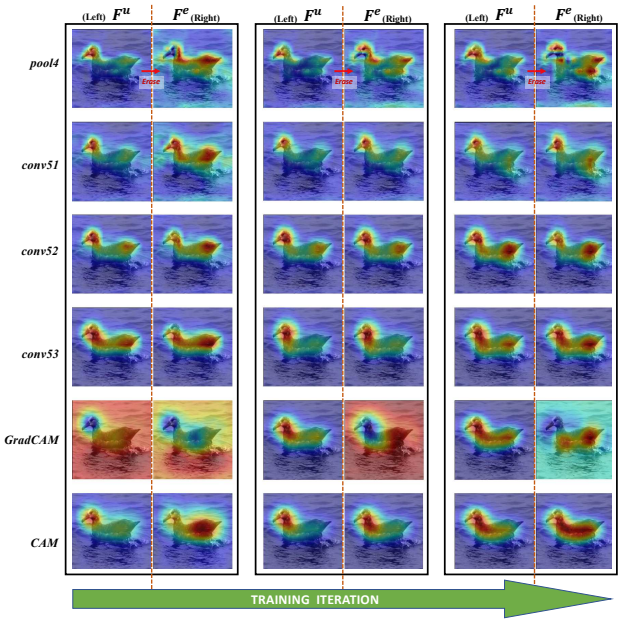
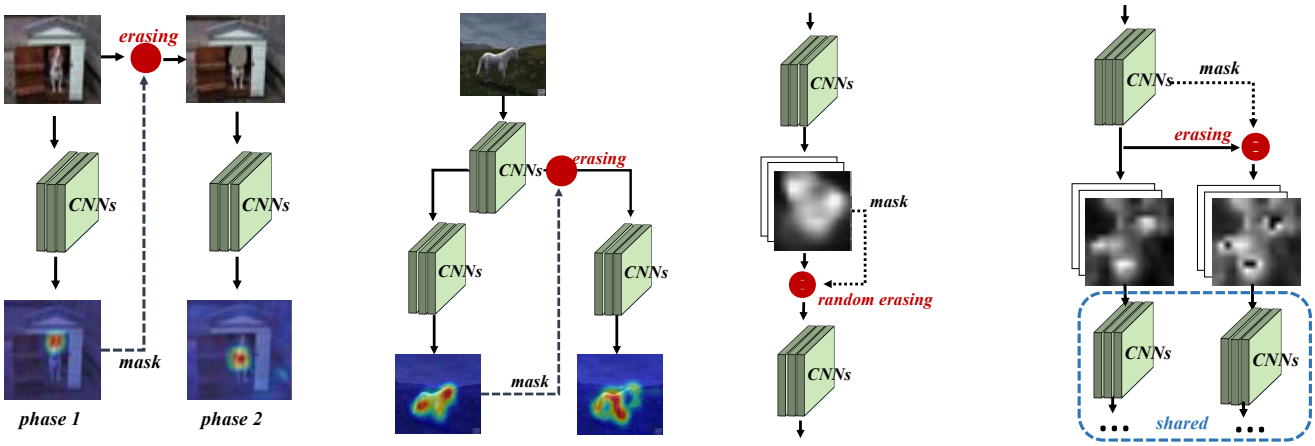


Figure 1: VGG16-EIL with erasing at *pool4*. Visualization of different layers as training proceeds. *pool4* to *conv53* is visualized using channel-wise average map. The left column of each box is visualization of different layers in unerased branch F^u during training, while the right for the erased branch F^e.

1. Introduction

Weakly Supervised Learning (WSL) aims to construct predictive models by learning only with weak supervision [42] like incomplete, coarse, or inaccurate labels. In the field of computer vision, as WSL doesn't require expensive manpower and efforts to obtain pixel-level annotations, weakly supervised object detection (WSOD) [41, 6, 5, 4, 34, 20, 12, 26, 23, 25, 32, 38, 1, 29, 15, 37] and segmentation [14, 10, 16, 25, 24, 18, 7] are attracting more and more attention.

Similar to WSOD, weakly supervised object localization (WSOL) also aims to localize object using coarse labels but for only one class. Recently, various methods [41, 43, 28, 13, 40, 39, 2] have been developed to handle this challenging task. Zhou et al. [41] proposed to replace top layers with Global Average Pooling[19] (GAP) in Convolutional Neural Network (CNN) trained for classification, making it feasible to mine the spatial location of the object. Although the modified CNN is able to generate Class Acti-



(a) Step-wise erasing[35] (b) Multi-branch erasing[39] (c) Random Erasing[2] (d) Ours

Figure 2: A comprehensive comparison of several popular adversarial erasing approaches.

vation Map (CAM) to locate the object, it always tends to mine the most discriminative class-specific regions instead of the full extent of object, resulting in the limited performance in object localization. To address this critical issue, adversarial erasing [35, 13, 33, 39, 8, 17, 2, 11] has emerged as a very popular approach to capture the entire object. The key idea is that, without the guidance of the most discriminative class-specific area, the network will be forced to classify the object by exploring more insignificant areas. On this basis, most of these techniques can be roughly divided into several classical types, as shown in Fig. 2.

