Inter-Image Communication for Weakly Supervised Localization

Xiaolin Zhang, Yunchao Wei, and Yi Yang

ReLER, AAII, University of Technology Sydney
{Xiaolin.Zhang-3@student.,Yunchao.Wei@,Yi.Yang@}uts.edu.au

Abstract. Weakly supervised localization aims at finding target object regions using only image-level supervision. However, localization maps extracted from classification networks are often not accurate due to the lack of fine pixel-level supervision. In this paper, we propose to leverage pixel-level similarities across different objects for learning more accurate object locations in a complementary way. Particularly, two kinds of constraints are proposed to prompt the consistency of object features within the same categories. The first constraint is to learn the stochastic feature consistency among discriminative pixels that are randomly sampled from different images within a batch. The discriminative information embedded in one image can be leveraged to benefit its counterpart with inter-image communication. The second constraint is to learn the global consistency of object features throughout the entire dataset. We learn a feature center for each category and realize the global feature consistency by forcing the object features to approach class-specific centers. The global centers are actively updated with the training process. The two constraints can benefit each other to learn consistent pixel-level features within the same categories, and finally improve the quality of localization maps. We conduct extensive experiments on two popular benchmarks, *i.e.*, ILSVRC and CUB-200-2011. Our method achieves the Top-1 localization error rate of 45.17% on the ILSVRC validation set, surpassing the current state-of-the-art method by a large margin. The code is available at https://github.com/xiaomengyc/I2C.

1 Introduction

Deep learning has achieved great success on various tasks, *e.g.*, classification [30,33], detection [13,26], segmentation [3,5,45,18,6] *et al.*. In this paper, we focus on the Weakly Supervised Object Localization (WSOL) problem. Briefly, WSOL tries to locate object regions within given images using only image-level labels as supervision. Currently, the standard practice for this task is to train a convolutional classification network supervised by the given image-level labels. The convolutional operations can preserve relative positions of the input pixels so that activations from high-level layers can roughly indicate the position of target objects. Some previous works [46,42,43] have already explored how to produce object localization maps effectively. These methods obtain localization maps by

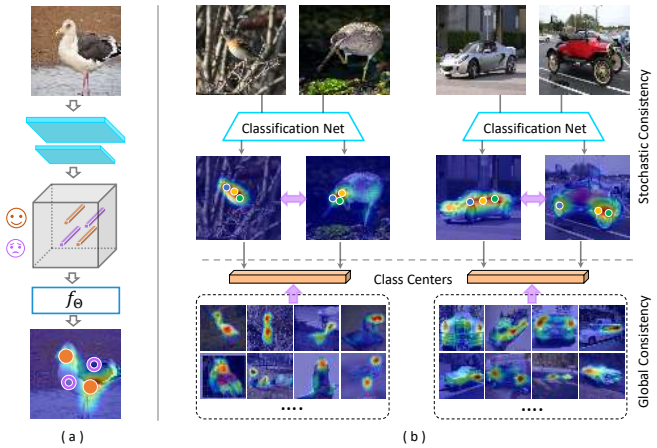


Fig. 1: (a) Convolutional operations preserve the relative pixel positions. Inconsistent response scores of different pixels on the class activation map are essentially caused by the inconsistent learned features. (b) The proposed Stochastic Consistency (SC) and Global Consistency (GC) are to align the object-related feature vectors. *Upper:* Pixel features of images sampled from the same category can reach consistency between the highly confident points within the same mini-batch. *Bottom:* We actively learn class-specific centers and push object features of the same category across different mini-batch towards the centers.

aggregating feature maps with a fully connected layer. Figure 1a shows the general pipeline for generating localization maps. Given an image of a bird, it is firstly fed into some convolution layers to yield the feature maps. These feature maps are then processed by a function f_Θ to get the localization maps of the right class using the methods in CAM [46], ACoL [42], ADL [8], *etc.*. Ideally, the localization maps are expected to highlight all the object regions and depress the background. Unfortunately, only some sparse parts are highlighted and cannot cover the entire target objects, which is a major problem in practice. We try to alleviate this issue based on the following intuition. Theoretically, for a specific output score of a pixel, the input features of the function f_Θ may be various. However, there are two observations in practice, *i.e.*, 1) convolution operations preserve the relative positions between the input and output feature maps [46,42]; 2) features of the same category objects tend to lie in the same clusters and similar input features produce similar output values [47]. Therefore, the root cause for this problem is that the network fails to learn consistent feature representations for pixels belonging to the object of interest. In Figure 1a, the low response scores in the localization maps (pink circles) of the bird are caused by the low-quality or non-discriminative features in the intermediate feature maps, while the high response scores (orange dots) are produced by the decent and discriminative features. Several efforts have been taken to alleviate this issue. MDC [39] attempt to learn consistent features of different parts by en-

larging receptive fields with dilated convolutions [4], but it is prone to draw into background noises. SPG [43] employs an auxiliary loss to enforce the consistency of object features using self-produced pseudo-masks as supervision. Nevertheless, the both methods only consider inter-pixel correlations within an image, and involve sophisticated network modules and many extra computational resources.

