

Self-Training and Adversarial Background Regularization for Unsupervised Domain Adaptive One-Stage Object Detection

Seunghyeon Kim
KAIST

Jaehoon Choi
KAIST

Taekyung Kim
KAIST

Changick Kim
KAIST

{seunghyeonkim, whdns44, tkkim93, changick}@kaist.ac.kr

Abstract

Deep learning-based object detectors have shown remarkable improvements. However, supervised learning-based methods perform poorly when the train data and the test data have different distributions. To address the issue, domain adaptation transfers knowledge from the label-sufficient domain (source domain) to the label-scarce domain (target domain). Self-training is one of the powerful ways to achieve domain adaptation since it helps class-wise domain adaptation. Unfortunately, a naive approach that utilizes pseudo-labels as ground-truth degenerates the performance due to incorrect pseudo-labels. In this paper, we introduce a weak self-training (WST) method and adversarial background score regularization (BSR) for domain adaptive one-stage object detection. WST diminishes the adverse effects of inaccurate pseudo-labels to stabilize the learning procedure. BSR helps the network extract discriminative features for target backgrounds to reduce the domain shift. Two components are complementary to each other as BSR enhances discrimination between foregrounds and backgrounds, whereas WST strengthens class-wise discrimination. Experimental results show that our approach effectively improves the performance of the one-stage object detection in unsupervised domain adaptation setting.

1. Introduction

Object detection is a fundamental and core problem in computer vision. Recent studies [26, 21, 25] have achieved remarkable improvements with the advances of deep neural networks and large-scale benchmarks [7, 19, 8]. Deep learning-based object detectors can be categorized as two-stage detectors or one-stage detectors. Two-stage detectors first extract object regions and then refine them through classification and regression [26, 6]. On the other hand, one-stage object detectors [25, 21, 35, 20, 39] directly estimate the coordinates and classes of objects without the Region of Interest (RoI) pooling procedure.

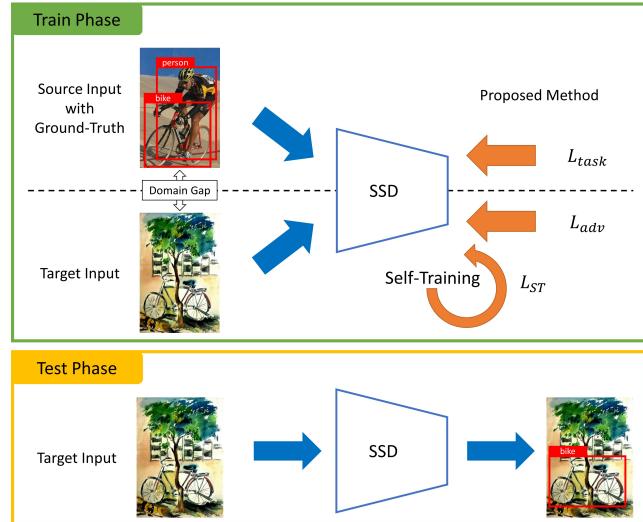


Figure 1. Illustration of unsupervised domain adaptive one-stage object detection. We train an object detector with labeled source images and unlabeled target images. Our method improves the performance of the network for target inputs.

One limitation of these supervised learning-based methods is the assumption that test data have the same distribution as train data. However, domain shift frequently occurs in many practical applications. For example, variances of object appearance, viewpoints, backgrounds, illumination, and weather condition can degenerate the performance of the network. One possible solution is collecting labeled data for a new domain, but it is usually expensive and time-consuming. To address this issue, domain adaptation transfers knowledge from the train data domain (source domain) to the test data domain (target domain). Especially, unsupervised domain adaptation assumes there are no labels available in the target domain. The goal of domain adaptation is to train the network that performs well on the target domain dataset.

Unfortunately, domain adaptive object detection has received less attention in contrast to classification [22, 10, 33,

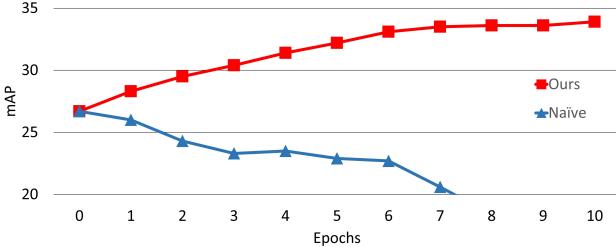


Figure 2. Trends of mAP on the target domain with training epochs. A naive self-training degenerates the accuracy without image-level labels and the regression loss (blue, triangle). Our weak self-training (WST) enables effective self-training under the same settings (red, rectangle).

32, 36, 23] and semantic segmentation [13, 14, 31, 40, 4]. For object detection, the authors of [3] propose a global feature alignment approach in an adversarial way. However, the global feature alignment is not sufficient for object detection since it will align non-transferable backgrounds. Besides, this work is designed only for a two-stage object detector, Faster R-CNN [26]. On the other hand, a recent work [15] presents cross-domain weakly-supervised object detection on SSD [21], which is a representative one-stage object detector. However, this work assumes that image-level labels are available on the target domain. Unlike prior works, we introduce one-stage object detection under unsupervised domain adaptation setting. Figure 1 summarizes the overall framework of our method.