As shown in Fig. 2a, a natural and straightforward idea [13, 33, 35, 17, 9] to train a convolutional network may firstly mark the most discriminative regions, then perform the erasing operation and retrain the entire network. Specially, [35] further introduces an iterative erasing approach. However, due to expensive computation overheads for step-wise training, multi-branch based erasing approaches [8, 39] introduce new branches into the network to perform erasing at the cost of extra parameters, as shown in Fig. 2b.

Most recently, [2] presents attention based dropout layer (ADL) shown in Fig. 2c, which stochastically erases the most discriminative regions in forward-propagation, saving quite a few computation and parameter overheads. However, ADL is still limited by classification degradation caused by the random dropout of informative regions.

To solve the aforementioned issues, we propose a brand new adversarial erasing method named *Erasing Integrated Learning (EIL)* . The proposed EIL roughly depicted in Fig. 2d is a simple yet more effective erasing solution. For EIL module, we integrate discriminative region mining and adversarial erasing in a single forward-backward propagation, instead of step-wise erasing requiring huge computation overheads. Different from typical multi-branch erasing approaches, adversarial erasing is integrated into the vanilla CNN directly without any additional parameters through sharing the weights after erasing. In this way, the proposed network can explore the integral extent of objects via the unerased data stream and the erased data stream simultaneously.

Moreover, we have observed a common limitation of existing erasing approaches that all of them only proceed adversarial erasing at particular positions like the input image [35, 17] or the feature map [8, 39, 2]. Such treatments can lead adversarial learning to only focus on mining visual patterns at a specific feature level. Thus, we push further to come up with an advanced Multi-EIL (MEIL) strategy. By plugging multiple EIL modules into different layers of CNN in a sequential manner, MEIL adversarially integrates semantic features from multi-level of the network and mines multi-scale informative regions of the object of interest. The proposed EIL and the advanced MEIL both achieve new state-of-the-art performance both on CUB-200-2011 [31] and ILSVRC [22] benchmark, improve the localization accuracy by a significant margin while maintain remarkable classification accuracy.

2. Related work

Erasing approach for weakly supervised learning. Step-wise erasing approaches [13, 33, 35, 17] roughly like Fig. 2a. usually perform erasing in additional training step. For example, Li et al. [17] propose guided attention inference networks (GAIN) implemented by introducing two streams. At the first stream, GAIN aims to find out discriminative regions and generate a trainable attention map as the erasing mask. Conversely, the second training stream is enforcing the network not to recognize the erased area, hence

the gradients will supervise the trained attention to cover the full object. [33] also adopts a two phase training strategy. They pretrain a classification network first and then erase the most discriminative regions to retrain the network, forcing the network to focus on the next most important part. [35] further introduces an iterative erasing approach, which repeatedly erases the most discriminative region in given image and finally combines attention maps in these steps to get a more complete attention map for the object.

For less computation overhead, multi-branch erasing approaches [8, 39] like Fig. 2b replace extra training steps with extra parameters for adversarial erasing. [39] leverages a dual-branch network, adversarial complementary learning (ACoL). ACoL applies two parallel classifiers on top to train the network, one is fed with the unerased feature map directly from the shared backbone and generates the erasing mask, while the other one is fed with the erased feature map by this mask. Further on, a three-branch SeeNet [8] proposed by Hou et al., introduces two self-erasing strategies both for the object and background cues, which can prevent the attention from transferring to background area thus excavate the object more accurately.

To further reduce computation and parameter overheads, attention-based dropout layer (ADL) [2] has been proposed, a lightweight module as shown in Fig. 2c. When ADL is plugged into the network, it stochastically choose to erase the most discriminative region or highlight the informative region in the feature map. But the random erasing would somehow drop the important information, leading to a performance loss in classification.

Other approaches for Weakly Supervised Object Localization (WSOL). Zhou et al. [41] employs CAM to identify the location of object of interest in an end-to-end manner through the global average pooling [19] module. Hide-and-Seek (HaS) [28] randomly hides patches of the given image to force the network to seek more relative part of the object, which can be also considered as a way of data augmentation. The soft proposal network [43] jointly optimizes the network parameters with the object proposal. Wei et al. [35] exploit segmentation confidence maps to discover tight object bounding boxes. [35, 39, 17, 2] all adopt the erasing mechanism to capture the integral extent of object, which have been discussed in the previous section. Self-produced guidance [40] (SPG) approach utilizes the supervisions from high confident regions and drives the attention to gradually spread to the whole object. [21] has proposed an advanced localization map generation strategy that combines the gradients of different convolutional layers to generate the localization map in a multi-scale manner. Most recently, Xue et al. [36] design a divergent activation (DANet) network. With the help of stronger supervision about objects' category hierarchy, DANet leverages cross-category semantic discrepancy and spatial discrepancy to learn complementary and discriminative visual patterns.