In this paper, we alleviate this issue by employing two constraints on the object pixels across images with the cost of negligible extra resources. Different from MDC and SPG, we argue that not only pixels within a object should keep close, but more importantly, object pixels of different images in the same category should also semantically keep consistent in the high-level feature space. In Figure 1b, different parts of the same category objects *e.g.*, heads and bodies of the birds, the wheels and bodies of the vehicles, are highlighted in different images. The corresponding features of these highlighted pixels do not necessarily very close, but we force them to communicate with each other to learn more consistent and robust features, and thereby, produce better localization maps. We propose to realize the pixel-level communications by employing two constraints *i.e.*, Stochastic Consistency (SC) and Global Consistency (GC). The two constraints act as auxiliary loss functions to train classification networks. With the training process, convolution networks are not only looking for discriminative patterns to support the classification purpose, but communicating between different object patterns for learning more consistent and robust features. Consequently, more accurate object regions will be discovered as we desired. Specifically, the proposed SC is to constrain the object features of the same category within a batch. We firstly feed training images through a classification network for obtaining localization maps. Then, we select several confident seed points among the pixels with high scores in the maps. Finally, features of these seed pixels can communicate with each other across images within the same category. We attain the inter-image communication by optimizing the Euclidean distance between high-level features of the seed pixels. In Figure 1b, given two images of the same class, different parts of the target objects are highlighted. Object features can reach consistencies by narrowing the distances between the seed points of different images. Notably, due to the lack of pixel annotations in our task, we firstly ascertain a portion of object pixels for each image. Pixels with high scores in localization maps usually lie in object regions with high confidence [42,43]. Meanwhile, it is also significant to avoid the influence from the background by only considering the confident object regions. Then, inter-image pixel-level communication can be explicitly accomplished.

Due to the limitation of the Stochastic Gradient Descent (SGD) optimization methodology, SC can only keep the semantic consistency within a batch. It cannot guarantee the class-specific consistency of the entire dataset. To tackle this issue, we further introduce the Global Consistency (GC) to augment SC by constraining images across the entire training set. As in Figure 1b, image features are forced towards their class-specific global centers. In detail, the class-specific centers are actively maintained. We adopt a momentum strategy to update the class centers. During each training step, class-specific centers are

updated using the seed features and memory centers. The ultimate goal of GC is to push object features approaching their global centers throughout the training set. GC constrains the object features across images. It is noticeable that the proposed SC constraint does not bring any extra parameters. GC brings a few parameters which is negligible compared to the backbone parameters.

We name the proposed method as Inter-Image Communication (I^2C) model. We conduct extensive experiments on the ILSVRC [9] [27] and CUB-200-2011 [34] datasets. Our main contributions are three-fold:

- We propose to employ inter-image communication of objects in the same category for learning more robust and reliable localization maps under the supervision of image-level annotations.
- We propose two constraints. Stochastic consistency can keep the features of the same semantic within a batch close in the high-level feature space. Global consistency can learn a class-specific center for each category and keep the features of the same category close throughout the training set.
- This work achieves a new state-of-the-art with the localization error rate of Top-1 45.17% on the ILSVRC [9,27] validation set and 44.01% on the CUB-200-2011 [34] test set.

2 Related Work

Weakly supervised segmentation has a strong connection with our WSOL task [38,37,36,21,25,39,16,19,2,1]. It normally and firstly applies similar techniques to obtain pseudo masks, and then trains segmentation networks for predicting accurate masks. DD-Net [28] studies the method of generating pseudo masks from localization maps, and proposes to improve the accuracy by removing noises via self-supervised learning. OAA [19] accumulates localization maps with respect to different training epochs with the optimization process. ICD [12] utilizes an intra-class discriminator to learn better boundaries between classes. SEAM [35] proposes to use consistency regularization on predicted activation maps of various transforms as self-supervision for network learning.

Weakly supervised detection and localization aims to apply an alternative cheaper way by only using image-level supervision [31,23,24,11,48]. Recently, ADL [8] borrows the adversarial erasing idea from ACoL [42] to provide a more neat and powerful approach without bringing much more parameters and computational resources. CutMix [41] also explores the techniques of erasing patches of images. In addition to the erasing operation, CutMix mixes ground truth labels along with image patches. SEM [44] and [7] reformulate the evaluation of WSOL problem. SEM also proposes an enhancement alternative approach to produce high-quality localization maps.

3 Methodology

Figure 2 depicts the framework of the proposed approach. Given a pair of images (I_i^y, I_j^y) sharing the same category y, we firstly forward them to obtain high-level

