Weakly Supervised Object Localization Using Self-Paced Pyramid Adversarial Learning

FuCheng Pan*, BeiLei Bian†, BinXu Wang†, YuePing Yang† and XiaoMing Ju*
*SoftWare Engineering Institute
East China Normal University, ShangHai, China
Email: {pfcqlj@stu,xmju@sei}.ecnu.edu.cn
†NingBo Electric Power Company, State Grid Corporation, NingBo, China

***Abstract*—Weakly Supervised Object Localization (WSOL) has increasingly attracted interests for only using image-level supervision instead of location annotations. Some common challenges for existing methods are that they cover only the most discriminative part of the object. And a substantial amount of noise in training causes ambiguities for learning in a robust manner. In this paper, we propose to address these drawbacks by Self-Paced Pyramid Adversarial Learning(SPAL). Specifically, for suppressing noise, we use self-paced learning (SPL) to training data from simple to complex and from coarse to fine. And our network divides two subnetworks: 1) coarse pyramid network(CPN), 2) fine pyramid network (FPN). In CPN, we aim to utilize pyramid adversarial erase mechanism to process the feature maps of different scale. Consequently, CPN can cover the entire object to generate initial object proposals. Then CPN builds the relevance score as pseudo labels of proposals. In FPN, object proposals and pseudo labels can be trained to locate precise object boundaries. Finally, We also propose adversarial loss function to fit our network. Detailed experimental results on the PASCAL VOC dataset demonstrate that SPAL performs promising against the state-of-the-art methods.**

***Index Terms*—weakly supervised learning, self-paced learning, object localization, pyramid adversarial erase, adversarial loss**

I. Introduction

OBJECT localization is a fundamental component of computer vision. It aims to locate all instances of particular object categories (*e.g.* person, cat, and dog) in images. Currently, object localization has made breakthrough progress [1]–[4] in a fully supervised object detection manner. However, strong supervision needs many object bounding boxes or segmentation masks annotations, which are time-consuming and labor-intensive. Besides, imprecise and ambiguous manual annotations also cause training instabilities. Try to alleviate these, recently, weakly supervised object localization (WSOL) has received widespread attention [5]–[8]. Despite this progress, WSOL still is a very challenging but promising task.

The gap of WSOL derives from high randomness of object location due to weakly annotated data. So the most usual methods for tackling WSOL is to formulate it as multiple instances learning problems (MIL) [9]–[11]. MIL treats each training data as a "bag" and iteratively selects high-scored instances from each bag when learning detectors. Many object proposals are chosen by conventional methods such as selective search [12], edge boxes [13], etc. When training large-scale datasets, however, MIL remains puzzled by random inadequate solutions. Especially for MIL's non-convexity, it causes MIL's methods to be sensitive for model initialization and prone to getting trapped into a local minimum. Although a series of efforts have been made to alleviate the problem by seeking better initialization and optimization strategies [14], [15] or empirically regularizing the learning procedure [16]. The issues of quantifying sub-optimal solutions and principally reducing localization randomness remain unsolved.

Recently, Class Activation Mappings (CAM) [17] gives WSOL a new perspective. CAM leverages Convolutional Neural Networks (CNN) classifier for learning the discriminative features. The key idea is that the classifier with a reasonable accuracy should observe the object region to decide the category label. In other words, the object region should co-occur with the discriminative features. Unfortunately, the classifiers always tend to focus only on the most discriminative features to decide final classification results. Therefore, the spatial distribution responses also manage to cover only the most discriminative part of the object, which leads to localization errors. For getting the extent of integrated objects, adversarial erase (AE) [18]–[22] has been proposed. The similarities between these technologies are that they prevent the model from relying solely on the most discriminative part for classification, instead encourages it to learn the less discriminative part as well.

Instead of MIL, in this paper, we focus on emerging approaches represented by AE. We investigate that the idea of adversarial erase only the most discriminative part is practical to capture the full extent of object. But some drawbacks can't be ignored. Firstly, whether it is MIL or AE, existing schemes always attempt to train object detectors directly from a large and noisy collection of data. It is challenging to get correct localization because the dataset contains many noisy results (*e.g.*, background clutter, object parts). Secondly, these AE methods always need additional classifier, which waste substantial computing resources. Finally, previous AE methods can not build links to multiple object locations and categories, which makes them difficult to put into practice. So our key observation is that effective AE methods should use progressive learning to locate objects and to learn in a robust manner.

For the status quo, in this paper, we propose a novel Self-Paced Pyramid Adversarial Learning(SPAL) for weakly