3. Erasing Integrated Learning

Erasing integrated learning aims to provide a more elegant and concise erasing solution for weakly supervised tasks, integrating adversarial erasing strategy into CNN directly without additional steps or classifiers. For this, we come up with EIL to integrate the unerased data stream and the erased data stream from the common backbone into a dual-branch network with shared weights. Furthermore, we propose Multi-EIL to introduce different scale semantic features into the network which further improves the localization performance through multiple adversarial erasing learning procedures.

3.1. Integrated with erasing

In this section, we give the detail of our proposed EIL, as shown in Fig. 3. In general, EIL is added between convolutional blocks of CNN in a sequential way. During training, taking the flow-in feature map as input, we simply follow [2] to generate the erasing mask, and remove the most discriminative regions on the feature map according to this mask. Then we feed both erased and unerased feature map into the next convolutional block, which will create two data flows. As such treatment can also be regarded as a dual-branch network with shared weights, it will produce two classification losses for the erased feature map and the unerased one, respectively. During testing, EIL is deactivated, thus the trained model is identical to the vanilla classification network. Through the *unerased loss*, the network can learn to classify the object by means of the most discriminative class-specific region. At the mean time, the *erased loss* drives the network to focus on the less discriminative part to explore the complementary object region, as shown in Fig. 1.

Detailed description is presented in Algorithm 1 and Fig. 3. Formally, we denote the training image set as $I = \{\{I_i, y_i\}\}_{i=1}^{N}$, where $y_i = \{1, 2, ..., C\}$ is the label of the image I_i, C is total classes of images and N is the amount of images. Let θ, lowercase f, and uppercase F denote network parameters, functions, and feature maps, respectively.

The network $f^1(I, \theta^1)$ before EIL is applied can produce the original unerased feature map, which is denoted as $F^u \leftarrow f^1(I_i, \theta^1)$ and $F^u \in \mathcal{R}^{K \times H \times W}$, where K stands for the number of channels, W and H for the width and height, respectively. We make use of F^u as self-attention to generate the erasing mask. Specifically, we compress F^u into an average map $M_{avg} \in \mathcal{R}^{1 \times H \times W}$ through channel-wise average pooling. Then we apply a hard threshold γ on M_{avg} to produce the erasing mask $M_e \in \mathcal{R}^{1 \times H \times W}$, in which the spatial locations of those pixels having intensity greater than γ are set to zero. We perform the erasing opera-

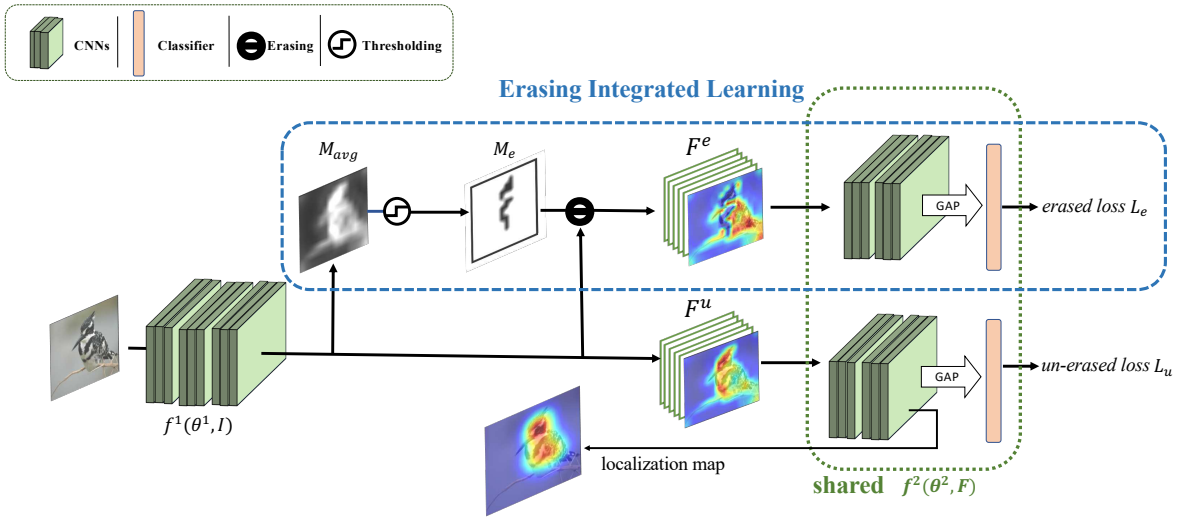


Figure 3: Overview of the proposed module EIL. When EIL is inserted at a feature map, an average map M_{avg} is first produced by channel-wise average pooling. With thresholding M_{avg} to obtain an erasing mask M_e, the erased feature map F^e by M_e and the unerased F^u are fed into the network again under a shared dual-branch treatment.

tion by doing spatial-wise multiplication between unerased feature map F^u and mask M_e, to produce the erased feature map $F^e \in \mathcal{R}^{K \times H \times W}$.