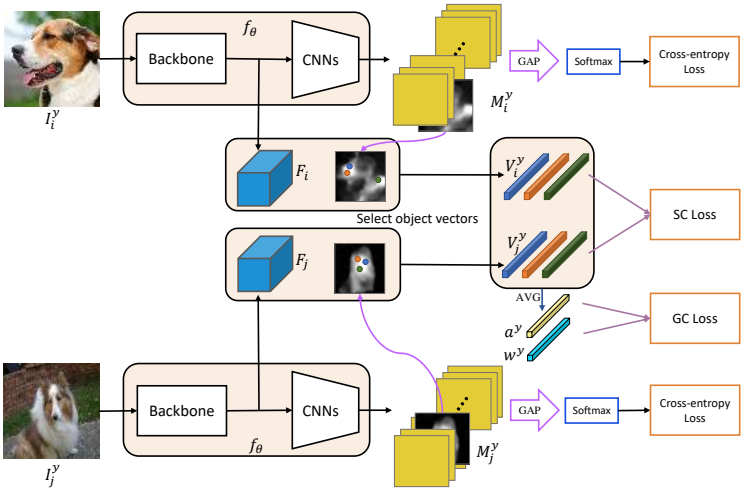


Fig. 2: The structure of the proposed approach. Given a pair of images (I_i^y, I_j^y) of the same category y, the localization maps (M_i^y, M_j^y) can be obtained by forwarding them through the classification network f_θ. Object seed vectors (V_i^y, V_j^y) are extracted from the high-level feature maps (F_i, F_j) according to the confident regions in the maps. Finally, the SC loss is employed on the object seed vectors. Also, the GC loss is employed on the averaged object feature a^y from a batch and the global class-specific center w^y. GAP refers to Global Average Pooling. AVG refers to the average operation.

feature maps (F_i, F_j) and localization maps (M_i^y, M_j^y). Then, we identify the reliable object regions according to the response scores in the produced localization maps. We randomly sample K representative seeds from the object regions. K object vectors of the seeds, denoted as (V_i^y, V_j^y), are extracted from the high-level feature maps (F_i, F_j) for each image according to the spatial localization. Next, we optimize the similarity between different object seed vectors using the Stochastic Consistency (SC) loss across images. Additionally, the Global Consistency (GC) loss is also employed to encourage the object representative vector of each minibatch to approach the global class-specific vector w^y throughout the training set. We will describe the details of constructing the object seed vectors, the SC and GC losses in the following sections.

3.1 Object Seed Vectors

Object seeds serve as the bridge for inter-image communication to narrow the distance between object pixels of the same category. Our target is to consistently high-light object pixels in localization maps, which is equivalent to get consistent high-level object features as the analysis in Section 1. The first obstacle we are facing is the lack of object pixel cues. Some previous works [36,42,43] employ a trade-off strategy of mining the confident object regions according to

the scores in localization maps. We further randomly select some seed pixels as the object representation, and optimize the distance of the seeds across images. Such that intact object regions are high-lighted consistently, and robustness can be improved by exchanging information with different objects.

In detail, given an input image $I \in \mathbb{R}^{3\times W_0\times H_0}$ and its class label $y \in \{0, 1, ..., Y-1\}$, we pick out the localization map from the output of the last convolution layer corresponding to the category y following the simplified method in [42], where W_0 and H_0 denote the width and height of the image, Y is the number of total categories. Suppose the normalized map is $M^y \in [0,1]^{W\times H}$ and the map before normalization is $M^{*y} \in \mathbb{R}^{W\times H}$, where W and H are the width and height of the localization map, we calculate the normalized map by following $M^y_{i,j} = \frac{M^{*y}_{i,j} - \min(M^{*y})}{\max(M^{*y}) - \min(M^{*y})}$, where i and j are the indices of the map. For each image in a batch, we extract $K \in [1, W \times H]$ object-related seeds within the object regions. The object regions can be identified according to the scores in the localization maps. To be specific, pixels whose corresponding values in localization maps are higher than a threshold $\delta \in (0, 1)$ are considered as reliable object regions. We randomly select K pixels from the object regions as the object representation seeds of the image I. We denote the object seed vectors as $V \in \mathbb{R}^{K\times D}$ which is extracted from high-level feature maps, where D is the dimension of the vectors, *i.e.*, the number of the channels in the feature maps. In this way, we finally obtain K vectors of the most discriminative object regions. Next, these vectors are employed to communicate with the other images in the same category.

3.2 Stochastic Consistency

Features of pixels belonging to the same category should be close. Although classification loss can drive networks to find the most discriminative patterns to produce correct classification scores, the learned features of the same objects are not necessarily similar due to the lack of pixel-level supervision. In other words, features between different pixels of objects are not consistent, which deviates from the requirement of highlighting integral object regions. We argue that more accurate localization maps can be obtained by keeping the consistency between features of objects in the same class. Object features of different images can communicate in a complementary approach, and thus the entire object regions can be consistently highlighted. We propose a Stochastic Constraint (SC) to drive the consistency of pixels from different objects of the same category in a minibatch. Given a batch $B = \{(I_i, y_i)|i = 0, 1, ..., N^B\}$ of randomly sampled images, we can find any two images I_i and I_j whose class labels are the same, namely, $y_i = y_j$, where N^B is the batch size. After forwarding them through the network, we obtain the object seed vectors according to Section 3.1. We denote the object seed features as $V_i \in \mathbb{R}^{K\times D}$ and $V_j \in \mathbb{R}^{K\times D}$, respectively. We expect the two images can communicate by explicitly constraining the pixels belonging to the same category. Particularly, we propose a constraint to optimize

the L2-norm between the object features of I_i and I_j according to Eq. (1).

$$L_{sc} = \frac{1}{K}\sum_{k=0}^{K-1}||v_i^k - v_j^k||_2^2, \tag{1}$$

where K is the number of randomly selected seeds, and $v^k \in \mathbb{R}^D$ is the k_{th} row of V. Thus, object features can communicate with each other across images in a batch, and influences from the background is not involved.