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

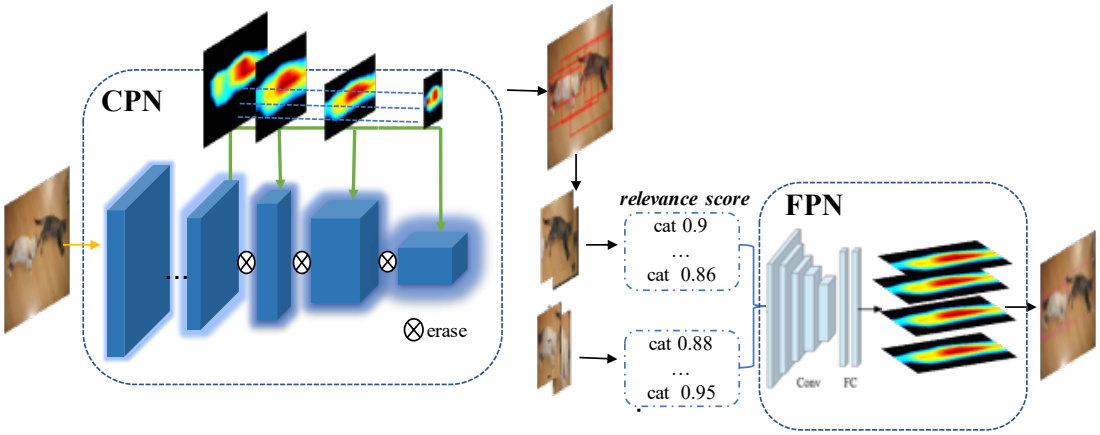

Fig. 1. Overview of our proposed SPAL methods, from coarse to fine, CPN generates initial object proposals, and FPN refines precise object boundaries

supervised object localization. The key ideas are inspired by self-paced learning (SPL) [23]–[25] and adversarial erase (AE) [18]–[22]. In particular, according to SPL strategies, the training data is divided into several levels from simple to complex and from coarse to fine (e.g., simple background and complex background, single object and multiple objects, etc). As shown in Fig. 1, SPAL trains two subnetworks: 1) Coarse Pyramid Network(CPN), 2) Fine Pyramid Network (FPN). In CPN, pyramid adversarial erase mechanism is used to process the feature maps of different scale. So that CPN can cover the entire object to generate initial object proposals. Another critical point is that we introduce adversarial multi-label image classification loss to train CPN, which can avoid noise as much as possible. To this end, we explore to score the collection of object proposals through relevance scores. Afterwards, we build pseudo labels of proposals by relevance scores. In FPN, we use selected proposal and pseudo labels to train FPN for refining more precise boundaries.

The proposed method is brief but efficient. And to validate the effectiveness of the proposed SPAL, we conduct a series of object localization experiments using the bounding boxes inferred from the generated localization maps. Detailed evaluations on the PASCAL VOC datasets demonstrate that our SPAL is as competitive as the state-of-the-art methods.

Briefly, we summarize our main contributions into three-fold:

- To the best of our knowledge, we first introduce self-paced learning(SPL) for AE methods, SPL effectively avoids the influence of noise and improves the robustness of adversarial erase learning for WSOL.
- We propose a novel pyramid adversarial erase mechanism. Different from previous AE methods, no additional branches needed, It can efficiently mine different discriminative regions in a weakly supervised manner, which discover integral target regions of objects for localization.
- Compare to existing methods, experimental results show strategies of SPAL are crucial for better performance of WSOL.

II. RELATED WORK

Several recent approaches adopt adversarial erase(AE) to facilitate learning to find integral objects of interest semantic with weak supervision. Singh *et al.* [19] propose Hide-and-Seek (HaS) to divide the input image into grid-like patches and randomly selects the patches to erase. While the random selection is fast and straightforward, it cannot effectively erase the most discriminative part. Meanwhile, Wei *et al.* [20] use adversarial erase (AE) to discover integral object regions by training additional classification networks on images whose discriminative object regions have partially been erased. One shortcoming that can not be ignored in this method is that it must cost more training time and computing resources to train several independent networks for obtaining integral object regions. Consider these issues, Zhang *et al.* [21] propose a novel Adversarial Complementary Learning (ACoL) approach for discovering entire objects of interest via end-to-end weakly supervised training, which can efficiently conduct end-to-end training. However, ACoL still needs additional classifiers. Look for erasing the most discriminative part effectively and efficiently, Choe *et al.* [22] propose an Attention-based Dropout Layer (ADL), a lightweight yet powerful method that utilizes self-attention mechanism to erase the most discriminative part of the object. Despite the difficulties, these efforts laid the foundation for our research.

Inspired by the cognitive science, Bengio *et al.* [26] first initialized the concept of curriculum learning (CL), in which a model is learned by gradually including samples into training from easy to complex. For more explanatory, Kumar *et al.* [23] substantially prompted this learning philosophy by formulating the CL principle as a concise optimization model named self-paced learning (SPL). Recently, several progressive and SPL algorithms in computer vision have been proposed. such as visual tracking [27]–[29], image search [25], object detection [30], [31]. These methods fully illustrate progressive methods that can get better performance at various computer vision tasks through decomposing complex problems into simpler ones. We note that progressive and self-paced learning

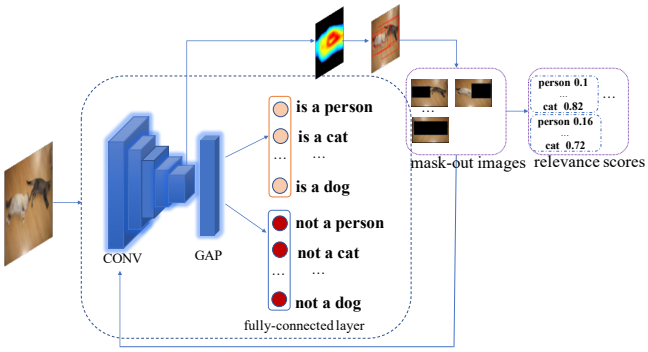

Fig. 2. Coarse Pyramid Network. Include CONV + GAP + classication+ adversarial branch. We note that CONV as convolution layer, GAP as global average pooling layer. Input a image, CONV+GAP use pyramid adversarial erase mechanism (See Fig. 3 for more details) to generate localization maps. then segment localization maps for initial proposal. proposals genearte a set of mask-out images to calculate relevance scores for every proposal.