In this paper, we propose weak self-training (WST) for stable learning procedure and adversarial background score regularization (BSR) to reduce domain shifts. Previous studies [32, 36, 9, 40, 4, 15] show the effectiveness of self-training for domain adaptation. However, naive self-training approaches without image-level labels harmful for object detectors as shown in Fig. 2. To achieve robust self-training, WST minimizes adverse effects of both false positives and false negatives occurred in pseudo-labels. WST does not require any label on the target domain contrary to weakly-supervised approaches [37, 30, 38, 15] that utilize image-level labels for a reliable choice of pseudo-labels. Besides, we add BSR at the training phase to reduce the domain shift. We point out that backgrounds of the source and the target data have less common features than foregrounds. From this motivation, BSR extracts discriminative features for target backgrounds. This objective is crucial for one-stage detectors since they do not have region proposal process. WST and BSR are complementary to each other as BSR considers discrimination between foregrounds and backgrounds, and WST provides category information of detections.

The contribution of our paper is as follows.

- We introduce weak self-training (WST) for domain

adaptive object detection which reduces negative effects of inaccurate pseudo-labels.

- We propose adversarial background score regularization (BSR) to reduce the domain shift by extracting discriminative features for target backgrounds.
- Experimental results show that our approach improves the performance of one-stage object detection under unsupervised domain adaptation setting.

2. Related Work

2.1. Object Detection

Two-stage detectors first extract foreground proposals and then refine the results in the second stage. R-CNN [12] utilizes selective search for region proposal and convolutional neural network for classification. For a faster inference, Fast R-CNN [11] shares the feature map from the same input images. Faster R-CNN [26] further improves the performance via replacing selective search with a fully convolutional network called region proposal network (RPN).

On the other hand, one-stage detectors also have been researched and show impressive performance on the inference speed. YOLO [25] is a fast object detector based on a fully convolutional network. SSD [21] improves the performance of accuracy by utilizing feature maps from various scales. The authors of [18] point out that one-stage detectors suffer from the class imbalance problem between foregrounds and backgrounds, and propose focal loss which focuses on hard examples rather than easy ones. Furthermore, recent studies [35, 20, 39] have improved the performance both in accuracy and inference speed maintaining the efficiency of one-stage detectors.

2.2. Domain Adaptation

Domain adaptation reduces the domain gap between train data and test data. For classification, most recent methods manage to reduce the discrepancy between feature distributions of the source and the target. Early work [22] uses Maximum Mean Discrepancy (MMD) as a metric for the discrepancy between two distributions. The authors of [10] propose adversarial learning by adding a gradient reversal layer (GRL) and a discriminator to extract domain invariant features. Recent works [33, 23, 32, 36] are based on the domain adversarial learning framework and further improve the discriminative property on the target domain.

For semantic segmentation, GAN-based domain adaptation approaches are shown to be effective. The authors of [13] point out that earlier feature-level distribution matching approaches fail to capture pixel-level domain shifts. They propose adversarial domain adaptation model, which aligns

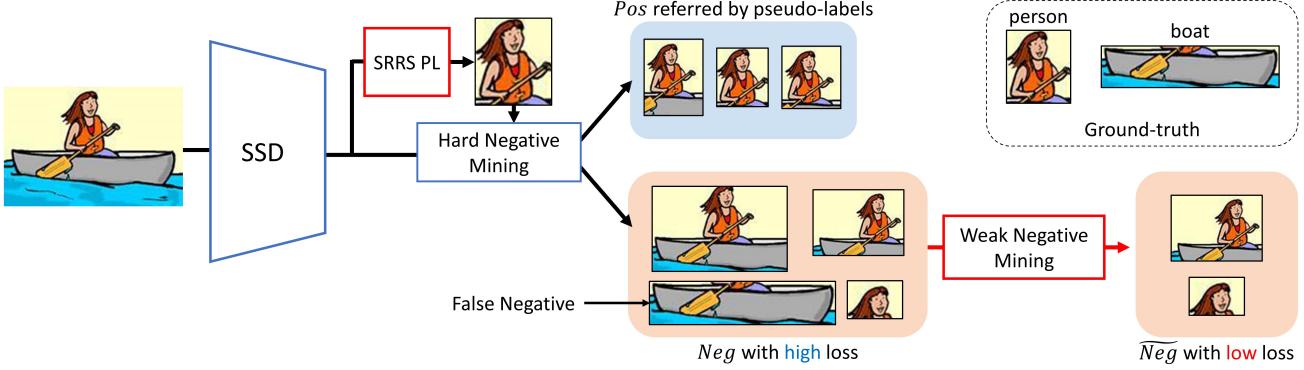


Figure 3. The framework of proposed weak self-training. First, we generate pseudo-labels using SRRS (Supporting Region-based Reliable Score) as a criterion. Second, by following traditional hard negative mining, we obtain two sets of positive examples and negative examples (*Pos* and *Neg* respectively). Finally, since the chosen hard negatives are risky for self-training, we select easy samples among *Neg* and construct $\widehat{N}eg$. We use examples in *Pos* and $\widehat{N}eg$ for weak self-training.

both pixel-level and feature-level distribution. In [14], conditional GAN was used to model the residual of the feature map between the source and target domain.