3.3 Global Consistency

Deep networks are trained with the SGD based optimization algorithms, *e.g.*, Adam [20], Adagrad [10] and RMSprop [14]. These methods construct mini-batches by randomly sampling images to perform training steps, which means SC can only constrain images within the same minibatch. To overcome this limitation, we propose to learn a global feature center for each category, so that features extracted from a batch can be optimized to approach the class-specific global centers. The object features of each class are hence gradually consistent with the global vectors.

We maintain a memory bank $W \in \mathbb{R}^{Y\times D}$. The y_{th} row of W denoted as w_y is the global center of category y. For each batch, we obtain one representation vector for each category by averaging the object seed vectors sharing the same class labels. Formally, we denote $B^y = \{(I_i, y_i)|y_i = y\}$ and $\{V_k^y|k = 0, 1, ..., K|B^y|\}$ as the class-specific subset and the object seed vectors corresponding to the category y in a batch , where $|B^y|$ is the image number in the subset B^y. We extract one representation vector a^y for category y from every minibatch. Formally, the representation vector a^y of class y in a batch is obtained according to Eq. (2).

$$a^y = \frac{1}{K|B^y|}\sum_{k=0}^{K|B^y|-1} V_k^y. \tag{2}$$

Global vector w_y *w.r.t.* class y in a minibatch is updated during each training step. We do *not* update the memory bank during the backward process. We propose a simple yet effective updating procedure for learning the memory matrix. For the class y, its global representation w_y is as in Eq. (3).

$$w^y = (1 - \eta_{t^y}^y)w^y + \eta_{t^y}^y a^y, 0 < \eta_{t^y}^y < 1, \tag{3}$$

where $\eta_{t^c}^c$ is the updating rate of the class y at step t^y. We adopt a class-specific updating rate as given in Eq. (4) for learning the global centers.

$$\eta_{t^y}^y = e^{-\alpha t^y}, 0 < \alpha < 1, \tag{4}$$

where t^y counts the updating steps of class y, and $\eta_{t^y}^y$ is the update rate of the class y at learning step t^y. Note the updating step t^y is also class-specific and maintained throughout the training process, because the updated counters of

different categories are not the same due to the random sampling procedure of each batch. α is a hyper-parameter for decaying the learning rate of the global representation w^y. Thereby, w^y will be gradually stable with the training process.

The features of each batch can be optimized to approach the global representation as in Eq (5).

$$L_{gc} = \frac{1}{|Y^B|} \sum_{y \in Y^B} ||a^y - w^y||_2^2, \tag{5}$$

where Y^B is the label set of the mini-batch.

We further apply a typical cross-entropy loss for classification and denote it as L_{cls}. In total, the optimization of our approach is a joint training process with three items following Eq. (6)

$$L = L_{cls} + \lambda_1 L_{sc} + \lambda_2 L_{gc}, \tag{6}$$

where λ_1 and λ_2 are for trading-off the three loss items.

In *testing*, we simply feed testing images into the network, and obtain the localization maps *w.r.t.* the predicted class labels. The extracted localization map is then normalized and resized to the original size of the input image by the bilinear interpolation. Following the baseline methods [42,43,46], we leverage the same strategy in CAM [46] for generating the bounding boxes of the target objects. Specifically, we firstly binarize the localization maps by a threshold for separating the object regions from the background. Afterward, we draw tight bounding boxes that can cover the largest connected area of the foreground pixels. The thresholds for splitting the object regions are adjusted to the optimal values. For more details, please refer to [46].

4 Experiments

4.1 Experiment setup

Datasets We evaluate the proposed method following the previous methods, *e.g.*, CAM [46], ACoL [42], SPG [43] and ADL [8]. Two datasets, *i.e.*, ILSVRC [9,27] and CUB-200-2011 [34] are applied to train classification networks for a fair comparison with the baselines. ILSVRC is a widely recognized large-scale classification dataset including 1.2 million images of 1,000 categories for training and 50,000 images for validation. Images in both the training and validation sets are well annotated with image categories and tight bounding boxes of objects. CUB-200-2011 [34] includes 11,788 images from 200 different species of birds. It is splitted into the training set of 5,994 images and the testing set of 5,794 images. Similarly, all images are annotated with class labels and tight bounding boxes. In our experiments, the proposed method is learned with the training sets using only image-level labels as supervision. The annotated bounding boxes on the validation set of ILSVRC and the testing set of CUB-200-2011 are employed for the evaluation.

Evaluation metrics We apply the recommended metric in [27] to evaluate the localization maps following the baseline algorithms [46,42,43,8]. To be specific, it calculates the percentage of the images that satisfy the following two conditions. First, the predicted classification labels match the ground-truth categories. Second, the predicted bounding boxes have over 50% IoU with at least one of the ground-truth boxes. In order to have a more explicit and pure comparison in localization ability, we calculate and compare the localization accuracy given ground-truth labels. We denote these results as the Gt-known localization accuracy. This Gt-known localization accuracy removes the influence of classification results and is much fairer in comparing the localization ability.