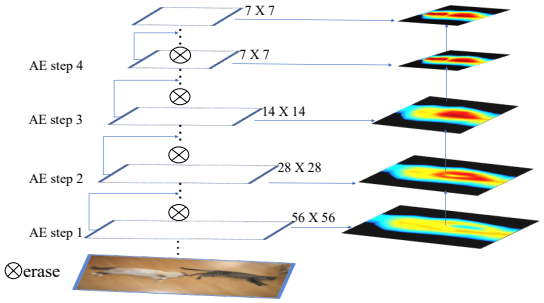

Fig. 3. Pyramid adversarial erase architecture. where AE step $\{1, 2, 3, 4\}$ erase and fuse progressively. From size $\{56 \times 56\}$ to $\{7 \times 7\}$, pyramid shape is built. Differnent scales could learn different discriminative object regions and fuse them through skip connections [32].

also is of particular importance to the weakly supervised object localization problem.

III. THE PROPOSED METHOD

In this section, we present details of our proposed method, Self-Paced Pyramid Adversarial Learning (SPAL). As mentioned earlier, SPAL has three parts. In Sec. III-A, we will introduce details of Coarse Pyramid Network (CPN). In Sec. III-B. We will present more details about Fine Pyramid Network (FPN). Finally, in Sec. III-C, Self-Paced Learning Mechanism will give more training details.

A. Coarse Pyramid Network

The goal of Coarse Pyramid Network (CPN) is to cover the entire object to generate initial object proposals. We use ResNet50 [32] as our network architecture and introduce pyramid adversarial erase mechanism to process the feature maps of different scales, CPN erases its discovered regions from different scales of feature maps and fuses them in some way. Progressive erasing and fusing help network to discover complete objects from localization maps. Meanwhile, unlike the problem in simple classification, CPN is trained as the multi-label image classification network to increase the specificity of the object categories of interest. In this case, we present adversarial multi-label classification loss to fit our task. Finally, we use relevance scores as the pseudo label to match class-specific proposals, whose ways remove substantial noise and potential confusion from similar objects. Then initial proposals and pseudo labels as data are inputted to fine pyramid network.

Pyramid adversarial erase mechanism. Motivated by ResNet [32] and FPN [33], a deep ConvNet computes a feature hierarchy layer by layer. Semantics from low to high levels build a pyramid shape. Different from previous AE methods, we propose pyramid adversarial erase. we erase feature maps from $\{56 \times 56,\ 28 \times 28,\ 14 \times 14,\ 7 \times 7\ \}$ size. There are often many layers producing output maps of the same size. we say these layers for a AE step. So we say our pyramid adversarial erase for AE step $\{1, 2, 3, 4\}$. As shown in Fig. 3, in every AE step, let M_i^{first} denotes the first layer of AE step i, M_i^{last} denotes the last layer of AE step i, Note that we normalize both maps to the range [0,1] and define them as $\widetilde{M}_i^{first}$, $\widetilde{M}_i^{last}$, the most discriminative region is identified as the set of pixels whose value is larger than the given threshold δ in $\widetilde{M}_i^{first}$. Afterwards, we erase the discriminative regions in $\widetilde{M}_i^{first}$through replacing the pixel values of the identified discriminative regions by zeros. Finally, we fuse them through $\max(\widetilde{M}_i^{first}, \widetilde{M}_i^{last})$ as M_i^{fuse}.

When testing an image, we get the fused localization maps and resize them to the same size with the original images by linear interpolation. For producing object bounding boxes, we segment the foreground and background by the fixed threshold. Afterwards, we seek the tight bounding boxes covering the largest connected area in the foreground pixels, which can generate initial proposals.

Adversarial multi-label classification loss. Real classification problems always assume only one single object exists per image. It is full biased and not objective. In this case, we describe multiple objects as multi-label image classification problems. However, conventional multi-label classification problems usually regard every label as an independent distribution. It ignores the contact between objects and so that inaccurate distribution will seriously affect the accuracy. Unlike traditional loss function. We introduce an adversarial multi-label classification loss to handle these problems.

We assume dataset has K categories and a collection of N training images. Training image set can be defined as $\ell = \{(I^{(1)}, L^1), ..., (I^{(N)}, L^{(N)})\}$, where I is the image data and L is the corresponding label. Formally, $L = [l_1, l_2, ..., l_K]^T$ is the K-dimensional label vector. Each l can be 1 or 0 indicating whether at least one specific object instance is present in the image. Adversarial multi-label classification loss can be described by the following steps.