Afterwards, both the unerased feature map F^u and the erased counterpart F^e are fed into the latter part of the network $f^2(F, \theta^2)$ together. As these two data streams are processed by the same function f^2 and parameters θ^2, such structure can be regarded as a dual-branch network of shared weights. More specifically, $f^2(F, \theta^2)$ produces class activation maps (CAM) [41], applies global average pooling [19] on CAM and utilizes a fully connected layer followed by softmax operation to get the prediction score p for each branch, with p^u and p^e for the erased and the unerased, respectively. In the end, the classification losses from the two branches will be added up to calculate the total loss $\mathcal{L}$. Note that we also introduce a loss weighting hyperparameter σ to control the relative importance between the unerased loss $\mathcal{L}^u$ and the erased loss $\mathcal{L}^e$.

Algorithm 1: Training algorithm for EIL

Input: Input image $I = \{\{I_i, y_i\}\}_{i=1}^N$ from C classes, erasing threshold γ, weighting hyperparameter σ

while *training is not convergent* **do**
- Calculate the feature map $F^u \leftarrow f^1(I_i, \theta^1)$;
- Calculate the average map $M_{avg} = \frac{\sum_{i=1}^K F_i^u}{K}$;
- Calculate the erasing mask $M_{e_{i,j}} = \begin{cases} 0, & if\ M_{avg_{i,j}} \geq \gamma \\ 1, & else \end{cases}$;
- Get the erased feature map $F^e = F^u \otimes M_e$;
- Calculate prediction of F^e: $p^e \leftarrow f^2(F^e, \theta_2)$;
- Calculate prediction of F^u: $p^u \leftarrow f^2(F^u, \theta_2)$;
- Obtain erased loss: $\mathcal{L}_e = -\frac{1}{C}\sum_c y_{i,c} log(p_c^e)$;
- Obtain unerased loss:$\mathcal{L}_u = -\frac{1}{C}\sum_c y_{i,c} log(p_c^u)$;
- Calculate the total loss: $\mathcal{L} = \mathcal{L}_u + \sigma\mathcal{L}_e$;
- Back-propagate and update parameters θ^1, θ^2 ;

end

3.2. Jointly mining the whole object

Firstly, considering the unerased loss $\mathcal{L}_u$ and the corresponding branch only, it is actually identical to a typical CNN without any difference. So this branch will surely learn to do as a network trained for classification supposed to do: highlight those class-specific discriminative regions for better object classification. In this way, the network parameter θ^1 can learn the ability to classify the object as well as the vanilla classification model. However, solely rely on the pure guidance of $\mathcal{L}_u$, CAM usually cover the small and sparse regions of object of interest, since $\mathcal{L}_u$ is overwhelmed by the most discriminative parts. As shown in the CAM map of Fig. 1, the network only focus on the most discriminative regions like the head of the bird at the initial stage.

Thus we integrate adversarial erasing to the network, through which the erased loss $\mathcal{L}_e$ can make a role play for the dense-pixel prediction task. With the prominent cross-class activations in F^e erased, the latter part of the network

$f^2(F^e, \theta^2)$ produces the loss from the activation units from the less discriminative area. Consequently, when the gradient $\mathcal{G}_e$ from the erased loss $\mathcal{L}_e$ flows back through θ^1 and θ^2, the neurons in them which are spatially corresponded to the distribution of the less discriminative regions in the object are updated with emphasis.

Once the erased loss $\mathcal{L}_e$ is optimized, the network θ^1, θ^2 can learn to mine the less discriminative and category independent visual patterns. As we have illustrated, these two data streams are exactly flowing in the shared network θ^2 based on the same backbone θ^1. So $\mathcal{L}_u$ and $\mathcal{L}_e$ are updating the same parameters θ^1, θ^2 but focusing on different specific units. Hence while the units for the most discriminative part are also fine tuned by $\mathcal{L}_u$, EIL can thus localize full object extent holistically through combining complementary and discriminative object patterns at the same time.

GradCAM[23] shown in Fig. 1 provides the visualization evidence for our explanation, shown in Fig. 1. It can be observed that the upcoming gradient $\mathcal{G}_e$ from F^e gradually leads the unerased branch F^u to cover the full extent of the object including the less discriminative regions, like the torso of the bird.