Implementation details We implement the proposed algorithm based on three popular backbone networks, *e.g.*VGG16 [30], ResNet50 [15] and InceptionV3 [33]. We make the same modifications on the networks to obtain localization maps following ACoL [42] and SPG [43]. We use the simplified method in ACoL [42] to produce localization maps, while the erasing branch is not applied. We enable the proposed constraints after finetuning the parameters for a few epoches to obtain a good initialization of the newly added parameters. In the ablation experiments, we compare a plain version network without SC and GC for comparison, named InceptionV3-plain. In order to assure that each batch contains images of the same category, we randomly draw 40 categories and then randomly sample two images from the selected categories, constructing the image batch size of 80. We apply multiple hyper-parameters *i.e.*, λ_1, λ_2, δ, α and K, and conduct extensive experiments for studying the impact of these variables. The global memory centers are randomly initialized. We set $\delta = 0.7$ following SPG [43].

4.2 Comparison with the state-of-the-arts

ILSVRC Table 1 compares the proposed method with various baselines on the ILSVRC validation set. I^2C surpasses all the baseline methods in Top-1 and Gt-known localization error. I^2C based on ResNet50 achieves the lowest error rate of 45.17%, significantly surpassing ADL by a large margin of 6.30%. It is notable ADL uses a stronger backbone network, *i.e.*, ResNet50-SE. I^2C based on InceptionV3 significantly surpasses the current state-of-the-art method, ADL, by 4.40%. The results based on VGG16 are also notably better than the currently reported results by 1.58% in Top-1. Additionally, the classification errors of our I^2C method are competitive with all the counterparts based on the same backbones. As demonstrated in Section 4.1, the Gt-known localization metric can reflect the pure localization ability regardless the affect from classification results. The proposed method achieves the best localization ability among the existing methods. The lowest Gt-known localization error of I^2C is 31.50%, outperforming the SPG approach by 3.81%. The I^2C model based on VGG16 achieves 36.10% in Gt-known localization error, which is also better than the counterparts.

CUB We implement the proposed method on the CUB-200-2011 dataset following the baseline methods, *e.g.*, ACoL [42], SPG [43] and CAM [46]. Incep-

Methods	Backbone	Loc Err.			Cls Err.	
		Top-1	Top-5	Gt-known	Top-1	Top-5
CAM [46]	AlexNet [22]	67.19	52.16	45.01	42.6	19.5
CAM [46]	GoogLeNet [32]	56.40	43.00	41.34	31.9	11.3
HaS-32 [31]	GoogLeNet [32]	54.53	-	39.43	32.5	-
ACoL [42]	GoogLeNet [32]	53.28	42.58	-	29.0	11.8
DANet [40]	GoogLeNet [30]	52.47	41.72	-	27.5	8.6
Backprop [29]	VGG16 [30]	61.12	51.46	-	-	
CAM [46]	VGG16 [30]	57.20	45.14	-	31.2	11.4
CutMix [41]	VGG16 [30]	56.55	-	-	-	-
ADL [8]	VGG16 [30]	55.08	-	-	32.2	-
ACoL [42]	VGG16 [30]	54.17	**40.57**	37.04	32.5	12.0
I^2C-Ours	VGG16 [30]	**52.59**	41.49	**36.10**	30.6	10.7
CAM [46]	ResNet50-SE [15,17]	53.81	-	-	23.44	-
CutMix [41]	ResNet50 [15]	52.75	-	-	21.4	5.92
ADL [8]	ResNet50-SE [15,17]	51.47	-	-	24.15	-
I^2C-Ours	ResNet50 [15]	**45.17**	**35.40**	**31.50**	23.3	6.9
CAM [46]	InceptionV3 [33]	53.71	41.81	37.32	26.7	8.2
SPG [43]	InceptionV3 [33]	51.40	40.00	35.31	30.3	9.9
ADL [8]	InceptionV3 [33]	51.29	-	-	27.2	-
I^2C-Ours	InceptionV3 [33]	**46.89**	**35.87**	**31.50**	26.7	8.4

Table 1: Comparison of the localization error rate on ILSVRC validation set. Classification error rates is also presented for reference.

tionV3 is chosen as the backbone network following [43,8]. Table 2 compares our method with the baselines. I^2C surpasses all the baseline methods on both Top-1 and Top-5 metrics, yielding the accuracies of Top-1 44.01% and Top-5 31.66% and significantly outperforming the current reported state-of-the-art errors by 2.95% in Top-1 and 6.38% in Top-5.

Visualization In Figure 3, we compare localization maps and the corresponding bounding boxes between the proposed method and ACoL [42]. We can observe that localization maps produced by I^2C can accurately highlight the object regions of interest. The proposed method can not only reduce the noises from the irrelevant objects or stuff in the background regions, but find more object-related regions accurately. As a result, it is easy to draw bounding boxes which can better match the target object regions as shown in Figure 3.

Methods	Top-1	Top-5
CAM [46]	56.33	46.47
ACoL [42]	54.08	43.49
SPG [43]	53.36	42.28
DANet [40]	47.48	38.04
ADL [8]	46.96	-
I^2C-Ours	**44.01**	**31.66**

Table 2: Localization error on the CUB-200-2011 test set. I^2C significantly surpasses all the baselines.