As shown in Fig. 2, first, a normal fully-connected layer is added, where label $L = [l_1, l_2, ..., l_K]^T$. Beyond that, we add a new fully-connected layer to K as adversarial branch, a new

label $L_a = [l_1^a, l_2^a, ..., l_K^a]^T$, where

$$l_i^a = \begin{cases} 1 & l_i = 0 \\ 0 & l_i = 1 \end{cases} (i \in (1, 2, ..., K)) \quad (1)$$

Specifically, each L represents whether the image contains the corresponding object. Similarly, each L_a represents whether the image does not contain the corresponding object. Next, we describe how to calculate final Loss. Input Image I, we get two K dimensional output $P(I)$ and $P_a(I)$ (adversarial branch). Computing the probabilities uses sigmoid function, in $P(I)$, $p^i(I)$ represents the probability that the image contains at least one object instance of i-th category. In $P_a(I)$, $p_a^i(I)$ shows the probability that the image does not contain objects of i-th category. We can define Loss of one image for category i as $Loss_i$,

$$Loss_i = -(l_i^a log p_a^i(I) + l_i log p^i(I)) \quad (2)$$

So the final Loss can be obtained by summing up all the training samples and losses for all the categories and averaging them.

$$Loss = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{K} Loss_i(I^{(n)}) \quad (3)$$

Compared with alternative loss functions(*eg.*, binary logistic regression loss, and label-wise cross-entropy loss), our proposed adversarial multi-label classification loss fully considers the relationship between different objects through adversarial branch. The network thus can remove substantial noise and potential confusion from different objects.

Relevance scores. Coarse Pyramid Network (CPN) can generate initial proposals. These proposals can cover full extent of objects. But they still can't ensure precise object boundaries. class-independent object proposals are difficult to refine. However, class-specific proposals are easy to improve. Considering these, in this case, we thus calculate relevance scores for class-specific proposals. Relevance scores define how relevant each proposal is to each class. Intuitively, a proposal has the most significant relevance score concerning i-th class. The proposal can consider the i-th class as pseudo label to match a specific proposal. We note that if the mask-out image by a proposal causes a significant drop in classification score for the i-th class, the proposal can be considered essential for the i-th class. So we exploit the degree of descent to define relevance scores

Formally, we denote relevance scores as K-dimensional vector $S = \{s_1, s_2, ..., s_K\}$ (assuming dataset has K categories), a set of initital proposals as P, input image I. Without loss of generality, we take $\widetilde{p}$ in P as an example. First, I_{mask} as mask-out image for $\widetilde{p}$, $I(\widetilde{p})$ as the output of original image (the network has two output,But $I(\widetilde{p})$ shows the probabilities of containing a object), $I_{mask}(\widetilde{p})$ as the output of mask-out image. Then we can calculate relevance scores for $\widetilde{p}$ as

$$S = SoftMax(\frac{|I(\widetilde{p}) - I_{mask}(\widetilde{p})|}{I(\widetilde{p})}) \quad (4)$$

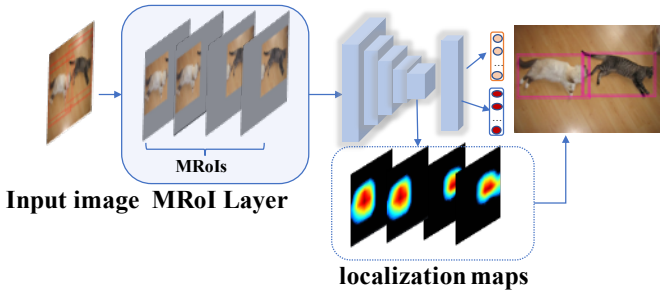

Fig. 4. Fine Pyramid Network. Input an image and a set of object proposals (proposals belong to class-specific). MRoI layer processes each proposal to generate a set of MRoIs (MRoIs reserve proposal's information but hide others). MRoIs are processed by pyramid adversarial erase mechanism to produce localization map for every proposal.

Here, SoftMax function ensures $s_1 + s_2 + ... + s_K = 1$, $I(\widetilde{p})$ and $I(\widetilde{p})$ all are K-dimensional vector, So All operations are element-wise.

B. Fine Pyramid Network

The purpose of Coarse Pyramid Network (CPN) is to mine a set of high confident but coarse class-specific proposals. Further, Fine Pyramid Network(FPN) can refine more precise object boundaries. Fig. 4 illustrates the FPN architecture. Although it adopts the same convolution architecture with CPN(we use ResNet50), the inputs of FPN are an entire image and a set of object proposals. These proposals use the largest relevance score as a pseudo label for training. Afterwards, for each object proposal, a mask region of interest (MRoI) layer keeps the proposal unchanged, for other regions MRoI uses mask-out strategy.

In other words, for each proposal, we reserve the proposal's information but hide others, these mask-out images (called MRoIs) constitute mini-batch data. MRoIs have same size with input image but only the knowledge of corresponding proposal. Next, MRoIs are processed by pyramid adversarial erase mechanism, So FPN will produce localization map for every proposal. Finally, These localization maps will segment more precise object boundaries.