Some prior works use self-training [1, 17] to compensate for the lack of categorical information for either classification [2, 24, 27, 34, 9, 36, 32, 5] or segmentation [40, 4].

2.3. Domain Adaptive Object Detection

Compared to classification and semantic segmentation, domain adaptive object detection has received less attention. The authors of [3] present two domain adaptation components, image-level adaptation and instance-level adaptation. They adopt domain adversarial approach using a discriminator for each component. Recently, the focal loss is utilized for a weak global alignment [28]. However, it is hard to apply their algorithm to one-stage detector since the algorithms are designed for Faster R-CNN [26]. The authors of [16] utilize style transfer method, but the method needs large amount of augmented data.

For one-stage object detection, the authors of [15] propose pseudo-labeling and pixel-level adaptation on a cross-domain weakly-supervised setting, which assumes that image-level labels are available for all target images. They first fine-tune the detector with style transferred source images. After that, they generate pseudo-labels by simply choosing the top-1 confidence detections considering image-level labels. They further improve the performance via fine-tuning the network with generated pseudo-labels. However, we confirmed that this method is not valid with the unsupervised domain adaptation setting and degenerates the detection performance.

On the other hand, proposed WST achieves stable learning without image-level labels and BSR extracts discriminative features for target backgrounds instead of aligning non-transferable features.

3. Proposed Method

In this section, we introduce details of WST and BSR. We adopt SSD [21] as our baseline.

3.1. Problem Setting

We assume that source data (x^s, y^s) is drawn from the source domain X_s , and target data (x^t, y^t) is drawn from the target domain X_t . Here, x is an image and $y = (b, c)$ is a corresponding label, where b is the coordinates of the bounding box and c is the class to which the object belongs. We denote the distribution of domain X as $P(X)$, and $P(X_s) \neq P(X_t)$. We assume that both source data and target data have $K + 1$ classes including the background. We set c to 0 for the background. We do not have access to target labels, y^t .

We denote layers before *conv5* of SSD by F , and the others by C . The output of the SSD is $O = \{r_i\}_{i=1}^n$, where r_i is the i^{th} detection and n is the total number of detections (e.g., $n = 8732$ for SSD300). We only take remaining detections after Non-Maximum Suppression (NMS) as final detections. We denote the final outputs as $O^* = \{r_l^*\}_{l=1}^{n^*}$, where r_l^* is the l^{th} detection and n^* is the number of detections in the final outputs.

3.2. Weak Self-Training

We propose a weak self-training scheme (WST) to compensate for the lack of categorical information on the target domain. Unfortunately, the base network often produces incorrect outputs with high confidence due to large domain shift. These misclassified outputs become false positives when we choose them for pseudo-labels. Also, false negative error occurs when the network fails to detect some objects in an image. To overcome such problems, WST is designed to omit unreliable examples from the training procedure. The framework of WST is shown in Fig. 3.

Reducing False Negatives. We minimize the effects of false negatives by modifying the training loss for supervised learning. As defined in [21], the original loss is

$$L_{task}(x^t, \hat{b}, \hat{c}) = - \sum_{i \in Pos} \log(p_i(\hat{c}_i|x^t)) - \sum_{i \in Neg} \log(p_i(0|x^t)) + L_{loc}(x^t, \hat{b}), \quad (1)$$

where Pos and Neg are sets of positive and negative examples respectively, \hat{b} is a pseudo-bounding box label, \hat{c}_i is a pseudo-class label of i^{th} detection, $p_i(\hat{c}_i|x^t)$ and $p_i(0|x^t)$ are the probability values of the class \hat{c}_i and the background of i^{th} detection, and L_{loc} is the localization loss. However, we observe that Eq. (1) is not effective for self-training. Especially, false negatives selected by hard negative mining are harmful to training. We reduce the false negatives by masking out the gradients of background examples during the training. However, it is undesirable to neglect all the background examples because the network will be biased to the foregrounds. Thus, we ignore background examples that have the potential of being foregrounds.

Negative examples in the Neg set have a large potential of being foregrounds since hard negative mining refers to incorrect pseudo-foregrounds and selects background examples with the highest confidence loss values. For example, in Fig. 3, conventional hard negative mining will select the false negative example of the boat as a background example. Thus, we choose $\|Neg\|/3$ examples that have the lowest confidence loss value among negative examples in Neg . We call this process as weak negative mining, and the obtained set from the process is denoted as \widetilde{Neg} . Additionally, we do not update the network for bounding box regression since pseudo-labels usually have inaccurate bounding box information. Finally, the modified loss function for weak self-training is defined as

$$L_{ST}(x^t, \hat{c}) = - \sum_{i \in Pos} \log(p_i(\hat{c}_i|x^t)) - \sum_{i \in \widetilde{Neg}} \log(p_i(0|x^t)). \quad (2)$$