In summary, our method can successfully obtain better localization maps and accurate bounding boxes. I^2C surpasses

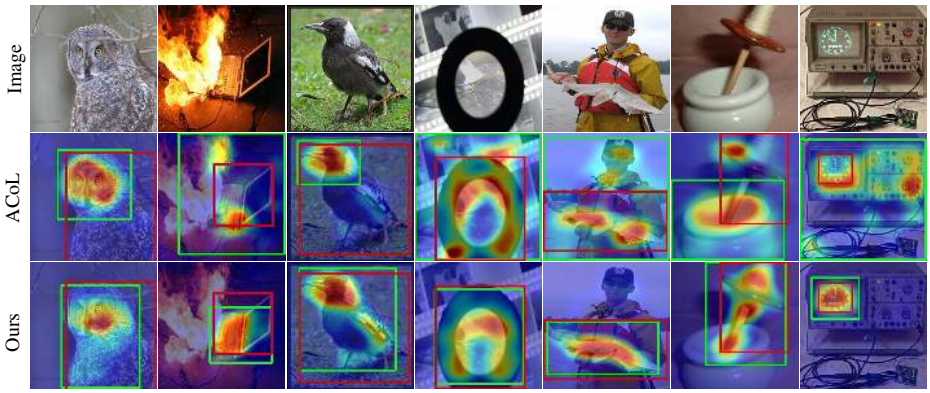


Fig. 3: Comparison of the predicted bounding boxes with ACoL [42]. Our method obtains better localization maps and better bounding boxes. *The predicted boxes are in green and the ground-truth boxes are in red.*

all the baseline methods on both ILSVRC and CUB-200-2011. We believe that there are two aspects accounting for the success of I^2C. First, the proposed two constraints can retain pixel-level consistent between object features of different images, so that images can benefit from each other to obtain better localization maps. Second, I^2 can not only increase the localization accuracy, but can get competitive classification results.

4.3 Ablation study

In this section, we analyse the insights and effectiveness of the different modules in the proposed method. We conduct the ablation experiments based on InceptionV3. The experiments are on the most convincing large-scale dataset *i.e.*, ILSVRC, with the input resolution of 320 by 320, unless specifically specified. When testing the classification and localization results, we do *NOT* apply the ten-crop operation to enhance classification accuracies for convenience.

DataSet	Resolution	Gt-known
ILSVRC	224×224	31.50
	320×320	31.04
CUB	224×224	27.40
	320×320	22.99

Table 3: Gt-known localization error with different input resolutions. Enlarging input resolution slightly improve the localization on ILSVRC, while dramatically boost the localization performance on CUB.

Does input resolution affect the localization ability? Table 3 depicts the Gt-known localization errors *w.r.t.* different resolutions of input images. We exclude the influences of classification results by comparing the Gt-known localization errors. We see that enlarging the resolutions can decrease the localization errors with other configuration unchanged. For ILSVRC, the gain is slight of only 0.46%, Dislike ILSVRC,

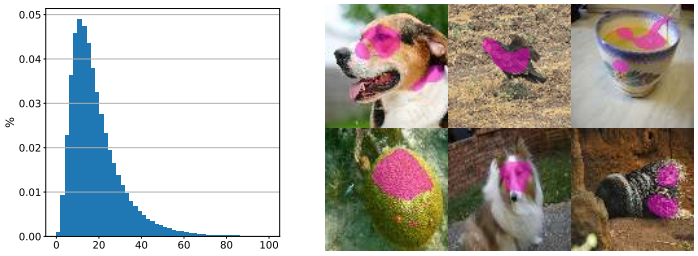
CUB is more sensitive to the resolution of input images and the Gt-known error drops significantly by 4.41%.

Are SC and GC really effective? Table 4 compares the localization errors of the proposed constraints to the plain version network. When using only the cross-entropy loss, the localization and the Gt-known localization errors are 53.71% and 37.32%, respectively. After adding the SC constraint, the localization accuracy is significantly improved and the error rate drops to 49.07%. The Gt-known localization error also decreases to 31.63% by a large margin of 5.69%. Furthermore, the localization performance can be boosted with the GC constraint. By adding the proposed two constraints simultaneously, the localization and Gt-known errors can finally be reduced to 48.45% and 31.30%, respectively. Moreover, to verify the superiority of the proposed updating strategy in GC, we conduct an experiment of updating the global memory matrix with the back-propagation process. The localization accuracies of such an updating method obtains 49.03% and 32.09% in localization and Gt-known error rates, respectively. The back-propagation updating strategy is worse than the proposed updating strategy. Moreover, although the global constraint introduces a vector for each class, these vectors only involve negligible computational resources. In particular, the number of parameters for the backbone network is 27M (InceptionV3). GC only increase 0.8M parameters by 2.9%. During the forward stage of training, the Flops is 21.42G. GC only gain 3K Flops which can be totally ignored. SC does not involve extra Flops nor parameters. During the testing phase, neither SC nor GC involve any extra Flops.