C. Self-Paced Learning Mechanism

The key point of self-paced learning mechanism is to rank our trainval sets from easy to hard. Then the training process should use progressive pyramid adversarial learning to learn weakly supervised object localization from easy object samples as far as possible. Our training images include ILSVRC 2012 [39] and Pascal VOC 2007 [40] trainval set. Consider the difference, and we extract ILSVRC 2012 20-classes corresponding to Pascal VOC classes[1] as training

[1]Note that the ILSVRC↔VOC class mapping is: bicycle↔bicycle, airplane↔aeroplane, bird↔bird, watercraft↔boat, wine bottle↔bottle, bus↔bus, tv or monitor↔tvmonitor, sheep↔sheep, sofa↔sofa, car↔car, domestic cat↔cat, chair↔chair, flower pot↔pottedplant, train↔train, cattle↔cow, table↔diningtable, dog↔dog, horse↔horse, motorcycle↔motorbike, person↔person.

Methods				aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	Avg
LCL [34]				80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	47.7
WSDDN [6]				65.1	63.4	59.7	45.9	38.5	69.4	77.0	50.7	30.1	68.8	34.0	37.3	61.0	82.9	25.1	42.9	79.2	59.4	68.2	64.1	56.1
TS2C [35]				84.2	74.1	61.3	52.1	32.1	76.7	82.9	66.6	42.3	70.6	39.5	57.0	61.2	88.4	9.3	54.6	72.2	60.0	65.0	70.3	61.0
C-WSL [36]				87.5	81.6	65.5	52.2	37.4	83.8	87.9	57.6	50.3	80.8	44.9	44.4	65.6	92.8	14.9	61.2	83.5	68.5	77.6	83.5	66.1
Ranking GAN [37]				85.5	75.0	66.9	47.5	43.6	67.4	83.6	61.7	36.8	75.1	29.8	55.9	70.4	80.6	29.0	52.9	71.0	31.2	66.9	58.1	59.4
SDCN [38]				85.8	83.1	56.2	58.5	44.7	80.2	85.0	77.9	29.6	78.8	53.6	74.2	73.1	88.4	18.2	57.5	74.2	60.8	76.1	79.2	66.8
PAE	AMCL	FPN	SPL																					
✓				60.4	40.0	54.6	26.2	30.8	39.6	56.9	52.8	20.3	52.0	30.5	50.4	45.2	60.6	26.7	36.5	29.2	36.6	45.6	48.0	42.1
✓	✓			62.6	60.4	63.5	36.0	32.0	66.8	64.8	56.9	36.8	66.0	32.2	57.6	42.6	70.3	29.0	39.6	32.8	42.5	52.8	52.3	49.9
✓	✓	✓		80.6	68.6	64.6	43.6	38.8	72.2	72.6	63.1	43.6	78.1	36.7	63.8	63.7	78.3	30.6	48.2	50.6	52.7	66.8	59.6	58.8
✓	✓	✓	✓	**83.7**	**70.6**	**68.0**	**56.5**	**42.6**	**74.8**	**76.3**	**66.7**	**46.8**	**80.2**	**42.8**	**68.2**	**66.9**	**81.5**	**33.8**	**50.3**	**56.8**	**58.3**	**70.3**	**62.6**	**62.9**

TABLE I

COMPARISONS OF SPAL WITH THE STATE-OF-THE-ART IN TERMS OF CORLOC (%) ON THE VOC 2007 TRAINVAL SET. OUR BEST NUMBER IS BOLDED, THE BEST NUMBER IN ALL METHODS IS MARKED IN RED.

Methods				aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
LCL [34]				48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
WSDDN [6]				46.4	58.3	35.5	25.9	14.0	66.7	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.2
TS2C [35]				59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
C-WSL [36]				62.9	68.3	52.9	25.8	16.5	71.1	69.5	48.2	26	58.6	44.5	28.2	49.6	66.4	10.2	26.4	55.3	59.9	61.6	62.2	48.2
Ranking GAN [37]				52.4	63.8	41.8	35.1	22.9	72.3	61.1	44.7	13.9	48.6	32.9	46.1	50.7	66.3	18.5	27.0	49.7	56.9	64.8	58.6	46.4
SDCN [38]				59.8	67.1	32.0	34.7	22.8	67.1	63.8	67.9	22.5	48.9	47.8	60.5	51.7	65.2	11.8	20.6	42.1	54.7	60.8	64.3	48.3
PAE	AMCL	FPN	SPL																					
✓				48.8	28.9	36.3	18.2	16.2	45.3	66.6	60.5	18.8	26.7	20.4	47.6	32.4	52.5	13.5	26.5	23.4	19.6	33.8	56.8	34.6
✓	✓			52.4	47.6	43.6	22.8	19.6	52.9	62.2	58.4	21.6	52.2	23.6	36.8	52.3	61.8	16.8	26.2	40.7	16.9	56.3	42.2	40.3
✓	✓	✓		–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	40.3
✓	✓	✓	✓	**62.6**	**58.3**	**50.2**	**36.8**	**26.0**	**68.9**	**70.2**	**56.5**	**28.8**	**62.4**	**38.2**	**36.5**	**56.3**	**65.5**	**20.6**	**26.4**	**48.8**	**30.9**	**62.6**	**49.8**	**47.8**

TABLE II

COMPARISONS OF SPAL WITH THE STATE-OF-THE-ART IN TERMS OF AP (%) ON THE VOC 2007 TEST SET. OUR BEST NUMBER IS BOLDED, THE BEST NUMBER IN ALL METHODS IS MARKED IN RED.

images. For selecting samples from easy to hard, we design a rank protocol through estimating the difficulty of visual search in an image.