Reducing False Positives. We propose a criterion for instance-level pseudo labeling based on supporting RoIs. Here, supporting RoIs denote examples having IoU value larger than some threshold δ with the final detection r^* . Rather than using only a single confidence score of a detected box, we consider all the boxes close to the final detection r^* . We define Supporting Region-based Reliable Score (SRRS) as

$$SRRS(r^*) = \frac{1}{N_s} \sum_{i=1}^{N_s} IoU(r_i, r^*) \cdot P(c^*|r_i), \quad (3)$$

where N_s is the number of supporting regions, $IoU(A, B)$ denotes the IoU value between region A and region B, c^* is

Algorithm 1 Generating Pseudo Labels

```

Input  $O, O^*, \epsilon, \delta$ 
Output  $\hat{Y} = \{\hat{y}\}$ 
1: for  $r_l^* \in O^*$  do
2:   for  $r_i \in O$  do
3:     if  $IoU(r_i, r^*) \geq \delta$  then
4:       Collect  $r_i$  from  $O$ 
5:     end if
6:   end for
7:   Calculate  $SRRS(r_l^*)$  by Eq. (3)
8:   if  $SRRS(r_l^*) \geq \epsilon$  then
9:     Add  $r_l^*$  into the set of pseudo-labels,  $\hat{Y}$ 
10:  end if
11: end for

```

a predicted class of r^* , and $P(c^*|r_i)$ is a probability for a region r_i to belong to the c^* . By thresholding the score with ϵ , we choose reliable detections. The pipeline of generating pseudo-labels is described in Algorithm 1.

3.3 Adversarial Background Score Regularization

We point out that backgrounds of the source domain and target domain share less common features compared to those of foregrounds. We claim that simple global feature alignment enforces to align non-transferable backgrounds, and makes the training procedure unstable. Motivated by [29], we propose background score regularization (BSR) in an adversarial way. The loss function for BSR can be defined as the binary cross entropy function as follows:

$$L_{adv}(x^t) = -t \sum_i \log(p_i(0|x^t)) - (1-t) \sum_i \log(1-p_i(0|x^t)), \quad (4)$$

where i is the index of a detection, and $t \in [0, 1]$ is a target value of $p_i(0|x^t)$. As we minimize the loss, the value of $p_i(0|x^t)$ becomes close to t . On the contrary, $p_i(0|x^t)$ should be close to 0 or 1 to maximize the loss.

At the training phase, both the classifier C and the feature extractor F minimize the supervised loss L_{task} for the source inputs. For the target inputs, we regularize C to predict the value of $p_i(0|x^t)$ close to t by minimizing L_{adv} . On the other hand, F is trained to maximize the adversarial loss to make $p_i(0|x^t)$ close to 0 or 1. Thus, F will learn discriminative features to deceive the classifier. The overall training objectives for BSR can be written as follows:

$$\min_F L_{task}(x^s, y^s) - L_{adv}(x^t), \quad (5)$$

$$\min_C L_{task}(x^s, y^s) + L_{adv}(x^t). \quad (6)$$

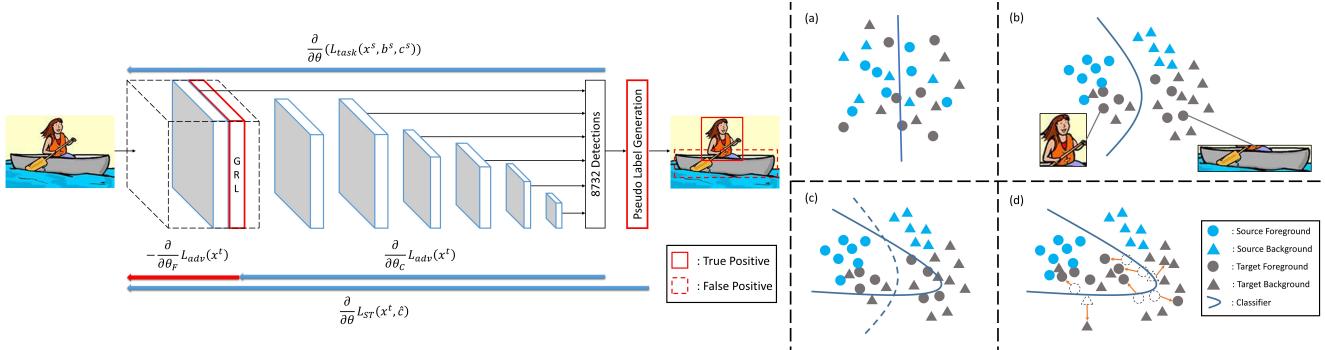


Figure 4. Left: The network architecture with the training losses. θ represents the parameters of SSD. We add Gradient Reversal Layer (GRL) after *relu4_3* of SSD300. Note that GRL is activated only when the target input x^t is used for BSR ($L_{adv}(x^t)$). The example input image is from the target domain. Right: We present feature space representations of the adversarial learning. Each point is a detection. (a): Initial state. (b): The network is trained with the source data. Source examples are well classified, while the boat object of the target data is misclassified as the background. (c): The classifier C minimizes the adversarial loss. The boundary passes through the clusters of the target data. (d): The feature extractor F maximizes the adversarial loss. The example points move away from the boundary.