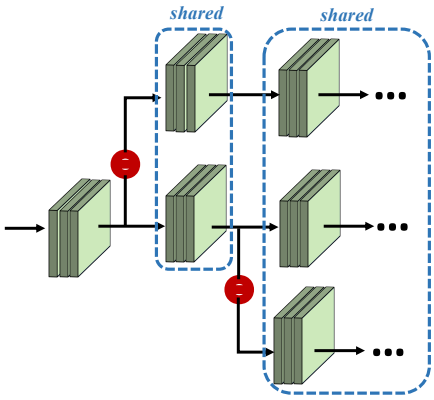


Figure 4: The structure of proposed MEIL I.

3.3. Multiple EIL for multi-scale features

While existing erasing approaches all choose to erase in a single location, we propose an advanced multiple EIL (MEIL) modules to perform erasing at multiple locations, through which multi-scale of visual patters can be learned adversarially and simultaneously.

A typical structure of MEIL, MEIL I, is shown in Fig. 4. Once after a single EIL is inserted into the vanilla CNN, another EIL is appended to the unerased stream. As a result, the network will produce three losses from the shared branches, which can lead the network to explore the object of interest from multi-level features, not just the finest discriminative features for classification.

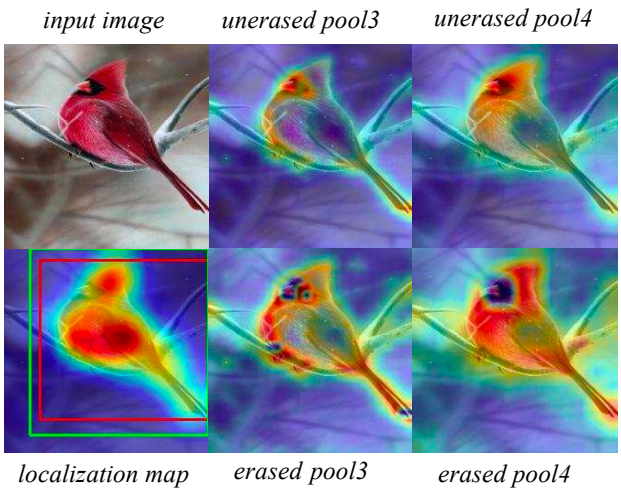


Figure 5: A visualization of VGG-MEIL with two EIL modules inserted in *pool3* and *pool4*. Average maps comparison of feature maps from different levels before and after erasing.

3.4. Discussion

Relation between the erased data stream and the unerased. As these two data streams are flowing forward on the shared θ^2 and backward on the same entire network θ^1, θ^2, one might worry that gradients of the two streams, $\mathcal{G}_e$ and $\mathcal{G}_u$ would make conflicts and counteract each other. But our experiment results do not support such hypothesis. As we have discussed, we believe that $\mathcal{G}_e$ and $\mathcal{G}_u$ are actually paying attention to different units in the network, while the former for the most discriminative part, and the latter for the less discriminative part. The visualization of EIL shown in Fig. 1 and 6 also supports our explanation. We can verify that, those high response area discovered in the raw CAM model, a classical CNN for classification, also keeps showing up in our EIL model. It means that EIL also learns the parameters to explore the most discriminative region *(e.g. the head of birds)* and retain the ability to well classify the object. Beyond that, we also notice that compared with CAM, the less discriminative region *(e.g. the body of birds)* has been given the equal highlighting treatment as well as the most significant part shown in CAM, which again proves our hypothesis. In other words, these areas of interest that are usually ignored in CAM are magnified in EIL.

Relation with existing erasing techniques for WSOL. Here we give a brief comparison with other typical adversarial erasing approaches similar to our EIL in WSOL tasks. For ADL[2] shown in Fig. 2c, which stochastically erases the most discriminative regions in a single forward-propagation, the random dropout of informative re-

gions could degrade its performance in classification. On the other side, our EIL also inherits the advantages of ADL. These include the flexibility to insert at arbitrary convolutional block, and the lightness of not requiring additional parameters. ACol[39] shown in Fig. 2b process two source with separate branches. The reason why our EIL works better may include three parts: 1) ACoL applies different erasing mask generation techniques from ours and ADL. They extract it at the top layer and resample it to perform erasing at the middle layer, where the resampling operation may blur dense-pixel information if the network hasn't learned the ground-truth class properly at that moment. 2) ACoL only shares parameters at the bottom layers of the network, from which the extracted low-level features like edges or texture is general and class-invariant. Consequently, the losses from two separate branches may not help the backbone to learn class-specific localization effectively. 3) ACoL fuses the CAM maps of two separate branches to produce the final localization map, which may be inconsistent as they might overwhelm each other.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate the proposed EIL on two popular benchmarks CUB-200-2011 [31] and ILSVRC 2016 [3, 22], which are all only annotated with image-level labels for training. There are 5,994 images for training and 5,794 for testing in CUB-200-2011 from 200 bird species. For ILSVRC 2016, there are approximately 1.3 million images in the training set and 50,000 images in the validation set come from 1,000 different classes.