Accurately, we describe the difficulty of an image using all kinds of image properties such as irrelevant clutter, their scale and position, their class type, occlusions and other types of noise. Without loss of generality, we use same evaluation criteria in [41], include number of annotated objects; mean area covered by objects normalized by the image size; number of different classes; number of objects marked as truncated; number of objects marked as occluded; number of objects marked as difficult. These criteria use Kendall's τ rank correlation coefficient [42]. Kendall's τ is a correlation measure for ordinal data based on the discrepancy between the number of discordant pairs and the number of concordant pairs among two variables, divided by the total number of pairs. As an effective measure, it can indeed be consistently measured in image difficulty. For more details please refer to [41].

Note that ILSVRC 2012 only contains an object, But Pascal VOC 2007 has multiple objects. So they are ranked separately. We firstly train ILSVRC 2012 for a single object and then learn various objects in Pascal VOC 2007. The process always is progressive.

IV. EXPERIMENTS

A. Experiment setup

Datasets and evaluation metrics. We train the network of SPAL on ILSVRC 2012 [39] and Pascal VOC 2007 [40] trainval set(Note that Pascal VOC 2007 trainval set consists of train and val splits. And ILSVRC 2012 only is chosen for 20-classes corresponding to Pascal VOC classes). We evaluate the localization performance on Pascal VOC 2007 dataset. For fair comparison, we apply general metric correct localization (called CorLoc) for measuring the performance of WSOL. Corloc calculates the percentage of the images whose bounding boxes have at least 50% IoU with the ground truth. Also, Average Precision(AP) in the test set is used to measure the performance of WSOL. Finally, we also visualize some state-of-the-arts localization performance on Pascal VOC 2007 test set.

Implementation details. We use ResNet50 [32] as our base model. Pyramid adversarial learning erases and fuses feature maps from $\{56 \times 56,\ 28 \times 28,\ 14 \times 14,\ 7 \times 7\ \}$ size. We test erase threshold δ from 0.5 to 0.9(we use 0.6 for final results). Two subnetworks CPN and FPN use the same resnet50 architecture. For the details in Sec. III-A, We add a GAP layer and two fully-connected layers on the top of the convolutional layers. We randomly crop the input image for 224×224. For the details in Sec. III-C, we train our train set on progressively.we don't use any pre-trained weights. First, we train ranked ILSVRC 2012 trainval set from scratch, then continue to train on ranked VOC 2007. All training and evaluation use Tesla P40 GPU with 24GB memory. We implement all codes using Keras and TensorFlow. The details of codes will be available soon.

B. Comparisons with the state-of-the-arts

We compare our SPAL with several state-of-the-art methods for weakly supervised object localization, including LCL [34], WSDDN [6], TS2C [35], C-WSL [36], Ranking GAN [37], SDCN [38]. For clear comparison, we use some abbreviations for representing each step of SPAL:

- PAE: Using pyramid adversarial erase to efficiently mine different discriminative regions

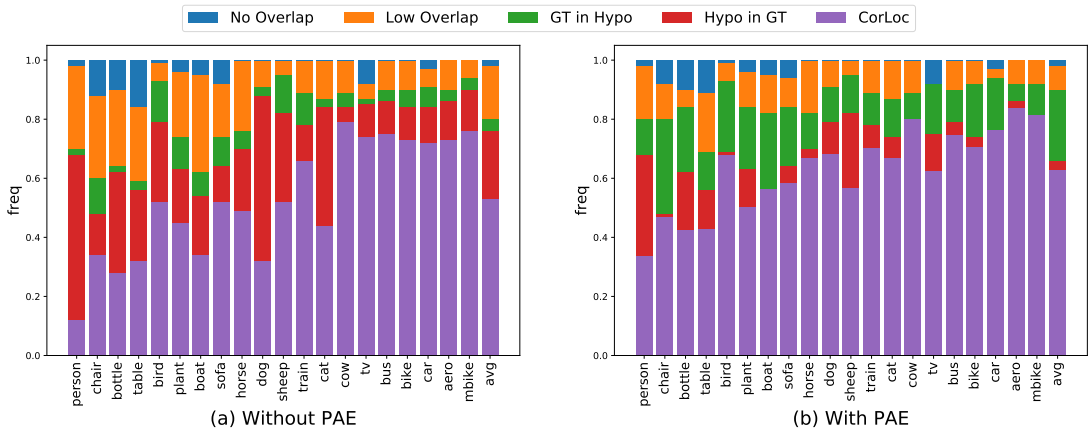

Fig. 5. The frequencies of error modes for five cases(include average result), Without PAE and With PAE are tested on the PASCAL VOC 2007 trainval set.

- AMCL: Using adversarial multi-label classification loss to train network.
- FPN: Using Fine Pyramid Network to refine object localization
- SPL: Using Self-Paced Learning Mechanism

Table I and Table II show all Comparisons of our SPAL with the state-of-the-art in terms of CorLoc and AP. our best results achieve an average of 62.9% CorLoc on VOC 2007 trainval set. Especially, our method achieve significant improvements than the state-of-the-arts on “bird”, “person”. Compared to the most competitive methods [6], [34]–[38], Our average Corloc is second only to [36] [38]. But it is worth noting that [36] [38] all use other image-level supervision not only label. [36] makes use of per-class object count supervision to identify the correct high-scoring object bounding boxes from a set of object proposals. [38] adds a segmentation branch, and uses a dynamic collaboration loop to complement both detection and segmentation for more accurate predictions. Different from these, our method only uses limited label information. But SPAL outperforms most state-of-the-art [6], [34], [37].