Methods	Plain	SC	SC + GC
Loc.	53.71*	49.07	**48.08**
Gt-known Loc.	37.32	31.63	**31.04**

Table 4: Localization error on ILSVRC validation set using different configurations of the proposed constraints. ($*$ indicates the numbers obtained with the classification results using the ten-crop operation.)

Is I^2C sensitive to λ_1, λ_2 and K? $\boldsymbol{\lambda_1}$ controls the relative importance of SC in training. In order to maximize the performance of the proposed method and study the robustness to λ_1, we test the localization accuracy *w.r.t.*various values of λ_1. We remove the GC constraint and only add the SC constraint. Figure 5a illustrates the classification error, localization error and Gt-known localization errors when λ_1 changes in 40 times of scale from 0.002 to 0.08. We obtain the best results of 29.24%, 49.07% and 31.63% and the worst results of 32.13%, 50.82% and 34.44% in terms of the classification, localization and Gt-known localization error, respectively. In general, with the increase of λ_1, the classification ability is getting better while Gt. localization is getting worse. The localization errors with respect to the predicted labels achieve the best value of 49.07% at 0.008, because only the generated bounding boxes are counted as correct hits when the predicted labels of classification meet the ground-truth labels in this criterion. It is notable that the gap between the best and worst



(a) (b)

Fig. 4: *(a)*: histogram of the number of object pixels with the threshold of 0.7 on the ILSVRC training set. *(b)*: identified object regions (in magenta) according to the localization maps

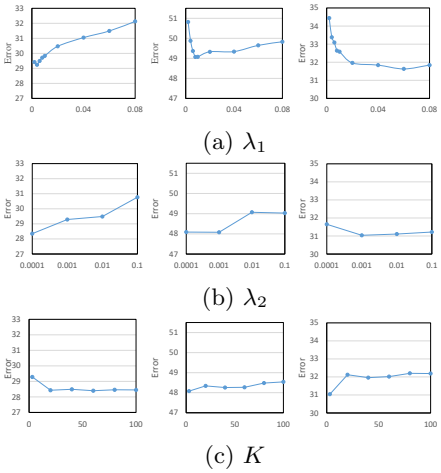
values of the localization error is *only* 1.75% *over the changes of* 40 *times in* λ_1, which means the proposed method is quite robust to the values of λ_1.

$\boldsymbol{\lambda_2}$ controls the relative importance of GC in training. According to Figure 5a, we leverage the best value of $\lambda_1 = 0.008$ to study the performance with λ_2 changing from 0.0001 to 0.1. Figure 5b shows the accuracies of the proposed model adding both SC and GC. The classification errors remain stable at around 30% after applying GC. The localization accuracy achieve lower error rates of 48.07% compared to only using SC, which reflects the effectiveness of the proposed global consistency strategy. As for the Gt-known localization metric, the error rates decrease to 31.04% by adding GC. It is also notable that the model is robust to the change of the hyper-parameter λ_2 in *a large range of 1,000 times from* 0.0001 *to* 0.1. The localization error *only fluctuates within a range of 0.96%*.

$\boldsymbol{K}$ is the number of the chosen object seeds. For an input image with the resolution of 320 by 320, the networks downsample the resolution by a factor of 8 and obtain the corresponding localization maps of 40 by 40 with 1600 pixels. Following the recommended threshold in SPG [43], we set the threshold δ to 0.7 for discovering the object regions in each image. Figure 4 illustrates the histogram of the number of the identified object pixels on the ILSVRC training set. We observe that most of the images contain object pixels in the range from 1 to 60. Particularly, we choose $K = \{3, 20, 40, 60, 80, 100\}$ to study the performance changes with the variant number of sampled representative pixels in each image. Figure 5c shows the classification and localization accuracies with different K values. The classification error is lower when the number of sampled pixels is relatively larger. The classification error rate reaches the lowest point of 28.40% at $K = 60$. On the contrary, the localization error increases when we adopt a larger value of K, and the localization error is lowest at $K = 3$. We believe the reason for this phenomenon is that the larger number may include more noises in the sampled object pixels, and more sampled pixels implicitly make the constraints stronger. The proposed I^2C is robust to the change of seed

numbers, and *the localization accuracy only fluctuates within a small range of 0.46% with K changing of 33 times from 3 to 100.*

In total, we can summarize that the improvement of the proposed method mainly attribute to three aspects. First, SC can semantically drive the consistency between the object features in a batch, which improves the localization maps. Second, GC can further drive the consistency throughout the entire dataset by forcing the object features towards their class-specific centers. Third, the proposed updating strategy can effectively learn the class-specific centers with the training process. Also, we have studied the robustness of the hyper-parameters. λ_1 and λ_2 are for balancing the costs of SC and GC. The experiment results show that the network performs superiorly in a very large range. In addition, according to the changes of the localization performance *w.r.t.*the change of K. We can obtain satisfied accuracies just selecting a small value of K, *e.g.*, $K = 3$.



(a) λ_1

(b) λ_2

(c) K

Fig. 5: Classification (*left*), localization (*middle*) and Gt-known Loc (*right*) error rates with the changes of hyper-parameters, *i.e.*, λ_1, λ_2 and K.