We enable the adversarial training using gradient reversal layer (GRL) right after *relu4_3* of SSD300.

Without BSR, the base network will produce incorrect prediction with high confidence. The proposed adversarial background score regularization can be thought of as training the classifier to predict with less certainty for target inputs and training the feature extractor to deceive the classifier. In the end, the feature extractor will extract discriminative features, which are easily classified. Figure 4 depicts the process of the adversarial learning.

However, it is not desirable to apply the adversarial loss for all examples because the output of the object detector has an enormous number of background examples. Thus, we sort all examples by their background scores in ascending order and choose low $3N$ examples. Here, N is the number of examples predicted as foregrounds. We found that applying this selection in a batch-wise helps to stabilize the learning. In other words, we combine all examples in a single batch and choose the samples among them. Furthermore, we add a focal term in Eq. (4) to make the loss numerically stable and still effective. The final adversarial loss is defined as

$$L_{adv}(x^t) = -t \sum_i |t - p_i(0|x^t)|^\gamma \cdot \log(p_i(0|x^t)) \\ - (1-t) \sum_i |t - p_i(0|x^t)|^\gamma \cdot \log(1 - p_i(0|x^t)), \quad (7)$$

where γ is a hyperparameter. We set t to 0.5 in our experiments.

Combining BSR and WST will complement each other. BSR reduces domain gaps by helping the network to extract discriminative features for the background class. On the other hand, the network learns category information through WST.

4. Experiments

In this section, we present details of the implementation and compare our results with other methods.

4.1. Datasets and Evaluation

In our experiments, we used Pascal VOC2007-trainval and VOC2012-trainval dataset [8] as a source domain dataset, and Clipart1k, Watercolor2k, or Comic2k dataset [15] as a target domain dataset.

Pascal VOC [8] is a real-world image dataset. It provides both instance-level bounding box annotations and pixel-level annotations. VOC2007-trainval and VOC2012-trainval set have total 16,551 images with 20 distinct categories. Clipart1k [15] is a dataset of graphical images which have a large domain gap with real-world images. It provides 1k images and has the same categories as Pascal VOC. We used all images as a target dataset both for training and evaluation. Watercolor2k and Comic2k [15] are also unrealistic datasets. Each dataset provides 2k of images, 1k for a train set and the other 1k for a test set. Both Watercolor2k and Comic2k have 6 classes which also exist in VOC. We used the train set for training and the test set for evaluation.

For all experiments, we evaluated methods using mean average precision (mAP). We set the confidence threshold to 0.05 and the IoU threshold to 0.5.

4.2. Implementation Details

In all experiments, we used SSD300 as a base network. Following the original paper [21], inputs were resized to 300×300 , and we applied all augmentations used in the original paper. SGD was used as an optimizer.

Base Network. For the base network, only source dataset was used. We trained the network for 120k iterations with

Method	BSR	WST	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Base [21]			27.3	60.4	17.5	16.0	14.5	43.7	32.0	10.2	38.6	15.3	24.5	16.0	18.4	49.5	30.7	30.0	2.3	23.0	35.1	29.9	26.7
ST			11.8	16.4	9.1	10.8	0.3	17.7	13.9	9.1	14.7	4.5	11.1	9.1	2.3	15.2	9.1	23.7	1.8	9.1	4.5	19.8	10.7
DANN [10]			24.1	52.6	27.5	18.5	20.3	59.3	37.4	3.8	35.1	32.6	23.9	13.8	22.5	50.9	49.9	36.3	11.6	31.3	48.0	35.8	31.8
Ours	✓		26.3	56.8	21.9	20.0	24.7	55.3	42.9	11.4	40.5	30.5	25.7	17.3	23.2	66.9	50.9	35.2	11.0	33.2	47.1	38.7	34.0
Ours		✓	30.8	65.5	18.7	23.0	24.9	57.5	40.2	10.9	38.0	25.9	36.0	15.6	22.6	66.8	52.1	35.3	1.0	34.6	38.1	39.4	33.8
Ours	✓	✓	28.0	64.5	23.9	19.0	21.9	64.3	43.5	16.4	42.2	25.9	30.5	7.9	25.5	67.6	54.5	36.4	10.3	31.2	57.4	43.5	35.7

Table 1. Comparison of various methods in terms of mAP. For all methods, the base network is SSD300 [21]. Pascal VOC2007 trainval and VOC2012 trainval is used for source dataset and Clipart1k is used for target dataset. Descriptions of each method is in Sec. 4.3.

Method	BSR	WST	bike	bird	car	cat	dog	person	mAP
Base [21]			77.5	46.1	44.6	30.0	26.0	58.6	47.1
ST			78.9	48.1	44.9	30.1	29.1	61.7	48.8
DANN [10]			73.4	41.0	32.4	28.6	22.1	51.4	41.5
Ours	✓		82.8	43.2	49.8	29.6	27.6	58.4	48.6
Ours		✓	77.8	48.0	45.2	30.4	29.5	64.2	49.2
Ours	✓	✓	75.6	45.8	49.3	34.1	30.3	64.1	49.9

Table 2. Comparisons on Watercolor2k test set.

an initial learning rate of 1.0×10^{-3} , momentum of 0.9, weight decay of 5.0×10^{-4} . We applied learning rate decay of 0.1 at 80k and 100k, so the final learning rate became 1.0×10^{-5} . From this setting, the base network shows an accuracy of 77.43% mAP on the VOC2007 test set.