As shown in Table II, the average precision (AP) performance on the VOC 2007 test set also performs promising against the state-of-the-art methods. The huge improvement on ”boat“, ”bottle“, ”car“, ”chair“, ”cow“, ”horse“, ”mbike“, ”person“ shows the effectiveness of SPAL. And the AP for every category all is close to the most advanced level. The best mAP of SDCN [38] is 48.3%, which exploits the segmentation cues. Our mAP arrives 47.8%. But we never use other supervision information. SPAL learns to detect more complete object boundaries throgh finite restrictions. the performance of SPAL has approached or even exceeded most MIL methods [9]–[11].

We also evaluate contributions of each step to the whole. in term of Corloc, The result of using PAE alone is worse. Because a substantial amount of noise in training causes ambiguities for learning in a robust manner. With the use of AMCL, the average Corloc gets a huge boost (from 42.1% to 49.9%). FPN also should be concerned. Object proposals can be refined to locate more precise coordinate through FPN. The results of Table I fully illustrate this argument. Corloc increases by 8% (from 49.9% to 58.8%). SPL has proven to be an effective learning strategy, we achieve a large improvement from 58.8% to 62.9% in terms of CorLoc, from 40.3% to 47.8% in terms of AP. Overall, SPAL significantly improves the performance of WSOL on CorLoc and AP, benefitting from progressive pyramid adversarial Learning. However, there are still several classes on which the performance is hardly improved as shown in Table I and Table II, *e.g.*“bottle”, “person”, ”plant“ . A main reason is the large portion of occluded and overlapped samples for these classes, which leads to incomplete or connected responses on the localization map. in addition, It is also one of the main reasons for the difficulty of localization that the object is too dense and object scale varies greatly. There’s no doubt these problems lead to more room for further improvements.

C. Ablation Study

We also study the influence of pyramid adversarial erase mechanism(PAE). Specifically, we focus on five types of localization metrics with PAE and without PAE, which include: (1) CorLoc(correct localization, whose overlap with the grouth-truth>=50%); 2) GT in Hypo(the ground-truth completely inside the hypothesis); (3) Hypo in GT(the hypothesis completely inside the ground-truth); (4) Low Overlap; (5)No Overlap.

The PAE’s performance is compared to without PAE in five cases. As shown in Fig. 5a and 5b, Different frequencies of five cases are shown. The main errors without PAE lies in the low overlap between the hypothesis and the ground-truth and the hypothesis completely inside the ground-truth. Because it always tend to focus only on the most discriminative features to decide final classification results. It is inevitable to locate object parts with real objects frequently. yet, The corresponding result for the PAE increases obviously. Especially for CorLoc, all classes have varying degrees of improvement. A typical error mode for the hypothesis inside the ground-truth decreases and ground-truth inside the hypothesis increases.

It indicate PAE can discover more integral object regions effectively.

D. Investigation of Hyper-parameters

The influences of Hyper-parameter about PAE are shown in Fig. 6, we keep other components same. Then we set different erase threshold $\delta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ and different erase times, range from $1 \sim 10$ to test. we obtain the best Corloc when the erase threshold $\delta = 0.6$ and $times = 4$. The results show the performance will be worse when the threshhold is larger and too many erase time will lead to CorLoc to plummet. We can also conclude too small threshold may bring background noises and too many erase times may erase important details. So it is imperative to choose the well-designed threshold and times in PAE.

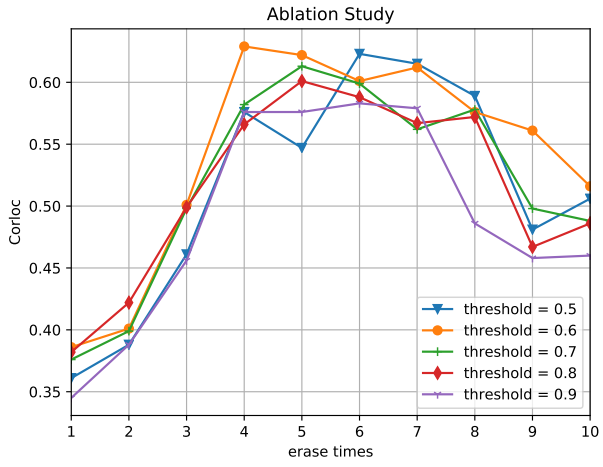

Fig. 6. The curves of the average CorLoc varying with different erase threshold and times on VOC 2007 trainval set.