Metrics. Following the setting of [22, 2], we adopt the Top-1 classification accuracy *(Top-1 Clas)*, Top-1 localization accuracy *(Top-1 Loc)* and localization accuracy with known ground-truth class *(GT Loc)* as our evaluation metrics. *Top-1 Clas* is the ratio of correct classification prediction. *Top-1 Loc* is the fraction of images with correct prediction of classification and more than 50% intersection over union (IoU) to the ground-truth bounding box. *GT Loc* is the accuracy that considering localization only regardless of classification result compared to *Top-1 Loc*.

Implementation details. We build the proposed EIL module upon two popular CNNs including VGGnet [27] and Google InceptionV3 [30]. Following the training settings of previous work [41, 40], we remove the top pooling layer and two fully connected layers for VGG16, and the layers after the second inception block for InceptionV3. Then we add two (one for VGG16) convolutional layers of kernel size 3×3, stride 1, pad 1 with 1024 filters a fully connected layer and finally a GAP layer on the top. Both networks are loaded with pretrained weights of ILSVRC. We insert the proposed EIL after the *pool4* layer for VGG16 and the first inception block for InceptionV3. We adopt SGD as optimizer with $momentum = 0.9$, $weight\ decay = 0.0005$. We set the initial learning rate as 0.001, and it is decreased by a factor of 10 at the decay points. The input images for training are resized to 256×256, then randomly cropped to 224×224 and flipped horizontally. We adjust the erasing threshold γ and the loss weighting parameter σ to fine tune the network.

For both backbones, we set $\gamma = 0.7$ and $\sigma = 2$ for a single EIL module, while optimizing these hyperparameters for specific dataset and backbone can further improve the performance. During testing, EIL is deactivated. For a fair comparison, we directly follow the localization map extraction method proposed by CAM [41].

4.2. Ablation study

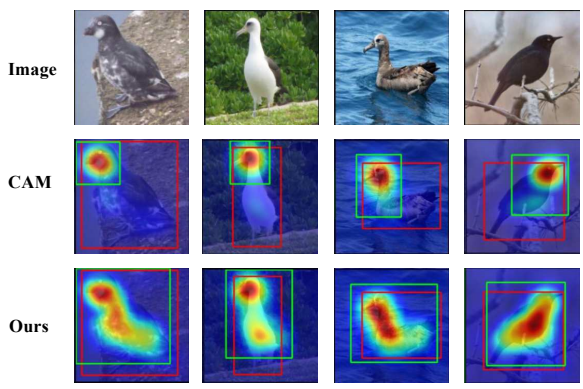
We utilize the modified VGG16 as backbone on CUB-200-2011 dataset for ablation study.

Location. Firstly, we examine the impact of where to insert EIL in the network. We fix $\gamma = 0.7$, and $\sigma = 1$ and then change the location selection of EIL, shown in Table 1. We can find that the best localization performance is achieved when EIL is applied in the middle of the network like *pool4*. There is a gap compared to adding it to the low-level like *pool3* or the top level like *conv 5-3*, which is also observed in existing works [28, 2]. We believe that this is because the low-level activation of the network is more about common basic features *(e.g. edge, texture)* in the whole image rather than regions of the object.

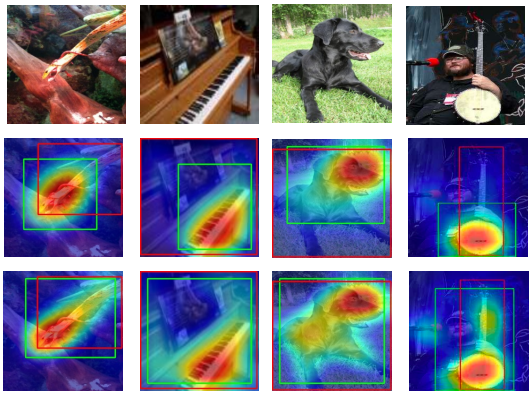
At the meantime, due to the smaller resolution at the high-level layer like *conv 5-3*, the larger receptive field can lead to inexact gradients for the bottom layers after upsampling, providing fuzzy guidance for dense-pixel object mining. Thus the improvement of localization is also limited. On the contrary, with the high-level layer closer to the *FC* layer, the classification performance is improved compared with other location, which can be regarded as a kind of regularization by suppressing the high response activation.

Hyperparameters. As illustrated in Algorithm 1, we introduce a necessary threshold σ for erasing operation and a weighting parameter γ to balance the erased loss $\mathcal{L}_e$ and the unerased loss $\mathcal{L}_u$. We plug the EIL module right behind *pool4* suggested by above discussion and change these two parameters respectively, as shown in Table 1. For γ, neither too high nor too low can yield the promising localization result. Because a low threshold could erase the activation of the entire object turning the network attention to background, while a high threshold could not erase the highest response area completely.