Adversarial Background Score Regularization. Both source and target data were used. Each batch is composed of 32 images, 16 from the source domain, and the other 16 from the target domain. We used $t = 0.5$ and $\gamma = 2.0$ in Eq. (4). We trained the network for 50k iterations with a learning rate of 1.0×10^{-3} , and reduced the learning rate to 1.0×10^{-4} for another 10k.

Weak Self-Training. We fine-tuned the base model with generated pseudo-labels. This method is different from PL in [15] because we update pseudo-labels for every iteration. The network was trained for 10 epochs with a batch size of 32 and a learning rate of 1.0×10^{-5} . In the process of generating pseudo-labels, we set ϵ to 0.8 in Algorithm 1 unless otherwise stated. Also, $\delta = 0.5$ was used since the evaluation metric regards an example as a positive when it has a value of IoU larger than 0.5 with the ground-truth.

BSR with WST. We followed all the settings of BSR. WST was employed after 50k iterations as the network is not reliable before then. The training was early stopped at 55k iterations since self-training is not helpful when it is overused. The threshold of $\epsilon = \frac{1}{1+e^{-3p}}$ was taken where $p = \frac{\text{current iteration}}{\text{max iteration}}$.

DANN. The same configuration with BSR was used to implement DANN for SSD. We aligned distribution of features extracted from *relu4_3*.

Method	BSR	WST	bike	bird	car	cat	dog	person	mAP
Base [21]			43.3	9.4	23.6	9.8	10.9	34.2	21.9
ST			27.3	9.1	17.3	1.5	9.1	20.8	14.2
DANN [10]			33.3	11.3	19.7	13.4	19.6	37.4	22.5
Ours	✓		45.2	15.8	26.3	9.9	15.8	39.7	25.5
Ours		✓	45.7	9.3	30.4	9.1	10.9	46.9	25.4
Ours	✓	✓	50.6	13.6	31.0	7.5	16.4	41.4	26.8

Table 3. Comparisons on Comic2k test set.

4.3. Results and Comparisons

We compared our method with the base network [21], DANN [10] and ST. Here, ST is the naive approach of self-training that utilizes pseudo-labels as ground-truth without localization loss. While PL in [15] generates pseudo-labels only once before training, ST and our method recreate pseudo-labels for every single iteration. By comparing DANN and ST with our method, we can directly confirm the effectiveness of the proposed algorithm.

Results on Clipart1k. As shown in Table 1, our weak self-training method can improve the accuracy of the object detector without any labels in the target domain. On the other hand, the naive approach (ST) degenerates the performance due to the effects of false positives and false negatives occurred in generated pseudo-labels. To validate each component of weak self-training, we did ablation study in Sec. 5.1. Applying BSR shows performance gaps of 8% and 2.2% mAP compared to the baseline and DANN respectively. Without any additional networks such as discriminator, adversarial background score regularization effectively improves the performance. Using both BSR and WST further enhances the performance as they complement each other.

Self-training approaches have inferior performance on the class of sheep, because of the poor performance of the base network, while the domain adaptation methods (DANN and BSR) show improvement of nearly 9% AP.

Results on Watercolor2k and Comic2k. Comparison of performances on VOC \rightarrow Watercolor2k and VOC \rightarrow Comic2k is shown in Table 2 and 3.

In the case of Watercolor2k, we set the learning rate to 1.0×10^{-6} for self-training methods, since most of the images contain single instance and thus the network is easily overfitted [15]. Furthermore, images in Watercolor2k have

Method	SRRS	Mask	Weak Mask	mAP
ST (A)				10.7
SRRS (B)	✓			10.5
Mask (C)		✓		16.8
SRRS+Mask (D)	✓	✓		29.2
Weak Mask (E)			✓	31.3
SRRS+Weak Mask (F)	✓		✓	

Table 4. Ablation study on WST. Mask denotes that no negative example is used for learning. Weak Mask indicates that weak negative mining is used for sampling negative examples. Method A is identical to ST in Table 1 and method F is the proposed weak self-training.

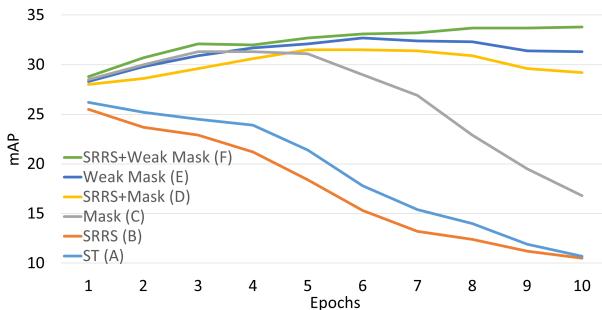


Figure 5. The change of accuracy with the training procedure. We provide all performances during training for each epoch. Different combinations of each method are shown in Table 4. The performances of method A, B and C decreased dramatically due to the adverse effects of false positives and false negatives in pseudo-labels. The proposed method F had the highest value with the stable learning process.

no hard backgrounds such as obstacles. From these reasons, our algorithms show less improvement compared to Clipart1k and Comic2k.