5 Conclusion

We propose an Inter-Image Communication approach (I^2C) to improve the accuracy of localization maps through training classification networks. First, we randomly select several object seeds according to the activated area of localization maps. These seed points are further employed to extract representation vectors of class-specific objects. Second, the extracted vectors are leveraged to communicate between different objects of the same classes. Concretely, stochastic consistency (SC) is proposed to optimize the objects within a mini-batch. Global consistency (GC) is designed to keep consistency across minibatchs. Also, a strategy is applied to update the global memory matrix. Finally, the proposed I^2C approach achieves the Top-1 localization error rate of 45.17% on the ILSVRC validation set, surpassing the current state-of-the-art method.

Acknowledgement.
This work is partially supported by ARC DECRA DE190101315 and ARC DP200100938. Xiaolin Zhang (No. 201606180026) is partially supported by the Chinese Scholarship Council.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: IEEE CVPR. pp. 2209–2218 (2019) 4
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: IEEE CVPR. pp. 4981–4990 (2018) 4
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. preprint arXiv:1412.7062 (2014) 1
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI pp. 834–848 (2017) 3
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611 (2018) 1
6. Cheng, B., Chen, L.C., Wei, Y., Zhu, Y., Huang, Z., Xiong, J., Huang, T.S., Hwu, W.M., Shi, H.: Spgnet: Semantic prediction guidance for scene parsing. In: IEEE ICCV. pp. 5218–5228 (2019) 1
7. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3133–3142 (2020) 4
8. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: IEEE CVPR (June 2019) 2, 4, 8, 9, 10
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE CVPR. pp. 248–255 (2009) 4, 8
10. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. JMLR **12**(Jul), 2121–2159 (2011) 7
11. Durand, T., Mordan, T., Thome, N., Cord, M.: Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: IEEE CVPR. pp. 642–651 (2017) 4
12. Fan, J., Zhang, Z., Song, C., Tan, T.: Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: IEEE CVPR (June 2020) 4
13. Girshick, R.: Fast r-cnn. In: arXiv preprint arXiv:1504.08083 (2015) 1
14. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013) 7
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR. pp. 770–778 (2016) 9, 10
16. Hou, Q., Jiang, P., Wei, Y., Cheng, M.M.: Self-erasing network for integral object attention. In: NIPS. pp. 549–559 (2018) 4
17. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE CVPR. pp. 7132–7141 (2018) 10
18. Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., Huang, T.: Ccnet: Criss-cross attention for semantic segmentation. IEEE TPAMI (2020) 1
19. Jiang, P.T., Hou, Q., Cao, Y., Cheng, M.M., Wei, Y., Xiong, H.K.: Integral object mining via online attention accumulation. In: IEEE ICCV. pp. 2070–2079 (2019) 4
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 7

21. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: ECCV. pp. 695–711 (2016) 4
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012) 10
23. Liang, X., Liu, S., Wei, Y., Liu, L., Lin, L., Yan, S.: Towards computational baby learning: A weakly-supervised approach for object detection. In: IEEE ICCV. pp. 999–1007 (2015) 4
24. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: IEEE CVPR. pp. 685–694 (2015) 4
25. Papandreou, G., Chen, L.C., Murphy, K., Yuille, A.L.: Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. IEEE ICCV (2015) 4
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015) 1
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y 4, 8, 9
28. Shimoda, W., Yanai, K.: Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: ECCV. pp. 218–234 (2016) 4
29. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013) 10
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015) 1, 9, 10
31. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. arXiv preprint arXiv:1704.04232 (2017) 4, 10
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE CVPR. pp. 1–9 (2015) 10
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE CVPR. pp. 2818–2826 (2016) 1, 9, 10
34. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 4, 8
35. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4
36. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: IEEE CVPR (2017) 4, 5
37. Wei, Y., Liang, X., Chen, Y., Jie, Z., Xiao, Y., Zhao, Y., Yan, S.: Learning to segment with image-level annotations. PR (2016) 4
38. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE TPAMI (2016) 4
39. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: IEEE CVPR. pp. 7268–7277 (2018) 2, 4

40. Xue, H., Liu, C., Wan, F., Jiao, J., Ji, X., Ye, Q.: Danet: Divergent activation for weakly supervised object localization. In: IEEE ICCV. pp. 6589–6598 (2019) 10
41. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: IEEE ICCV. pp. 6023–6032 (2019) 4, 10
42. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.: Adversarial complementary learning for weakly supervised object localization. In: IEEE CVPR (2018) 1, 2, 3, 4, 5, 6, 8, 9, 10, 11
43. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: ECCV. Springer (2018) 1, 3, 5, 8, 9, 10, 13
44. Zhang, X., Wei, Y., Yang, Y., Wu, F.: Rethinking localization map: Towards accurate object perception with self-enhancement maps. arXiv preprint arXiv:2006.05220 (2020) 4
45. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE CVPR. pp. 2881–2890 (2017) 1
46. Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. IEEE CVPR (2016) 1, 2, 8, 9, 10
47. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE TPAMI (2018) 2
48. Zhu, Y., Zhou, Y., Ye, Q., Qiu, Q., Jiao, J.: Soft proposal networks for weakly supervised object localization. arXiv preprint arXiv:1709.01829 (2017) 4