V. CONCLUSION

In this paper, we proposed SPAL, a novel self-paced pyramid adversarial learning method to improve the localization accuracy for WSOL. Our method incorporates progressive adversarial learning into two subnetworks, including Coarse Pyramid Network(CPN) and Fine Pyramid Network (FPN). Combined with Self-Paced Learning Mechanism, They can progressively improve the localization accuracy. In CPN, pyramid adversarial erase can consistently mine different object parts and discover integral but coarse object regions. Simultaneously, adversarial multi-label classification loss helps the training more robust. These steps can generate coarse initial object proposals. In FPN, we develop a mask region of interest (MRoI) layer to process a set of initial proposals. MRoI can generate a collection of mask-out images that reserve proposal's information but hide others. Finally, every proposal can be refined more precise object boundaries. Experiments conducted on PASCAL VOC benchmarks demonstrate the effectiveness of SPAL.

ACKNOWLEDGMENT

This work is partially supported by State Grid Corporation Science and Technology Project–Research on Key Technologies of Intelligent Preprocessing and Visual Perception of Power Transmission and Transformation Equipment (NO.SGMDDK00SPJS1800022)

REFERENCES

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

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[4] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[5] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 914–922.

[6] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2846–2854.

[7] Z. Shi, Y. Yang, T. M. Hospedales, and T. Xiang, "Weakly-supervised image annotation and segmentation with objects and attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2525–2538, 2016.

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[9] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2016.

[10] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1081–1089.

[11] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-mil: Continuation multiple instance learning for weakly supervised object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2199–2208.

[12] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[13] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*. Springer, 2014, pp. 391–405.

[14] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell, "On learning to localize objects with minimal supervision," *arXiv preprint arXiv:1403.1024*, 2014.

[15] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1297–1306.

[16] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2843–2851.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[18] D. Kim, D. Cho, D. Yoo, and I. So Kweon, "Two-phase learning for weakly supervised object localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3534–3543.

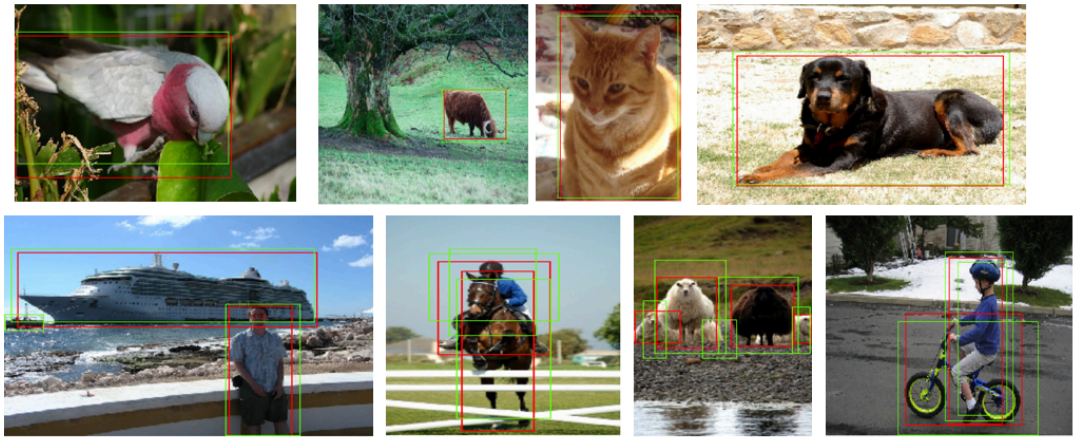

Fig. 7. Sample localization performance on VOC 2007 test set. we use bounding boxes to show results(ground-truth bounding boxes are in red and the predicted are in green).

[19] K. K. Singh and Y. J. 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)*. IEEE, 2017, pp. 3544–3553.

[20] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. 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*, 2017, pp. 1568–1576.

[21] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1325–1334.

[22] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2219–2228.

[23] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems*, 2010, pp. 1189–1197.

[24] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[25] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 547–556.

[26] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[27] J. S. Supancic and D. Ramanan, "Self-paced learning for long-term tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2379–2386.

[28] C. Kerdvibulvech, "Hand tracking by extending distance transform and hand model in real-time," *Pattern Recognition and Image Analysis*, vol. 25, no. 3, pp. 437–441, 2015.

[29] T. Siriborvornratanakul, "An automatic road distress visual inspection system using an onboard in-car camera," *Advances in Multimedia*, vol. 2018, 2018.

[30] Y. J. Lee and K. Grauman, "Learning the easy things first: Self-paced visual category discovery," in *CVPR 2011*. IEEE, 2011, pp. 1721–1728.

[31] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, "Self paced deep learning for weakly supervised object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 712–725, 2018.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[34] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 431–445.

[35] Y. Wei, Z. Shen, B. Cheng, H. Shi, J. Xiong, J. Feng, and T. Huang, "Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 434–450.

[36] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis, "C-wsl: Count-guided weakly supervised localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 152–168.

[37] A. Diba, V. Sharma, R. Stiefelhagen, and L. Van Gool, "Weakly supervised object discovery by generative adversarial & ranking networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[38] X. Li, M. Kan, S. Shan, and X. Chen, "Weakly supervised object detection with segmentation collaboration," *arXiv preprint arXiv:1904.00551*, 2019.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[40] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[41] R. Tudor Ionescu, B. Alexe, M. Leordeanu, M. Popescu, D. P. Papadopoulos, and V. Ferrari, "How hard can it be? estimating the difficulty of visual search in an image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2157–2166.

[42] M. G. Kendall, "Rank correlation methods." 1948.