For Comic2k, $\gamma = 3.0$ was used in Eq. (7). Proposed methods improve the accuracy about 5% mAP from the base, while DANN method seems no improvement.

5. Analysis

We conducted ablation studies on WST and parameter sensitivity experiments on BSR. All experiments in this section use Clipart1k as a target dataset.

5.1. Ablation Study on WST

To validate each component of proposed weak self-training, we provided ablation study. In these experiments, we set ϵ to 0.9 without SRRS, and 0.8 with SRRS in Algorithm 1. In Table 4, we present several combinations of three components with their method names and performances. Method A is the naive self-training and method F is the proposed weak self-training. In Fig. 5, we provide the performance trends with training epochs to validate learning stability.

t	γ	mAP	γ	t	mAP
	0.0	not converge		0.25	28.5
	1.0	not converge		0.33	19.8
0.5	2.0	34.0	2.0	0.5	34.0
	4.0	32.9		0.67	21.2
	5.0	31.1		0.75	20.8

Table 5. Parameter sensitivity on BSR. The table shows the performance of BSR with various values of γ and t in Eq. (7).

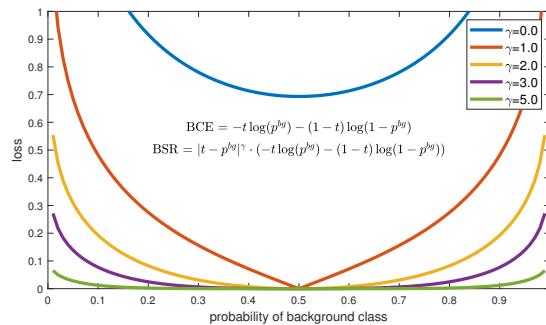


Figure 6. Visualization of background score regularization (BSR) with different values of γ . p^{bg} denotes the probability of background class. BCE stands for binary cross entropy loss, and it is identical to BSR with $\gamma = 0$. Note that $t = 0.5$ for all graphs.

From the results of method A, B, and C, self-training cannot be accomplished without either SRRS or Mask. The accuracy of the three methods rapidly decreases as training proceeds. This implies that reducing both false positives and false negatives is crucial for object detection.

Comparing with Mask, proposed Weak Mask showed remarkable performance improvement by 14.5% mAP without SRRS and 4.6% mAP with SRRS (E compared to C and F compared to D). Comparing methods A and E, only adding weak negative mining can improve the performance dramatically. From these results, we confirm that selecting reliable negative samples is even more effective than selecting positive samples. More specifically, since pseudo-labels commonly omit some foreground instances, hard negative mining tends to select those missing instances as backgrounds. Masking all the gradients of the negative examples can stabilize the learning, but the network will be biased to foregrounds as it never learns about backgrounds. Thus, the proposed weak negative mining is critical for a self-training under domain adaptation setting.

We observed that SRRS is not valuable when it is used alone. Combined with either Mask or Weak Mask, SRRS succeeded to enhance both learning stability and accuracy (D compared to C and F compared to E). Although SRRS can reduce not only false positives but also true positives, we experimentally confirmed that reducing false positives is crucial even though less true positives are used for training.