Interestingly, we find that making the erased loss $\mathcal{L}_e$ occupy a larger weight by setting σ higher can even make a better localization result. Our explanation is from two parts. Firstly, as the most discriminative region is small



(a) CUB-200-2011



(b) ILSVRC 2016

Figure 6: Visualization comparison with the baseline CAM method. The groundtruth bounding boxes are in red, while the predictions are in green. EIL is putting more attention to the object and thus providing more accurate prediction.

and sparse, $\mathcal{L}_u$ is overwhelmed by the activation of just a few neurons. Instead, the less discriminative region is usually larger than the former. So neurons corresponding to the less discriminative region *(e.g. area close to the object edges)* are actually making relatively little contribution to $\mathcal{L}$. Therefore, to magnify $\mathcal{L}_e$ several times can make these "less discriminative" neurons get a more equal treatment in backward-propagation. Our visualization in Fig. 6 also supports that both the most and the less discriminative regions are getting comparable attention from the network through applying EIL.

Location	GT Loc (%)	Top-1 Clas (%)	Top-1 Loc (%)
N/A	55.32	71.24	44.15
conv 5-3	60.75	**73.37**	46.77
pool4	**72.37**	72.99	**55.44**
pool3	67.48	70.04	51.06
pool2	63.27	68.43	47.51
pool1	62.74	71.19	46.89

Table 1: The result upon the selection of location.

		γ		
		0.5	0.7	0.9
σ	0.5	52.57 / 67.59	53.23 / 70.61	50.72 / 71.61
	1	52.41 / 66.97	55.44 / 72.99	51.41 / 72.20
	2	50.34 / 66.00	**56.21** / 72.26	52.13 / 73.11
	4	52.14 / 68.05	55.64 / 72.52	51.34 / **74.61**

Table 2: The affection of hyperparameters, Top-1 Loc (%) / Top-1 Clas (%)

Structure of MEIL. We also evaluate the performance when multiple EIL modules are inserted in different ways. In a case that EIL already exists in the network, one may choose to plug another EIL whether in the unerased branch (Fig. 4) or the erased one (Fig. 7). After trying various combination of training settings, we observe that the effect of MEIL II in Fig. 7 is usually worse than MEIL by a margin about $2\% \sim 5\%$. Because MEIL II may sometimes erase too much regions of the object of interest on the feature map, drive the attention to background and lead to a worse performance. Additionally, when performing erasing again just after a few convolutional layers, the next most important part may not have been excavated yet. On the other hand, for MEIL I shown in Fig. 4, training the network with erased stream from multi levels can drive the network to learn multi-scale features, as we have discussed in Section 4.2. Also, such approach is similar to increase σ in single EIL, which enhances the importance of erased loss $\mathcal{L}_e$.

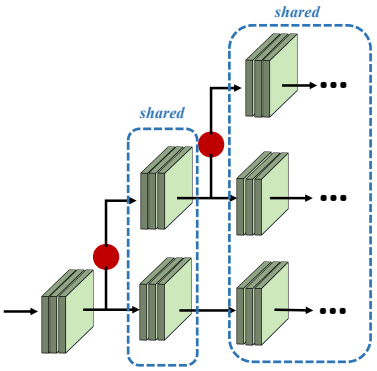


Figure 7: MEIL II, a variant of MEIL shown in Fig. 4

Next, we push further to apply MEIL I at the combination of different layers in VGG16. Results shown in Table 3

indicates that multiple EIL modules have outperformed the best performance of single EIL shown in Table 2. So the employment of EIL and MEIL can be a trade-off between training resources and testing accuracy. As the combination of multiple EIL is numerous, we advocate that the performance of MEIL can be further improved by setting the optimal localization of insertion tuning the hyperparameters or even introducing more than two EIL modules.

Location	GT-Loc	Top-1 Clas	Top-1 Loc
N/A	55.32	71.24	44.15
pool3+pool4	**73.84**	74.77	**57.46**
pool4+conv53	62.21	**74.87**	47.62
pool3+pool4+conv53	65.52	74.80	50.54

Table 3: Influence of the location selection with MEIL I.

4.3. Comparison with State-of-the-art Methods

We compare our result with other state-of-the-art techniques on CUB-200-2011 and ILSVRC 2016 in Table 4 and Table 5 respectively. From the results, we observe that our EIL has outperformed all the existing methods on localization accuracy.

Methods	Top1-Loc(%)	Top-1-Clas(%)
InceptionV3-CAM [41]	43.67	73.80
InceptionV3-SPG [40]	46.64	-
InceptionV3-ADL [2]	**53.04**	**74.55**
InceptionV3-DANet [36]	49.45	71.20
VGG-CAM [41]	44.15	71.24
VGG-ACoL [39]	45.92	71.90
VGG-ADL [2]	52.36	65.27
VGG-DANet [36]	52.52	**75.40**
VGG-EIL (ours)	56.21	72.26
VGG-MEIL (ours)	**57.46**	74.77

Table 4: Quantitative result on CUB-200-2011

On the CUB-200-2011 test set, we insert MEIL I at *pool3+pool4* of VGG16. As a result, VGG-MEIL indicates 13.31% localization boost on the baseline CAM approach, which is a very impressive improvement. Compared with the current state-of-the-art DANet [36], which has introduced extra supervision about category hierarchy, VGG-MEIL is in a narrow margin that only 0.63% lower for classification. But for localization, it reports a significant performance gain of 4.94% over DANet. Also, even VGG16 with single EIL can achieve 56.21% / 72.26% accuracy in classification and localization respectively. In conclusion, the proposed EIL can promote the quality of object localization by a big step while maintaining high performance in classification.