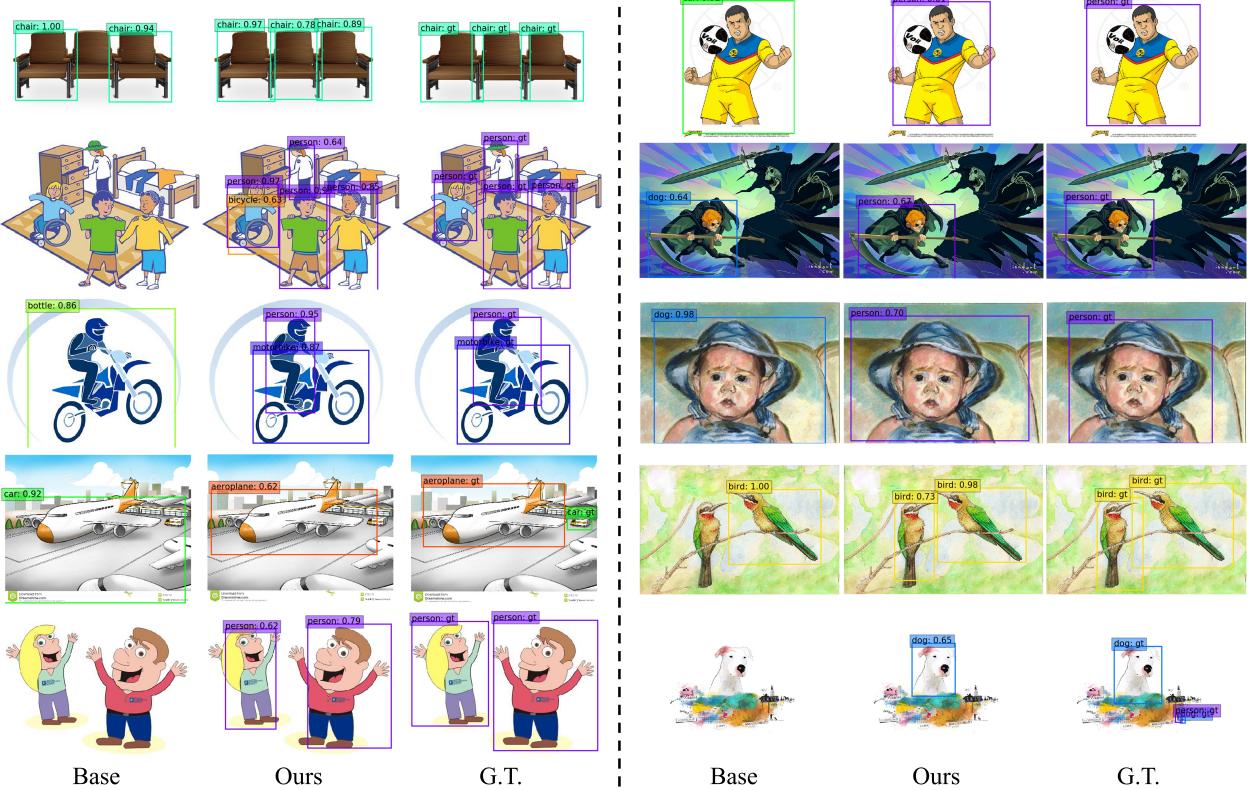


Figure 7. Qualitative results on Clipart1k, Watercolor2k, and Comic2k. We present the results of the base network, our method, and ground-truth from left to right.

5.2. Parameter Sensitivity on BSR

Table 5 shows the results for parameter sensitivity of both γ and t in Eq. (7). The focusing parameter γ controls the strength of the loss. As γ gets smaller, more detections contribute to the adversarial training. More specifically, the network will have too strong regularization on backgrounds with $\gamma = 0$ and $\gamma = 1$. On the other hand, the regularization is relaxed with large values of γ . The network will ignore examples which have the probability of the background around 0.5 with large γ . See Fig. 6 for visualization of BSR with various values of γ . The performance is insensitive to the value of γ unless it is too small.

We conducted experiments on $t \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}\}$ to verify trends of the learning according to t . The network trained with $t = 0.5$ shows better performance than the others. For the other values, the network is easily over-regularized and the performance rapidly drops after the learning rate decay.

5.3. Qualitative Results

We compare the qualitative results of the base network, the proposed method, and ground-truth as shown in Fig. 7. We found that the proposed method detects objects with less confidence but correctly, compared to the base network due

to BSR. As shown in top left example in Fig. 7, the probabilities of two chairs that are detected by the base network are decreased while the chair between them is only detected by our method.

6. Conclusion

In this paper, we have addressed unsupervised domain adaption for one-stage object detection. We enable self-training for object detection by reducing the adverse effects of inaccurate pseudo-labels. Proposed weak self-training (WST) effectively reduces false negatives and false positives by masking the gradients of hard negative examples and utilizing SRRS as a criterion for pseudo-labeling. We have also present adversarial background score regularization (BSR) to reduce the domain shifts by enhancing discrimination between foregrounds and backgrounds of the target data.

Acknowledgements This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIP) (No. 2018-0-00198), Object information extraction and real-to-virtual mapping based AR technology)

References

- [1] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.
- [2] Minmin Chen, Kilian Q Weinberger, and John Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, pages 2456–2464, 2011.
- [3] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.
- [4] Kashyap Chitta, Jianwei Feng, and Martial Hebert. Adaptive semantic segmentation with a strategic curriculum of proxy labels. *arXiv preprint arXiv:1811.03542*, 2018.
- [5] Jaehoon Choi, Minki Jeong, Taekyung Kim, and Changick Kim. Pseudo-labeling curriculum for unsupervised domain adaptation. *arXiv preprint arXiv:1908.00262*, 2019.
- [6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.
- [7] 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.
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [9] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.
- [10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- [11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.
- [14] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018.
- [15] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018.
- [16] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.
- [17] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.
- [18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [20] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018.
- [21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.
- [23] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.
- [24] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.
- [25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [27] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2988–2997. JMLR. org, 2017.
- [28] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive

- object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.
- [29] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018.
- [30] Enver Sangineto, Moin Nabi, Dubravko Culibrk, and Nicu Sebe. Self paced deep learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):712–725, 2018.
- [31] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018.
- [32] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.
- [33] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.
- [34] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5419–5428, 2018.
- [35] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4203–4212, 2018.
- [36] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018.
- [37] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4262–4270, 2018.
- [38] Yongqiang Zhang, Yaicheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. Weakly-supervised object detection via mining pseudo ground truth bounding-boxes. *Pattern Recognition*, 84:68–81, 2018.
- [39] Peng Zhou, Bingbing Ni, Cong Geng, Jianguo Hu, and Yi Xu. Scale-transferrable object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 528–537, 2018.
- [40] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.