Methods	Top1-Loc(%)	Top-1-Clas(%)
VGG-CAM [41]	42.80	66.60
VGG-ACoL [39]	45.83	67.50
VGG-ADL [2]	44.92	69.48
VGG-EIL (ours)	46.27	**70.48**
VGG-MEIL (ours)	**46.81**	70.27
InceptionV3-CAM [41]	46.29	68.1
InceptionV3-HaS-32 [28]	45.47	-
InceptionV3-SPG [39]	48.60	-
InceptionV3-ADL [2]	48.71	72.83
InceptionV3-DANet [2]	47.53	72.50
InceptionV3-EIL (ours)	48.79	**73.88**
InceptionV3-MEIL (ours)	**49.48**	73.31

Table 5: Quantitative result on ILSVRC

In the ILSVRC 2016 experiments, which is a more larger scale dataset, both EIL and MEIL achieve new state-of-the-art performance in all the metrics upon all the backbones. Specifically, VGG-MEIL obtains an localization accuracy of 46.81%, 0.89% improvement compared to ACoL [39]. In addition, on the InceptionV3 backbone, EIL and MEIL not only obtain the best localization performance, but also improve the classification accuracy by 5.78%/5.21% over the baseline CAM approach.

5. Conclusion

We come up with a simple yet effective adversarial erasing approach, Erasing Integrated Learning (EIL), which integrates the stream of erased feature map into the classification network. Without introducing any extra parameters both in training and testing, this is the first time that the network learns to explore the full extent of the object via concurrent data streams with and without erasing in a single forward-backward propagation. Further on, to the best of our knowledge, this is also the first time that multi-scale and multi-level object features are explored through integrating erasing based learning. In the end, the proposed EIL and its variant Multi-EIL have achieved the new state-of-the-art performance for weakly supervised object localization.

6. Acknowledgement

This work is partially supported by National Natural Science Foundation of China (Grants no. 61772568), Natural Science Foundation of Guangdong Province under Grant 2019A1515012029, and the Guangzhou Science and Technology Program (Grant no. 201804010288).

References

[1] Aditya Arun, CV Jawahar, and M Pawan Kumar. Dissimilarity coefficient based weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9432–9441, 2019.

[2] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 914–922, 2017.

[5] Xuanyi Dong, Deyu Meng, Fan Ma, and Yi Yang. A dual-network progressive approach to weakly supervised object detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 279–287. ACM, 2017.

[6] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 642–651, 2017.

[7] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[8] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In *Advances in Neural Information Processing Systems*, pages 549–559, 2018.

[9] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018.

[10] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018.

[11] Zhaoyang Huang, Yan Xu, Jianping Shi, Xiaowei Zhou, Hujun Bao, and Guofeng Zhang. Prior guided dropout for robust visual localization in dynamic environments. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[12] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1377–1385, 2017.

[13] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017.

[14] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016.

[15] Satoshi Kosugi, Toshihiko Yamasaki, and Kiyoharu Aizawa. Object-aware instance labeling for weakly supervised object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[16] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019.

[17] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223, 2018.

[18] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[19] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[21] Pei Lv, Haiyu Yu, Junxiao Xue, Junjin Cheng, Lisha Cui, Bing Zhou, Mingliang Xu, and Yi Yang. Multi-scale discriminative region discovery for weakly-supervised object localization. *arXiv preprint arXiv:1909.10698*, 2019.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[24] Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L Rosin, and Liang Wang. Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia*, 2019.

[25] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 697–707, 2019.

[26] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3381–3390, 2017.

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

[28] Krishna Kumar Singh and Yong Jae Lee. 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)*, pages 3544–3553. IEEE, 2017.

[29] Krishna Kumar Singh and Yong Jae Lee. You reap what you sow: Using videos to generate high precision object proposals for weakly-supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9414–9422, 2019.

[30] 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.

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[32] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019.

[33] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6):2868–2881, 2017.

[35] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[36] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[37] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[38] Zhenheng Yang, Dhruv Mahajan, Deepti Ghadiyaram, Ram Nevatia, and Vignesh Ramanathan. Activity driven weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2917–2926, 2019.

[39] 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.

[40] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 597–613, 2018.

[41] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[42] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.

[43] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1850, 2017.